Universitat Politècnica de Catalunya
Electrical Engineering Department

PhD Thesis

Advanced data–driven fault diagnosis schemes for active distribution grids

Author: **Paschalia Stefanidou-Voziki**

Advisors: **Jose Luis Dominguez-Garcia**
Roberto Villafafila-Robles

Barcelona, November 2022

Catalonia Institute for Energy Research (IREC)
Electrical Engineering Research Area
Jardins de les Dones de Negre 1 2nd floor,
08930 Sant Adrià de Besòs, Barcelona, Spain

To my family, always

Abstract

Reliable and uninterrupted power supply is crucial for modern societies. Therefore, the minimization of the power outages' duration is a priority for the power system operators. This thesis aims to contribute to the efforts to increase the power restoration speed by proposing automatized solutions for the fault diagnosis process applied in distribution grids. The fault diagnosis process comprises the fault detection, the fault classification and the fault location; all three of these steps are addressed in this study. The traditional fault diagnosis methods that are currently applied to the grids are rapidly becoming obsolete due to the smart grid transition. Renewable energy sources, electric vehicles (EVs) and other smart devices may pose a challenge to conventional fault diagnosis techniques but they also provide opportunities for the application of advanced technological solutions. Hence, in order to take advantage of the grid's digitalization and the subsequent growing data availability the proposed methods are all data–driven, with emphasis given to the use of machine learning (ML).

The fault detection method presented in this thesis refers to active low voltage (LV) grids and specifically the ones with EV fast charging (FC) and ultra fast charging (UFC). The proposed ML–based algorithm utilizes a CatBoostClassifier for the detection of faults and manages to efficiently train the model with static simulation data. These data cor-

respond to the grid's normal and faulty operation when the loads are operating at nominal power and the EV charging is ignored, i.e. to a grid's static loading state. In this way the algorithm is independent of the EV charging and the intermediate operating states. When tested on unseen data corresponding to potential intermediate states the algorithm achieved an accuracy of 97.61% with only 6000 examples included in the training data.

Regarding the fault classification, a data–driven improved version of the popular threshold–based techniques is presented here. The method includes an algorithm that studies the grid's current values before and after a fault under various fault resistances and outputs the most suitable threshold values for the criteria describing each fault type. Thus, it combines increased accuracy with high adaptability to any distribution grid. The proposed technique achieves an accuracy of nearly 100% regardless of the fault's resistance. This constitutes an approx. 20% higher accuracy compared to traditional threshold–based techniques with fixed criteria.

Finally, a complete and practical ML–based fault location method for active LV grids is proposed in this thesis. The method first identifies the faulted branch with the use of a Random Forest (RF) classifier and then locates the faulted point with the use of a regression model. Since the second part of the process is the most crucial and complicated one two tree–based predictive models are tested here, a RF regressor and an XGBoost regressor. In order to improve the models' and, as an extend, the method's performance a thorough data management strategy is included in the algorithm. This includes among others the application of a smart data storage strategy, the comparison of two data minimization approaches and an efficient re–training scheme. Both models lead to mean absolute errors (MAEs) of less than 2 m, nevertheless, the XGBoost proves the most suitable model for this application. When paired with the SelectFromModel dimensionality reduction algorithm it leads to a MAE of 0.49 m and is robust against the major influencing parameters such as the fault resistance, the bidirectional power flow, the data loss and more. The research addresses all the aspects of the method's application, offers an insight on the designing process of ML–based algorithms and presents an efficient, easily–applicable and generalizable solution.

keywords: active low voltage grid, data analysis, data management, fault diagnosis, fault location, fault resistance, machine learning

Resumen

Un suministro eléctrico fiable y continuo es crucial para las actividades de la sociedad moderna. Por lo tanto, la minimización de los cortes de energía y su duración es una prioridad para los operadores de sistemas eléctricos. Esta tesis pretende contribuir a los esfuerzos para aumentar la velocidad de restablecimiento de la energía proponiendo soluciones automatizadas para el proceso de diagnóstico de averías aplicado en las redes de distribución. El proceso de diagnóstico de averías comprende la detección, la clasificación y la localización de ésta. Cada una de estas etapas son estudiadas en esta tesis. Los métodos tradicionales de diagnóstico de averías que se aplican actualmente en las redes eléctricas se están quedando obsoletos debido a la transición a las redes inteligentes. Las energías renovables, los vehículos eléctricos (VE) y otros dispositivos inteligentes pueden suponer un problema para las técnicas convencionales de diagnóstico de fallos por sus comportamientos (bidireccionalidad, picos de corriente, etc), pero por otro lado ofrecen oportunidades para la aplicación y desarrollo de soluciones tecnológicas avanzadas. Por lo tanto, para aprovechar la digitalización de la red y la consiguiente disponibilidad de datos de ésta, en la tesis se desarrollan y presentan nuevos procesos basados en datos y su tratamiento, con énfasis en el uso del aprendizaje automático (ML).

Primero, el método de detección de fallos presentado en esta tesis se

centra en las redes eléctricas de baja tensión (BT) activas y, en concreto, a las redes que incluyen renovables, sistemas de carga rápida (FC) y carga ultrarrápida (UFC) de VE. El algoritmo propuesto, basado en ML, utiliza un CatBoostClassifier para la detección de fallos y consigue entrenar eficientemente el modelo con datos de simulación. Estos datos se corresponden con el funcionamiento normal y defectuoso de la red cuando las cargas funcionan a la potencia nominal y se ignora la carga del VE, es decir, al estado de carga estática de la red. De este modo, el algoritmo es independiente de la carga del VE y de los estados de funcionamiento intermedios (modos de carga y estado de esta). Cuando se validó considerando datos de casos con estados intermedios, el algoritmo alcanzó una precisión del 97,61% con sólo 6000 ejemplos incluidos en los datos de entrenamiento.

En la clasificación de fallos, se presenta una versión adaptada y mejorada basada en datos de las técnicas convencionales y basadas en umbrales. El método incluye un algoritmo que estudia los valores de corriente de la red antes y después de una falta, bajo diferentes resistencias de falla, y emite los valores de umbral más adecuados para los criterios que describen cada tipo de falta eléctrica. De este modo, combina una gran precisión con una gran adaptabilidad a cualquier red de distribución. La técnica propuesta logra una precisión de casi el 100%, independientemente de la resistencia de la falta. Esto constituye una precisión aproximadamente un 20% mayor en comparación con técnicas tradicionales basadas en umbrales con criterios determinados.

Por último, en esta tesis se propone un método completo y práctico de localización de fallos basado en ML para redes activas de BT. El método identifica primero la rama con falta a través del uso de un clasificador Random Forest (RF) y luego localiza el punto de avería mediante un modelo de regresión. Dado que la segunda parte del proceso es la más crucial y complicada, aquí se prueban dos modelos de predicción basados en árboles, un regresor RF y un regresor XGBoost. Para mejorar el rendimiento de los modelos y, por tanto, del método, se incluye en el algoritmo una estrategia de gestión de datos. Esto incluye entre otras cosas, la aplicación de una estrategia inteligente de almacenamiento de datos, la comparación de dos enfoques de minimización de datos y un esquema eficiente de reentrenamiento. Ambos modelos conducen a errores medios absolutos (MAE) de menos de 2 m, sin embargo

el XGBoost resulta ser el modelo más adecuado para esta aplicación. Cuando se combina con el algoritmo de reducción de la dimensionalidad SelectFromModel, se obtiene un MAE de 0,49 m y es robusto frente a los principales parámetros que influyen en el comportamiento como la resistencia al fallo, potencia bidireccional, la pérdida de datos y otros. La investigación aborda todos los aspectos de la aplicación del método, ofrece una visión del proceso de diseño de los algoritmos basados en ML y presenta una solución eficiente, fácilmente aplicable y generalizable.

Palabras clave: análisis de datos, aprendizaje automático, diagnóstico de fallos, gestión de datos, localización de fallos, red de baja tensión activa

Acknowledgements

This thesis was carried out at and financially supported by the Catalan Institute for Energy Research (IREC), and it was completed within the Electrical Engineering doctoral program of the Polytechnic University of Catalonia (UPC).

I would like to express my sincere gratitude to my IREC supervisor Dr. Jose Luis Dominguez for his constant support and guidance. He was always there, patiently offering valuable advice and assistance even when I was too full of myself to take it. My time in IREC has been transformative and I will always be grateful to my supervisor and my colleagues for that. I would also like to thank my UPC supervisor Prof. Roberto Villafafila for his useful comments and observations throughout this research.

Moreover, I would like to thank the person that influenced this research the most and offered invaluable help, Mr. David Cardoner. Both his professional assistance and his personal encouragement were decisive for the completion of this thesis. Another important person that I am grateful to for his mentoring and excellent collaboration is Dr. Nikolaos Sapountzoglou. I could always count on him to offer the right advice and the best graphics whenever I needed them.

I also want to sincerely thank my colleagues at the KTH for mak-

ing me feel at home during my stay there. Special mention to Prof. Nathaniel Taylor for accepting me as a visiting PhD student and taking the time to advise and guide me in my research endeavors.

Finally, I would like to thank my family and friends for their love and encouragement all these years. Without their support I would not have found the strength and determination to finish this work. They always inspire me to do my best and try harder.

Contents

Abstract I

Resum IV

Acknowledgement IX

Table of Contents XI

List of Tables XV

List of Figures XVII

Glossary XXIII

1 Introduction 1
1.1 Motivation 1
1.2 Objectives 3

2 Fault diagnosis in distribution grids: State–of–the–art 5
2.1 The LV and MV grid differences 6
2.2 Fault detection in MV grids 7

2.3 Fault classification in MV grids 9
2.3.1 Artificial intelligence–based methods 10
2.3.2 Hybrid methods 12
2.3.3 Other methods 13
2.3.4 Observations 15
2.4 Fault location in MV grids 16
2.4.1 Impedance–based methods 17
2.4.2 Traveling–wave–based methods 22
2.4.3 Artificial intelligence–based methods 26
2.4.4 Sparse and distributed measurements methods . 30
2.4.5 Hybrid methods 32
2.4.6 General comparison 34
2.5 Fault detection and classification in LV grids 35
2.6 Fault location in LV grids 36
2.7 Observations 38
2.8 Research gaps 39
2.8.1 Fault location 39
2.8.2 Fault classification 41
2.8.3 Fault detection 42
2.8.4 LV grids 42
2.9 Problem definition and selected methodology 43

3 Artificial intelligence algorithms 45
3.1 AI method comparison 46
3.2 Tree–based models 47
3.3 Random Forest 49
3.4 Gradient boosting 50
3.5 RF vs XGBoost 52
3.6 AI metrics 52
3.6.1 Regression metrics 52
3.6.2 Classification metrics 53
3.7 Conclusions 54

4 Fault detection in active low voltage grids with fast and ultra fast charging 55
4.1 EV charging stochasticity and fault detection 56

4.2 Fault detection method description 58
4.2.1 The ML model 59
4.2.2 Training process 61
4.3 Case study 62
4.3.1 Test grid 62
4.3.2 Simulation conditions 63
4.4 Results 67
4.4.1 Algorithm's performance on grids with EV's FC and UFC 68
4.4.2 Sensitivity analysis 69
4.4.3 Conclusions 75

5 Fault classification in distribution grids 77
5.1 Theoretical background 77
5.2 Proposed method 78
5.2.1 Threshold selection process 80
5.3 Case study 86
5.3.1 Test grid and results 86
5.3.2 Conclusions 91

6 Fault location methods for low voltage grids 93
6.1 General method description 93
6.2 Faulted branch classification 96
6.3 Faulted point location 96
6.4 Data management 96
6.4.1 Data collection 98
6.4.2 Data analysis 99
6.4.3 Data storage 101
6.4.4 Data pre–processing 102
6.4.5 Data requirement minimization 103
6.4.6 Hyperparameter tuning 108
6.4.7 Topology change adaptation strategy 109
6.5 Conclusions 110

7 Fault location: Case studies 111
7.1 Case study 1 112
7.1.1 Test grid 112

7.1.2 Data generation 112
7.1.3 Dataset evaluation 114
7.1.4 Data minimization using dimensionality reduction 123
7.1.5 Hyperparameters 125
7.1.6 Fault location accuracy 126
7.1.7 Sensitivity analysis 127
7.2 Case study 2 133
7.2.1 Test grid 134
7.2.2 Data generation 134
7.2.3 Magnitude vs phasor utilization 135
7.2.4 Feature selection method comparison 136
7.2.5 Hyperparameters 139
7.2.6 Results 139
7.2.7 Sensitivity analysis 143
7.3 Comparison of the possible data minimization and predictive model combinations 146
7.4 Conclusions 148

8 Conclusions **151**
8.1 Future work 155

Bibliography **157**

A List of Publications **183**
A.1 Journal articles 183
A.2 Conference articles 184

List of Tables

4.1 Hyperparameter values for the CatBoostClasiffier 62
4.2 Grid element values 64
4.3 The algorithm's performance results for a LV grid without EV charging for different dataset sizes. 68
4.4 The algorithm's performance results for the different dataset characteristics. 68

5.1 The Final Fault Classification Criteria 79
5.2 Selected thresolds for each fault resistance range 88

7.1 Grid element values 113
7.2 Voltage and current values in each analyzed dataset. . . 115
7.3 Comparative table of the dimensionality reduction methods 125
7.4 Hyperparameter values 126
7.5 Fault location prediction model results 127
7.6 Comparison of the model's performance with and without the angles of the V, I included in the dataset 136
7.7 Comparison of feature selection methods 137
7.8 Hyperparameter values 139
7.9 Fault classification results 140

7.10 Fault location algorithm's performance 140
7.11 Retraining scheme performance for the tested topologies 142
7.12 Comparison of feature selection methods 148

List of Figures

2.1 Single–phase fault 6
2.2 Fault location methods published each year for MV and for LV grids [1]. 8
2.3 Fault location methods published each year for active MV and active LV grids [1]. 9
2.4 Number and methodology of the yearly published fault classification papers for MV grids [1] 10
2.5 Overview of the most used fault classification methods for MV grids [1] 11
2.6 Number and methodology of the yearly published fault location papers for MV grids [1] 17
2.7 Circuit with multiple branches, containing a fault with fault current I_f and fault resistance R_f. The fault's distance from the branch preceding it is f_d and from a following bus is $(l_{3x} - f_d)$, where l_{3x} is the distance between buses 3 and x. 18
2.8 Traveling–wave method representation. 23
2.9 Neural network diagram. 27
2.10 Comparative analysis of fault location methods [1]. . . . 34

3.1 The structure of a decision tree [2]. 48
3.2 Illustration of a Random Forest prediction model. 50
3.3 Illustration of the XGboost model's sequential training and weight assignment [2]. 51

4.1 Pre– and post–fault current in the case of an a–phase–to–ground fault in a grid with full EV charging. 57
4.2 Pre– and post–fault current in the case of an a–phase–to–ground fault in a grid with no EV charging. 57
4.3 The training and implementation phases of the proposed fault detection algorithm. 60
4.4 Modified CIGRE European LV benchmark 63
4.5 Charging curves of residential and public chargers. . . . 64
4.6 Weekday occupancy rate curve. 65
4.7 Weekend occupancy rate curve. 65
4.8 Weekday load curve. 66
4.9 Weekend load curve. 66
4.10 Tesla Model 3 charging curve. 67
4.11 Peak current values collected from meter 1 in relation to the target value for the no EV load case. 70
4.12 Peak current values collected from meter 1 in relation to the target value for the full EV load case. 70
4.13 Peak current values collected from meter 1 in relation to the target value for the test case. 71
4.14 The algorithm's accuracy and F1 score in relation to the fault resistance value. 72
4.15 The features' importance. 73
4.16 The algorithm's accuracy and F1 score in relation to the number of meters. 74
4.17 The algorithm's accuracy and F1 score in relation to various PV generation levels. 74
4.18 The algorithm's accuracy and F1 score in relation to various PV generation levels. 75

5.1 The relation between the $\Delta I_a/\Delta I_b$ ratio and the fault resistance for an a–phase–to–ground fault. 81

5.2 The relation between the $\Delta I_a/\Delta I_c$ ratio and the fault resistance for an a–phase–to–ground fault. 81
5.3 The relation between the $\Delta I_a/\Delta I_c$ ratio and the fault resistance for an a–b fault. 82
5.4 The relation between the $\Delta I_b/\Delta I_c$ ratio and the fault resistance for an a–b fault. 82
5.5 The relation between the I_{a_f}/I_{a_p} ratio and the fault resistance for a three phase fault. 83
5.6 Flowchart of the implementation of the proposed fault classification algorithm 85
5.7 Flowchart of the overall method's application process . . 86
5.8 IEEE 13-node test feeder - single line diagram 87
5.9 The method's accuracy with the use of the optimum ratios vs the use of static thresholds in relation to the fault resistance. 89
5.10 Accuracy of the tested algorithms for one and four measurement points [3]. 90
5.11 Accuracy of algorithm 1 and the proposed algorithm in relation to the number of measurement points [3]. 91
5.12 The mean accuracy of the algorithm for all the cases, tested with the updated version of the algorithm. 92

6.1 Flow chart of the method's training process. 95
6.2 Flow chart of the method's implementation after a fault is detected. 97

7.1 Modified CIGRE European LV benchmark used in case study 1. 113
7.2 Kendall's coefficient calculated for the first dataset [2]. . 117
7.3 Kendall's coefficient calculated for the second dataset [2]. 117
7.4 Kendall's coefficient calculated for the third dataset [2]. 118
7.5 Spearman's coefficient calculated for the first dataset [2]. 118
7.6 Spearman's coefficient calculated for the second dataset [2].119
7.7 Spearman's coefficient calculated for the third dataset [2].119
7.8 Pearson's coefficient calculated for the first dataset [2]. . 120
7.9 Pearson's coefficient calculated for the second dataset [2]. 120
7.10 Pearson's coefficient calculated for the third dataset [2]. 121

7.11 Features' importance for the first dataset [2]. 122
7.12 Features' importance for the second dataset [2]. 122
7.13 Features' importance for the third dataset [2]. 123
7.14 The MSE in relation to the CT of the algorithm for a range of [20, 90] dimensions with the use of T–SVD [2]. 124
7.15 The algorithm's MSE in relation to the dataset composition [2] 125
7.16 MPE of the algorithm in relation to the fault's distance from the feeder [2]. 128
7.17 The MSE and CT of the algorithm in relation to the number of utilized examples [2]. 129
7.18 The MSE of the algorithm in the case of data loss from one, three or five meters [2]. 131
7.19 The accuracy of the algorithm in relation to the fault resistance value ranges [2]. 132
7.20 The ratio between the current before and after the fault in relation to the fault resistance as measured by meter 8 [2]. 132
7.21 The accuracy of the algorithm in relation to the PV power generation levels [2]. 133
7.22 Modified CIGRE European LV benchmark used in case study 2. 135
7.23 The percentage of features coming from each meter that were selected by the feature selection algorithm. 138
7.24 The ratio of magnitude vs angle features in the selected features list. 138
7.25 Confusion matrix [4] 140
7.26 The values and weighted average of the MAE and the CT for different amount of utilized examples for Topology 1. 142
7.27 PME in relation to the fault distance from the main feeder [4]. 143
7.28 Test accuracy of the predictive model in relation to the fault resistance [4]. 144
7.29 MSE of the predictive model in relation to the PV power generation [4]. 145
7.30 Model's test accuracy in relation to the nominal value percentage the grid's loads are operating at. 146

7.31 Model's test accuracy in relation to the percentage of noisy data included in the dataset. 147

Glossary

AFIR	Alternative Fuels Infrastructure Regulation
ANFIS	Adaptive Neuro–Fuzzy Inference Systems
AI	Artificial Intelligence
ANN	Artificial Neural Networks
CT	Computational Time
CWT	Continuous Wavelet Transform
DT	Decision Tree
DNN	Deep Neural Network
DFPI	Directional Fault Passage Indicator
DWT	Discrete Wavelet Transform
DG	Distributed Generation
DSO	Distributed System Operator
EV	Electric Vehicle
EMTR	Electromagnetic Transient
ESS	Energy Storage Systems
FC	Fast Charging
FFT	Fast Fourier Transformation
FastICA	Fast Independent Component Analysis

FL	Fuzzy Logic
GE	Great Example
HHT	Hilbert–Huang Transformation
ICT	Information and Communication Technology
IED	Intelligent Electronic Device
ISOMAP	Isometric Feature Mapping
KNN	K–Nearest Neighbors
LAMDA	Learning Algorithm for Multivariable Data Analysis
LV	Low Voltage
ML	Machine Learning
MAE	Mean Absolute Error
MPE	Mean Percentage Error
MSE	Mean Squared Error
MV	Medium Voltage
O&M	Operation and Management
PMU	Phasor Measurement Unit
PV	Photovoltaic
PCA	Principal Component Analysis
RF	Random Forest
RES	Renewable Energy Sources
SAIDI	System Average Interruption Duration Index
SSC	Source Short–Circuit
SVM	Support Vector Machine
TDR	Time Domain Reflectometry
T–SVD	Truncated Singular Value Decomposition
UFC	Ultra Fast Charging
WT	Wavelet Transform
XGBoost	eXtreme Gradient Boosting

1 Introduction

1.1 Motivation

Electricity is one of the most basic utilities, inextricably linked to all parts of modern life. Major sectors such as the industry, health and transportation rely heavily on constant and uninterrupted power supply, and society's progress and prosperity is highly influenced by its access to reliable and affordable electricity. Hence, apart from the construction of a well-designed grid, emphasis should be also given to its protection, maintenance and repair.

In order to ensure that the electricity providers pay the necessary attention to the quick restoration of power supply, various indexes have been established as a means to measure the reliability of each provider [5]. Among the most important ones is the System Average Interruption Duration Index (SAIDI), which is defined as follows:

$$SAIDI = \frac{\sum U_i N_i}{N_T} \tag{1.1}$$

where U_i is the annual power outages' duration and N_i is the number of customers whose power supply was interrupted in location i. N_T is the sum of the company's customers. In the case of exceeding the SAIDI's set limit, the provider is charged with the respective fine. Therefore, the minimization of power outage times is in the best

interest of both the utility companies and the general public.

So far, however, the location of faults has relied on customer calls and the visual inspection of the lines, which are both time consuming. This can change with the implementation of automatized and fast fault diagnosis schemes. Fault diagnosis includes the detection of the fault, then, optionally, the classification of the type of fault and in the end the location of the faulted point, branch or sector. Over the years, multiple methods have been developed on this topic, especially for the detection and location of faults in medium voltage (MV) networks. The most important ones are: the impedance-based technique [6–10], the travelling wave method [11–13], the artificial intelligence (AI) [14–17], the sparse measurements method [18–21] and the hybrid techniques [22–24]. All the methodologies are analyzed in detail in Chapter 2.

As new technologies emerge and the electricity grid is being transformed in accordance with the clean energy goals, the grid's complexity is rising significantly. The part of the grid that is currently undergoing the most radical changes is the distribution grid. The vast integration of Renewable Energy Sources (RES), Energy Storage Systems (ESS) and Electric Vehicles (EVs), among others, together with the increasing flexibility of the electricity markets and the expanding role of prosumers have affected remarkably the characteristics of the distribution grid, e.g. the grid's inertia, the direction of the power flow, the short-circuit level and more. At the same time, the majority of disruptions, approximately 80%, in the distribution grid's power supply are caused by faults [25]. Thus, traditional fault diagnosis methods are becoming obsolete and efficient fault diagnosis is deemed more crucial and challenging than ever.

The grid changes, however, and the associated developments in fields like the ICT (Information and Communication Technology) also offer increased grid observability, growing data availability and various powerful tools for the optimization and modernization of the fault diagnosis processes. The increased data availability facilitates the automation of the fault diagnosis process, however, it also constitutes a source of skepticism towards the applicability and practicality of data–dependent algorithms. Therefore, there are both opportunities as well as challenges associated with the development of novel and fast fault diagnosis schemes, in particular those that take advantage of the new

ICTs.

This applies in particular to the low voltage (LV) side of the distribution grid. Due to its prior lack of complexity, there is not extensive research that takes into consideration the particularities of this part of the grid [26–28]. However, with the major ongoing changes in the LV grid's characteristics and its unbalanced nature that is frequently overlooked, the development of sophisticated fault diagnosis methods adapted to it is imperative.

1.2 Objectives

The aim of this thesis is to present a complete solution to the problem of fault diagnosis in distribution grids with the use of modern technologies. More specifically, emphasis is given on the use of data–driven solutions for the development of fault diagnosis methods for the low voltage (LV) grid. The LV is a part of the electricity grid that is starting to gain attention due to its radical transformation. The transition to the smart grid era reveals new challenges and renders the review and redesign of the traditional fault diagnosis techniques necessary. At the same time, the technological advancements in the field of ICT facilitate the development of improved fault diagnosis solutions. The increased observability over the grid particularly favors the utilization of AI–based methods; AI has been establishment as a powerful tool in data pattern recognition.

Therefore the main goal of this research is the presentation of a complete, fast, accurate and practical fault diagnosis method that optimizes the use of the grid data that either are already available to the grid operator or are expected to be available in the near future. More specifically, this study addresses all the steps of the fault diagnosis process, however, due to the fact that the challenges related to each part of the process are different, the detection, classification and location algorithms presented here do not refer to the same case study. The individual solutions provided for each step are adapted to the particularities of the active LV grids and address the effect of potential accuracy–influencing parameters such as the photovoltaic (PV) penetration, the EV fast charging (FC) and ultra fast charging (UFC), the

topology changes and more. Moreover, this research aims to not only reap the benefits of Machine Learning (ML) algorithms but also to optimize their implementation with the development of advanced data management techniques. Thus, important issues related to the applicability of ML–based methods, such as the required data volume and storage, the computational time and the complexity are addressed.

To sum up, the concrete research objectives of this thesis are:

- The analysis of the state–of–the–art in the field of fault diagnosis and the identification of the related research gaps. (Chapter 2, related publication [1])
- The analysis of the EVs' FC and UFC effect on a ML–based fault detection algorithm and the development of a suitable solution for active LV grids that is robust against the charging stochasticity.(Chapter 4, presented at the CIGRE Session 2022)
- The development of an easily–adaptable fault classification method for distribution grids. (Chapter 5, related publication [3])
- The development of a turnkey ML–based fault location method for active LV grids that analyzes all the aspects related to the application of ML and provides appropriate solutions to the challenges arising. Emphasis is given to the management of the input data in order to achieve optimum predictive accuracy. (Chapters 6 and 7, related publications [2, 4])

2

Fault diagnosis in distribution grids: State–of–the–art

This thesis is focused on the diagnosis of shunt faults in distribution grids. Shunt faults can occur between one or more conductors and the ground or just between two or three conductors. Examples of shunt faults are a broken conductor in contact with the ground, a short circuit between two lines and a contact between a tree branch and the line leading to a short circuit. From here onward the term fault will be referring to the following types of shunt faults:

- single–phase fault
- double–phase fault
- double–phase–to–ground fault
- three–phase fault
- three–phase–to–ground fault

Figure 2.1 illustrates a single phase fault. The other types of faults develop accordingly. It should be noted here that this thesis does not

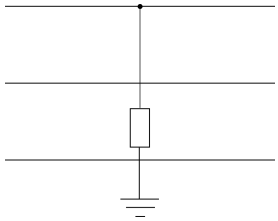

Figure 2.1: Single–phase fault

refer to arc or high impedance faults. The methods could apply to those faults as well, however, no such faults were simulated or included in the testings.

Regarding the fault diagnosis process, this consists of the detection, classification and location of a fault. The presented state–of–the–art analyzes the existing research for all the parts of the process both for the MV and the LV. In order to avoid confusion the fault diagnosis steps are defined as follows:

- Fault detection: It is the information that a fault has occurred in the grid.
- Fault classification: It is the identification of the type of the fault according to the ten types of shunt faults described above.
- Fault location: There are different kinds of fault location. The most common are: i) the calculation of the distance between the feeder and the faulted point, ii) the identification of the node closest to the faulted point, iii) the estimation of the faulted zone, iv) the identification of the faulted branch. The exact type of fault location is defined in each method.

2.1 The LV and MV grid differences

Before reviewing the available fault diagnosis literature on the MV and the LV parts of the distribution grid, the characteristics that differentiate the two should be discussed. In general, the LV grid is characterized

by a more complex structure compared to the MV one [29]. The multiple laterals and the unbalanced loads and RES that can either be connected with a single–phase or a three–phase conductor create an asymmetry that does not exist in the MV grid. This results also in the appearance of negative and zero components during the normal operation as well. Additionally, the conductors in the LV grids can either be overhead or underground and they are highly resistive, thus they should be simulated and treated differently. Finally, the effecy of the consumers' transformation to prosumers has a much more significant effect to the operation of the LV grid.

All these factors pose serious challenges to the application of fault diagnosis methods at the LV grid and render the development of fault diagnosis methods adapted to the characteristics of this part of the grid imperative. More specifically, the differences in the utilized conductors affect the voltage and current values during the fault, which in turn affect the accuracy of the methods. Moreover, the different size and operation mode of the RES, the EVs and other smart grid devices such as the batteries command the utilization of techniques with specific attributes. E.g. the stochasticity of the EV charging and the operation of residential RES deriving from the diversity of the human activity, requires the development of highly generalizable fault diagnosis methods.

In Fig. 2.2 it can be observed that so far the bulk of the fault location literature refers to the MV grid. Thus, it only addresses the particularities of the MV and not those of the LV part. In the recent years there has been an increasing trend in the study of the LV grid and specifically of active LV grids as it can be seen in Fig. 2.3. Nevertheless, there are still significant research gaps in the field of fault diagnosis in LV grids. These will become clearer in the literature review presented in sections 2.5 and 2.6.

2.2 Fault detection in MV grids

Typically the detection of faults in MV grids is performed by over–current relays. The broad integration of RES in the grid, however, has posed a serious challenge to the traditional protective mechanisms [30].

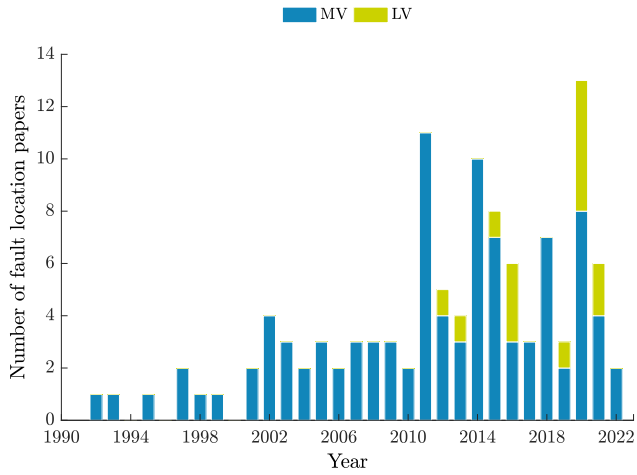

Figure 2.2: Fault location methods published each year for MV and for LV grids [1].

On one hand the bidirectional power flow can mislead the protection devices and on the other hand the low inertia of RES can lead to low and thus undetectable fault currents.

Nevertheless, the Distributed System Operators (DSOs) have not yet reported any significant difficulties in detecting faults, thus there is very limited research on the detection of faults in MV grids. The existing literature mainly refers to the new devices that could be installed on the grid, new strategies for the coordination and setup of the existing devices and a combination of the two. More specifically, [31] focuses on the communication between the existing protection devices and the analysis of the traditional circuit breaker tripping and subsequent re–closing process, without, however, proposing any significant innovation. Then, in [32] the installation and utilization of alternative measuring devices is proposed for the detection of faults. The method is offering a new perspective on the traditional fault detection methods, however, it assumes an important investment.

Another method that is based on specialized equipment is the one extracting high frequency signatures from the recorded voltage and current signals. These signatures could be used among others for the

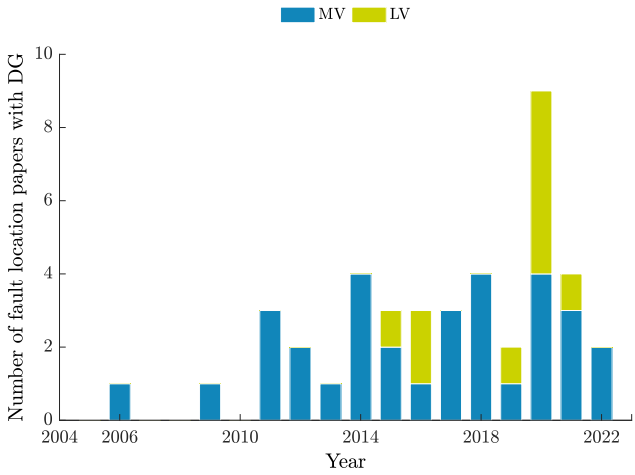

Figure 2.3: Fault location methods published each year for active MV and active LV grids [1].

detection of faults. Emphasis is given also in the different signal characteristics between the overhead and underground lines. Finally, a specific type of MV grids, a MV microgrid, is studied in [33]. The method presented in the paper is based on the injection of harmonic current by the inverter and the subsequent analysis of the harmonic circuit. The latter is less affected by the fault resistance, thus its analysis leads to more accurate results. The method, however, is tested only for three–phase faults and a maximum resistance of 2 Ω.

2.3 Fault classification in MV grids

The detection of a fault is usually followed by the classification of the fault's type. Even though this is not a mandatory step in the fault diagnosis process and it can often be omitted, it is a prerequisite for the implementation of certain fault location methods such as the impedance–based ones. As it can be seen in Fig. 2.4, even though the number of published research papers on the topic is higher than that of the fault detection it is still rather low. Most past research was focused on the HV grid [34–38]. Nevertheless, the distribution grid presents signifi-

cantly different characteristics in relation to their topology and the type of lines, loads, protection systems, measuring devices, RES and more. Therefore, the development of specialized fault classification methods for the distribution grid is necessary.

Figure 2.5 presents the most popular techniques in the field. These can be broadly categorized into AI–based methods, hybrid methods and miscellaneous methods. Figure 2.4 shows a clear trend towards the employment of AI algorithms for the classification of faults in the recent years. Nevertheless, the overall number of papers utilizing each methodology is comparable. The available literature corresponding to each methodology is analyzed in the following sections.

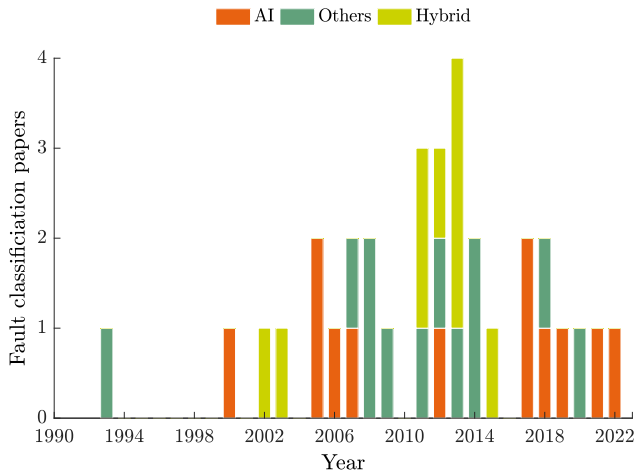

Figure 2.4: Number and methodology of the yearly published fault classification papers for MV grids [1]

2.3.1 Artificial intelligence–based methods

The use of AI techniques in all scientific fields has skyrocketed over the last few years and the fault diagnosis field is no exception. The most frequently used AI techniques in fault classification methods are the Artificial Neural Networks (ANNs), the Support Vector Machine (SVM) and the Fuzzy Logic (FL).

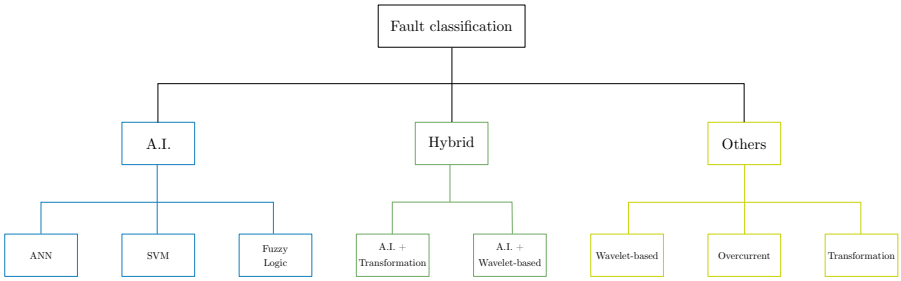

Figure 2.5: Overview of the most used fault classification methods for MV grids [1]

Literature overview

The ANNs have been used for the development of methods such as the [17, 39, 40]. In [39] the ANN is trained using as inputs the post–fault voltage and current measured at one point in the grid. A different approach is implemented in [40], where the authors utilize the ANN to identify the thresholds of the positive– and zero–sequence current that define the type of the occurring fault. This is an evolution of the, more traditional, observation of the network variables during a fault event for the establishment of current thresholds. Furthermore, in [17] the collected features for the training of the ANN are the voltages during the fault measured by micro Phasor Measurement Units (μPMUs) or smart meters placed at the end of the lines/branches for increased accuracy [23].

In a variation of ANN, [41] utilizes artificial immunological systems in a process that combines the fault type identification with the calculation of the fault resistance and the location of the fault.

Then, in [24] and [14] a SVM is used for the classification of the fault type. In the first case, the SVM serves multiple purposes as it is also used for the configuration of the Source Short–Circuit (SSC) level. In the second one, the use of a radial basis function kernel is studied for the maximization of the SVM's efficiency. On the other hand, in [42], a SVM is compared with an ANN with respect to its ability to classify disturbances on the grid including voltage sags. The SVM proved more accurate, however only two– and three–phase faults

were simulated. Finally, in [43] the SVM is combined with the K–Nearest Neighbors (KNN) and a Random Forest (RF) in a stacked architecture for optimum results.

Regarding FL–based methods, [44] proposes a technique that measures the three–phase current in the substation, converts the magnitudes and angles to fuzzy variables, and then compares them with the set fuzzy current values that characterize each type of fault. Reference [45] studies the subject of fault classification from the point of substations. The process followed is the same as before; the faults classification process consists of the transformation of the crisp current and voltage values into fuzzy ones and the subsequent evaluation of their magnitude. Additionally, [46] is also based on the variables' fuzzification and takes advantage of it by training five Learning Algorithm for Multivariable Data Analysis (LAMDA) nets, one for each type of fault, for the classification of faults.

Finally, in [47] a new method based on usupervised learning is proposed. The method aims to identify the different types of faults using mostly unlabeled data. Thus, it offers a practical solution for real datasets.

2.3.2 Hybrid methods

The majority of hybrid methods combine an AI model, often one of the aforementioned, with a signal transformation technique that is used for the feature extraction. The most frequent combination is that of a Wavelet Tranform (WT) with an ANN.

Literature overview

References falling under this category include [48–56]. In [48] the collected current at the feeder is processed with the Hilbert–Huang Transformation (HHT) in order to extract a signature component based on its frequency. Then, an ANN is trained using each fault type's encoding as a comparative measure. The same logic is applied also in the methods proposed in [49, 50]. The first one combines a Discrete Wavelet Transform (DWT) with an ANN, while the second [50] combines a DWT with FL. In [51], in addition to the WT and the FL,

an ANN is used not only for the classification of shunt faults, but also for the distinction between the different categories of faults (e.g. open circuits, high impedance etc.). Finally, in [52] the ANN is trained based on the fault identifiers produced with the use of the WT and the Fast Fourier Transformation (FFT) of the zero–sequence voltage and the three–phase fault current. Moreover, three Adaptive Neuro–Fuzzy Inference Systems (ANFIS) are deployed with fuzzy variables used as fault identifiers.

Other hybrid methods include [53] and [54] in which the Clarke–Concordia transformation is employed for the processing of the phase currents and the extraction of the eigenvectors. These are then compared with the fault patterns deriving from the theoretical basis of eigenvectors. Even though in [54] the sensitivity analysis is more thorough and there is a slight differentiation in the theoretical analysis, the two papers present the same idea. Once again, the transformed variables are used as inputs for an ANN. An identical procedure is also followed in [55]. Furthermore, [56] employs the Clarke modal transformation but only for the distinction between grounded and ungrounded faults. The phase angle shift method is used for the exact classification. The shift in each phase's voltage phasor before and after the fault is calculated and based on the degree of deviation the faulted phase(s) is/are determined.

2.3.3 Other methods

Multiple other methods have also been used for the classification of the type of faults occurring in electricity networks, including purely wavelet–based methods, over–current methods utilizing criteria and thresholds, the Park transformation etc.

Literature overview

As discussed above the wavelet transform has been used a lot as part of hybrid methods. Nevertheless, there are papers presenting only wavelet–based methods [12, 57, 58]. Reference [58] proposes the utilization of the sub–band information of the current signals in the substation, which contain fault signatures that point to the type of fault.

For the extraction of these information, the wavelet multi–resolution analysis (MRA) of the input data is employed. Taking into consideration also the changes caused in the network due to the connection of Distributed Generation (DG), [12] proposes a method based on the information exchange between relay agents spread out on the grid and their analysis through a complex wavelet transformation. The sum of the non–normalized Shannon absolute entropies of wavelet coefficients of the Clarke components and the three–phase currents are used, first, for the identification of the general type of fault, and then, for the identification of the affected phases. Moreover, in [57], after the wavelet transformation the current signal's phase energy signatures are compared with threshold values for the determination of the type of the fault. The thresholds were formed based on critical fault case simulations, nevertheless they are case–sensitive.

Another traditional fault classification approach that is based on thresholds is the over–current technique. References [6, 59–62] emphasize on the presentation of fault location techniques, however, they also include fault classification methods which utilize mainly the over–curent technique. Paper [6] is one of the early, fundamental references of this technique, which compares the change in the current's magnitude due to the fault with a set threshold. What differentiates the method in [59] is that instead of a set threshold, it uses constant analogy factors, creating in this way criteria that describe each type of fault. In the case of the other three papers [60–62] the comparison is made between the values of the current phasors after the fault and specific constant threshold values. The simulation results of the aforementioned paper, however, refer only to the fault location methods, not enabling the validation of the fault classification algorithms.

A variation of the over–current method is analyzed in [63]. In this method, the fault's type is determined with the use of the three–phase normalized current at the substation and the maximum current. Along the same path, in [13] the filtered, normalized three–phase voltage and the zero–sequence voltage are utilized for the classification of faults.

Finally, among the proposed techniques for the classification of faults in MV grids is also the use of the mathematical morphology [64].

2.3.4 Observations

After reviewing the available research on fault classification some general conclusions can be drawn regarding the existing methods. First of all, the AI methods outweigh the rest when it comes to their accuracy, however models such as the ANN require a large amount of data and significant computational power for their training. On the other hand conventional methods such as the over–current and the wavelet–based ones also have strong points and those are, in this case, their simplicity and speed respectively. Hybrid methods reflect the attempt to combine the advantages of the individual methods. They are more complex overall, but they opt for excellent results.

More specifically, considering that the dataset size is one of the main drawbacks related to the AI's application, some methods aim to tackle this problem. However, the solution proposed in fault classification methods so far only consist of the training and testing of the algorithm with a small dataset, without the employment of any advanced data processing techniques. In this case the high test accuracy is probably the result of overfitting and not of the algorithm's potential. The overfitting makes the algorithm vulnerable to the slightest changes in the grid, especially to topology changes.

Nevertheless, the AI's capability to learn complex relations between the data presents significant benefits. It constitutes a great solution to problems that have proved challenging for conventional methods, such as the classification of faults in grids with bidirectional power flow. Moreover, in the majority of cases AI–based algorithms have proven robust against grid parameters such as the fault distance, the fault impedance, the capacitive effect and the noise in the measurements. In addition they can also be used for the computation of more variables, e.g. the fault resistance and the source short–circuit level. Finally, it is the most suitable tool for the handling of real grid data.

Regarding the wavelet–based techniques, these are often used as part of a hybrid method. They offer the advantages of the high efficiency for the detection of singularities in the grids' signals [65] and the high speed. However, these propertis alone do not present any practical value. Without any further processing steps the anomaly recognition would require the evaluation of the extracted signatures from trained

personnel. Furthermore, the whole process requires the utilization of specialized equipment and it is less efficient in grids with multiple laterals such as the distribution grid.

The other traditional method, the over–current technique, bases its applicability on its simplicity. Since it depends entirely on threshold–based criteria dependent to the specific grid though, it requires re–calibration every time a major topology change occurs. In addition, its accuracy in the presence of bidirectional power flow has not been tested.

Finally, as previously commented, the hybrid methods are characterized by higher complexity compared to the rest of the methods, nonetheless, they aim also at higher accuracy. Combining, in their majority, an AI model with a data transformation technique, the hybrid methods inherit the positive characteristics of both methods while reducing their negative aspects.

2.4 Fault location in MV grids

The final part of the fault diagnosis process is the localization of the faulted branch, node and/or point of the line. This is the most challenging part and, possibly because of that, also the most studied part. Figure 2.6 presents the analyzed papers according to the year of their publication and the utilized method. Similarly to the fault classification methods, the main methodologies used for the location of faults in MV grids are the impedance–based ones, the traveling wave–based ones, the AI–based ones, the sparse and distributed measurements ones and the hybrid ones.

First of all, the need for novel fault location algorithm is depicted in the growing number of relevant publications. Moreover, it can be observed that up until 2010 impedance–based methods constituted the most used methodology. However, in the more recent years the trend has shifted towards the AI–based and subsequently the hybrid methods. The characteristics of each method and the reasons behind these research trends are analyzed in detail in the following sections.

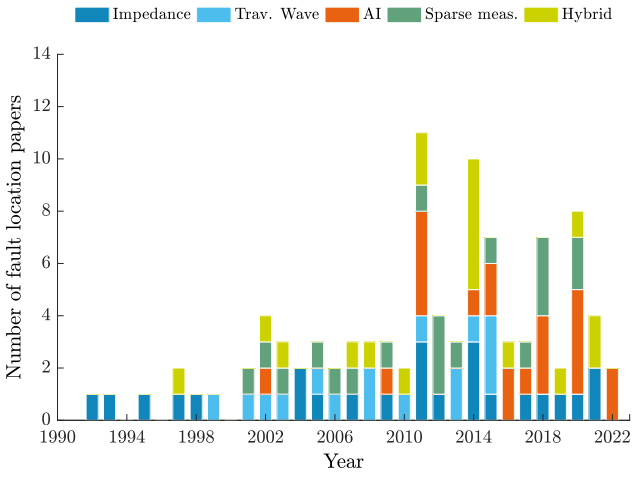

Figure 2.6: Number and methodology of the yearly published fault location papers for MV grids [1]

2.4.1 Impedance–based methods

As commented above, the impedance–based methodology is one of the first and most applied methodologies in the fault location field. This method is used to calculate the exact faulted point based on the Kirchhoff's laws. It analyzes one section of the line at a time until it locates the faulted point. The section in this case is defined as a part of the line that does not contain any laterals, as illustrated in Fig. 2.7.

The method's application begins with the assumption that the analyzed section contains the fault. It utilizes pre– and post–fault voltage and current values on both sides of the section as well as the line impedance matrix in order to form equations that calculate the overall impedance between the beginning of the section and the faulted point, and subsequently the fault's distance from the beginning of the section. If the calculated distance is bigger than the section's length then it is concluded that the fault is not located in that part of the line and the analysis continuous with the next section. The voltage and current values are either measured or estimated, depending on the measurements' availability. The minimum measurement requirement is one measurement point. Equation 2.1 is based on the aforementioned methodology

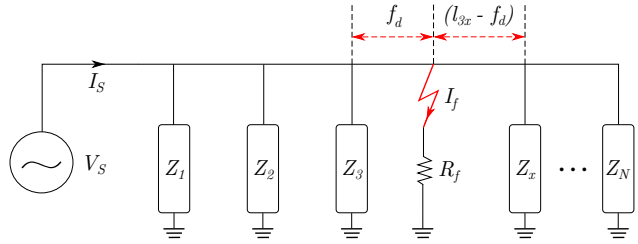

Figure 2.7: Circuit with multiple branches, containing a fault with fault current I_f and fault resistance R_f. The fault's distance from the branch preceding it is f_d and from a following bus is $(l_{3x} - f_d)$, where l_{3x} is the distance between buses 3 and x.

and is used for the calculation of the voltage at the fault point (V_F). The (V_F) is then utilized for the calculation of the faulted–point current (I_F) which is used in Eq. 2.1 for the calculation of the exact fault location.

$$\begin{bmatrix} V_{F_a} \\ V_{F_b} \\ V_{F_c} \end{bmatrix} = \begin{bmatrix} V_{S_a} \\ V_{S_b} \\ V_{S_c} \end{bmatrix} - d \cdot \begin{bmatrix} Z_{aa} & Z_{ab} & Z_{ac} \\ Z_{ba} & Z_{bb} & Z_{bc} \\ Z_{ca} & Z_{cb} & Z_{cc} \end{bmatrix} \cdot \begin{bmatrix} I_{S_a} \\ I_{S_b} \\ I_{S_c} \end{bmatrix} \tag{2.1}$$

where V_S, I_S are the voltage and the current at the local end respectively, d is the distance between the fault and the beginning of the line and Z_{xy} is the impedance between phases x, y.

An important variable in the equations is the fault resistance (R_F). The latter is a real value meaning that it does not have an imaginary part. Thus, in order to eliminate this unknown from the equations their imaginary part is separated and set equal to zero. Equation 2.2 is based on [9] and represents the fundamental equation utilized for the location of a single–phase fault occurring at phase n in a line section.

$$\begin{bmatrix} d \\ R_F \end{bmatrix} = \frac{1}{M_{1_n} I_{F_{n_i}} - M_{2_n} I_{F_{n_r}}} \cdot \begin{bmatrix} I_{F_{n_i}} & -I_{F_{n_r}} \\ -M_{2_n} & M_{1_n} \end{bmatrix} \cdot \begin{bmatrix} V_{S_{n_r}} \\ V_{S_{n_i}} \end{bmatrix} \tag{2.2}$$

where subscripts r, i refer to the real and imaginary part of the variables respectively, V_{S_n} is the voltage at the local end and I_{F_n} is the

fault current.

Regarding the variables M_{1_n}, M_{2_n}, they are defined as:

$$M_{1_n} = \sum_k (Z_{nk_r} I_{S_{k_r}} - Z_{nk_i} I_{S_{k_i}}) \tag{2.3}$$

$$M_{2_n} = \sum_k (Z_{nk_r} I_{S_{k_i}} + Z_{nk_i} I_{S_{k_r}}) \tag{2.4}$$

where k equals to one of the three phases each time, Z_{nk} is the impedance between the phases n, k and I_{S_k} is the current at the local end.

The unknown variables can either be directly calculated or they can be estimated through an iterative process. Although direct calculations are more complex, the iterative process may result in an unacceptable accumulation of errors [66]. The final result is the distance f_d between the fault and the bus preceding it. The fault's location is the f_d added to the cumulative length of the line sections between the feeder and the bus preceding the fault.

In most cases, a different set of equations corresponds to each type of fault. Therefore, as commented in section 2.3 the employment of a fault classification method is usually required before the application of an impedance–based method.

Literature overview

The foundations of most impedance–based methods for distribution systems were laid in the 1990s [6,7,67]. Based on these the first patent on both direct and iterative impedance–based methods as well as a fault classification method was established [59].

Implementing the direct circuit analysis, researchers in [68,69] use a distributed parameter line model for the development of the system's equations for a single–phase [68] and a line-to-line [69] fault respectively. Both methods use the matrix inverse lemma for the simplification of the calculations. For simplification purposes again, in [66] the modal transform is employed to decouple the impedance matrices of the circuit's phases. Hence, every phase is analyzed individually and

the direct computation of the distance is facilitated. The direct circuit analysis is also tested on non-radial grids with encouraging results [8].

On the other hand, methods described in [9, 70–73] utilize iterative solvers for the estimation of the fundamental equations' unknown parameters. In one of the early studies following this approach, the use of the symmetrical components is incorporated in the proposed fault location method [70]. Later studies omitted the use of the symmetrical components, offering more solid equations first for single–phase faults [71] and then for all the types of faults [9]. Regarding the practical side of the problem, [72] describes the field test done to a prototype device, manufactured to locate faults, that was installed in a distribution substation in Brazil. The device proved capable of identifying transient faults as well. In other iterative–based methods, in [73] the location of line–to–line and three–phase faults in ungrounded systems is discussed and in [74] a backward–forward sweep is proposed for the method's adaptation to active grids.

In [10, 62, 75, 76] the apparent power is used instead of the fault resistance for the formation of the required equations. In this way the need for the fault's classification is eliminated. The fundamentals of this approach are presented in [62]. Following that, [75] suggests an implementation of the method through the use of synchronized measurements for the fault location in grids with DG. In [76] the same problem is solved with the use of non–synchronized measurements. Then, researchers in [10] present an equation for the location of all shunt faults in AC microgrids.

Other variations of the method include the use of a wide–band frequency analysis and the Clark's transformation in a distributed parameter line mode [77]. Furthermore, the formation of fifth–order polynomial equations are applied, with emphasis given to the inclusion of the capacitive effect in the equations [78]. This effect is studied also in [79], where, for simplification purposes, the fault are only categorizes as ground or line–to–line faults, and in [80], with the method tested on an underground cable. Moreover, in another attempt to adjust the impendance–based method to active grids and achieve faster and more accurate results, the golden section technique is proposed as an analysis tool in the place of the traditional fixed step technique [81].

Finally, regarding the inputs, the majority of the references utilize

the phasors of both the voltage and the current for the location of the fault. Nevertheless, [82] proposes two separate approaches, one using the current phasor and one using only the current magnitude; the first one lead to higher accuracy. Moreover, in [83] an impedance–based method utilizing only the voltage measured in two points of the line is presented. The location of the measuring devices is not important for the application of the method, which is suitable also for grids with bidirectional power flow.

Advantages and disadvantages

The main drawback of the impedance–based methods is the multiple location estimation. It is a frequent problem in grids with laterals, especially larger grids. Another parameter that can significantly affect the method's performance is the changes in the loads during the fault interval [66]. The load changes affect the calculated currents and voltages and can lead to inaccuracies in the final result. Furthermore, methods utilizing current measurements may also be affected by the current transformer's saturation. Finally, there is very limited research regarding the application of the method in grids with bidirectional power flow.

On the other side, impedance–based methods are low-requirement, easy-to-implement methods that are based on simple physical relations. This is one of the main reasons that has lead to the continuous development of such methods and the research on ways to improve them. As it will be shown in the hybrid methods' subsection, many ancillary techniques have been deployed for the elimination of the multiple estimations' problem. Further optimization efforts include the utilization of transformations that speed up the process, the reduction of the equations and the input variable and the use of non–synchronized measurements. Another advantage of the impedance–based methods is that each section is analyzed individually, thus the methods take into account the heterogeneity of the line in the calculations. Finally, the method also allows the calculation of the fault resistance.

2.4.2 Traveling–wave–based methods

As commented in the previous sections, a popular method in the whole of the fault diagnosis field is the traveling–wave–based one. This method is based on the study of the high frequency wave that appears after a fault occurrence. The wave travels towards the ends of the line and the time that it takes for it to arrive at each end depends on its location. Since the wave's speed is known and the arrival times can be measured, the distance f_d from the beginning of the line can be calculated using eq. 2.5,

$$f_d = \frac{l - v(t_A - t_B)}{2} \tag{2.5}$$

where l is the length of the line, t_A, t_B the time that it takes for the wave to arrive at the two ends and v the velocity of the wave.

More recent methods rely on the reflection of the traveling wave at the faulted point after it is reflected at one of the line's endings. In this way it is not required to measure the signal at both sides of the line. The wave is reflected in the extremities and the time that it takes to travel from one end to the place of the fault and back can lead to the estimation of its location by using eq. 2.6,

$$f_d = \frac{v(t_2 - t_1)}{2} \tag{2.6}$$

where f_d is the distance from one end of the line, v the wave's velocity and t_1, t_2 the time that it takes for the wave to travel to the faulted point and back respectively. The moment the wave leaves the end of the line is considered as the inception time. The term "end of the line" in this case can also refer to a measurement point and not the actual ending point of the line. An illustrative example of this approach is provided in Fig. 2.8.

Literature overview

Most traveling–wave–based methods include as a first step the modal transform. In [58, 84–86] the wavelet transform is used for the extraction of the signals' unique signatures. The signals are recorded at the

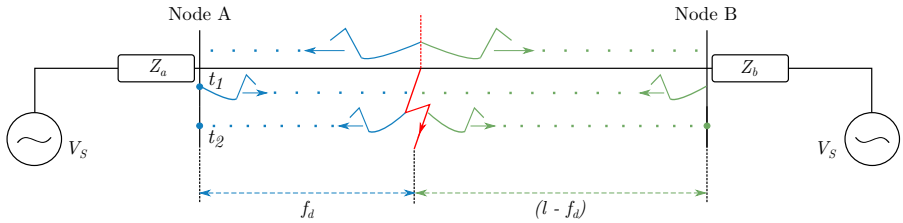

Figure 2.8: Traveling–wave method representation.

substation. These signatures are compared to those stored in the respective simulation–generated database. The quality and subsequently the usefulness of the fault signatures though is affected by the amount of junctions existing between the fault location and the substation; the more junctions the worse the recorded data. Thus, various studies examine the application and adaptation of wavelet–based methods to grids with multiple laterals. In [84] only the faulted section is identified and another method is required for the calculation of the exact distance to the fault. In [85], which uses only voltage measurements, the frequency–distance curve of the faulted lateral is used for the exact location of the fault. Moreover, [58] is based on the distinctive characteristics of the third level output of the current signal after the wavelet transform. Finally, reference [86] presents a new tool called "Time Trees" for the faster creation of the signature database. This tool, however, does not take into consideration parameters such as the attenuation and the dispersion.

Following the same modus operandi and correlating the fault signals' frequency with the possible fault paths, researchers in [87] use a traditional mother wavelet (Morlet–wavelet) as a filter. However, not all frequencies and subsequently faulted paths can be identified with this technique. Therefore, an improved version of the method that dismisses the traditional mother wavelets and constructs new ones based on the admissibility criteria of the Continuous Wavelet Transform (CWT) was later presented [88]. In another attempt to optimize the method, the faulted section identification was complemented by a criterion based on the time span between the points of the signal coefficients' maximum locals [11].

Further literature on methods utilized single–end measurements includes [89, 90]. In [89] an ANN is used for the estimation of the wave's zero mode velocity, which is then used along with the aerial mode component for the calculation of the fault distance through eq. 2.7,

$$X_C = \frac{v_1 \cdot v_0(T_{C2} - T_{C1})}{v_1 - v_0} \tag{2.7}$$

where, X_C is the estimated fault distance, v_1 and v_0 are the velocities of the aerial and zero mode components respectively, and T_{C1}, T_{C2} are the arrival times again of the aerial and zero mode components. Reference [90] proposes a method specifically designed for unearthed compensated systems. According to this method, the problem of the multiple reflected signals faced in one–end measurement systems can be overcome by tracking the arrival times of the aerial mode one and utilizing eq. 2.6.

Some research has also been done on the comparison of single– and double–end methods [91, 92]. The first paper has certain problematic points as it presents methods that are based on techniques developed for transmission networks. Moreover, no conclusions can be drawn, as the test results presented are insufficient. The second paper constitutes a more solid contribution as a field test is conducted for the verification of the theory. The test shows that a combination of the methods leads to optimum results.

Furtheremore, there are techniques that use multiple devices placed on the grid. In [93] signal recording devices for digital wavelet transform (DWT) are placed, apart from the substation, also at the load terminals. The time of the wave's arrival is recorded and after reviewing the topology of the network, the location of the fault is estimated. This method requires synchronization of all the fault transient detectors with a GPS clock. The same technique and requirements but with a different mathematical analysis are also presented in [94]. Then, El-Zonkoly [12] also employs distributed measurements but in the form of relay agents that sum the absolute entropies of the wavelet coefficients of the measured current's Clarke transform and form criteria according to its value.

Additionally, DWT and CWT can be also combined [13]. In the first step the DWT is utilized for the identification of the faulted section.

Then, if the fault is located between two measuring devices, DWT is, again, used for the location of the exact faulted point, otherwise CWT is employed.

Finally, [95] proposes the exploitation of more properties of the Electromagnetic Transients (EMTR) generated by the fault. According to it, through the back injection of the recorded signal after having been reversed in time, the fault point can be detected by a process of trial and error of possible fault locations. The variables that are randomly chosen and changed each time are the fault location and impedance.

Advantages and disadvantages

Traveling–wave can be categorized as a rather complex yet fast method, that in most cases is independent of the network parameters, but requires specific equipment able to record the transient waves of voltage and/or current and, usually, their arrival times. This last factor leads to a significant increase in the method's cost. Apart from that, the techniques utilizing current measurements have the drawback of an accuracy drop when the DC component is high and the current transformer saturated. In single–ended methods, the isolation of the wave of interest can be obstructed by unwanted reflections and noise added by other components and junctions of the network. Furthermore, methods like EMTR work only when the topology of the system remains the same during the studied transient phenomenon. Moreover, most traveling–wave–based techniques refer only to single–phase faults, since they are the most frequently encountered.

Nevertheless, single–ended methods are widely developed mainly because they do not depend on any kind of communication and synchronization between devices. Additionally, EMTR methods lead to a minimization of the measurement points and at the same time they are a good fit for complex heterogeneous systems. Finally, between DWT and CWT, which are the most commonly used transforms, CWT has the advantage of the more detailed analysis of the spectrum of energy of the recorded transients [87].

2.4.3 Artificial intelligence–based methods

AI has found vast application also in the location of faults in MV grids. As illustrated in Fig. 2.6 it has been turned into the most used method in the recent years. Unlike the fault classification AI–based methods though, in this case regression models are used for the prediction of the faulted point as the target value is a numerical one. Nevertheless, some of the most commonly used models are also in this case the ANN [15, 17, 39, 40, 60, 61, 63, 96–98], the SVM [14] and the FL [19, 99, 100]. Other utilized AI solutions are the genetic algorithms [101,102] and the tree–based models [103, 104].

Literature overview

So far ANNs have been the most utilized AI–based fault location method as they were among the first models to be studied and optimized by the developers. They have strong pattern identification capabilities thus they are highly accurate when trained properly. An ANN comprises various layers which contain neurons with different weights. There are different techniques for the calculation of weights, however, their common goal is to replicate as accurately as possible the relations between the input features and the target value. given to the An illustrative example of the ANN concept is presented in Fig. 2.9.

Most methods applying ANNs have the same structure with the main differences between them laying on the selection of certain parameters of the algorithm such as: the selected features, the pre-processing of the data, the training method, the activation functions, the number of neurons and the number of layers. Regarding the last point, all but two papers [61,97] use just one hidden layer, which is the simplest solution. Then, regarding the activation function, the most popular ones are: the sigmoid [17,61,96], the gaussian radial basis [15] and the hyperbolic tangent [63]. Furthermore, for the training of the ANNs, two algorithms have been mainly used: the Levenberg–Marquardt [40, 60, 61, 63] and the back propagation [17,97]. Finally, as far as the utilized features are concerned, the three most frequent types of inputs are: a) those based on current, b) those based on voltage and c) those that use both or more.

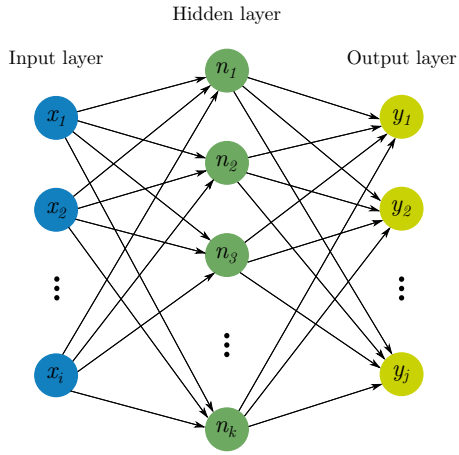

Figure 2.9: Neural network diagram.

In a more traditional approach, [61] considers just the fundamental frequency current phasors at the substation level. An extension of this approach is found in [15] which utilizes both the three–phase currents at the feeder and those measured at the DG points as input for the ANN. Furthermore, in [63] the only utilized feature is the ratio of the grid's short–circuit current to the short–circuit current of the DG. On the other hand, in [17] the use of just voltage measurements from smart meters placed along the grid is proposed. Most ANN-based methods, however, use at least both current and voltage measurements. Some of these methods try to improve the quality of the measurements with the use of Butterworth filters that remove the high-frequency components [39]. Others propose the addition of more features for the increase of the accuracy. In [40, 60] the pre–fault and post–fault three–phase current, voltage and active power phasors measured at the substation level are used. Then, the largest set of inputs (17 in total) is proposed by [97] including voltage under a short–circuit, active and reactive power under healthy and faulty operation at the feeder level as well as circuit breaker and motor status. This implies a significantly bigger training dataset which can lead to higher computational times and does not guarantee an increase in the accuracy.

An alternative approach is proposed in [96] and considers as input

to the ANN the status of circuit breakers and relays. Then, in [105] a Deep Graph Convolutional Network was employed, with the voltage and current phasors utilized as features. Finally, a variation of the ANN, the stacked auto encoder, is applied in [106] for the location of single–phase faults. First the faulted section is identified with the use of the Pearson's correlation on the zero sequence current and then the auto encoder locates the exact faulted point by spotting the changes in the voltage and current waveforms. The method is accurate, however, it requires the installation of μPMUs.

Another popular AI model is the SVM. SVMs create hyperplanes based on the data characteristics in order to split them into homogeneous groups. Therefore, it can also be used as a pre-processing step. In [14], for example, it is used for the simplification of the relation between the features and the target value. The outcome is then used to train an ANN. For this reason, SVM models have been mostly used in hybrid methods; these will be analyzed in a following section.

The third subcategory of AI techniques is the FL. FL translates the measured values into a degree of possibility between 0 and 1, making it possible to identify the parts of the grid that show some abnormality. Such an application is proposed by [99] where fuzzy reasoning identifies the faulty bus by comparing the measured voltage sag patterns with the saved fault patterns from different buses. Another application is proposed by [19] where Fuzzy–c clustering is implemented to determine possible fault points. Finally, the combination of FL with an ANN forms the ANFIS. With this tool the current signals can be used to locate the faulty zone of the grid [100].

Apart from these three algorithms that have been used the most until now, there multiple other AI models that have also been employed in fault location applications for MV grids; among them the genetic algorithms which are basically optimization techniques capable to overcome limitations of conventional methods in their search for a global minimum point [101, 102]. Another family of models that is growing in popularity are the tree–based ones. Tree–based techniques such as the RF [103,104] combine versatility with low variance and are able to perform well with a variety of data types. In the case of [104], the method is not only able to locate the faulted point but also to predict the duration of the fault. Nevertheless, the overall accuracy is not particularly

high and the method requires real–time data streaming which is not feasible with the current measuring devices.

Furthermore, not only ANN have been inspired by natural systems. Models such as the artificial immunological system have been used to solve pattern recognition and optimization problems. Their application is similar to those of the ANN. In [41] the model estimates the fault location utilizing the three–phase voltage at the substation level and the DGs. Additionally, a method based on the calculation of the variation of the spatio–temporal correlation of the data (three–phase voltage at the substation level) is presented in [107]. Finally, in [43] multiple classification models are stacked in order to accurately locate the faulted node. More specifically, the proposed method aims to combine the benefits of a SVM, a KNN model and a RF for optimum results.

Advantages and disadvantages

As discussed before, the employment of AI models is accompanied by certain drawbacks. First of all, they usually require a large amount of data for their training. Secondly, the quality of the data can significantly impact the algorithm's performancy and reliability. Currently it is rather hard to obtain real data and most methods do perform a thorough data analysis to ensure the quality of the generated data that are used for the algorithm's training and testing. At the same time, the constant feed of the models with real data from the grid is not yet feasible and it would require infrastructural investments. Furthermore, the models' accuracy when handling unseen data is not guaranteed; e.g. topology changes could lead to a significant drop in the algorithm's accuracy.

Nevertheless, when it comes to the overall method's efficiency, knowledge–based methods outperform conventional ones. Their ability to detect non–linear patterns in the data is unique and enables the accurate and fast resolution of complex problems. Moreover, thanks to the latest technological advancements, the practical application of AI–based methods can be greatly facilitated with the use of algorithms optimizing the data management and the computational processes. Finally, the algorithm's performance with unseen data can be easily improved with the use of algorithms that increase its generalizability and smart

re-training techniques.

2.4.4 Sparse and distributed measurements methods

Due to the growing grid monitoring possibilities and the technological advancements various techniques based on the use of sparse and distributed measurements have been developed. The development of smart devices specifically has inspired a lot of methods and potential future applications [108]. This section presents an overview of the available methods.

Literature overview

The most popular method falling under this category is the one based on the voltage sags recorded by meters spread across the grid, when these are available. This method locates the fault by comparing the collected fault measurements with the data generated by the simulation of all the possible faults on the grid. The faulted point is the one presenting the smallest difference between the measured and the simulated values. An index ϵ is usually calculated for every measurement point, as shown in eq. 2.8; the highest index corresponds to the fault's location.

$$\epsilon_i = \frac{1}{\Delta V^m - \Delta V_i^c} \tag{2.8}$$

Methods relying on this techniques include: i) [20, 109], which only locate the node closest to the fault, ii) [99] with a fuzzy approach of the results, iii) [110], which considers a non–linear voltage sag profile and iv) [111], which is characterized by increased complexity compared to the other methods but locates the exact faulted point. Similarly, in [112] the maximum voltage drop is calculated and is followed by the study of the voltage profile. An important drawback of this last method, however, is that it is unable to locate faults in the first and last sectors of the grid.

Apart from the classic use of the voltage measurements there are also methods proposing a different approach for their utilization. In [18] the voltage is used for the calculation of the nodal fault currents and the

estimation of their mean values which then point to the location of the faulted bus. An improved version of this method is presented in [19]. Among the improvements are the reduction of the required input data and the method's extension to all types of shunt faults. Additionally, another extension of the methodology, this time targeting the method's adaptation to grid's with DG, is proposed in [113].

The particularities of grids with changing power flow directions are also studied in [21, 114]. In [114] two different approaches cover the cases of both radial and non–radial grids. Similarly, in [21] a new approach to fault location in all types of networks including ring and meshed ones is presented. With the utilization of Directional Fault Passage Indicators (DFPIs) and Intelligent Electronic Devices (IED), the algorithm first isolates the faulted section, and then estimates the exact fault location. The devices employ the function ANSI 21FL, and the algorithm interprets the data obtained from them. The installation of DFPIs is also proposed by [115], which aims at optimizing the current trial and error process followed by DSOs with the employment of the Markov decision process. Then, regarding microgrids with DG, a recursive least–square method is proving to be an accurate and fast solution [64].

Furthermore, there are also studies utilizing as inputs the zero–sequence values of the variables [116, 117]. In [116] the zero–sequence voltage and current are used for the construction of the fault factor and then a genetic algorithm, that lead to an accurate fault location. In [117], however, the zero–sequence voltage was excluded from the method in order to avoid the error introduced by the use of the zero–sequence voltage criterion.

Finally, a different perspective is presented in [118], where real–time state estimation is used for the location of the faulted line. The fault is treated as an added bus that absorbs the fault current. Based on that parallel scenarios that are based on the different possible topologies are run. The result that are the closest to the measurements point to the fault location. For the implementation of this method, however, every node is assumed to be monitored by a PMU.

Advantages and disadvantages

The sparse measurements techniques are taking advantage of the smart grid transition in order to provide simple and fast fault location solutions. They are accurate in grids with different layouts, e.g. radial and meshed grids, and they are not affected by the bidirectional power flow. In fact, their performance seems to improve in the presence of DG in the grid, as the inverters are a potential source of additional measurements. Most procedures take place offline and combine the fault location with the estimation of the fault resistance. Moreover, the methods utilizing state estimation are normally independent of the nature of loads/generation [118]. Finally, more recent methods have reduced requirements, including their ability to locate the fault without knowing the grid's topology [21]. Overall, the development of novel smart devices and control systems leaves plenty of room for innovation in fault location methods utilizing sparse measurements.

The biggest drawback of the sparse measurement methods is the investment cost that is usually implied for their application. For the minimization of this cost an optimization of the meter placement should be included in the respective studies. Moreover, they require extensive simulations of all the possible fault scenarios and large databases for the storage of the results. Additionally, in most cases, the methods based solely on sparse measurements suffer from the multiple estimation problem and are thus often combined with other methods. Then, methods like the one in [116, 117] include complex calculations, which increase the difficulty of their implementation.

2.4.5 Hybrid methods

Similarly to the fault classification methods, in an effort to minimize the drawbacks of the individual fault location methods, new hybrid techniques trying to combine their benefits have emerged. The most common combinations are, once again, an AI model with a wavelet transform and in the fault location field also an impedance–based method with a sparse measurements method.

Literature overview

There is an abundance of hybrid fault location methods combining AI models with a wavelet transform. The wavelet transform is used for the extraction of useful information from the voltage and/or current signals. The extracted information is used as features by the AI model for the prediction of the fault location. The most commonly used AI models are: i) the ANN [51, 119–122] and the ii) SVM [22].

Another less common case is the combination of AI with another data processing method such as a rule–based scheme [123] or the more modern edge–computing [124]. Furthermore, apart from the wavelet transform there are other types of transformations used in combination with an AI model. A popular one is the Clark–Concordia transformation [53–55].

Moreover, the combination of impedance–based methods with sparse measurements methods mainly aims to resolve the multiple estimation problem. These methods are usually structured as follows: the impedance–based methods are used for the calculation of the fault's distance from the feeder and the sparse measurements are used for the estimation of the voltage sags throughout the grid in order to locate the faulted branch/zone/node [23, 125–127]. Other approaches to the multiple estimation problem include the use of transient analysis [128], the combination of impedance–based methods with AI [24, 57] and the combination of voltage sag methods with AI models [129], mathematical methods [130] or state estimation [131].

Advantages and disadvantages

The main goal of hybrid methods is the improvement of the available fault location solutions. They aim to increase not only their robustness and accuracy but also their efficiency. This includes the minimization of the input data and the reduction of the required investment cost. Nevertheless, these benefits come at a cost which usually translates to more complex calculations, increased implementation times and loss of physical meaning. In general, they combine the benefits of the individual methods used but they also inherit some of their drawbacks.

2.4.6 General comparison

Due to the higher number of available methods in the fault location field it is important to evaluate their fundamental properties and see how they compare to each other before selecting a methodology to work with. The established criteria for their comparison are their accuracy, complexity, number of inputs, speed and cost. The desired method characteristics depend on the specific application and define the method selection. In Fig. 2.10 the evaluation of the methods based on the established criteria is presented. The best combination (outer layer Fig. 2.10) refers to the case with the highest accuracy, lowest complexity, smallest number of inputs, highest speed, lowest cost.

As commented before, although the hybrid methods lead to the highest accuracy, their cost and complexity are also important factors to be considered. On the other hand, the impedance–based methods lack in accuracy, nevertheless, depending on the case, this drawback can be compensated by their low requirements for input data. Overall, the AI methods could be considered the best alternative, constituting the best trade–off between a) accuracy and speed, and b) complexity, cost and number of inputs.

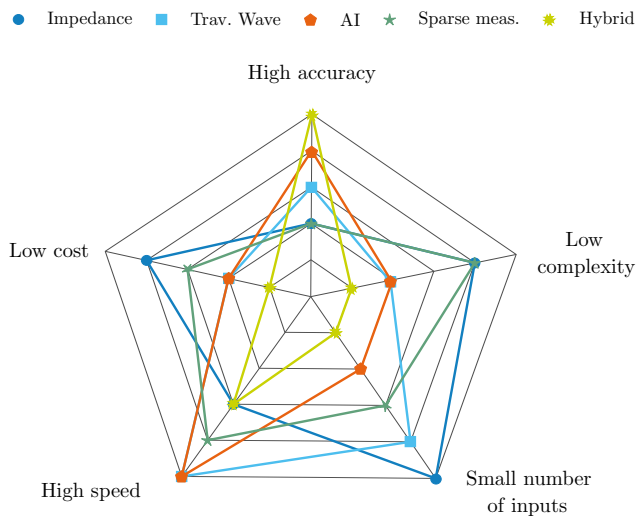

Figure 2.10: Comparative analysis of fault location methods [1].

2.5 Fault detection and classification in LV grids

The detection of faults in the LV grid is usually performed by the fuses installed on the grid. Nevertheless, the smart grid transition has created both the need and the opportunity for smarter fault detection methods.On one hand the vast integration of RES could cause phenomena such as the blinding of the protective devices. This case is thoroughly studied in [132] with the utilization of different protection devices and settings being proposed for the problem's resolution. On the other hand the appearance of new measurements sources such as the inverters or various sensors could increase the quantity and versatility of the available data, thus facilitating the accurate detection of faults.

A conventional measurement–based method is presented in [133], which studies the rms current increase and the voltage drop in relation to the fault resistance in order to set fault–detecting thresholds. Another method that utilizes thresholds is presented in [134] which is suitable for low frequency measured data collected from the secondary side of the MV/LV transformer. It uses the Kalman filter for the estimation of the grid's state and calculates the error between the estimated state and the measured one. Even though the study includes a thorough analysis of the threshold selection process, it is a grid–specific method that is easily affected by the grid's parameters, e.g. the loads. Thresholds are also used in [135,136], where the measured values are collected and analyzed either by advanced sensors or the distributed transformer controller. Moreover, more recent fault detection methods take advantage of the increased data availability by applying AI algorithms. More specifically, a deep learning [137] and a gradient boosting [28] algorithm have been used for the detection of faults in LV grids.

Finally, various fault detection methods have also been developed for LV microgrids. There are both threshold–base approaches [138, 139] and more advanced ones [140, 141]. The method proposed in [138] is distinctive from the other threshold–base methods since the value that is compared to set thresholds is the resistance in various points of the grid and not the current or the voltage. Nevertheless, it is also characterized by high uncertainty and grid dependency. Another modified threshold–base approach is presented in [139], which monitors the tran-

sient inverter current in certain time windows. The recorded current is then compared with its estimated fundamental frequency value. If their difference exceeds a set thresholds then it indicates a fault occurrence. Even though this is a more elaborate method, apart from the typical drawbacks of threshold–based methods, it also includes the problem that in order for the fault to be detected the monitoring window needs to contain normal operation data as well.

Regarding the more advanced methods, these are both hybrid methods. In [140] a wavelet transform is applied to the collected signals and the energy signatures extracted are used as inputs to an eXtreme Gradient Boosting (XGBoost) classifier. Similarly, in [141] a wavelet transform is combined with a Deep Neural Network (DNN). Both methods require the usage of specialized equipment and trade-off complexity for accuracy. It should be noted, though, that in [141] some test results show a higher accuracy than the corresponding training results, which causes concerns for the paper's credibility.

The fault classification methods in LV grids are even more limited than the fault detection ones. These are mainly inspired by the corresponding methods developed for the MV. Specifically, in [26] a threshold–based method is used, that compares the zero–sequence current phasor with a threshold set as the maximum possible current amplitude during normal operation multiplied with a safety factor. Then, in [142] Park's transform is used for the analysis of the voltage's zero–sequence component. Moreover, another conventional method is presented in [133] which is based on the monitoring of the per phase current increase. Its reliability, however, is limited to low–impedance faults (lower than 10 Ω). Finally, an AI–based method utilizing gradient boosting trees is proposed in [28]. This is the most advanced of the available techniques and is characterized by the highest accuracy and reliability.

2.6 Fault location in LV grids

The part of the fault diagnosis process for LV grids that has been the most studied is, once more, the fault location. Similarly to the method categorization presented in section 2.4, most methods developed for the

LV grid belong to one of the following families of methods: i) traveling–wave–based methods, ii) sparse measurements methods, iii) AI–based methods and iv) hybrid methods.

Regarding the traveling–wave–based methods, these vary slightly from the ones proposed for the MV grid. The use of underground cables inspired the application of Time Domain Reflectometry (TDR) [143,144]. TDR's results are considered reliable, however, the method's implementation requires specific hardware and trained personnel. Over the years the use of wavelet–based methods has advanced, in parallel with the methods developed of the MV grids, and it includes the use of test signals [145] and of the Park's transformation [142] for the location faults.

The infrastructure improvements in the LV grid resulting in increased observability over the grid have also prompted the use of advanced measurement–based techniques for the location of faults in this part of the grid as well. Hence, also in the case of LV grids, an increase was observed in fault location methods relying on smart devices. Among the proposed smart devices to be used for the data recording required are fault indicating devices [27] and local sensors [135, 136]. On the other hand, there are techniques that use conventional well-known devices such as the PMUs, but propose either the increase of their installation points [26] or the use of alternative measurements such as the negative–sequence–voltage [146].

Furthermore, the utilization of sparse measurements for the improvement of the impedance–based methods' accuracy is also observed in fault location techniques developed for the LV grid. In [147] novel devices operating in very high frequencies are employed for the collection of data to be used in the location of arc faults. Then, in [23] the obtained voltage measurements are compared with thresholds that once again require adjustment based on the specific application. In an effort to counterbalance this negative characteristic and increase the flexibility of the method, a zonal division of the grid was proposed.

Moreover, the growing data availability in the LV grid has also lead to an increasing interest in the development of AI–based methods. Over the past decade, AI has established itself as a powerful computational tool and has found multiple applications in various sectors including in the power systems [148, 149]. Regarding the location of faults in LV

grids different models have been trained and tested for the location of either the faulted node or the exact faulted point. In [150] the statistical basis of machine learning (ML) was put in the forefront and used to locate the exact point of the fault. However, the overall accuracy was not particularly high and only three phase faults with a maximum fault resistance of 1 Ω were considered in the study. The method used in [28] employs a gradient boosting tree model for the detection and identification of faults as well as the location of the faulted branch for single phase and three phase faults. The utilized model is very efficient and leads to very high accuracy, however, the study focuses only on two types of faults and, more importantly, there is no location of the exact fault point. Finally, in [137] a deep learning algorithm is presented that interpolates the input measurements in order to become independent of the number of measurement points. The method achieves a mean accuracy of more than 88.2% in the prediction of the fault's distance from the feeder for a wide range of fault resistances. Nevertheless, the training of deep learning models is highly demanding both for the machine and for the developer as it contains more complex processes than other AI models and a large number of parameters that need to be tuned. Moreover, the main characteristic of deep learning is the requirement of a large amount of data, which is one of the scientific community's major concerns regarding its practical application.

2.7 Observations

The advantages and disadvantages of the fault diagnosis methodologies used in LV grids are similar to those of the MV grid. The traveling–wave–based methods are fast, however, they are expensive and hard to implement. At the same time, their accuracy decreases significantly in grids with multiple laterals and junction points. Therefore, their practical value is low for LV grids. On the other hand, sparse measurement methods are more practical and robust compared to the travelling–wave–based ones, nevertheless, they also present important drawbacks, such as the need for infrastructure investments, the fact that they are grid–specific and the lack of thorough performance validation in the existing research. Then, hybrid methods are usually more accurate,

however, they are also more complex and carry some of the combined methods' negative characteristics. This last part is especially noticeable when threshold–based methods are used. These are very vulnerable to any topology changes and are very grid–specific. Finally, AI–based methods also require the collection of large datasets, which implies potential infrastructure investments, and they usually lack generalizability. Furthermore, they can be computationally expensive and be treated as black–box methodologies. However, there are multiple techniques that can minimize AI's negative characteristics and highlight its unmatched ability to recognize complex patters in the data and its robustness against potential accuracy-decreasing parameters.

2.8 Research gaps

Based on the literature review on the fault diagnosis methods developed for the distribution grid it can be concluded that the ongoing energy transition has posed a great challenge to the traditional methods and has created the need and opportunity for novel techniques. The vast integration of RES, ESS and EVs, among others, together with the increasing flexibility of the electricity markets and the expanding role of prosumers have altered remarkably the topology of the distribution grid. At the same time the new smart measuring devices and the advanced communication systems allow for the automatization of the fault diagnosis process and the development of faster and more accurate methods. Especially in the LV grid, the available research is very limited and the ongoing changes are very radical, thus the need for smarter fault diagnosis methods is growing. The following section discusses in detail the advantages and disadvantages of the available methodologies in order to determine the most suitable methodology for the resolution of the main challenges faced in the field of fault diagnosis for active LV grids.

2.8.1 Fault location

Regarding impedance–based methods, the question remains as to how efficiently they can be adjusted to a meshed grid with bidirectional

power flow or a grid in islanded mode. Some hybrid methods that incorporate the impedance–based technique are able to tackle this matter, however, when it comes to purely impedance–based methods the number of publications on this subject is extremely low. Moreover, a common characteristic of the active grids is the frequent changes in the operational state or the topology of the grid. These can affect significantly the performance of an impedance–based fault location method, however, are not considered in most studies.

A research gap concerning the effect of the new smart grid components can be found also in the traveling–wave–based methods; from the sixteen papers studied here only three take into account the integration of RES in the simulations. Additionally, more in-depth research is required on the effect of the growing number of laterals and tapped loads in the distribution grid. These lead to multiple signal reflections which make the detection of the faulted point harder. Furthermore, in this case as well, the topology changes should be taken into consideration during the method's development, as they could lead to the misinterpretation of the recorded signals.

In the field of AI, a major talking point in relation with its practical applicability is the required data volume and the existing legislation related to the data collection and usage. Regarding the data volume and the need for large data storage systems, the development of efficient data management schemes to accompany the fault diagnosis method is imperative. Moreover, every AI–based method should specify the data collection process that needs to be followed in order to ensure its applicability. Finally, emphasis should be given on the generalizability of the AI–based methods, so that they become more robust against topology changes.

Furthermore, considering that the hybrid methods are a combination of other techniques including the aforementioned ones, most challenges related to the other methods apply to the hybrid ones as well. Due to their versatile nature, however, there is plenty of room for novelty and improvement on their efficiency. This could be easily achieved with the use of novel measuring and feature extraction techniques. Nevertheless, as the use of smart devices is intensified and their functions are multiplying, the ramifications of possible device malfunctions become more critical. Smart devices are becoming an essential part of the grid's

Operation and Management (O&M) and the impact of external signal interference or cyberattacks should be thoroughly evaluated. Thus, the effect of missing or distorted measurements needs to be considered in future studies.

Finally, the biggest challenge yet, is the large scale practical implementation of automatic fault location methods by the DSOs. Even though there is an abundance of fault location methodologies in literature, in practice most DSOs have not yet implemented an automatized process for the location of faults. The practices vary between countries and even between DSOs operating in the same country, however, the majority relies on customer calls followed by the visual inspection of the line by a technical crew in order to locate the faults in the distribution grid [151]. Another common tactic is the manual or remotely controlled test switching of the pole–switches [152]. Both of these techniques, however, are time-consuming and can be both costly and dangerous for the general public and the equipment; e.g. if the isolated part of the grid becomes energized before the fault is cleared [153]. There are also devices that have been developed in order to facilitate the detection of faults, such as the FPIs. Nevertheless, they constitute significant investments that in most cases are considered uneconomical due to the fact that they serve only a single purpose. Furthermore, even in the case where an already installed device has an integrated fault location function, this might be insufficient for a grid with multiple laterals due to the multiple estimation problem [154]. Thus, the main focus of the DSOs should be either a) the application of elaborate fault location methods that are based on simple measurements or b) the utilization of the advanced functions of new multi–purpose smart devices, such as the last gasp messages [155], that are expected to be installed on the grid in the upcoming years.

2.8.2 Fault classification

As already discussed, the amount of methods focusing on the identification of the type of faults occurring in distribution grids is quite small and allows for more research and experimentation to be conducted. The challenges concerning these methods are similar to the ones already mentioned in the previous subsection. On top of that,

however, the data concerning the accuracy of fault classification algorithms are almost non–existent. Fault classification techniques are usually part of a general fault isolation or fault diagnosis method, focusing on fault location, thus not including tests and simulation results specifically referring to the classification process; and even when they do present some results, the sample is often quite small and the outcome, although highly promising, potentially untrustworthy. To that end contributes also the fact that important parameters, such as the noise in the measurements, are frequently ignored in the simulations conducted specifically for the fault classification methods.

2.8.3 Fault detection

In line with the research gaps identified in the fault location and classification methods, further analysis is required on the effect of the bidirectional power flow on the fault detection methods. Moreover, the vast integration of FC and UFC stations on the LV grids is expected to increase significantly the grid's load and affect the current values. The current is a crucial parameter in the fault detection process and big fluctuations of the current's values could lead to high inaccuracies. Most available methods are based on conventional techniques, which are unsuitable for active grids. At the same time, the technologically advanced solutions do not take include a thorough sensitivity analysis on the emerging influencing parameters.

2.8.4 LV grids

The LV grid is a research gap in itself. As described in the previous section, only a very limited number of papers has been published on the subject, thus leaving plenty of room for the development of new methods; e.g. the efficiency of AI algorithms has not been examined in depth yet. Additionally, although the effect of RES has been considered in some studies, an analysis of the effect of batteries or the integration of EVs on fault diagnosis methods is missing. Then, the effect of parameters like the smart devices and their novel features, the data availability, the load and the line characteristics on the fault diagnosis algorithms requires also further research on the LV side.

2.9 Problem definition and selected methodology

In line with the discussed research gaps and research methodologies, this PhD research focuses on the development of practical and efficient fault diagnosis methods for active distribution grids. In the case of the fault detection and location steps of the process, the research gaps are identified in the active LV grids, therefore, the developed algorithms are tailored to the characteristics of a representative LV grid. Moreover, the effect of bidirectional power flow is taken into consideration in both algorithms and in the case of the fault detection also the effect of FC and UFC. Regarding the fault classification part, a research gap is identified in the bibliography concerning both the MV and the LV parts, hence, the developed method is applied to the whole distribution grid. Finally, the AI is selected as the main work methodology as it encapsulates the most benefits and shows the greatest potential among the available methods. Its application, however, still poses significant challenges. Therefore, this research attempts to address the main challenges related to the application of AI in fault diagnosis algorithms and to present fast, accurate and generalizable solutions.

3

Artificial intelligence algorithms

As commented in Chapter 2, the main AI–based models used in fault diagnosis algorithms are the ANN, the SVM, the FL and the tree–based models. Each model has its own characteristics and the selection of the most suitable one depends on the application at hand. The main factors that influence the model selection process are the type of the target value, the relationships between the data, the size of the dataset, the outliers/noise in the data and the system's computational power.

The target value is the parameter that defines the type of the problem. Based on the latter the problems and subsequently the AI models types are split into two major categories. In the case of categorical target values, classification models are utilized, whereas in the case of continuous target values regressions target values are used. The detection of faults constitutes a binary classification problem, since the target value is one of two possible states. The identification of the faults' type is a multi-class classification problem, since the target value is one of ten possible classes. Finally, the fault location problem constitutes a regression problem since the target value is the distance between the fault and the main feeder.

Regarding the rest of the parameters defining the model selection

process, these need to be carefully and individually evaluated for each model depending the specific application. A brief analysis of the most significant models' characteristics in relation to these parameters is presented in the next section.

3.1 AI method comparison

The ANN is the most widely used model in fault diagnosis applications so far. There are multiple variations of ANNs, nevertheless, they are all more suitable to applications where large datasets are available. They have increased computational abilities compared to simpler models such as the linear regression and are able to identify complex patterns between the data when the appropriate training method is followed, i.e. when appropriate activation functions are applied. Nevertheless, this also means that they are computationally expensive, requiring significant training time and computational power. The technological advances both in hardware (memory) and in software (optimization functions) have allowed for faster ANNs implementations and the development of more deep learning methods. Overall, however, ANNs constitute rather demanding ML models whose performance relies heavily on the fine tuning of their hyperparameters, nevertheless, they can accurately learn complex patterns.

The SVMs have also been popular in the past, however, their use is less frequent nowadays. Due to their nature they perform well in high dimensional spaces and they are memory efficient. Nevertheless, similarly to the ANNs, SVMs are characterized by high computational times. Moreover, they require an expensive corss–validation process and their accuracy is significantly affected by the noise in the data. Regarding this last factor, a more robust alternative is the FL. It constitutes a more accurate representation of the real world compared to the classical logic and usually has low hardware requirements. However, FL–based models require extensive validation and are highly dependent on the developer's expertise.

Finally, the last largely used family of models in fault diagnosis problems are the tree–based models. The advanced tree–based models such as the RF are characterized by low bias, low dependence on the hyper-

parameters' fine–tuning and low overfitting [156]. They evaluate each feature independently, thus, there is no need for data scaling. Moreover, they are effective with non–linear data and with smaller datasets. Nevertheless, they lack in extrapolation abilities and their training could be long. To that end contributes also the fact that certain tree–based models perform internal cross–validation. This, however, results also in increased predictive accuracy and eliminates the need for a validation dataset.

Taking into consideration their aforementioned characteristics, tree–based models are selected as the most suitable for the fault diagnosis problem in this study. The main problem–related factors that dictated this decision were the lack of real data, the size and type of the training dataset and the generalizability, flexibility and adaptability requirements. The properties and operation principles of tree–based models are analytically described in the following sections.

3.2 Tree–based models

The decision–making process of tree–based models is already described by their name: the models' development resembles that of a tree. As commented above, tree-based prediction models offer high flexibility and efficiency, with reduced computational demands [156] and have been proven accurate in fault location problems for the MV [103]. The simplest tree–based model is the decision tree as illustrated in Fig. 3.1. Even though it is a weak learner, i.e. its predictive capabilities are limited, it constitutes the basis of all the tree–based models. Due to the different aspects of the fault diagnosis process both classification and regression tree–based models are utilized in this study.

The tree's training process starts from a single node containing the initial dataset. This could be either the original dataset or a bootstrapped version of it. The dataset is then split based on the attribute with the highest entropy or with the use of other information gain scores in order to create subsets with the high homogeneity. Each subset corresponds to a new sub–node and each attribute is only evaluated once. The process is repeated until the set depth of the tree is reached. This is usually followed by the tree's pruning which assists the minimization

of overfitting by removing terminal nodes that do not provide useful information related to the prediction result. The selection of the attributes, the number of features considered in each split and the number of splits differ for the different tree–based models. Some of these parameters are defined during the models' hyperparameter tuning while others depend on internal processes of the model. The hyperparameter tuning is discussed in each individual application presented in the following chapters, while the analysis of the models' internal algorithms falls beyond the scope of this research.

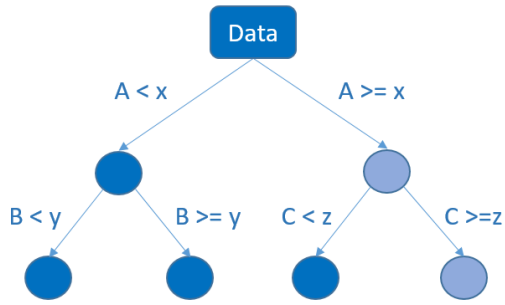

Figure 3.1: The structure of a decision tree [2].

The model's predictive power is determined with the use of the objective function, which is defined as the sum of the training loss and a regularization term:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{3.1}$$

where θ is the vector of the weights added to each feature in order to predict the target value. In tree–based models the predicted target value can be expressed as follows:

$$\widehat{y}_i = \sum_{k=i}^{K} f_k(x_i) \tag{3.2}$$

where $f_k \in F$ is a function containing the tree structure and leaf scores for all the possible trees, which form the functional space F. K is the number of trees employed.

Usually the training loss is calculated with the use of the mean squared error (MSE) and can be expressed as:

$$L(\theta) = \sum_i (y_i - \widehat{y}_i)^2 \tag{3.3}$$

Thus, the form of the objective function for a tree–based model is the following:

$$obj(\theta) = \sum_i^n (y_i - f_k(x_i))^2 + \sum_{k=1}^K \Omega(f_k) \tag{3.4}$$

The models' training goal is the optimization of the objective function, in this case of Eq. 3.4. It should be noted here that the exact form of the objective function differs according to the trained model.

3.3 Random Forest

One of the tree–based models used in this research is the RF. Both a RF classifier and a regressor are employed in the proposed methods. A RF is an ensemble of decision trees (DTs) trained in parallel with a different bootstrapped subset of the training data, as illustrated in Fig. 3.2. Each DT makes a decision based on the provided data and attributes. The decisions are then combined according to the type of model: in classification problems the technique used is major voting whereas in regression problems it is averaging. This prediction approach is called bagging and contributes to the minimization of the model's overfitting, which is the biggest drawback of weak learners such as DT. Moreover, a RF model does not require cross-validation since the algorithm applies out–of–bag error estimation throughout the forest development process [157]. Overall, the RF is characterized by lower variance, higher robustness against missing or unprocessed data, smaller precision requirements in hyperparameter tuning and faster training. On the other hand, the bootstrapping used in the training dataset and the randomly selected attributes evaluated in each split are sources of arbitrariness for the model.

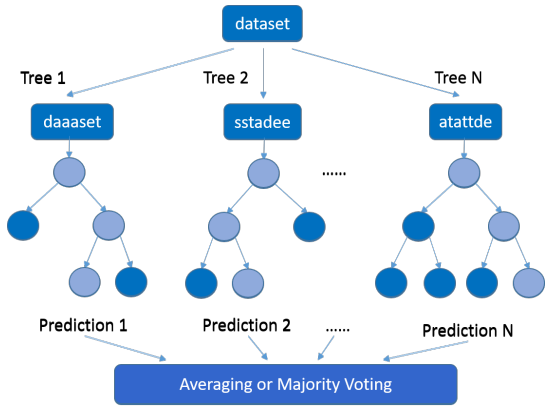

Figure 3.2: Illustration of a Random Forest prediction model.

3.4 Gradient boosting

A subcategory of tree–based models that is growing in popularity and predictive capabilities is the one based on gradient boosting. Gradient boosting also constitutes an ensemble method since it combines multiple weak learners for the prediction of the target value [158]. However, contrary to the RF here the DTs is trained sequentially. More specifically, the gradient boosting targets the cases that have been wrongly predicted, instead of focusing equally on all the examples. In order to achieve that it trains sequentially multiple DT and penalizes each tree's correct predictions by assigning them a smaller weight while assigning a bigger weight to the incorrect predictions. So new weak learners are added in order to help the stronger learner. An example of the process is illustrated in Fig. 3.3. In this research two advanced gradient boosting models have been utilized: the *CatBoostClassifier* [159] and the *XGBoost* [160].

As it is indicated by the name, the CatBoostClassifier is a classification model. In this study it is used as a binary classifier that performs fault detection in the LV grid. CatBoostClassifier has not been previously used in power system–related problems, nevertheless, its unique characteristics and exceptional performance [159, 161] suggest its suitability to the fault detection problem. Specifically, its two most impor-

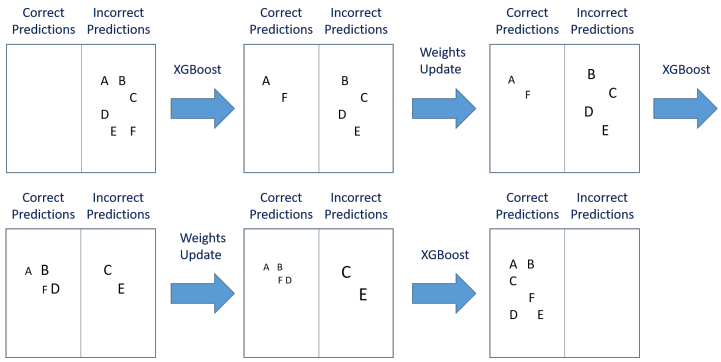

Figure 3.3: Illustration of the XGboost model's sequential training and weight assignment [2].

tant attributes, the high computational speed and the low overfitting render it perfect for the fault detection application.

On the other hand, the XGBoost model used here for the location of the exact faulted point is a regression model. Even though there are no existing applications of a regression XGBoost model in fault location problems, it has been successfully used for the classification of the faulted branch as well as a variety of other grid–related applications [162–164] and its multifarious characteristics could provide a robust solution to a complex problem such as the fault location. Among its noteworthy attributes that distinguish it from other tree–based and boosting models are: i) the performance of regularization, ii) the parallel processing and subsequent computational speed, iii) the backward pruning and iv) the optimization possibilities.

According to Eq. 3.2, the additive function of the boosting trees can be expressed by the prediction model equation as follows:

$$\begin{aligned} \widehat{y}_i^{(0)} &= 0 \\ \widehat{y}_i^{(1)} &= f_1(x_i) = \widehat{y}_i^{(0)} + f_1(x_i) \\ &\dots \\ \widehat{y}_i^{(n)} &= \sum_{k=i}^{n} \widehat{y}_i^{(n-1)} + f_n(x_i) \end{aligned} \tag{3.5}$$

where n is a step of the prediction process. The objective function of the model is formed in accordance with Eq. 3.4.

3.5 RF vs XGBoost

As stated above, both the RF regressor and the XGBoost are used in this research for the location of faults. In this way the performance of two powerful tree–based models with different training principles can be evaluated. XGBoost constitutes a better option for unbalanced datasets, since it is less biased than RF, it is more robust against overfitting in datasets with multiple similar examples due to its pruning approach, it is computationally faster and is less affected by the hyperparameters' values. On the other hand, the RF is very robust against overfitting when the training dataset is properly pre–processed and its hyperparameters are easier to tune. Overall, the XGBoost represents the latest, more advanced generation of tree–based models while RF a vastly used and validated generation of tree–based models.

3.6 AI metrics

For the evaluation of the AI model's performance certain metrics are being used. These metrics differ depending on whether the model is a classification or a regression one because of the difference in the target values' nature.

3.6.1 Regression metrics

The main metrics used for the evaluation of regression models are the mean squared error (MSE), the mean absolute error (MAE) and the R^2. The first two metrics are widely known and do not need further analysis. The coefficient of determination is the degree of predictability of the target value from the features. The highest and best possible value is 1, which corresponds to the perfect depiction of the feature–target relations by the model, hence, to the model's ability to make accurate predictions at all cases. The metrics' mathematical definitions are described by the following formulas:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i^*)^2 \tag{3.6}$$

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - Y_i^*|}{n} \tag{3.7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2} \tag{3.8}$$

3.6.2 Classification metrics

Regarding the evaluation of classification models the most important metrics are the accuracy and the F1 score. These are defined as follows:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{3.9}$$

$$F1 \quad score = 2 * \frac{precision * recall}{precision + recall} \tag{3.10}$$

Precision and recall are described by the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{3.11}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.12}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives.

Finally, a confusion matrix is also frequently used in order to present the model's predictive capacity. The matrix shows the model's number of correct predictions for each class as well as the number of incorrect predictions and which class was falsely predicted instead.

3.7 Conclusions

In this chapter a general overview of the technical characteristics of the most popular ML models used in fault diagnosis applications is presented. Based on that, the tree–based models are selected as the most suitable tool for the development of the fault detection and location solutions presented in this thesis. This decision was made after taking into consideration the specific parameters of the problem at hand, such as the dataset size and nature, the available computational power and time and the final solution's desired properties. The selection of the exact tree–based models used in each step of the fault diagnosis process was made again in accordance with the type of problem, e.g. regression or classification problem, and after reviewing the characteristics of the various models. The selected models are analyzed in the corresponding sections. Finally, the chapter is concluded with the presentation of the metrics used to evaluate the models' performance.

4

Fault detection in active low voltage grids with fast and ultra fast charging

The first part of the fault diagnosis process is the detection of the fault. As commented in chapter 2, the detection of faults in LV has been studied in the literature for both traditional and active grids. Nevertheless, no fault detection method has taken into consideration the effect of FC and UFC, therefore, this study focuses on this particular case of fault detection. Over the past few years a rapid increase in the number of EVs has been recorded. In 2020, 3.3 million EVs were circulating in the roads of Europe, a 73,7% more than 2019, despite the pandemic [165], rendering Europe the world's largest EV market for the first time. Furthermore, the recent revision of the Alternative Fuels Infrastructure Regulation's (AFIR) implementation in 2021 reinforces the predictions of an even more significant increase in the EV circulation in the coming years. This implies the multiplication of the EV charging points and an additional stress to the electricity grid.

Among the grid's parameters that can be affected by the extensive EV charging are the voltage stability and the current levels. Thus, unexpected grid behavior may be observed both under normal or faulty

operation [166]. So far, however, the existing studies analyze the effects of EV charging only on the grid's normal operation [166–169]. Moreover, the role of EVs in critical situations is analyzed mainly in relation to their effectiveness as voltage regulators [170–172]. Hence, as commented in chapter 2, EVs constitute another stochastic variable that needs to be taken into consideration during the study of development of fault diagnosis solutions as well, especially when it comes to the detection of faults in LV grids. To the author's knowledge there is no existing literature on the topic.

4.1 EV charging stochasticity and fault detection

Even though it is thought by many that the effect of EVs on the grid is similar to that of the RES and the ESS, the size, charging phases, location, use and operation of EVs distinguish them from the other grid elements. Their unique characteristics make the simulation and evaluation of their effect on the grid's operation and, more specifically, on the fault detection methods much more challenging. Probably the biggest challenge is the stochasticity stemming from the unpredictable behaviour of the EV drivers. There are various load predicting algorithms, however, the size and technology of EV chargers as well as the nature of the EV as a means of transportation make the development of accurate fault detection methods much more complex. Moreover, the significant load differences between the quickly spreading FC (50 kW), UFC (150 kW) and the slow charging [173, 174] could easily lead to the false identification of charging events as faults.

Figures 4.1 and 4.2 illustrate the pre– and post–fault current for a phase–to–ground fault occurring in the benchmark presented in section 4.3. In the first case all the EV FC and UFC stations are simultaneous occupied and charging at nominal power. In the second case all the EV charging stations are free. It can be observed that in the first case both the pre– and post–fault currents are much higher. This difference is expected to have a direct impact on the accuracy of the fault detection methods.

The methods that are expected to be affected the most by the charging stochasticity are the conventional threshold–based ones. AI–based

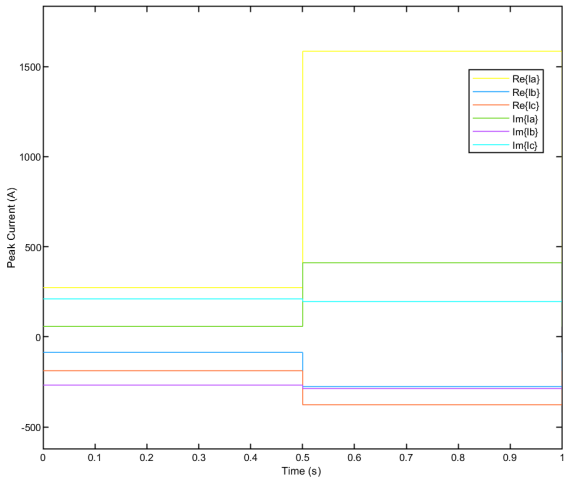

Figure 4.1: Pre– and post–fault current in the case of an a–phase–to–ground fault in a grid with full EV charging.

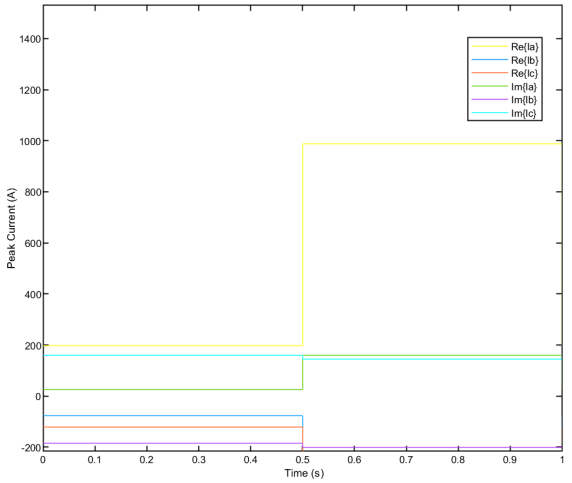

Figure 4.2: Pre– and post–fault current in the case of an a–phase–to–ground fault in a grid with no EV charging.

methods may prove more robust with the proper training, however, this would require the examination of numerous normal and faulty operation scenarios, and the collection of especially large datasets. Apart from time consuming and computationally expensive, this approach would also increase the chances of overfitting the algorithm to the data. This would make the algorithm vulnerable to even the slightest topology changes, since the model would be fitted so perfectly to the training data that it would not be able to make accurate predictions for new unseen cases.

In an attempt to address the aforementioned challenges and take advantage of AI's benefits, a novel ML–based fault detection algorithm was developed. The algorithm's target is twofold; it aims to cope with the lack of real data and also to provide accurate results despite the changes in the grid's operation caused by unpredictable big loads such as the EV FC and UFC. In order to achieve these goals the use of static simulation data is proposed. In this way, the algorithm becomes generalizable and independent of the grid's elements' state. Additionally, the algorithm's application is facilitated, as the required simulation parameters for the training of the algorithm are known to any system operator. In the following sections the optimum way to accurately simulate an active LV grid containing EV charging points without requiring any knowledge of the intermediate charging states is presented.

4.2 Fault detection method description

The proposed fault detection algorithm falls under the category of ML–based methods. Contrary to the other available methods, however, emphasis is also placed on the selection of the training data and not only on the model's characteristics. In order to ensure the algorithm's effectiveness in grids with multiple FC and UFC stations and its robustness against the EVs' stochasticity without requiring constant retraining or the collection of enormous datasets, the training data are selected to be static simulation data. This means that no intermediate loading states are included in the training data. The static states that are considered in the case of the big, unpredictable loads such as the EV chargers are the zero loading state and the full loading state. In the first case all the

EV chargers are considered to be disconnected while in the second case all the EV chargers are occupied and charging at their nominal value. The algorithm's performance when trained with data generated from the simulation of each of the two cases is studied. The rest of the loads are simulated with their nominal values. The only variable elements are the PV penetration level – only 5 specific penetration levels are simulated though – and the fault characteristics.

The features that are utilized by the algorithm are the three-phase voltage and current phasors. These are provided by various devices already installed in the grid, such as the PMUs, and can be easily acquired from any grid simulation software. Once the ML model is trained and saved, the algorithm can be used at any moment for the detection of faults with the input of real data collected from the grid. The algorithm's detection speed depends on the measuring devices' data transmission rate. After receiving the data, the algorithm instantly makes a decision regarding the existence of a fault. The method's training and implementation phases are illustrated in Fig. 4.3. The presented method could be applied either in a digital twin, or directly by a DSO that supports the constant feed of basic grid data, such as the voltage and current phasors at the LV substations to the model.

4.2.1 The ML model

Fault detection constitutes a binary classification problem since there are only two possible outcomes of the process, the detection of a fault or the absence of faults. Hence, a binary classification ML model is used for the realization of the predictions. More specifically, the selected model is the *CatBoostClassifier* [159]. The CatBoostClassifier is a gradient boosting tree-based model; its basic principles are explained in chapter 3. The model's main benefits over other boosting and tree–based models are its higher accuracy [159,161] and computational speed as well as the fact that it is characterized by decreased bias and overfitting. The latter is especially useful in the analyzed application and it is one of the main reasons that led to the selection of the specific predictive model.

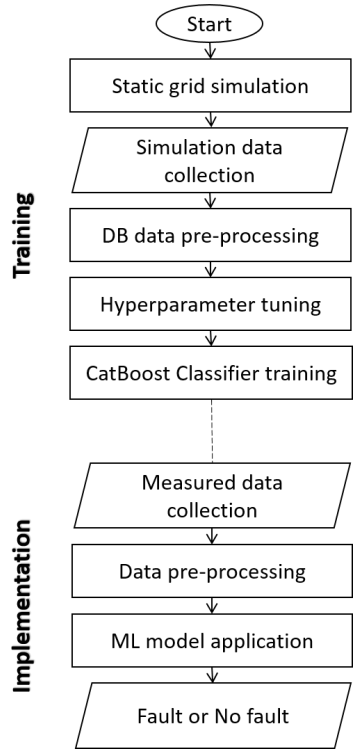

Figure 4.3: The training and implementation phases of the proposed fault detection algorithm.

4.2.2 Training process

The model's training commences with the data pre–processing. The typical data pre–processing includes the scaling, shuffling and splitting of the data. As commented in chapter 3 in the case of tree–based models there is no need to scale the data or perform cross–validation. Hence, the data pre–processing proposed here consists of the data shuffling followed by their split in training and test set with an 80/20 ratio. These steps are important for the formation of uncontaminated datasets which is prerequisite for the overfitting's minimization. Moreover, in order to increase the randomness and heterogeneity of the dataset its sampling is suggested. The sampling process denotes the bootstrapping of the dataset. The final test phase, that comprises the algorithm's testing in out–of–sample data, is performed in a separate dataset that undergoes the same pre-processing.

Once the data pre–processing is completed the model's hyperparameters are tuned. These are the model's parameters that define its development and learning process, e.g. the depth of a decision tree, the learning rate etc. Their selection has a direct impact on the model's performance thus they should be carefully tuned. The most common tuning process and the one proposed here consists of two parts. First, the hyperparameters to be tuned and the value ranges for each one are selected; usually, the models have tens of hyperparameters and an effort to fine–tune them all can be time consuming and lead to overfitting. Thus, the selection of the most significant hyperparameters is an important part of the process that can be rather challenging as it relies heavily on the developer's experience. Following that, the final hyperparameter values are selected with the help of a meta–estimator. In this case the applied algorithm is the *RandomSearchCV* as was deemed the most efficient one. It tests the model's performance for different hyperparameter values' combinations and saves the best one. The number of the tested combinations is set by the developer and the combined values are chosen randomly by the meta–estimator. In this study case the number of iterations was set at 100 and the hyperparameters chosen to be tuned, their ranges and their final values are presented in Table 4.1.

Table 4.1: Hyperparameter values for the CatBoostClasiffier

Hyperparameters	**Value range**	**Final value**
Subsample	[0.5, 0.7]	0.5
No of estimators	[100, 700]	500
Max depth	[4, 10]	7
Learning rate	[0.03, 0.3]	0.1
L2 regulation	[1, 30]	1
Border count	[5, 200]	50

4.3 Case study

In order to evaluate the algorithm's accuracy when trained with static simulation data and tested in various unseen scenarios a case study was performed.

4.3.1 Test grid

Since the method is based on the utilization of static simulation data, a modified version of the CIGRE European LV benchmark was simulated for the generation of the training dataset. Due to the lack of real life data, the same benchmark was used for the generation of the intermediate operation scenarios as well. The simulation conditions were different in the two cases though in order to ensure the reliability of the test results. The benchmark's general layout and characteristics are included in [175]. The modifications done in this case pertain to the addition of PVs, residential FC and UFC points at the first and third feeder, a public UFC point at the second feeder and a fourth feeder simulating a public FC and UFC station. Moreover, multiple measuring devices were spread out in the grid, even though, as it is later discussed, very few measurement points are required by the algorithm. The addition of the new elements was done according to the lines' thermal ratings. The modified grid's topology is illustrated in Fig. 4.4. The squared Ms follow by a number depict the added meters.

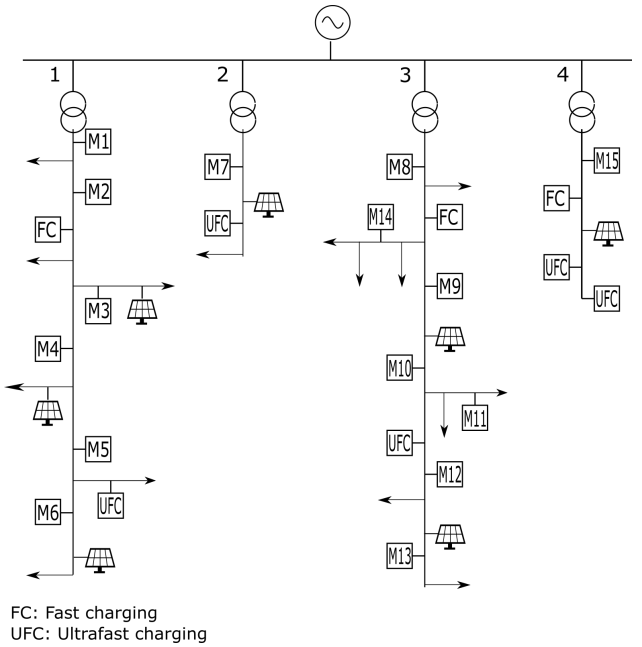

Figure 4.4: Modified CIGRE European LV benchmark

4.3.2 Simulation conditions

During the generation of the training dataset the only modified parameters were the PV generation, the fault resistance, the fault location and the fault type in order to create a diverse dataset that contains enough examples. The residential loads were simulated as constant power loads consuming their nominal power. For the static simulation of the EV charging points two different strategies were followed and then compared. The first strategy, corresponding to the first generated dataset, considered that none of the EV charging points were occupied. The second strategy, used for the generation of the second dataset, considered that all the EV charging points were operating simultaneously at their nominal power. In this way, the training dataset is quickly and easily generated for any grid. The exact simulated values used here are presented in Table 4.2.

Table 4.2: Grid element values

Parameters	Number of scenarios	Values
Fault resistance	Training datasets: 75262 Test datasets: 16236	$(0, 100)\Omega$
Locations	20	$[35, 315]m$ from the source
PV generation levels	5 for each feeder	1st branch: 0, 800, 1700, 3000, 5000 W 2nd branch: 0, 4000, 7000, 10000, 13000 W 3rd branch: 0, 2500, 5000, 7500, 10000 W

Due to the lack of real life data, the same grid was also used of the generation of the test dataset. In this case, however, both the EV chargers and the rest of the loads were simulated operating in various possible scenarios corresponding to intermediate power values. The charging curves and subsequently the occupancy rate used for the simulation of the residential and public charging points as well as the load curves used of the simulation of the residential loads were based on existing literature [176–178] and are illustrated in Fig. 4.5, 4.6, 4.7, 4.8 and 4.9 respectively. The exact simulated points are marked on the load and occupancy rate curves. The EV charging was simulated using a Tesla Model 3 curve as shown in Fig. 4.10.

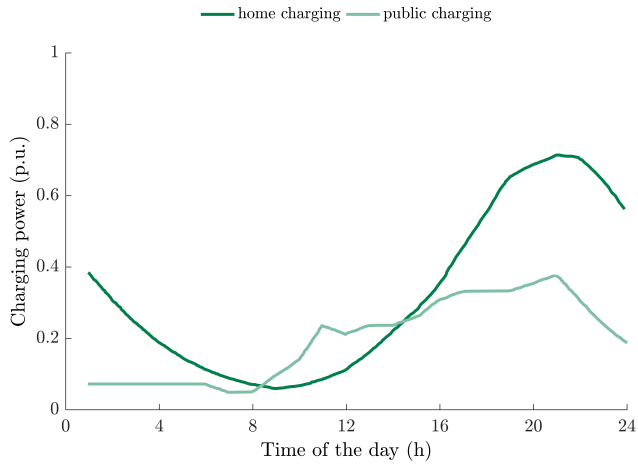

Figure 4.5: Charging curves of residential and public chargers.

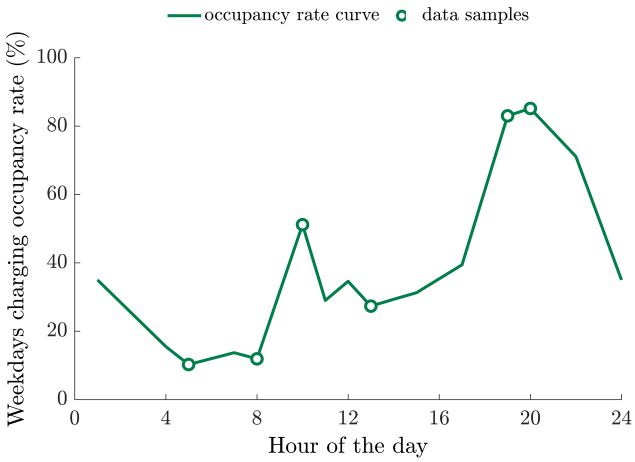

Figure 4.6: Weekday occupancy rate curve.

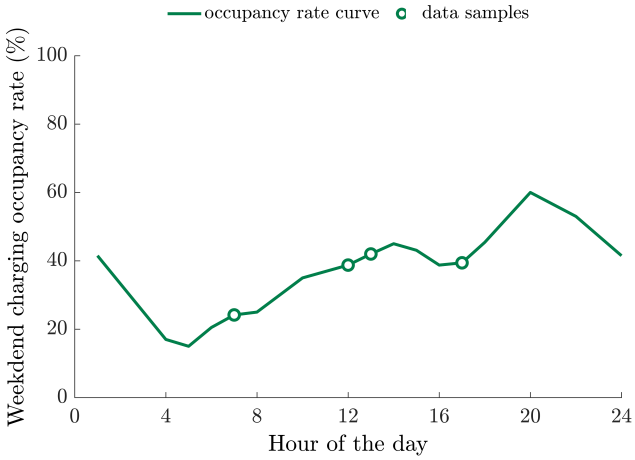

Figure 4.7: Weekend occupancy rate curve.

The measured data were recorded both under normal and faulty operation and consisted of the three-phase voltage and current phasors. The faulty operation measurements were recorded within half cycle of the fault, preceding the activation of the grid's protections. During

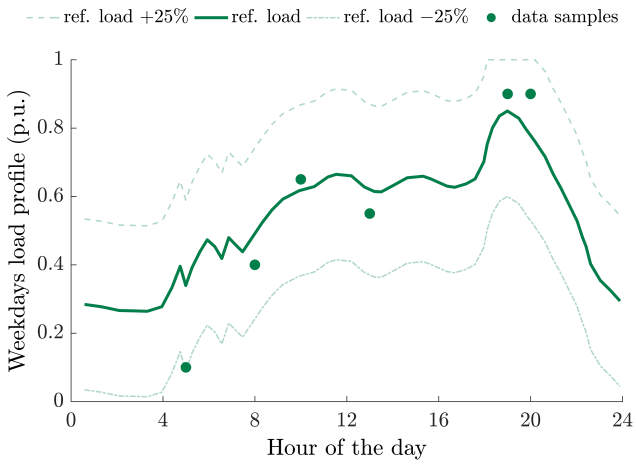

Figure 4.8: Weekday load curve.

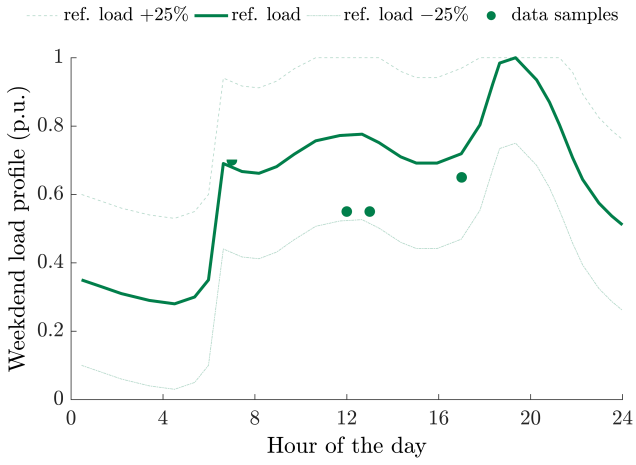

Figure 4.9: Weekend load curve.

the simulations the grid was considered to be operating in steady state and any transient phenomena related to PVs, EVs and their converters were ignored. Furthermore, any measurement noise was omitted. This does not compromise the algorithm's accuracy as the test dataset

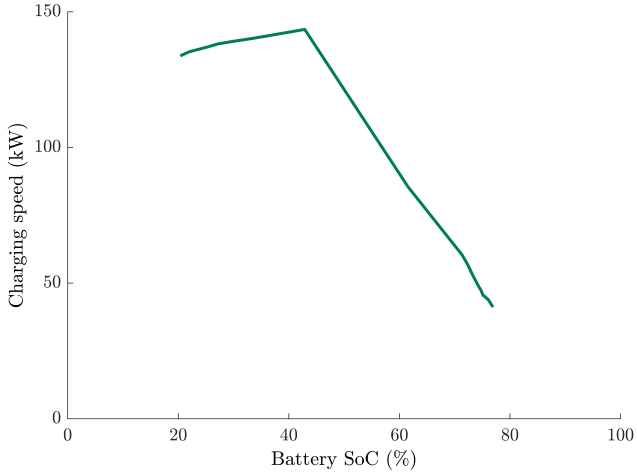

Figure 4.10: Tesla Model 3 charging curve.

includes scenarios produced under different operation states than the ones included in the training datasets. The features saved were the PV generation, the fault resistance, type and location and the measured data. The target value was the existence or not of a fault. The size of the generated training datasets was (37631, 28) each, which after sampling was reduced to (10000, 28). The size of the test dataset was (16236, 28). The processor used during the study was an Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz.

4.4 Results

First, in order to assess the model's predictive capabilities in fault detection, the algorithm's performance was tested on the grid without any EV charging. Table 4.3 presents the accuracy and F1–score results for various dataset sizes. It can be observed that the model's performance remains exceptionally high even when small datasets are used for its training.

Table 4.3: The algorithm's performance results for a LV grid without EV charging for different dataset sizes.

No of examples	Accuracy (%)	F1 score (%)
10000	100	100
6000	99.75	99.74
2000	99.75	99.74
1600	99.68	99.66

After the validation of the model's suitability to fault detection applications, its evaluation on the more complex case of a grid with EV charging is performed in the following sections.

4.4.1 Algorithm's performance on grids with EV's FC and UFC

The performance of the two datasets is compared on the basis of the algorithm's detection accuracy, the F1 score, the training times and the required dataset size. The results are presented in Table 4.4. It can be observed that both datasets lead to highly accurate results, however, the first dataset leads to an overall better performance. It is characterized by a slightly better accuracy, lower training times and, also, a greater robustness against the number of examples required to accurately train the model. It should be noted here that due to the random sampling applied in the datasets the number of examples does not have a linear relation with the model's accuracy.

Table 4.4: The algorithm's performance results for the different dataset characteristics.

	No load dataset			Full load dataset		
No of examples	Accuracy (%)	F1 score (%)	CT (min)	Accuracy (%)	F1 score (%)	CT (min)
10000	97.61	98.79	35.53	97.57	98.77	38.35
6000	97.61	98.79	33.56	92.92	96.33	34.67
2000	97.55	98.76	33.52	97.57	98.77	33.72
1600	97.61	98.79	30.02	96.00	97.96	32.72

This similarity in the overall performance of the two training approached is attributed to the fact that, due to the lines' hosting capacity limits, the added EV charging points account only for 15% or

less of the feeder's maximum loading, with the exception of feeder 4 that simulates a charging station. This is a representative percentage for most LV grids, since many residential feeders are already operating close to their limit. Nevertheless, compared to conventional residential loads, FC and UFC can alter significantly the grid's load flow and its operating state, therefore, the study of their effect on the grids is imperative.

Furthermore, in order to really understand and validate the performance of an AI model a closer look at the training data is always required. Hence, a representative feature, in this case the current's magnitude values collected by a meter in the beginning of the line (meter 1), in relation to the target value was studied. As illustrated in Fig. 4.11, 4.12 and 4.13, the value range of the first training dataset is closer to that of the test dataset than the value range of the second training dataset, even though the actual data points vary considerably. Therefore, it can be concluded that omitting the EV loads during the generation of static simulation data can lead to the training of a highly accurate and generalizable AI–based fault detection method for an active LV grid. Furthermore, the algorithm constitutes a global solution that can be applied to any grid, since it does not require the prior knowledge of the number and location of the EV chargers.

The high performance of the final algorithm trained with no EV charging can be summarized by the fact that it misclassified only 388 out of the 16236 test cases and they all corresponded to false positives. This means that even though the algorithm may cause a few false alarms, it is always able to detect the occurrence of an actual fault, ensuring the safety of people and equipment.

4.4.2 Sensitivity analysis

In order to reinforce the validity of the algorithm's exceptional performance a sensitivity analysis was performed. The effect of the most important influencing parameters, i.e. the fault resistance, the number of utilized meters, the PV penetration and the occupancy rate of the EV chargers, on the fault detection method was thoroughly analyzed.

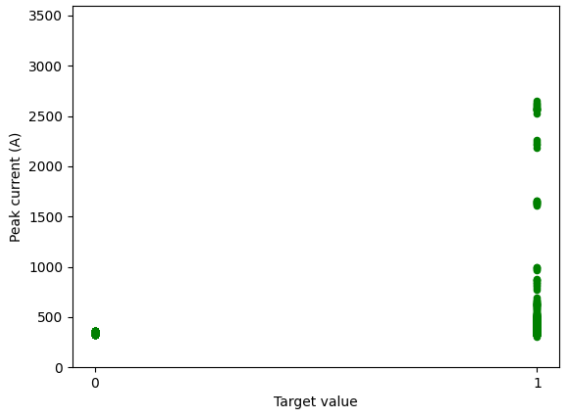

Figure 4.11: Peak current values collected from meter 1 in relation to the target value for the no EV load case.

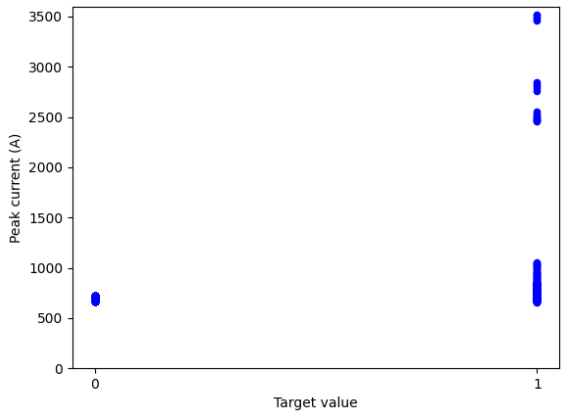

Figure 4.12: Peak current values collected from meter 1 in relation to the target value for the full EV load case.

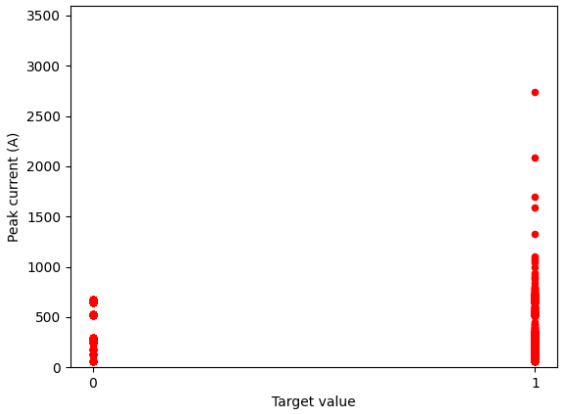

Figure 4.13: Peak current values collected from meter 1 in relation to the target value for the test case.

Fault resistance

The most influential parameter on any fault diagnosis method is the fault resistance. Due to its effect on the fault current it can affect significantly the performance of the method. Especially in the case of the fault detection, if the fault resistance is too high and the fault current too low then it is very difficult for the system to detect the fault. Thus, a big range of fault resistance values was used for the testing of the algorithm. More specifically, 8 resistance values were randomly selected in the range of (0, 100) [127]. In Fig. 7 it can be observed that despite the randomness in the fault resistance selection and the wide range tested, the algorithm is still leading to an accuracy higher than 96% regardless of the fault resistance's size. The small error increase (approx. 2%) observed in the [30, 40) range usually appears when a lower amount of examples from a specific range is included in the training dataset.

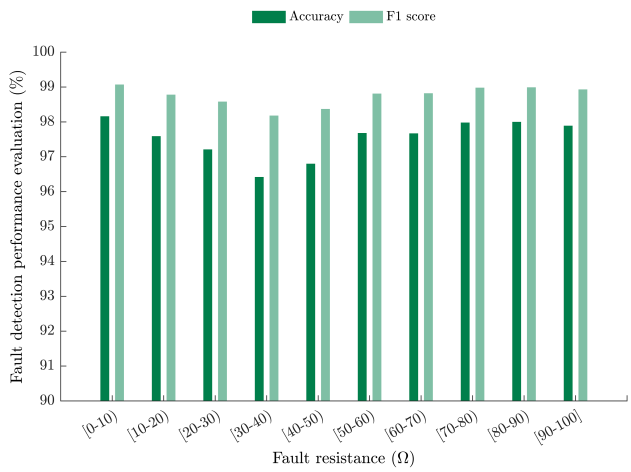

Figure 4.14: The algorithm's accuracy and F1 score in relation to the fault resistance value.

Metering requirements

Even though the measuring devices on the grid are multiplying, the number of measurements required for the application of a fault detection algorithm is an important factor. Especially when it comes to AI–based methods, there is a widespread skepticism regarding their large data requirements. Therefore, a detailed analysis of the algorithm's measurement requirements was performed.

In order to minimize the measuring requirements in an optimum way an features' importance analysis was first performed. The results are presented in Fig. 4.15. Based on the results, six meters were selected as the most important for the recording of the training data: no. 1, 3, 6, 7, 8, 15. These meters are placed in strategic points on the grid, in their majority at the beginning or the end of the feeders, thus they are normally available in every grid. As it can be seen in Fig. 4.16 , the algorithm's performance is the same when the data from 15 meters are collected and when the data from these 6 meters are collected. Thus, it is concluded that the rest of the meters were providing redundant information.

The algorithm's performance was then tested in the case that one

or two of the six meters were removed. The results of this stress test are also illustrated in Fig. 4.16. It can be observed that the accuracy of the algorithm remains high even when only four measuring devices are used. The most extreme case, which is the only one leading to an accuracy bellow 90% corresponds to the simultaneous elimination of meters no. 6 and 7. This is, however, a highly unlikely scenario.

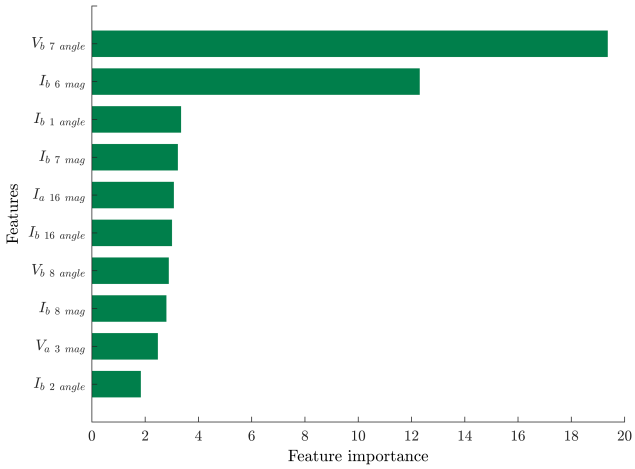

Figure 4.15: The features' importance.

Distributed generation

As established in chapter 2, the bidirectional power flow is one of the main challenging parameters that affect the fault detection methods. As discussed in section 4.3.2, the proposed method was tested in 5 different PV generation levels, corresponding to the whole range of possible generation values. In Fig. 4.17 it is shown that the algorithm remains unaffected by the injected power and maintains its high accuracy regardless of its magnitude.

Occupancy rate

When studying the EV charging stochasticity's effect on a method, the occupancy rate of the charging points is also a significant influencing

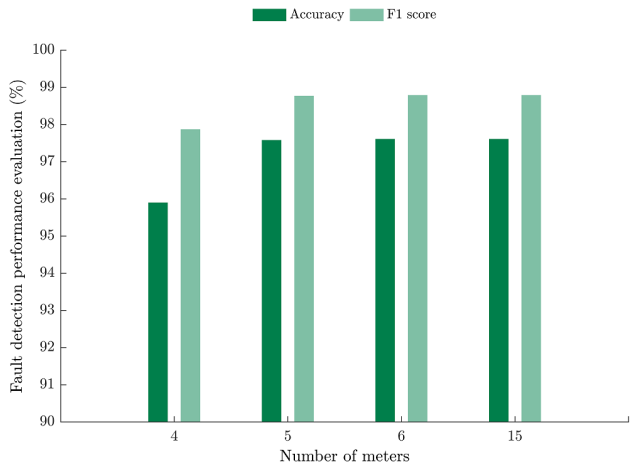

Figure 4.16: The algorithm's accuracy and F1 score in relation to the number of meters.

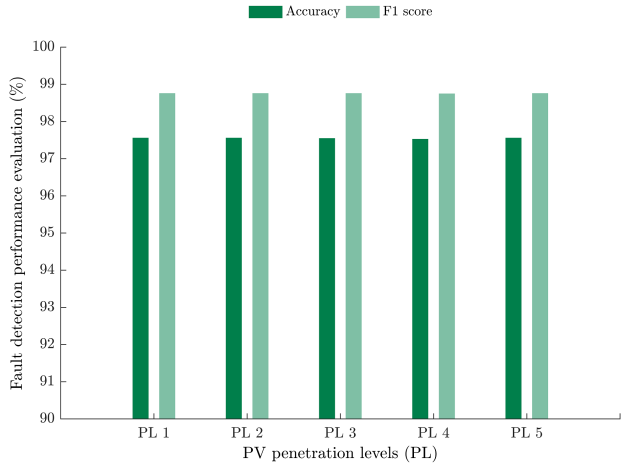

Figure 4.17: The algorithm's accuracy and F1 score in relation to various PV generation levels.

parameter. The occupancy rate refers to the cumulative power that the chargers consume in relation to their cumulative nominal power. In Fig. 4.18 it is illustrated that the rate of misclassified cases for each

occupancy rate in relation to the total cases misclassified is low and independent of the occupancy rate. Thus, the algorithm is indeed robust against the EV stochasticity and highly generalizable.

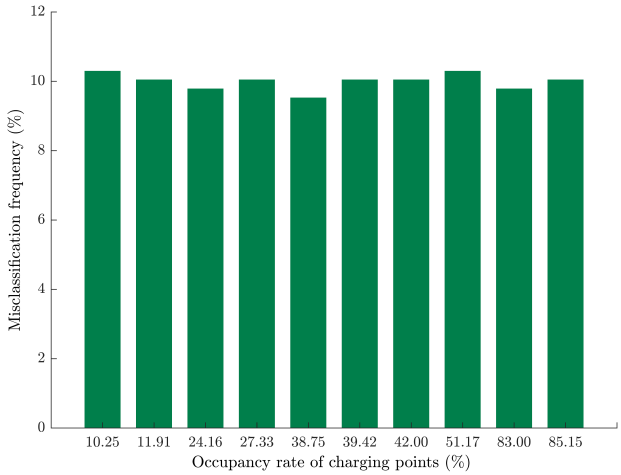

Figure 4.18: The algorithm's accuracy and F1 score in relation to various PV generation levels.

4.4.3 Conclusions

In this part of the doctoral research a fault detection method for active LV grids with EV charging was developed and evaluated. The goal of this study was the development of an AI–based method that could be easily implemented in any LV grid without being affected by the stochasticity of EV charging. For this purpose the proposed algorithm utilizes only static simulation data and omits the simulation of the EV loading. This approach results in high fault detection accuracy in unseen data and independence from the EV charging state and other major influencing parameters such as the fault resistance, the PV penetration and the measuring points.

5

Fault classification in distribution grids

The second step of the fault diagnosis process is the classification of the fault type. The literature review presented in chapter 2 revealed a major research gap in the fault classification methods developed both for the MV and the LV grids. Since threshold–based methods seemed to be among the most widely used and easily applied methods, the fault classification part of this fault diagnosis solution focuses on the development of an optimized threshold–based fault classification method for a distribution grid.

5.1 Theoretical background

Threshold–based methods study the current and voltage patterns during a fault in order to establish certain criteria that determine the type of the fault based on the phase measurements. Due to the nature of the method, however, these criteria are usually grid–specific. There are very few methods that propose generic criteria that could be implemented in any grid. The most noteworthy studies on the topic are presented in [179] which refers to the transmission grid and in [59] which refers to the distribution one.

The first one constitutes a practical guide on the behavior of transmission grids under faults. More specifically, the analysis of the protection systems' operation related to a fault's appearance led to the discovery of patterns on the current and voltage values. These were, utilized for the development of classification criteria for phase–a faults that were more complex and, thus, more accurate compared to other similar methods. The research, however, applies solely to transmission grids since the characteristics of the transmission and the distribution grid differ greatly. Furthermore, only faults related to phase a are included in the formulas and there is no sensitivity analysis performed. Parameters like the management of different meters placed on the grid or the fault resistance, which affect fault diagnosis methods significantly, are ignored in the study.

The fault classification formulas presented in [59] are developed for distribution grids, they are more modern and they refer to all the phases. Nevertheless, the method does not distinguish between line-to-line and double line-to-ground faults and relies on the lack of positive classification of another type of fault for the classification of three phase faults. This has limited applicability, especially in complex grids with various measuring points. Moreover, the criteria were formed based on the observation of the current's values on the meter placed at the feeder of the line, with no other data sources taken into consideration. Finally, there was no analysis of the effect of the fault resistance on the method's accuracy.

Overall, the available literature is rather outdated and inadequately tested. Therefore, this method aims to develop a modernized threshold–based fault classification solution that is easily adaptable to any distribution grid and under varying influencing parameters.

5.2 Proposed method

Motivated by the existing research as well as the research gaps, the developed fault classification method expands the applicability and efficiency of similar over-current techniques by presenting a threshold–based solution that is easily tailored to any conventional distribution grid. The form of the generic criteria presented in Table 5.1 is based

on current patterns recorded in the aforementioned literature as well as new data analysis.

Table 5.1: The Final Fault Classification Criteria

Fault Type	**Criteria**	
1ph	$\Delta I_x > \mathbf{n} * \Delta I_y$ AND $\Delta I_x > \mathbf{m} * \Delta I_z$	
2ph	$\Delta I_x > \mathbf{k} * \Delta I_y$ AND $\Delta I_z > \mathbf{l} * \Delta I_y$	$I_{0_f} - I_{0_p} < I_{0_p}$
2phg		$I_{0_f} - I_{0_p} > I_{0_p}$
3ph	$I_{i_f} > \mathbf{g} * I_{i_p}$	

In Table 5.1 subscript f stands for values after the fault, subscript p stands for values before the fault, subscripts $i = \{x, y, z\} and \{x, y, z\} \in \{a, b, c\}$, and ΔI_x is defined as I_{xp}/I_{xf}. With the exception of the three phase faults, the used criterion evaluates the current variation of each phase before and after the fault and how it compares to that of the other two phases. This technique increases the robustness of the criteria since it simultaneously checks the current variation of each phase and the relation between the different phases. In the case of three phase faults it is rather redundant to compare the current behavior between the phases, as it is enough to establish the variation magnitude of each phase current.

Regarding the criteria used to distinguish the double–phase faults from the double–phase–to–ground one these are based on the principals of the symmetrical component's theory. Even though the theory does not apply to unbalanced grids, it was observed that the values of the zero sequence after the two phase faults were approaching the ones expected according to the theory. More specifically, in the occurrence of line–to–line faults, the zero sequence current values were particularly low, almost zero. On the contrary, during double line–to–ground faults, the values of the zero sequence were much higher both compared to those of the line–to–line faults as well as to those of the zero sequence current before the fault. Therefore, the criterion selected as more accurate was that comparing the difference in the zero sequence current before and after the fault. To ensure that the symmetrical component

analysis will be possible, all the meters utilized in this method should be placed in three phase lines.

5.2.1 Threshold selection process

The most important step in threshold–based methods and the method's greatest source of error is the selection of thresholds. Current values are grid dependent, thus it is impossible to establish universal thresholds for the classification of faults. Hence, in order to ensure that the thresholds used in the proposed criteria (values n, m, k, l, g in Table 5.1) are effective for all grids, the automatization of the threshold–selection process is proposed here. The first step for that is the study of the current's behavior under various fault scenarios. The analysis presented here is performed for a benchmark grid whose characteristics are described in the following section. It includes the behavior of the criteria variables for phase a–to–ground faults and a–b faults as well as the behavior of the phase a during three phase faults, as representative of the three general types of shunt faults. The analysis revealed that the most influencing parameter on the current values, as measured in a conventional distribution grid, is the fault resistance.

Figures 5.1 and 5.2 present the criteria values in relation to the fault resistance for an a–phase–to–ground fault, Fig. 5.3 and 5.4 present the same for an ab fault and Fig. 5.5 presents the ratio of the phase–a pre– and post–fault currents (I_{i_f}/I_{i_p}) in relation to the fault resistance. It can be observed that the ratio between the change in the current values before and after a fault for two phases ($\Delta I_x/\Delta I_y$) presents a certain pattern that differs for the various fault resistance value ranges and the meters. The same is true also for the waveform corresponding to the three phase fault. Thus, this research proposes the establishment of an automatized threshold–selection process based on the study of the current in relation to the fault resistance ranges and the utilized meters. In this way the accuracy of the method is maximized and the method can be quickly and easily applied to all kinds of grids.

Similarly to the fault detection method, the threshold selection process utilizes static simulation data for the establishment of the thresholds. These data can be easily generated for any grid. The data generation process includes the simulation of the grid under various fault

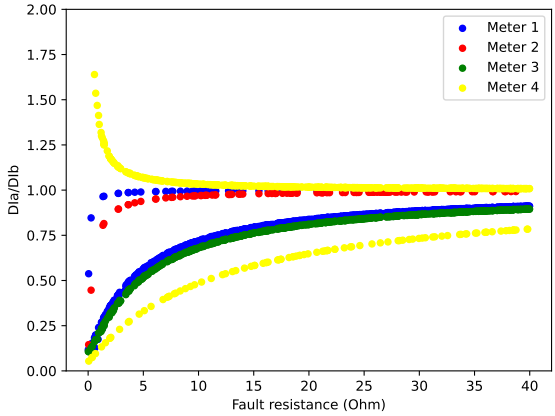

Figure 5.1: The relation between the $\Delta I_a/\Delta I_b$ ratio and the fault resistance for an a–phase–to–ground fault.

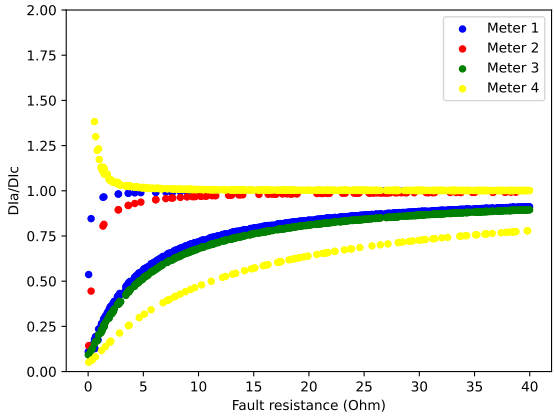

Figure 5.2: The relation between the $\Delta I_a/\Delta I_c$ ratio and the fault resistance for an a–phase–to–ground fault.

scenarios. Here, 1000 faults were simulated for each fault location, summing up to a total of 9000 simulated cases. A random fault resistance with a value between 0 and 40 Ω was added to each simulated fault and

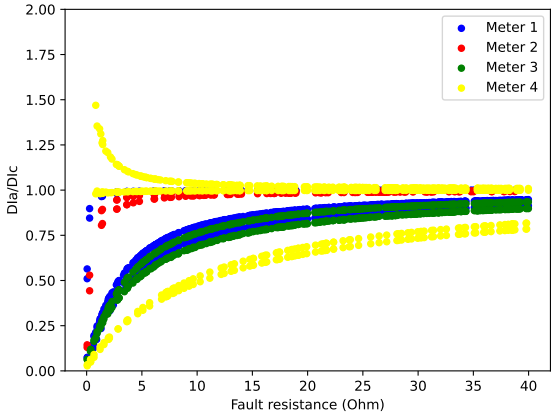

Figure 5.3: The relation between the $\Delta I_a/\Delta I_c$ ratio and the fault resistance for an a–b fault.

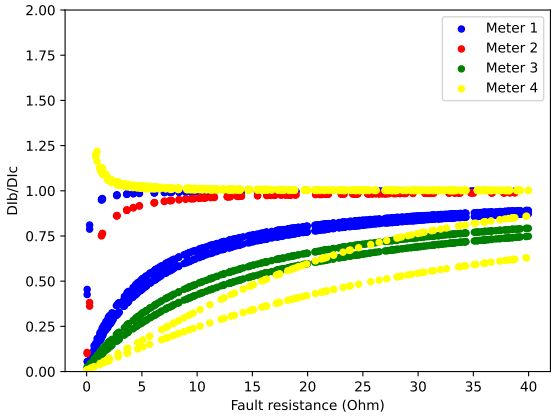

Figure 5.4: The relation between the $\Delta I_b/\Delta I_c$ ratio and the fault resistance for an a–b fault.

the three phase and zero sequence currents of each meter were recorded before and after the fault. All ten types of short circuits were simulated and stored as a number between 1 and 10.

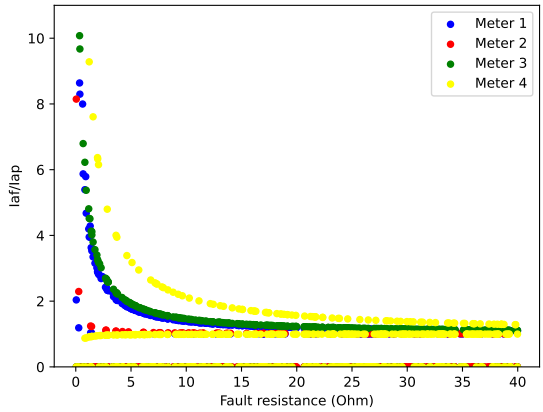

Figure 5.5: The relation between the I_{a_f}/I_{a_p} ratio and the fault resistance for a three phase fault.

The proposed algorithm assesses these generated data and selects the appropriate numerical values, i.e. thresholds, that mark the limit between the fault and no-fault cases for each phase. First, the values of the criteria variables, e.g. the $\Delta I_x/\Delta I_y$, are computed based on the input data. Then, these values are split into smaller subsets based on their fault resistance and the meter they were collected from. The characteristics of each subset are analyzed and then their average properties are combined leading to the selection of the appropriate thresholds for each type of fault. The exact process is analytically described in algorithm 1.

Regarding the assignment of weights in the algorithm, in the case of the previously presented data, the minimum weight would be assigned to the $0 - 10\Omega$ subset that presents the biggest heterogeneity, while the rest of the subsets would be assigned bigger weights. In this case the weight assignment was 0.4, 1.2, 1.2, 1.2 for each of the four subsets respectively. After the selection of the thresholds and when a fault is detected, the algorithm evaluates the percentage of the criteria fulfillment for each type of fault for the data collected from each meter. The criterion that has the highest percentage, i.e. that is fullfilled for most

Algorithm 1 Threshold–selection algorithm

Require: $maxM, minR, maxR \in Z^*, I_p, I_f \in N, fault\_types, w_i \in R*, with \sum w_i = i$

function Threshold_selection($maxM, maxR, fault\_types$)
for fault in fault_types **do**
1. Read the pre and post fault current, I_p, I_f for each phase and meter
2. Calculate the corresponding criteria variables
3. Establish equally ranged fault resistance groups based on the $(maxR - minR)$ range
4. Create a data subset $Sr_i, i \in Z^*$ for each fault resistance range
5. Inside each subset Sr_i create a data subset $Sm_j, j \in Z^*$ for each meter up to $maxM$.
for i in Sr **do**
for j in Sm **do**
Calculate $avgM(j) = avg(Sm_j)$
end for
1. Calculate $avgR(i) = avg(avgM)$
2. Assign weight w_i based on the subset's heterogeneity.
if heterogeneity = small **then**
$w_i > 1$
else
$w_i < 1$
end if
end for
Calculate $avg\_all = avg(w * avgR)$
Set threshold_fault as avg_all
end for
return thresholds

meters indicates the type of the fault. The method's grid application is illustrated in Fig. 5.6. Finally, the algorithm's overall implementation process is presented in Fig. 5.7.

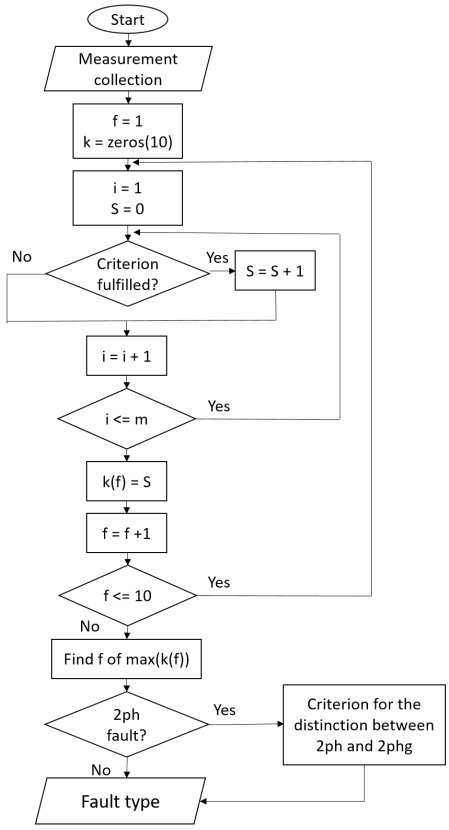

Figure 5.6: Flowchart of the implementation of the proposed fault classification algorithm

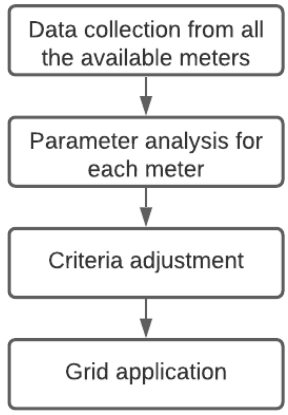

Figure 5.7: Flowchart of the overall method's application process

5.3 Case study

For the validation of the proposed algorithm's high performance and novelty, the case study evaluates not only the proposed algorithm but also three other threshold–based algorithms; two of them (algorithms 1, 2) were based exclusively on [179] and one of them (algorithm 3) on [59]. The point of differentiation between the first two algorithms is only the source of input data. Algorithm 1 processes the data collected from all the measurements for the calculation of the final result, while algorithm 2 processes only the data collected from the meter closest to the fault.

5.3.1 Test grid and results

All four algorithms were tested regarding their performance under different conditions. The grid chosen for the simulations and extraction of the results was the IEEE 13 node test feeder, illustrated in Fig. 5.8; the simulations were run in Simulink. The choice of the grid was dictated by the variety of different elements included in it. Specifically, it contains four meters placed at nodes 632, 633, 671, 692, marked in

red in the respective figure, highly unbalanced lines and loads and a LV branch. The data were collected before the breaker's activation. The different cases included the simulation of all kinds of shunt faults occurring in nine different locations of the grid, the utilization of data acquired from one up to four meters installed in different locations of the grid as well as a range of fault resistance values.

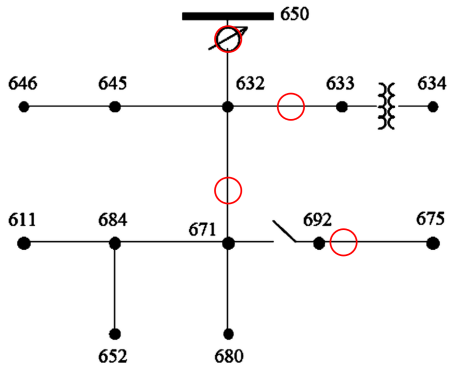

Figure 5.8: IEEE 13-node test feeder - single line diagram

Table 5.2 presents the criteria set by the algorithm for the classification of the different types of faults.

Table 5.2: Selected thresolds for each fault resistance range

Fault type	**Thresholds**	
a–g	$\Delta I_a/\Delta I_b$	$\Delta I_a/\Delta I_b$
	3.56	3.54
b–g	$\Delta I_b/\Delta I_a$	$\Delta I_b/\Delta I_c$
	2.6	3.16
c–g	$\Delta I_c/\Delta I_a$	$\Delta I_c/\Delta I_b$
	3.52	3.55
ab(–g)	$\Delta I_a/\Delta I_c$	$\Delta I_b/\Delta I_c$
	3.57	3.24
bc(–g)	$\Delta I_b/\Delta I_a$	$\Delta I_c/\Delta I_a$
	2.97	3.17
ca(–g)	$\Delta I_c/\Delta I_b$	$\Delta I_a/\Delta I_b$
	3.56	3.57
abc(–g)	I_f/I_p	-
	3.45	-

In order to verify the superiority of this method against non–optimized threshold–based methods, the algorithm's accuracy with and without the application of the automatized threshold–selection process is compared. In Fig. 5.9 it can be seen that the algorithm presents an overall increased accuracy in the case of the algorithm–selected thresholds. The figure also presents the algorithm's effectiveness on the MV and the LV side separately.

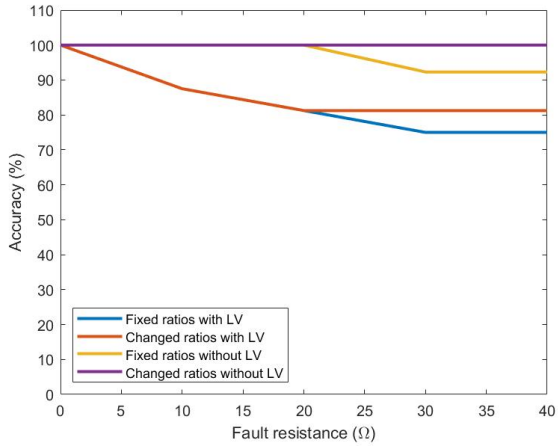

Figure 5.9: The method's accuracy with the use of the optimum ratios vs the use of static thresholds in relation to the fault resistance.

Measurements' Effect

In order to isolate the measurements' effect on the algorithm's accuracy, the grid was simulated without the use of fault resistance. Taking into consideration the fact that in [179] there is no reference to the process of data collection and that in [59] the data are considered to be collected only from a meter on the feeder (node 632 here), different approaches were tested for the selection of the appropriate data collection method. More specifically, the utilization of one meter in the feeder was tested against the utilization of all four meters available in this grid. As seen in Fig. 5.10, the proposed algorithm performs greatly in both cases, with an accuracy of 100%. It should be noted here, that during the analysis of the measurements' effect the fault resistance was considered to be zero and the measurements obtained assumed to be synchronized.

Algorithm 3 also appears to be leading to highly accurate results, however, these results do not include the classification of three phase faults, as this algorithm is not capable of classifying those. Another interesting observation regarding the performance of algorithm 3 is the fact that it is more accurate when data from only the meter in the

feeder is used rather than when data from all four meters are used. The explanation lays on the original establishment of the criteria. Since the thresholds were selected based on the current's behavioral patterns on the feeder, the current measurements in other parts of the grid that, due to the unbalances of the grid, do not follow the same patterns, lead to false results.

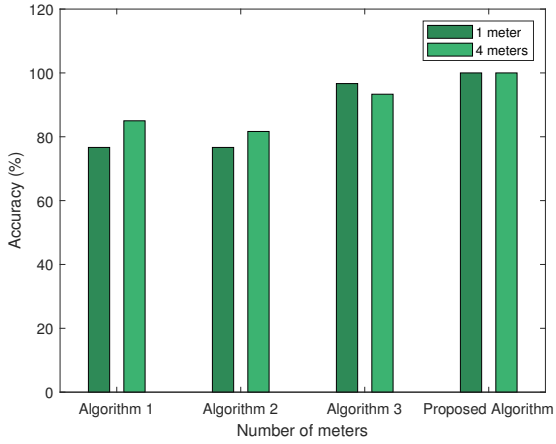

Figure 5.10: Accuracy of the tested algorithms for one and four measurement points [3].

The second case that was studied was the effect of the meters' location on the algorithm's accuracy. The performances of both algorithm 1 and the proposed algorithm were tested and compared. The measurements collected from the grid's meters were progressively added to the dataset starting from the top of the grid. As illustrated in Fig. 5.11, algorithm 1 presents an increased accuracy with the addition of the second and the third meter, while the fourth one does not cause any changes due to its isolated position. The proposed algorithm on the other hand remains unaffected by the number and location of meters and even one meter could be adequate for accurate results. Therefore, it constitutes a more efficient alternative both in terms of accuracy and of cost.

Finally, the algorithm's performance for the MV and the LV part

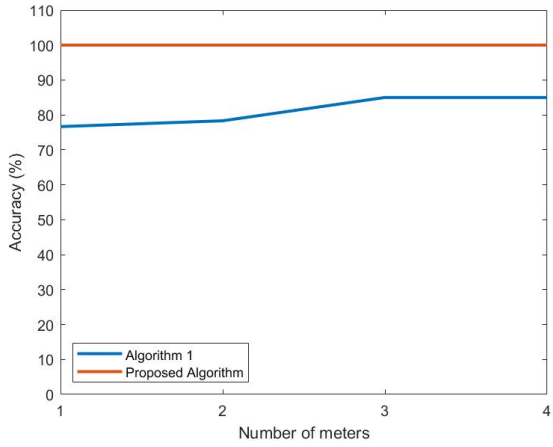

Figure 5.11: Accuracy of algorithm 1 and the proposed algorithm in relation to the number of measurement points [3].

of the grid was tested in relation to the number of utilized measuring devices. In this case a random resistance was also added to the fault. As it can be observed in Fig. 5.12, the algorithm is highly accurate for the MV part. Regarding the LV part, the classification of faults has proven more challenging there and it constitutes a source of inaccuracy.

5.3.2 Conclusions

In this part of the research the development of a simple, automatized and efficient fault classification method for distribution grids is discussed. The method is based on the utilization of sets of versatile, flexible criteria, capable of processing data from all the meters available on the grid. The required inputs are minimized and include only the three phase and zero sequence current rendering the method economical and easy–to–implement. The main novelty of this method is its adaptability to different grid topologies. The presented algorithm allows for the provided criteria to be easily adjusted and directly applied to every conventional distribution grid with minimum data and highly accurate results. Practical guidelines are proposed for this purpose.

For the verification of the method's performance, an asymmetrical,

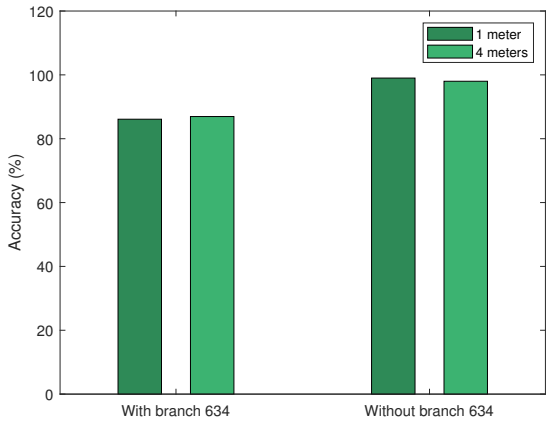

Figure 5.12: The mean accuracy of the algorithm for all the cases, tested with the updated version of the algorithm.

heavily loaded grid was used. The tests comprised the simulation of the ten kinds of shunt faults in nine different locations spread through the grid, including one on the Low Voltage (LV) branch of the grid. An analysis on the effect of the most important parameters such as the amount and location of meters and the fault resistance, was performed. The mean accuracy of the algorithm when applied to the Medium Voltage (MV) side reaches 99%, with data imported from only one meter.

6

Fault location methods for low voltage grids

The last step of the fault diagnosis process is the location of the fault. As discussed in chapter 2 the location of faults in LV grids is another scarcely explored field. From the limited available research, the AI–based methods have shown the most promising results [137], nevertheless, their applicability is still questioned. Parameters such as the required input data volume and computational time remain significant sources of skepticism regarding the use of AI. Hence, this research focuses on the application of advanced data management and processing techniques for the optimization of ML–based fault location algorithms.

6.1 General method description

Artificial Intelligence pipelines can have many variations depending on the algorithm's main objective and application. A general layout consists of the data collection, the data pre–processing, the hyperparameter tuning and the model's training. The proposed method has enriched this process with various advanced techniques that aim to increase the algorithm's applicability. More specifically, the main goals of this method are the minimization of the algorithm's data require-

ments and computational time, and the increase of its generalizability and applicability. In order to fulfill these objectives, the algorithm is structured as follows:

1. First, the algorithm's data requirements are addressed. This includes the data collection process, the analysis of the collected data, the feature extraction and a data storing strategy.

2. Once the datasets are stored, the data are pre–processed and transformed in order to have a suitable format for the employed ML techniques.

3. Then, the application of feature selection and dimensionality reduction techniques is explored. These techniques aim to minimize the model's inputs and increase its generalizability.

4. After that, the model's hyperparameters are tuned for maximized accuracy.

5. A stacked architecture is adopted for the accurate location of faults. More specifically, in order to resolve the multiple location estimation problem two predictive models are trained. First, a classification model identifies the faulted branch and then a regression model locates the exact faulted point.

6. Finally, in order to ensure the method's accurate performance in all possible operation scenarios, while maintaining the training time and data low, a smart retraining topology change strategy is proposed.

Figure 6.1 illustrates the the preparatory steps for the algorithm's training. These are the ones defining the algorithm's accuracy, thus the ones that emphasis should be given to.

Once the data have been appropriately processed and organized, the algorithm's implementation is straightforward, as it can be seen from the method's implementation flowchart in Fig. 6.2 . All the steps of the data processing and the algorithm's training and implementation are analytically described in the following sections. The test results are presented in the next chapter. Overall, the proposed method can be

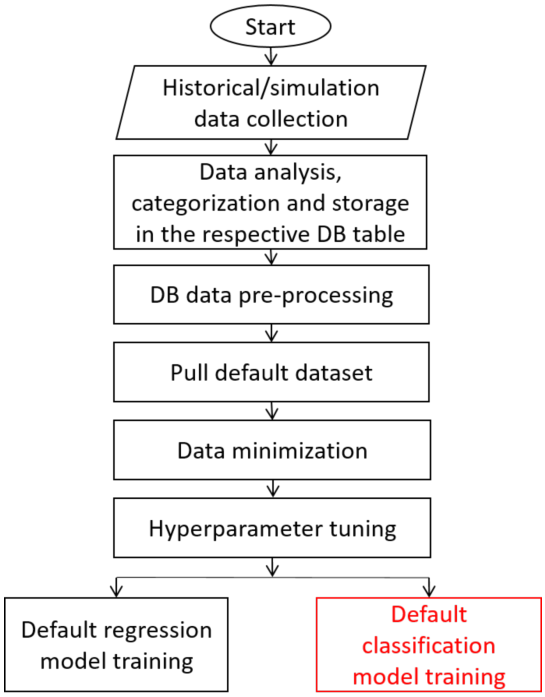

Figure 6.1: Flow chart of the method's training process.

easily implemented to any active LV grid and provides a practical and accurate solution to the fault location problem.

6.2 Faulted branch classification

As commented above, the fault location process comprises two parts: i) the faulted branch identification and ii) the faulted point location. The classification model used in the first part is a RF Classifier as this is described in Chapter 3. The target value in this case is the name of the faulted branch's last node. The latter is encoded as an integer between 1 and the total number of branches, based on the position of the branch in relation to the feeder. Due to the fact that this is a simple classification problem no data minimization techniques were applied.

6.3 Faulted point location

The location of the exact faulted point constitutes a more complex problem. The target value in this case is a continuous number and, more specifically, the distance between the fault point and the secondary of the MV/LV transformer, which depend heavily on the grid's variables. Thus, this part of the algorithm requires the application of advanced data management techniques for the increase of the method's reliability and the decrease of its CT. For the prediction of the fault's distance two different tree–based predictive models were trained and compared, a RF regressor and an XGBoost model. The benefits and characteristics of tree–based models in general as well as of the ones utilized here are all analyzed in chapter 3.

6.4 Data management

As ML–based methods traditionally depend heavily on the quality and the quantity of the available data, the collection of all the required data, their storage and their processing is considered challenging by many. It is a common perception that the application of ML–based algorithms implies the collection of large amounts of data from devices with increased capabilities that are not yet installed on the grid and

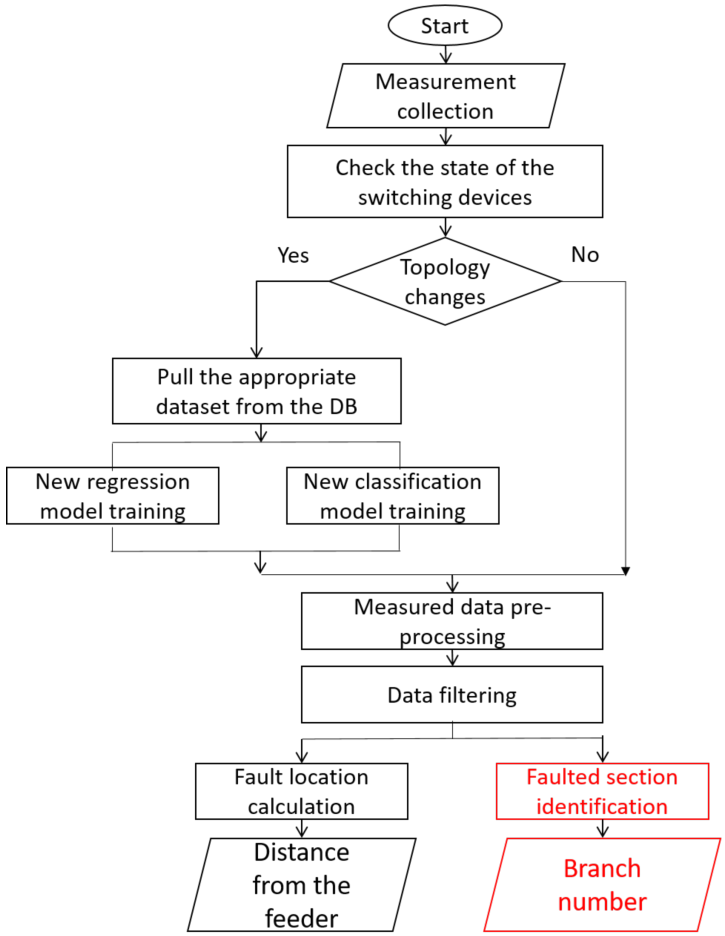

Figure 6.2: Flow chart of the method's implementation after a fault is detected.

the subsequent occupation of vast storage spaces. Nevertheless, most of the above are not prerequisites for the successful employment of ML models and, as suggested in [180], there are practices that have the potential to limit any heavy data–related requirements. Moreover, there are various data minimization techniques that can significantly limit the size and subsequent computation time of the problem. The following sections provide various solutions for the efficient management of a ML–based algorithms' data requirements. The steps are analyzed in implementation order. It is important to highlight here that the proposed techniques are universal and their application does not depend on any grid–specific characteristics, as long as the feature type requirements are respected.

6.4.1 Data collection

The source of utilized data constitutes a discussion point for all fault location methods. The collected measurements define the amount of information that can be harvested from the grid and the potential investment costs the the method's implementation could imply. Therefore, the type e.g. voltage and current phasors, only current or voltage magnitudes etc., and the transmission frequency of the measured data to the control center should be carefully reviewed. This study presents a a comprehensive data collection analysis covering various possible measurement methods that may be used on the grid in order to make the algorithm's application possible in the majority of grids.

First of all, regarding the measured values, these can be only the variables' magnitudes or they can be phasors; a comparison of the algorithm's response in these two cases is presented in chapter 7. Moreover, the algorithm utilizes both voltage and current measurements, nevertheless, based on the data analysis process that follows, the current measurements are of higher importance for the algorithm, therefore, their collection should be prioritized. Regarding the transmission of the data, the algorithm does not utilize time series data, hence the exact times of the data collection and transmission are not important. However, one measurement before and one after the fault are required and the measurements after the fault should be collected before the operation of the grid's protections. Thus, the method supports all bulk,

scheduled and on demand data collection configurations. Overall, the measurement requirements of the method are similar to those of the simpler fault location methods and can be easily collected from various devices that are already installed on the grid or are expected to be installed in the near future.

6.4.2 Data analysis

Once the data collection process has been determined, the analysis of the collected data is required. This process is performed only once, during the training stage of the algorithm, and it is one of its most important parts. It is a step that should be included in all fault location methods, regardless of the utilized techniques, nevertheless it is not encountered in the existing literature. Both real and simulation data should be analyzed and studied as they directly affect the reliability of a method's test results. Moreover, a deep understanding of the available data and their relations can lead to the development of innovative solutions and new methods.

Especially for ML–based methods, the data analysis could point to the most suitable predictive model or to an effective data transformation leading to higher accuracy. Furthermore, the data analysis can point to the features that are the most informative for the prediction of the target as well as, in the case of electricity grids, the position of the most important meters. Overall, the data analysis facilitates the researchers' methodology–related decisions and the interpretation of the algorithms' outcome, and it leads to the formation of an efficient dataset and the development of an accurate model.

So far, the majority of the existing fault diagnosis methods rely on data generated by simulations, as there is a general lack of real data, especially in relation to fault events on the grid. Even though with the evolution of O&M solutions for electricity grids the DSOs' have an increased observability over the grid, there is still a long way to go until real data become vastly available to the research community. Hence, the evaluation of data generated from benchmarks like the one used in this study is of great value. The associated parameters analyzed in this research are the data normality, the correlation between the features and the target value, and the feature importance.

The data normality is tested with the use of the Shapiro–Wilk normality test [181]. This test assumes that the provided data are normally distributed, then compares them with an actual normal distribution and finally calculates the probability of similarity between the provided data and those belonging to a normal distribution. If the probability is higher than 0.05 then the data is categorized as normal.

The evaluation of the data correlation is a more complicated process, therefore, three different methods are used for its calculation. The first one is the Pearson's correlation coefficient r [182], which applies to normally distributed data. It studies the linearity between the features and the target value by calculating the r coefficient. This is defined as the ratio between the covariance of a random pair of samples (x_i, y_i) and their respective standard deviations:

$$r = \frac{cov(x_i, y_i)}{\sigma_{x_i}\sigma_{y_i}} \tag{6.1}$$

For non–Gaussian data two ranking coefficients are mainly used. The first one is Kendall's correlation coefficient τ [183], and it is utilized for the identification of concordant and discordant pairs in the dataset. The τ coefficient is calculated as follows:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \tag{6.2}$$

where (x_i, y_i) and (x_j, y_j) are two random pairs of observations and n is the total amount of observations. The other ranking coefficient is Spearman's ρ correlation coefficient [184] which is used for the calculation of the strength between two variables. More specifically, the ρ coefficient examines the existence of a monotonic relation between two variables, i.e. if one decreases/increases with the decrease/increase of the other, and is defined as follows:

$$\rho = \frac{cov(rank(x_i), rank(y_i))}{\sigma_{rank(x_i)}\sigma_{rank(y_i)}} \tag{6.3}$$

It can be observed that the definition of the ρ coefficient is similar to that of the r; their difference lies in the use of the variables' values rank in the calculation of ρ, instead of the actual sample values. Hence,

the Spearman correlation is less affected by outliers compared to the Pearson correlation.

The data analysis is concluded with the examination of the features' importance. The features' importance is an indication of each feature's contribution to the prediction of the target value. It should not be confused with the correlation methods as it could be more accurately described as a measure of each feature's influence on the model's decision–making process. Therefore, it can serve as a starting point for the employment of feature selection or dimensionality reduction techniques. Moreover, it provides information regarding the response of the prediction model to the specific type of data, since the calculation of the feature importance depends on the estimator fitted to the data. In decision trees the feature importance indicates which values are mostly used for the splitting of trees into branches. The method applied here for the calculation of the features' importance is the computation of the mean and the standard deviation of the impurity decrease accumulation in each tree. It was selected as an appropriate method as it is less computationally expensive than other alternatives and its use is only discouraged in the case of datasets containing high cardinality features. This does not apply to the generated dataset, as it does not contain a large number of unique values.

6.4.3 Data storage

To the author's knowledge, there are no data storage strategies proposed in the existing literature for fault location methods. Thus, in order to address potential concerns regarding the applicability of the presented method, a data storage plan is proposed. First of all, regarding the size of the required storage space, the employment of data minimization techniques as discussed in the following sections ensures that this does not constitute a problem for the method's application. Furthermore, in order to increase the algorithm's ability to generalize to unseen data, particularly in the case of topology changes, a novel database construction plan is proposed. The plan entails the categorization and storage of the collected data in tables according to the status of the grid's switching devices. Each combination of the devices' statuses corresponds to a specific topology and subsequently to a table

in the database. This unique combination is saved as the table's characterization number and it is used for the selection of the appropriate dataset when the topology is changed. The size and identifiers of the database are defined as follows:

$$length(DS) = count(unique(C_j = \{B_1 * 1_{S_1}, ..., B_i * 1_{S_i}\})_{i=\{1,...,S_{max}\} \in \mathbb{Z}_{>0}}) \tag{6.4}$$

where C_j is the j-th combination of the switching devices B_i with $j \geq 1$, i is the number of switching devices, S_max is the total of switching devices and $1_{S_i} = \{0, 1\}$ is a boolean variable corresponding to the state of each device, with 0 denoting that the device is open and 1 that it is closed. When $S_i = 1$ the $1_{S_i} = 1$ and when $S_i = 0$ the $1_{S_i} = 0$. The overall database size is $2^{S_{max}} - 1$ - it is assumed that at least one switching device is closed at all times.

The database data can either be historical data, when available, or simulation data. In the case of the simulation data, though, noise should be added in the measurements in accordance with the real–life devices as the extrapolation capabilities of tree–based models are limited. The size of the additional stored datasets and the noise effect are both analyzed in the next Chapter; it is verified that the accuracy of the method can be maximized without any additional computational or storage expense.

6.4.4 Data pre–processing

As briefly discussed in chapter 4, the pre–processing of the data refers to their transformation into a form suitable for the predictive model. It is an important step for the increase of the model's accuracy. The main processes include the data scaling, shuffling and splitting. The data scaling ensures that all the features are within the same value range, thus they are all considered to be equally important by the model. In other words, it removes the models' bias against features with smaller numerical values. In the case of tree–based models this step is not required as each feature is evaluated individually, nevertheless, it is a mandatory step for the application of the data minimization techniques presented in the following section. In this research two scaling functions were used, the *StandardScaler()* and the *RobustScaler()*,

both by the *scikit–learn* library. They both place the mean of the data on point zero. This is a prerequisite for dimensionality reduction methods such as the Principal Component Analysis (PCA); point zero is the common cross point of all the linear subspaces formed by the PCA. The RobustScaler is used in the cases were the outliers need to be removed from the dataset as they do not constitute useful training examples. Regarding the shuffling of the data, this is also performed by the scaling functions.

Finally, the splitting of the data into subsets is crucial for the minimization of the algorithm's overfitting and the reliability of the test results. If the test data are included in the algorithm's training dataset then the model is fitted to the specific data too accurately and will be unable to generalize to unseen data. This constitutes a major problem for the algorithm's implementation. Therefore, the data are split into three distinct subsets, the training dataset, the validation dataset and the test dataset. Here, the original dataset was only split into training and test datasets, since the employed models perform internal cross–validation. For this purpose, the *train_test_split()* function by *scikit–learn* was used, with a 80/20 ratio between the training and the test data.

6.4.5 Data requirement minimization

A major contribution of this research and an important parameter for any ML–based method is the minimization of the method's data requirements. The assumption that all ML–based methods require large datasets in order to be implemented constitutes a point of skepticism regarding the applicability and practicality of ML–based methods. Hence, the application of data minimization techniques can be a powerful tool for the practical applications of ML–based fault diagnosis methods. Especially as the data sources are multiplying and the data influx is growing, the optimization of the data management process is of utmost importance. At the same time, fewer inputs lead to lower overfitting and, thus, greater generalizability of the algorithm. Therefore, in this study the two major approaches to the data minimization problem, the feature selection and the dimensionality reduction, are deployed and compared. Furthermore, the most important algorithms

pertaining to each approach are analyzed, tested and compared. The comparison results are presented in the following chapter.

It should be noted here that the two data minimization approaches and the related algorithms have different properties which should be taken into consideration in each individual application. This research does not intent to prove that there is a superior one between them. It merely presents some of the currently popular and high performing options suitable for this application and compares their performance with respect to the particularities of the fault location problem.

Feature selection

The first approach proposed for the minimization of the input data requirements is the feature selection. Feature selection methods remove the redundant features from the dataset. Thus, they clean it from the inputs that not only are not important for the prediction of the target value, but may also be adding noise to the prediction process. At the same time, fewer features mean also lower computational times. Finally, the application of a feature selection algorithm in fault diagnosis problems points to the most important measurements, which is also an indication of the measurement points offering higher observability over the grid. This could be a useful guide for infrastructure investments in electricity grids.

The feature selection methods tested in this study are the SelectFromModel meta–transformer [185] and the Boruta algorithm [186]. Both of them constitute sophisticated algorithms with high performance [187,188], nevertheless neither of them has been tested in distribution grid data. Their selection among the various available feature selection techniques was based on the type of data and the nature of the problem as well as on their computational speed and reliability.

1. **SelectFromModel**

   SelectFromModel is a meta–transformer developed by *scikit-learn*. It calculates the features' importance for a specific estimator, i.e. a specific predictive model, and ranks them accordingly. The most important ones are selected based on the set threshold as

the output of the algorithm. The selection of the threshold constitutes the main source of arbitrariness in the process. Thus, here the default threshold value was used, i.e. the mean value of the features' importance.

2. **Boruta**

   Boruta also performs an importance–based feature ranking. The selection of the important features, however, is a more complex process. First the algorithm generates the so–called shadow features. These are created by shuffling the examples for each feature while maintaining the target column unchanged, thus forming new random features. The shadow features are added to the original dataset, which is then fitted with the estimator. The importance of the shadow features is also calculated and is then used as a reference for the selection of the importance threshold for the original features; this is set as the value of the shadow features' highest importance. Hence, a feature is considered relevant only if it performs better than the best performing shadow feature.

   In order to increase the method's robustness and minimize the stochasticity stemming from the shadow features' generation and the model's training, the process is repeated as many times as indicated. As a result, the features that were selected by the algorithm in most iterations are categorized as the best ones, while there is also a middle zone for features about which the algorithm did not make a clear decision. Even though this iterative process adds to the reliability of the method is also results in a significant increase in its computational time and complexity.

Dimensionality reduction

The second approach used to minimize the method's data requirements is the dimensionality reduction. In this case the features are suitably transformed in order to reduce the problem's space while retaining the important information. Some times, however, the combination of these two goals is hard to be achieved. Therefore, a data management strategy that aims to improve the application of dimensionality reduction

in the fault location problem is proposed. More specifically, in order to conserve the physical meaning of the grid's most informative measurements, a features' imporance analysis is performed. The most informative features are stored in the dataset as they are. Then, in order to reduce the size of the problem the rest of the features are transformed by the selected dimensionality reduction algorithm. In this way the model's accuracy and generalizability remain high while the computational time decreases significantly. The time added by the dimensionality reduction process is insignificant in comparison with the drop in the training time. This strategy is independent of the number and location of the meters, thus it is applicable to all grids regardless of their topology.

Similarly to the case of the feature selection algorithms, the features' importance can be automatically calculated by the predictor's library. Then, regarding the dimensionality reduction technique applied, there are various available methods. Based on the goals set above, i.e. the simultaneous minimization of the input data and the information loss, the techniques tested in this study are the ones favoring a dimensionality space higher than the three–dimensional one. These are the PCA [189], the Kernel PCA [190], the Fast Independent Component Analysis (FastICA) [191], the Truncated Singular Value Decomposition (T–SVD) [192] and the Isometric Feature Mapping (ISOMAP) [193].

1. **PCA**

   PCA is a well–known statistical method that aims to compress a dataset's correlated features and extract only the useful information. This results in the generation of a more compact dataset that is easier to manage during the model's training. It is the most popular dimensionality reduction technique with applications that include the location of faults [14]. The theory behind PCA lies in the projection of the data points onto a lower feature space, whose every axis is perpendicular to the rest, thus uncorrelated with them. Each axis corresponds to one eigenvector, i.e. principal component. The eigenvectors are sorted based on their eigenvalues; in this case the higher the eigenvalue the more important the eigenvector. The first principal component is the vector representing the line with the minimum squared distance from all

the data points, hence it is characterized by the highest possible variance. The generation of each new component is dictated by the data points and aims to find the best fit to the data while remaining orthogonal to the rest of the principal components. The total number of the principal components created by the PCA is the same as that of the original features, however, the biggest amount of information is compressed into the first components. Thus, the utilization of only the first few components can lead to the successful training of the ML model. Usually, the principal components used for the model's training are those with an explained variance adding up to 80% of the original dataset's.

2. **Kernel PCA**

   Kernel PCA is a variation of the traditional PCA that, contrary to the original method, is able to perform non–linear dimensionality reduction. Thus, before the application of the PCA's linear operations, the utilized kernels project the dataset's dimensions to a linear space. There are various available kernels; the selection of the most suitable one depends on the shape of the data. In this study the cosine kernel was used.

3. **FastICA**

   ICA methods aim at isolating the independent components of the dataset by finding the matrix that maximizes the non–gaussianity of the original features. The non–gaussianity metric is a means of measuring the statistical independence of the components. The difference between the traditional ICA methods and the FastICA lies in the calculation of the non–gaussianity. In the first case it is calculated with the use of the kurtosis whereas in the second case with the use of the negentropy. Due to that the FastICA is faster and more reliable.

4. **Truncated SVD**

   T–SVD is a dimensionality reduction method based on the factorization of the data matrix. Its operating principle is similar to that of the PCA with the exception that it does not center

the data before performing the computations. While PCA transforms the covariance matrix, SVD transforms the data matrix. Thus, the computational complexity and time of the T–SVD are significantly lower. The truncating nature of the method that differentiates it from the SVD method pertains to its dimensionality reduction abilities.

5. **ISOMAP**

   ISOMAP is another non–linear dimensionality reduction method. It can be considered an extension of the Kernel PCA. The main goal of ISOMAP is the projection of the data into a lower–dimensional space where the geodesic distances between the data points are maintained unchanged. This is achieved with the application of the nearest neighbors methodology in order to distinguish the various manifolds of the dataset.

6.4.6 Hyperparameter tuning

Another decisive part of the model's training process is the tuning of the hyperparameters. The latter are the model's properties that define its characteristics and its learning process, e.g., the depth of a decision tree, the learning rate etc.. Thus, it is clear that hyperparameters can affect significantly the outcome of the training process and they should be chosen carefully. Since the models have multiple hyperparameters, however, not all of them can be tuned since that would be time consuming and computationally expensive. Hence, the developer needs to select the most important hyperparameters to be tuned and the suitable range values for each hyperparameter. This constitutes the more delicate part of the tuning process.

The selection of the final values is automatized with the use of meta–estimators such as the *RandomizedSearchCV()*. This one was selected since it constitutes the best trade–off between low computational time and high tuning performance. It evaluates the model's performance under the use of different hyperparameter values' combinations, which are selected based on the ranges provided by the developer. The number of different combinations that are formed and tested is also selected by the developer. The hyperparameter combination leading to the best

result is subsequently used for the training of the model. Due to the inherent randomness of the process, the best combination may differ between different runs of the algorithm. Nevertheless, the final accuracy of the trained model does not present significant differences. The tables including the hyperparameter values selected for each test case are presented in the next chapter.

6.4.7 Topology change adaptation strategy

The goal of all fault location methods is to be accurate for every possible fault case. This, however, can prove challenging, for certain methodologies such as the sparse measurements or the ML–based ones. Changes in the topology due to a switching event or a grid reconfiguration could affect significantly the performance of these methods. Therefore, a strategy for the adaptation of the proposed ML–based algorithm to potential changes in the grid's topology is also presented in this research. This strategy aims to facilitate the algorithm's retraining by omitting certain parts of the training process and have immediately available the required retraining dataset. It is applicable to changes due to switching events. A switching event may be due to the operation of protection devices or to a scheduled maintenance operation. A permanent grid reconfiguration that is not related to the operation of a switch is not covered by the proposed scheme and may require the repetition of the algorithm's entire training phase, depending on the magnitude of the change.

The application of the proposed topology adaptation strategy assumes that the appropriate data storage and pre–processing have been performed, as described above, for all the potential switch states combinations. In this way the data pertaining to each case are immediately available to the model whenever a switching event occurs. Moreover, the use of a data minimization technique and a tree–based predictive model increase the generalizability of the algorithm, as stated in the previous sections. Hence, the minimization of the problem's size and the hyperparameter tuning do not need to be repeated every time the grid's topology changes. These are only performed once for the default grid topology. Thus, the algorithm is able to adapt fast to a new topology and maintain its high accuracy.

Whenever the algorithm is being employed due to the detection of a fault the status of the switching devices on the grid is checked. Based on that it is concluded whether there is a topology change or not. If no topology change is detected then the default model is being loaded for the prediction of the faulted point. In the case that a topology change has indeed occurred the corresponding dataset is pulled from the database and the prediction model is retrained with the new data. By retraining the model each time a topology change is detected instead of having a model for each case already trained, the method requires less storage space and allows for the renewal of the stored data whenever more recent data become available.

6.5 Conclusions

In this chapter a complete fault location algorithm for active LV grids is presented. The algorithm is based on the use of an ensemble of tree–based ML models, for the prediction of the faulted branch and the exact fault location. In order to provide a turnkey solution the main aspects related to the application of a fault location algorithm, and particularly a ML–based one, are addressed. The performance of ML–based algorithms is mostly defined by the data quality and their management. Therefore, this study emphasizes on the data analysis, collection, storage, pre–processing and minimization. Regarding the latter, two strategies are proposed: one based on the feature selection methodology and one based on the dimensionality reduction methodology. The most noteworthy algorithms within each methodology are then explained and compared; their comparison forms part of the case studies presented in the next chapter. The method is concluded with the presentation of a smart retraining strategy that ensures the algorithm's adaptation to potential topology changes. This is the first ML–based fault location method that provides an effective solution to the data storage, minimization and retraining problems. The method's validation is performed in the next chapter.

7

Fault location: Case studies

The evaluation of the fault location method proposed in chapter 6 was split into two parts based on the followed data minimization technique and the employed predictive model. Case 1 tests the version of the method that utilizes dimensionality reduction techniques and an XGBoost model while case 2 the version of the method that utilizes feature selection techniques and a RF. Apart from some common influencing parameters whose effect was tested in both cases, the two case studies examined different aspects of the method in order to avoid repetition while providing a global evaluation of the fault location method. Once more, due to the lack of real life data the test grid used in both case studies for the generation of the required datasets was a modified version of the CIGRE European LV benchmark [175]. The grids used in the two case studies are almost identical and include only slight differences related to the different test cases.

It should be noted that the CTs mentioned throughout this chapter refer to the algorithm's training phase and not its implementation. The algorithm's implementation times depend on the data collection speed. Once the data have been collected the algorithm instantly returns the result.

7.1 Case study 1

Case study 1 evaluates the performance of the fault location algorithm employing dimensionality reduction for the minimization of the input data and utilizes an XGBoost regressor as the predictive model. The aspects of the method that were analyzed in this case are the dataset composition, the performance of the dimensionality reduction techniques described in chapter 6, the accuracy of the algorithm and the effect of the size of the dataset, the loss of data, the fault resistance and the PV penetration level on it.

7.1.1 Test grid

In this case study the modifications on the benchmark grid pertain only to the addition of PVs and measuring devices in the preexisting topology, as these are illustrated in Fig. 7.1. More specifically, three PVs were added to the first feeder, one to the second and two to the third. The PVs placed in the first feeder have a nominal power of 5kW each, the one in the second feeder has a nominal power of 13kW and the two PVs placed in the third feeder both have a nominal power of 10kW. Furthermore, various meters were added across the grid; these are marked in the grid with a squared M followed by the number of the meter. The meters were added in abundance, almost in every load and node of the grid, in order to provide useful information for the grid's behavior. These were only used to draw conclusions on a theoretical basis. The method's actual measurement requirements are analyzed in section 7.1.7.

7.1.2 Data generation

For the generation of the training and testing datasets various scenarios of the grid's normal and faulty operation were simulated. The parameters that were modified during the simulations were the type, location and resistance of the fault and the PV penetration levels. Table 7.1 presents the simulated value ranges.

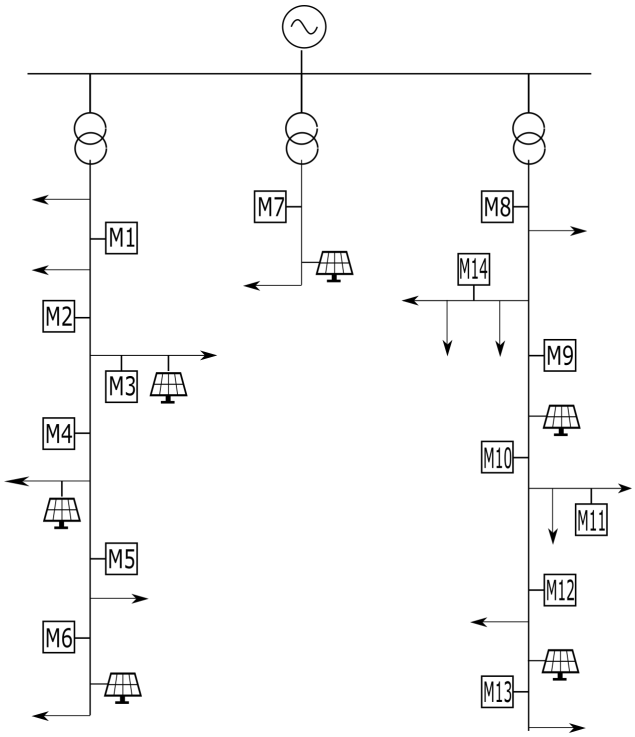

Figure 7.1: Modified CIGRE European LV benchmark used in case study 1.

Table 7.1: Grid element values

Parameters	Number of scenarios	Values
Fault resistance	17	[0, 40]Ω
Fault locations	25	$[35, 315]m$ from the source
Fault types	10	All types of shunt faults
PV generation levels	5 for each feeder	1st branch: 0, 800, 1700, 3000, 5000 W 2nd branch: 0, 4000, 7000, 10000, 13000 W 3rd branch: 0, 2500, 5000, 7500, 10000 W

The generated dataset comprises the modified variables, the node at the end of each faulted branch and the three phase current and voltage phasors measured before and after the fault. The post-fault measure-

ments were recorded within half cycle from the fault's occurrence, before the activation of the protection devices. The exact measurement time is not important and it can be after the first half cycle, as long as the recorded data contain disturbance information. This is a common requirement for all fault location methods. Prior to the fault the grid was set to operate in steady state without transient phenomena. The number of simulation scenarios was 21250. The processor used during the simulations and the execution of the algorithms was an Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz.

In order to make the generated dataset as realistic as possible this was sampled and various examples were randomly discarded, making the final dataset non-homogeneous. The dataset size used for the evaluation of the data and the comparison of the dimensionality reduction techniques after the sampling was randomly set at (10200, 344), where 10200 is the number of examples and 344 is the number of features. The number of features corresponds to the maximum amount collected. The different datasets analyzed in section 7.1.3 had a different (smaller) number of features. For the fault location part, the optimum dataset size is studied in section 7.1.7 as part of the sensitivity analysis of the algorithm.

7.1.3 Dataset evaluation

As commented in chapter 6, the utilization of any dataset, either real or generated, should be accompanied by the analysis of its elements. An effective analysis ensures the quality and validity of the data and subsequently of the test results, points to the suitable techniques to be used, in this case the suitable predictive model and defines the design choices that shape the method. Thus, here a statistical analysis of the generated dataset is presented. This dataset and the respective analysis can be considered representative of the basic variables measured in a small LV grid with RES and different kinds of loads. The analysis examines the normality of the data, the correlation between them and their importance for the predictive process of the utilized model.

Regarding the analyzed data, these include the features as described in the previous section and the target value, which in this case is the distance between the fault and the main feeder. The recorded variables,

i.e. the voltage and current, are split into magnitude and angle values and these are treated as different features. Based on the use of the recorded variables three different datasets were formed. As shown in Table 7.2, the first dataset contains the voltage and current measurements before and after the fault, the second contains only the measurements recorded after the fault and the third the ratio of each variable's values before and after the fault $\Delta I, \Delta V$. These datasets correspond to the ones used more frequently in fault location methods.

The variables' names in the first two datasets contain the letter a or b in the second position, with a denoting a measurement after the fault and b denoting the measurement before the fault. The letter in the third position - or second position in the case of the third dataset - corresponds to the measured phase. All the names include also a number which points to the meter that performed the measurement. Finally, the letters m and a at the end of each name denote whether it is a magnitude or an angle variable respectively.

Regarding the analysis results, it should be noted that due to the large number of features contained in the datasets the corresponding plots illustrate only the features with the highest scores in each case.

Table 7.2: Voltage and current values in each analyzed dataset.

1st dataset	**2nd dataset**	**3rd dataset**
$\overline{I}^b_{ph} = I^b_{ph} < \theta$	$\overline{I}^a_{ph} = I^a_{ph} < \theta$	$\Delta I_{ph} = \frac{\overline{I}^a_{ph}}{\overline{I}^b_{ph}}$
$\overline{V}^b_{ph} = V^b_{ph} < \theta$		
$\overline{I}^a_{ph} = I^a_{ph} < \theta$	$\overline{V}^a_{ph} = V^a_{ph} < \theta$	$\Delta V_{ph} = \frac{\overline{V}^a_{ph}}{\overline{V}^b_{ph}}$
$\overline{V}^a_{ph} = V^a_{ph} < \theta$		

b: value measured before the fault, a: value measured after the fault, ph: each of the three phases.

The first test performed in all three datasets was the Shapiro normality test. As a rule of thumb it is set that a probability of similarity smaller than 0.05 means that the dataset does not follow a normal distribution. This was the case for all three datasets studied here. As discussed in chapter 6, the correlation coefficients computed for datasets

that do not follow Gaussian distribution are the Kendall's coefficient and the Spearman's coefficient. The Pearson's correlation coefficient is also calculated here, but only for validation purposes, as it assumes a normal data distribution.

The first observation that can be made after the calculation of the first two coefficients is in line with the theoretical background and confirms that Kendall's coefficient is smaller than that of Spearman's. The analyses results can be seen in Fig. 7.2, 7.3 and 7.4 and in Fig. 7.5, 7.6 and 7.7. Based on the same figures, it can be concluded that the rank of the features and that of the target value in the first dataset present the strongest correlation of the three. This signifies that the variables in the first dataset have stronger monotonic relationships with the target value. Nevertheless, the different types of faults lead to an overall weak monotonic relationship between the features and the target. During the evaluation of the analysis results the physical side of the features should also be taken into consideration. Here, for example, the high correlation between features such as the currents before the fault and the target value can be misleading. These values do not have a direct physical relationship since the current values before the fault are independent of the fault location. These pre–fault current values can be useful but only when evaluated together with the current values after the fault.

Regarding Pearson's correlation coefficient, as illustrated in Fig. 7.8, 7.9 and 7.10, in a small LV grid such as the one examined, the linear correlation between the features and the target value is rather weak. This is to be expected as the non–linearity of these relations is obvious already from the analytical equations presented in the impedance–based methods and is enhanced by the tree–shaped topology of LV grids and the RES added to it.

Another important metric that can provide a valuable insight to the dataset and the entire predictive process is the features' importance. In Fig. 7.11, 7.12 and 7.13 it can be observed that the third dataset is the one containing the most useful features for the prediction of the target value. This means that the features contained in the third dataset assist the most in the split of a tree–based predictive model. The other two datasets also includes features of high importance, but not as high as those in the third dataset. Taking into account the

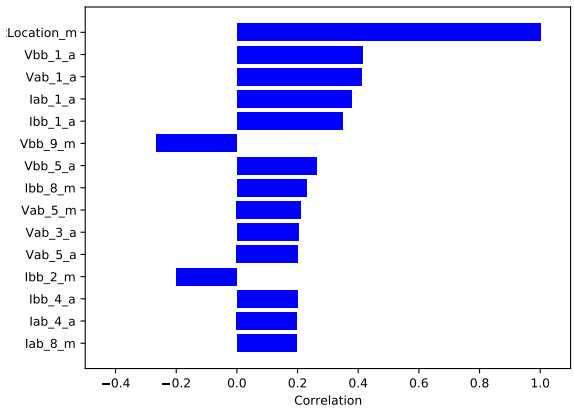

Figure 7.2: Kendall's coefficient calculated for the first dataset [2].

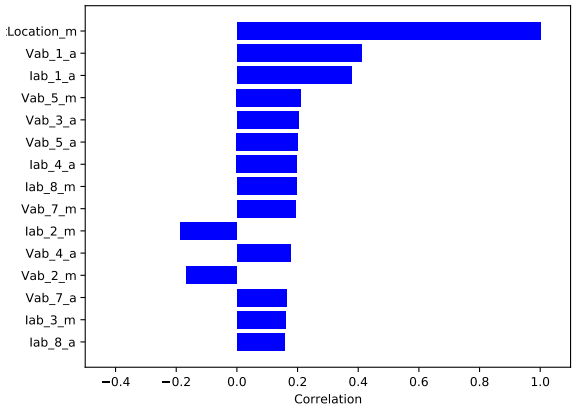

Figure 7.3: Kendall's coefficient calculated for the second dataset [2].

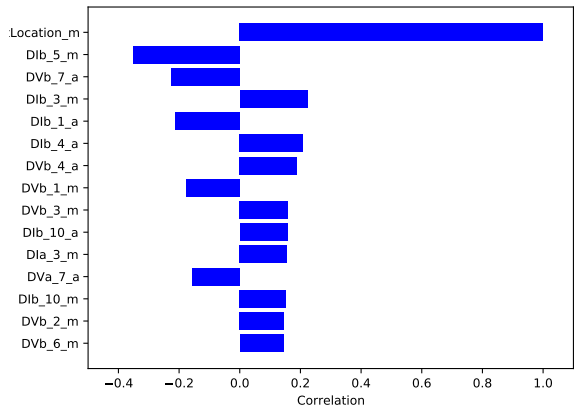

Figure 7.4: Kendall's coefficient calculated for the third dataset [2].

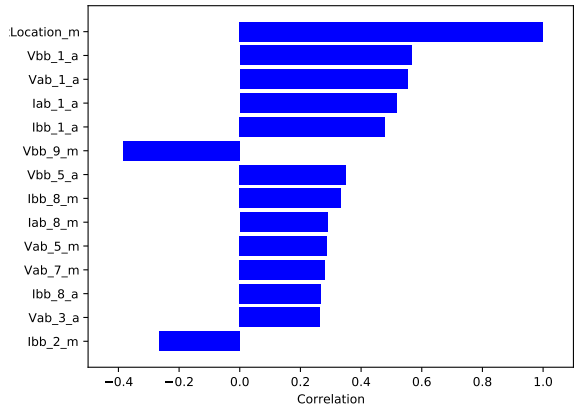

Figure 7.5: Spearman's coefficient calculated for the first dataset [2].

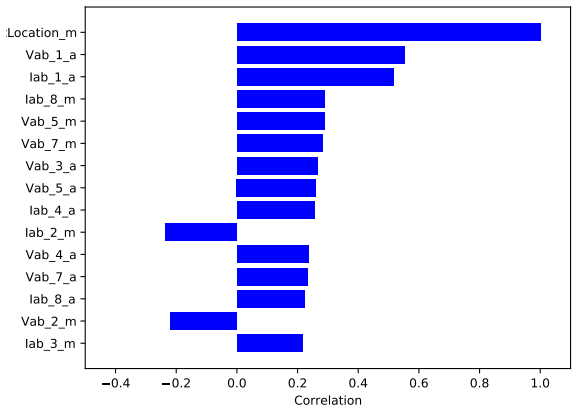

Figure 7.6: Spearman's coefficient calculated for the second dataset [2].

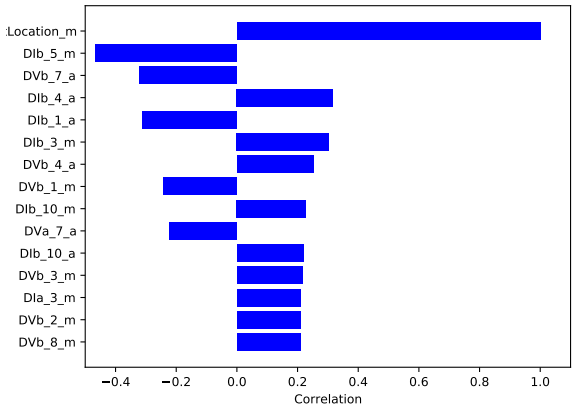

Figure 7.7: Spearman's coefficient calculated for the third dataset [2].

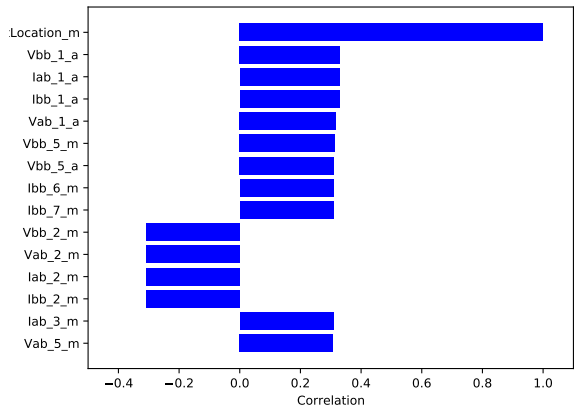

Figure 7.8: Pearson's coefficient calculated for the first dataset [2].

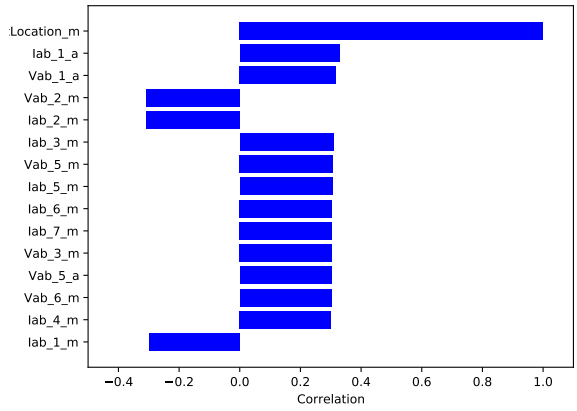

Figure 7.9: Pearson's coefficient calculated for the second dataset [2].

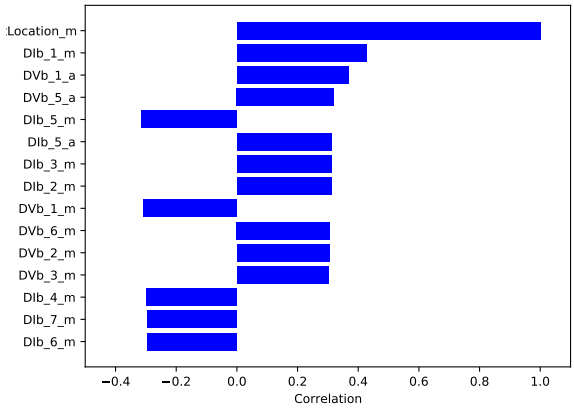

Figure 7.10: Pearson's coefficient calculated for the third dataset [2].

weak correlations between the features and the target value in all the datasets, the features' importance is a crucial criterion for the selection of the best dataset.

After evaluating the correlations, the features' importance and of course the physical relationships between the variables, the dataset that was selected for the training of the ML model was the third one. The aforementioned dataset contains all the information regarding the state of the grid before and after the fault in a compact form and leads to better tree splits, hence more accurate results with less features.

Overall, the analysis results show the difficulty of conventional methods to generalize the complex relations between the voltage/current values and the fault location for different LV grids. On the other hand, it appears that tree–based ML models are capable of providing accurate predictions for these datasets; this needs to be verified by the metrics of the final model. Finally, the data analysis does not only bring out the most informative dataset but also to the most important meters. Both the correlation coefficients and the features' importance point to the same meters as the most informative ones. These are the meters found in the beginning of each feeder and a meter placed in the middle/end of the first feeder. This is a useful observation for the installation of

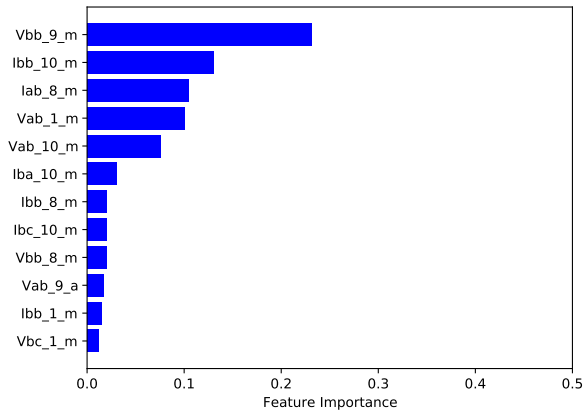

Figure 7.11: Features' importance for the first dataset [2].

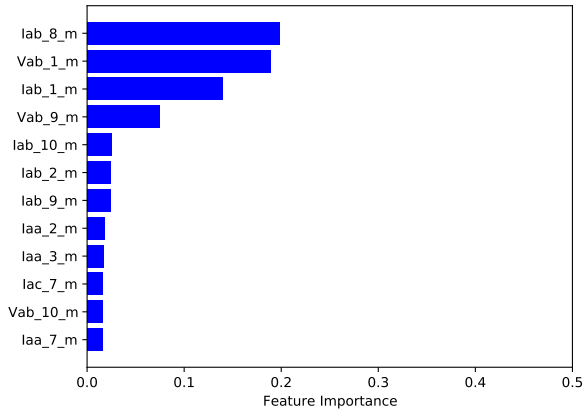

Figure 7.12: Features' importance for the second dataset [2].

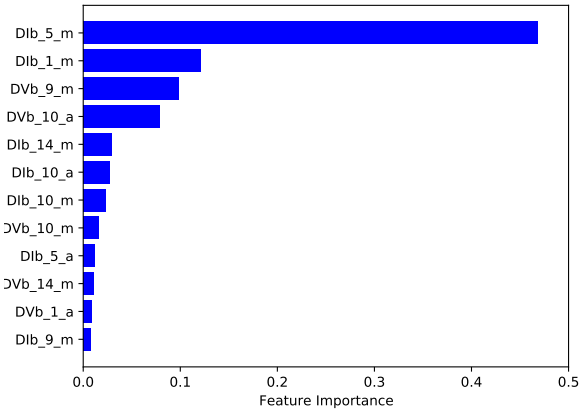

Figure 7.13: Features' importance for the third dataset [2].

new meters and the collection and analysis of only the necessary data.

7.1.4 Data minimization using dimensionality reduction

After the appropriate dataset is selected a good practice is to reduce its content to only the necessary data. In this way both the algorithm's computational speed and generalizability are increased. One of the data minimization approaches presented in chapter 6 is a data management strategy that combines the use of the most important features in their original form and the application of a dimensionality technique on the rest of the dataset. The most important features are selected according to the features' importance analysis presented in the previous section. In this case the number of the selected features was 10. For the selection of the most suitable dimensionality reduction technique the five methods presented in chapter 6 were tested and compared. The criteria used for their comparison were: a) the CT of the dimensionality reduction process, b) the CT of the predictive process with the use of the reduced dataset and without the hyperparameter tuning, and c) the predictive accuracy.

Before applying the dimensionality reduction techniques the size of

the reduced dataset needs to be defined. Similarly to the criteria established for the selection of the dimensionality reduction method, the selection of the new dimensionality space was done based on the optimum trade–off between the CT and the accuracy. Fig. 7.14 illustrates the mean square error (MSE) in relation to the CT for the T–SVD method for a range of [20, 90] dimensions. It can be observed that there is an almost inversely exponential relation between the two variables up until a point where the MSE starts to increase with the increase in the number of features. This indicates the model's overfitting to the training data when too many features are used. Based on these results and in line with the multi–objective optimization process adopted, the size of the reduced dimensionality space was set at 30. Thus, for a 30–dimensional space the performance of each technique is presented in Table 7.3. According to these results, the selected dimensionality reduction technique is the T–SVD. Even though Kernel PCA leads to the lowest MSE, T–SVD combines a comparable error with a much lower execution time.

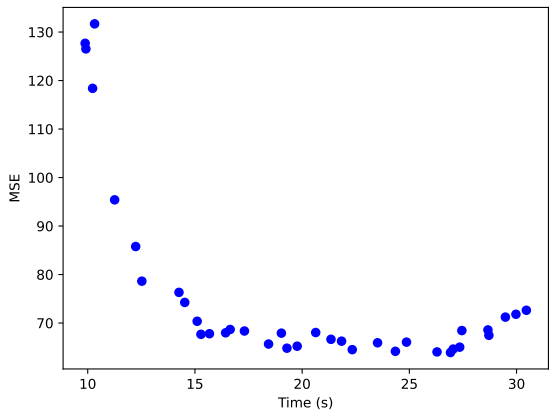

Figure 7.14: The MSE in relation to the CT of the algorithm for a range of [20, 90] dimensions with the use of T–SVD [2].

Table 7.3: Comparative table of the dimensionality reduction methods

Method	CT of dimensionality reduction (s)	MSE (m^2)	Total CT (s)
PCA	0.38	119.25	10.08
KPCA	76.6	87.8	85.8
FastICA	1.3	155.44	11.1
T–SVD	0.18	95.61	11.65
ISOMAP	102.78	696.66	111.7

The application of the proposed data minimization strategy resulted in the reduction of the dataset's features to 40. This mix of original and transformed data not only reduces the method's CT but, as it can be seen in Fig. 7.15, it also leads to a much higher accuracy compared to the use of a higher number of dimensions in the dimensionality reduction process. It manages to compress the most informative parts of the original dataset in less than a third of its size.

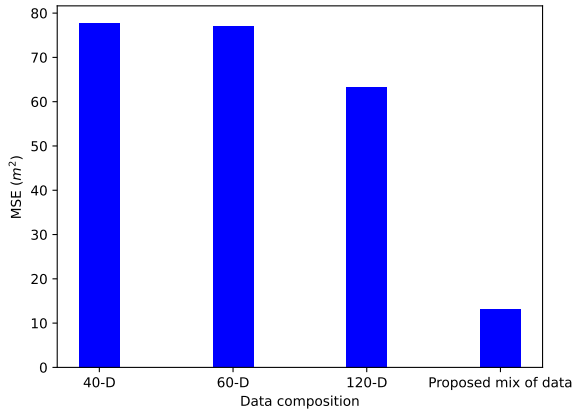

Figure 7.15: The algorithm's MSE in relation to the dataset composition [2]

7.1.5 Hyperparameters

The last step before the training of the algorithm is the tuning of the model's hyperparameters. Table 7.4 presents the final hyperparameter

values that were selected for the training of the XGBoost Model, based on the process described in chapter 6. These values refer to the specific application and the tuning of the hyperparameters should be repeated every time the method is to be applied to a different grid.

Table 7.4: Hyperparameter values

No. of gradient boosting trees	700
Maximum tree depth for base learners	7
Subsample ratio of the training instance	0.7
Min sum of instance weight needed in a child	3
Boosting learning rate	0.1
Subsample ratio of columns when constructing each tree	0.7
Objective	"reg:squarederror"

7.1.6 Fault location accuracy

After the model's training its predictive performance on the test data is evaluated. The results of the test are presented in Table 7.5. The hyperparameter tuning and the use of the mix dataset have a direct impact on the method's MSE which is significantly smaller than the one shown in Table 7.3. Additionally, a further decrease on the algorithm's CT can be observed. As a reference point it should be noted here that the algorithm's CT when trained with the original dataset, without the use of the dimensionality reduction but with the tuning of the hyperparameters, was 1970,1s or 32 min. This is more than twice the CT of the method after the use of dimentionality reduction. Furthermore, the train and test accuracy are both high and almost identical, indicating the lack of overfitting.

Table 7.5: Fault location prediction model results

MSE (m^2)	13.26
MAE (m)	1.69
CT without hyperparameter tuning (s)	11.89
CT with hyperparameter tuning (min)	12.27
Train accuracy (%)	99.9
Test accuracy (%)	99.8

7.1.7 Sensitivity analysis

A fault location method's accuracy can be affected by various parameters. In order to test the its robustness under different scenarios a sensitivity analysis was performed. The studied parameters were chosen based on the characteristics of the fault location problem and the properties of a ML–based algorithm.

Fault distance

The first parameter that was studied was the effect of the fault's location on the algorithm's accuracy. In Fig. 7.16 it can be seen that the Mean Percentage Error (MPE) of the method is less than 6% in all cases. The MPE is defined as follows:

$$MPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{(Y_i - Y_i^*)}{Y_i} \tag{7.1}$$

where Y is the real value of the fault distance, Y^* is the predicted value, $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and n is the total number of examples.

The spikes in the plot correspond to locations whose distance from the feeder is the same but are placed on different branches. The more branches with equally distant points from the feeder, the higher the algorithm's error. This is particularly noticeable for shorter distances where the voltage values are similar in all the branches.

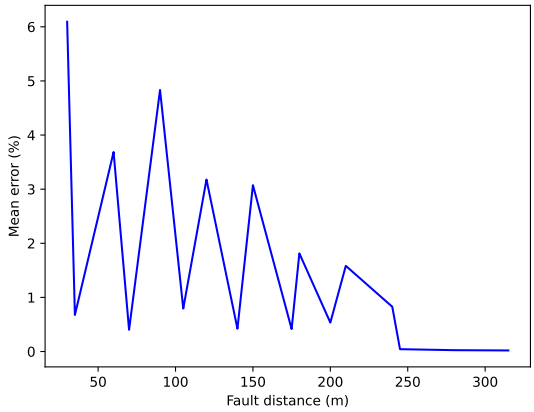

Figure 7.16: MPE of the algorithm in relation to the fault's distance from the feeder [2].

Dataset size

When it comes to the size of the dataset used for the algorithm's training this is determined not only by the number of features, which has been previously analyzed, but also by the number of examples. This is also an important factor affecting the algorithm's predictive capabilities. A training dataset that is too large may lead to overfitting or storage problems while a dataset that is too small may lead to underfitting. Therefore, the dependence of the proposed algorithm on the number of training examples is analyzed here. In Fig. 7.17 the MSE and the training time of the model in relation to the amount of utilized training examples is presented. As expected, the more the utilized examples the lower the MSE and the higher the CT. The MSE may be high when few data are used, however, it drops significantly for more than 6800 examples. This is a small number of examples that is easily collected and stored. Thus, the presented algorithm combines high accuracy with low data and storage requirements. It should be stressed here that the different types of predictive models respond differently to the amount of utilized training data. The results of this sensitivity analysis support the claim that tree–based models perform well even

with smaller datasets.

The optimum dataset size selected for the study is approached as another case of multi-objective optimization, with the target being the balance point between the lowest CT and the lowest MSE. Both parameters are considered almost equally important with CT being given a slightly higher importance factor in order to partly counterbalance its lower numerical values. Specifically, the weights selected were 1.1 for the CT and 0.9 for the MSE. The red line in Fig. 7.17 illustrates the weighted average curve of the CT and the MSE in relation to the number of examples. The minimum of the curve constitutes the optimum point and corresponds to 11900 examples. Out of these 80% was used for the training of the algorithm and the other 20% was used for its testing.

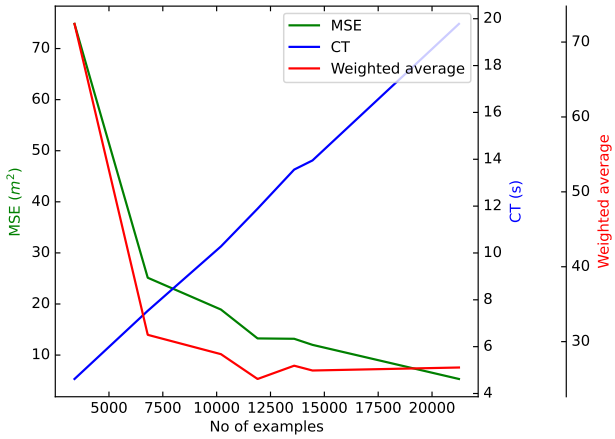

Figure 7.17: The MSE and CT of the algorithm in relation to the number of utilized examples [2].

Loss of data

Another factor that could have an important impact on the performance of a fault location method is the loss of data due to a communication error or the malfunction of one or more measuring devices. For this purpose, this research analyzes the effect of a potential data loss

from one, three or five of the grid's meters. The missing values were replaced with 0. Regarding the failing meters in each scenario, in order to ensure the validity of the analysis the MSE of five failing meter combinations was calculated each time and the mean value of the MSEs was computed as the final result for each scenario. Even though the selection of the failing meters was mostly random, a parameter that needed to be taken into consideration during this process due to the applied data management strategy was the importance of the meters.

Hence, three separate cases were tested regarding the failing meters included in each data loss scenario. The first case explored the possibility that all the failing devices were providing data that were classified as important. These are the data utilized by the algorithm in their original form; the corresponding devices are characterized as primary devices. In the second case it was considered that the missing measurements came from devices providing the less important data, thus characterized as secondary devices. Finally, in the third case the data loss originated from both the primary and secondary devices. Meters 1, 4, 5, 9, 10 were selected as the failing primary devices and meters 2, 6, 8, 13, 14 as the failing secondary devices. In the third case, when the scenario of the three failing meters was tested, one of them was considered to be a primary meter and two secondary meters. In the same case, for the scenario with the five failing meters the ratio primary to secondary devices was 2/3.

As expected, Fig. 7.18 confirms the algorithm's high dependence on the primary measuring devices. Thus, the simultaneous failure of three or more can result in high errors. This is, however, a very improbable scenario that does not characterize the algorithm's average performance. Regarding the other two meter failure cases, the MSE is almost the same as that under normal conditions. Therefore, the test results verify the algorithm's robustness against potential data loss.

Fault resistance

A parameter that affects all fault diagnosis methods and should not be omitted from a related study is the fault resistance. Depending on the consistency of the parts involved in the fault the value of the fault resistance can vary significantly. This is directly reflected in the fault

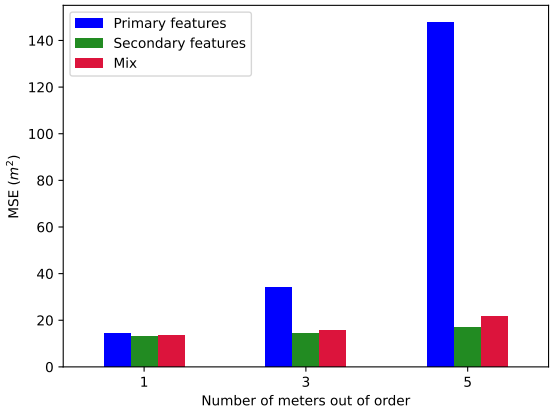

Figure 7.18: The MSE of the algorithm in the case of data loss from one, three or five meters [2].

current measurements, whose value fluctuates according to the different fault resistances. These notable fluctuations can impact the efficiency of fault location methods and limit their generalizability. Thus, in this study a broad range of fault resistances was simulated in order to identify the algorithm's sensitivity to it. More specifically, the tested fault resistances were in the range of [0, 40] Ω. Higher impedance faults are out of this research's scope, therefore, this range was selected as the most representative for detectable faults in LV grids [133].

In Fig. 7.19 the algorithm's response for the different fault resistance value ranges is illustrated. Even though the algorithm's accuracy remains above 99% for the entire analyzed spectrum, it is lower for the lower values of the fault resistance. The reason behind this is the greater variation of the current's values in this range of fault resistances, as it can be observed in Fig. 7.20 which depicts representative current measurements recorded by meter 8. Hence, it is more challenging in this case for the algorithm to distinguish between the different cases and predict the correct target value. Overall, however, the fault resistance's effect on the proposed algorithm can be considered negligible.

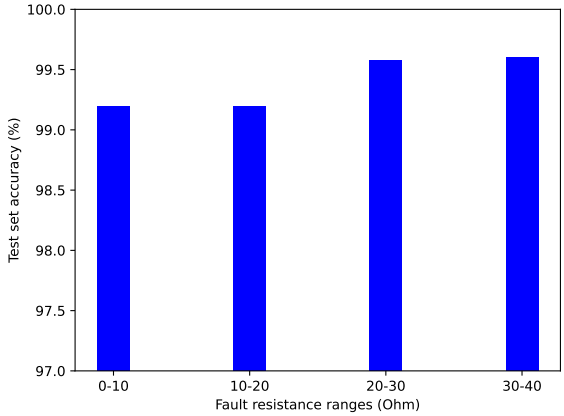

Figure 7.19: The accuracy of the algorithm in relation to the fault resistance value ranges [2].

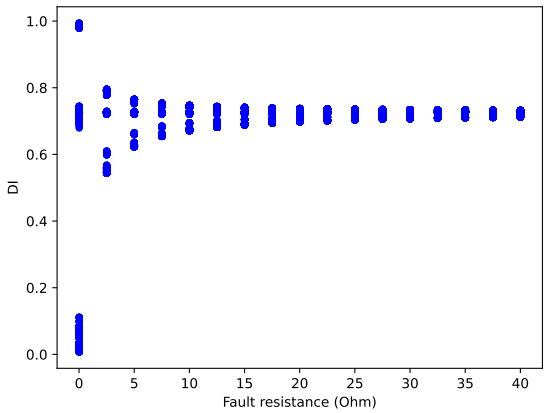

Figure 7.20: The ratio between the current before and after the fault in relation to the fault resistance as measured by meter 8 [2].

PV penetration level

Finally, a major parameter to be considered in all power systems studies in the smart grids era is the integration of RES. The power injected by the RES can also affect the fault current and subsequently the fault location accuracy. The bidirectional power flow can cause phenomena such as the blinding of the protection and measuring devices which alter the values of the grid's variables and can confuse the algorithm. Therefore, in this study the effect of five different PV power generation levels was simulated and analyzed. The generation levels are presented in Table 7.1. As illustrated in Fig. 7.21 there is no clear pattern between the generated power and the algorithm's accuracy. Nonetheless, the consistently high accuracy indicates that the algorithm remains practically unaffected by the various PV penetration levels.

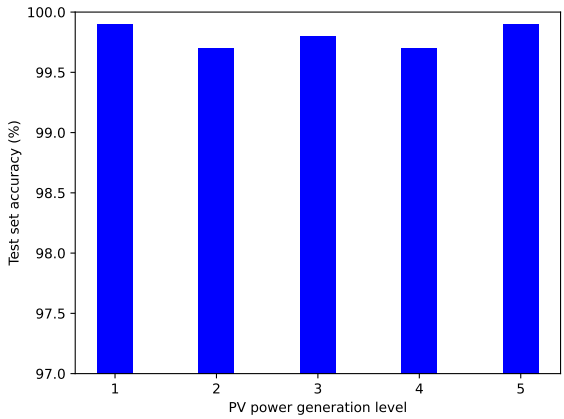

Figure 7.21: The accuracy of the algorithm in relation to the PV power generation levels [2].

7.2 Case study 2

In case study 2 the utilized data minimization technique is the feature selection and the predictive model is a RF. Apart from that, this

case study tests the proposed re–training strategy in case of topology changes and analyzes the load changes and measurement noise effects on the algorithm. Moreover, the difference in the algorithm's performance when trained with phasors or only magnitude values is discussed.

7.2.1 Test grid

In this case study the grid is identical to that of case study 1 with the addition of 4 switching points and one reserve connection between feeders 1 and 3. These are marked with red Xs and a red dashed line respectively in Fig. 7.22. They are accompanied by the identification codes indicating the topology change scenario they are used at. The default grid topology is the one illustrated in black.

7.2.2 Data generation

The simulations performed for the generation of the training and testing datasets for each topology included the modification of the fault's parameters and the grid's elements as in Table 7.1, with the exception of the fault resistance whose range was expanded to a maximum of 100 Ω. Moreover, in this case study the loading levels were also modified and noise was added to the collected data at a later stage. These modifications are discussed in the corresponding parts of the sensitivity analysis in section 7.2.7.

The data generation and collection process is the same as in case study 1. The size of the sampled dataset in this case for the default topology is (13600, 344). The datasets used for the sensitivity analyses tests have a size of (10000, 344) for each level of data contamination and of (23800, 344) for each loading level in order to examine as many potential scenarios as possible and ensure the reliability of the results. The number of features in all the datasets decreased after the application of the feature selection method. The size of the datasets that need to be stored for each topology change is discussed in section 7.2.6. It should be noted here that the dataset used was the one identified as dataset 1 in case study 1. This selection was mandated by the fact that the maximum number of features in their natural form were chosen to be utilized in the testing of the feature selection methods.

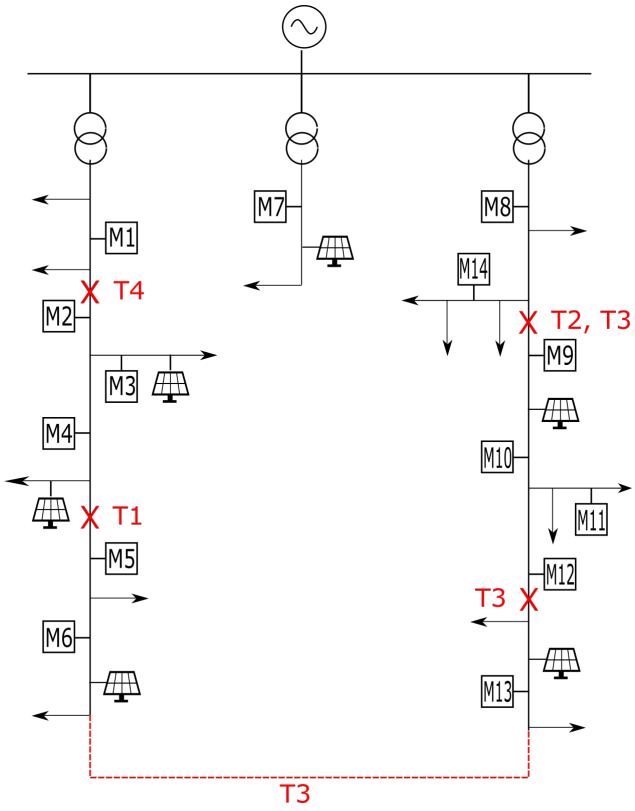

Figure 7.22: Modified CIGRE European LV benchmark used in case study 2.

7.2.3 Magnitude vs phasor utilization

As commented in chapter 6 the measuring devices installed on the grid vary between the different countries and DSOs, meaning that the recorded variables differ as well. The most commonly measured variables are the voltage and current magnitudes. As the installation of smart meters, PMUs and other devices providing synchronized measurements grows, however, the phasors' angles are also available in most cases. Therefore, the performance of the algorithm with and without the use of the angles is tested and compared in this research.

Table 7.6 presents the results of the tests, which verify that the use

of the variables' angles lead to higher accuracy but also to higher CTs. Nevertheless, the algorithm's error in the case when only magnitude values are utilized remains very low. Based on these results and after taking into consideration the research goals of this thesis, the data that were selected to be used for the rest of the study were the three phase current and voltage phasors before and after the fault. In this way, the full potential of the algorithm can be explored.

Table 7.6: Comparison of the model's performance with and without the angles of the V, I included in the dataset

	Dataset with angles	**Dataset without angles**
MSE (m^2)	11.34	23.68
MAE (m)	0.51	1.01
CT without the hyperp. tuning (min)	9.74	4.81

7.2.4 Feature selection method comparison

After the selection of the model's training dataset its minimization is once again implemented. In this case study the SelectFromModel and the Boruta feature selection algorithms are compared. The estimator used for the calculation of the features' importance was in both cases a RF. The criteria for the algorithms' comparison were their CT, the CT of the overall algorithm's training with and without the minimization of features, the number of the selected features and the MSE of the trained model. The last point of comparison is thoroughly analyzed in subsection 7.2.6 as it coincides with the evaluation of the proposed method. Nevertheless, it is presented in Table 7.7 together with the other results for comparative purposes. The table also includes the algorithm's MSE without the deployment of feature selection, as a reference point.

The results confirm its method's theoretical characteristics. More specifically, the Boruta algorithm leads to the method with the highest accuracy, nevertheless due to the complexity and iterative nature of the algorithm, the CT is much higher than that of the SelectFromModel. The SelectFromModel algorithm may lead to a higher MSE, however, it

Table 7.7: Comparison of feature selection methods

	SelectFromModel	**Boruta**	**No feature selection**
CT (s)	23.13	2731	-
No of selected features	23	112	-
MSE (m^2)	11.65	2.29	4.95
ML model's CT (min)	20.78	43.59	100.16

is very fast and selects very few features, thus it significantly decreases the method's data requirements and CT. The same observations can be made also regarding the relation between the SelectFromModel and the algorithm without feature selection. Even though the difference in the MSE is not great, the differences in data requirements and CT are important. Thus, it is concluded that the SelectFromModel algorithm constitutes the best feature selection option and it is the one applied to the training dataset.

SelectFromModel features

The application of the SelectFromModel algorithm has indicated the utilization of 23 features out of the 344 that were originally included in this dataset. Thus, the final size of the default dataset utilized for the training and testing of the algorithm was (13600, 23). The selected features comprise only of current and voltage values both before and after the fault, mostly from phase b. The percentage of selected features corresponding to each measuring device is presented in Fig. 7.23. In accordance with the data analysis presented in case study 1, the most important meters are the ones in the beginning and middle of the branches. Additionally, in Fig. 7.24 it is shown that almost 2/3 of the selected features correspond to magnitude values. This supports the results presented in section 7.2.3 stating that the algorithm's performance remains high also when trained only with magnitude values.

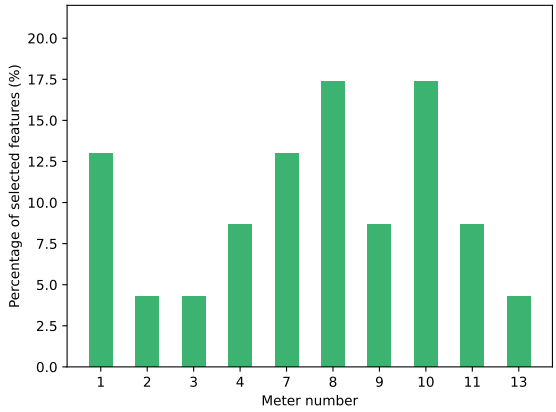

Figure 7.23: The percentage of features coming from each meter that were selected by the feature selection algorithm.

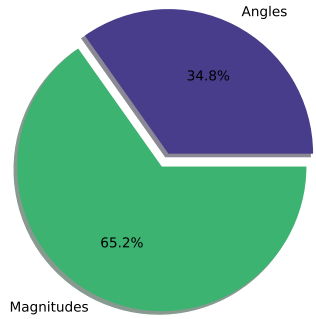

Figure 7.24: The ratio of magnitude vs angle features in the selected features list.

7.2.5 Hyperparameters

Finally, before the presentation of the method's performance, the hyperparameter ranges and the selected values used for the training of the RF regression and classification models are presented in Table 7.8.

Table 7.8: Hyperparameter values

	No. of trees	**Maximum depth of the tree**	**No. of features considered in each split**	**Min No. of samples required to split a node**	**Min No. of samples required at each node**
	[100 : 1200 : 100]	[5 : 30 : 5]	"auto", "sqrt"	[2, 5, 10, 15, 100]	[1, 2, 5, 10]
R	800	30	"sqrt"	2	1
C	100	20	"sqrt"	2	1

7.2.6 Results

After the selection of the appropriate data processing techniques and the suitable training dataset, the method's overall performance is evaluated. This includes the accuracy of the classification model used for the identification of the faulted branch and the accuracy of the regression model used for the calculation of the distance between the faulted point and the main feeder.

Faulted branch classification accuracy

As analyzed in chapter 6, the faulted branch classification is performed by a RF classifier. The target variable is each branch's last node. There were 6 different target values in the studied grid. The model's test results are presented in table 7.9. It can be seen that both the accuracy and the F1 score are particularly high. The excellent predictive capabilities of the model are also verified by the confusion matrix shown in Fig. 7.25.

Table 7.9: Fault classification results

Accuracy (%)	**F1 score (%)**
99.79	99.79

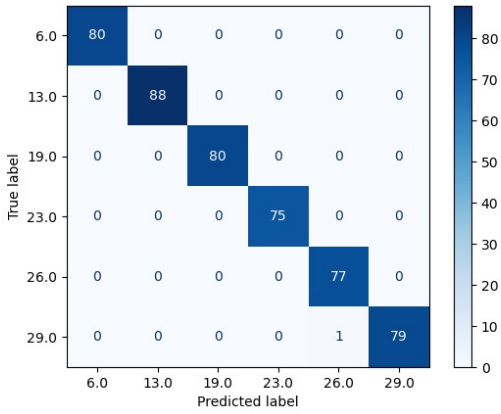

Figure 7.25: Confusion matrix [4]

Fault location accuracy

The regressor used for the location of the faulted point is a RF regressor. The algorithm's performance in the location of faults after the selection of the most informative features is summarized in Table 7.10. It can be observed that thanks to the data management techniques embedded in the algorithm both the error and CT are maintained low.

Table 7.10: Fault location algorithm's performance

MSE (m^2)	**MAE (m)**	R^2	**CT with hyperp. tuning (min)**	**CT without hyperp. tuning (s)**
11.65	0.83	0.9983	20.78	36.69

Topology change scheme – retraining performance

In order to avoid a performance loss during topology changes an efficient retraining scheme was presented in chapter 6. The scheme's testing was divided into two parts. The first part focuses on the retraining data volume and the corresponding CT. As commented also in case study 1, the dataset size affects both the CT and the accuracy of the method. Hence, different dataset sizes were used for the algorithm's retraining after the occurrence of the topology change 1 scenario. This is illustrated in Fig. 7.22 and corresponds to the disconnection of a part of branch 1.

Fig. 7.26 shows the MAE and the CT in relation to the number of examples included in the dataset. As expected once again, the more the utilized examples the lower the MAE and the higher the CT. Similarly to the multi–objective optimization performed in section 7.1.7 the weighted average of the two parameters was calculated pointing to the dataset size that leads to the best trade–off between the MAE and the CT. Based on the results it is concluded that the optimum number of examples that should be contained in each dataset is 7000. This is a small number that can be easily collected and stored. Regarding the re–training scheme's actual CT, it can be observed that this is much lower compared to the one required for the repetition of the whole training process, which includes the time–consuming hyperparameter tuning. It should be noted here that the CTs do not include the time required for the loading of the data into the algorithm.

The second part of the retraining scheme's efficiency evaluation tests the algorithm's accuracy in the case of four different topology changes. All the cases are illustrated in Fig. 7.22; three of them include the disconnection of a part of the grid while one includes the isolation of a part of branch 3 and the re–connection of another part of it to branch 1. Moreover, in order to make sure that the algorithm is able to perform well even in situations were fast retraining is the priority or there is a lack of data, only 5500 examples were included in each retraining dataset instead of 7000. The results for each new topology scenario are presented in Table 7.11 and verify the efficiency of the proposed re–training scheme. Despite the small size of the utilized datasets the error is maintained low and the accuracy high in all the cases, while

the CT is only a few seconds. For comparative reasons it should be stated that the MAE in the case of the topology change 1 without the implementation of the retraining scheme is 113m, while the R^2 has a negative value which indicates the big differences between the default and the changed topology dataset variables. Overall, the incorporation of the proposed re–training scheme maximizes the reliability of the power supply even in extreme scenarios.

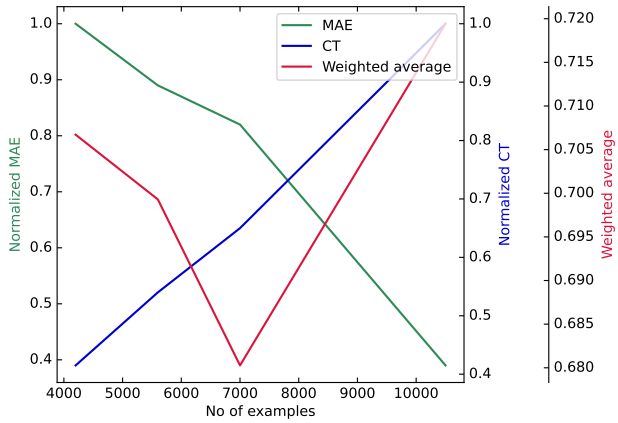

Figure 7.26: The values and weighted average of the MAE and the CT for different amount of utilized examples for Topology 1.

Table 7.11: Retraining scheme performance for the tested topologies

	Topology 1	**Topology 2**	**Topology 3**	**Topology 4**
MSE (m^2)	10.35	0.485	0.957	14.49
MAE (m)	0.76	0.032	0.05	0.74
R^2	0.9975	0.999	0.999	0.997
CT without hyperp. tuning (s)	21.2	9.75	9.56	13.05

7.2.7 Sensitivity analysis

The sensitivity analysis performed in order to verify the robustness of the fault location algorithm tested in case study 2 includes once more the study of the location of the fault, the fault resistance and PV penetration as influencing parameters. Additionally, in order to broaden the validation spectrum of AI–based fault location methods, the effect of load changes and measurement noise on the algorithm's performance is also analyzed.

Location of the fault

As it can be seen in Fig. 7.27, the effect of the fault's location on the algorithm is similar to that observed in case study 1. This is expected since the grid presents the same characteristics and the two algorithm's share similar properties. More specifically, here the mean error is less than 5% for the faults closer to the main feeder, and less than 1% for the rest of the distances.

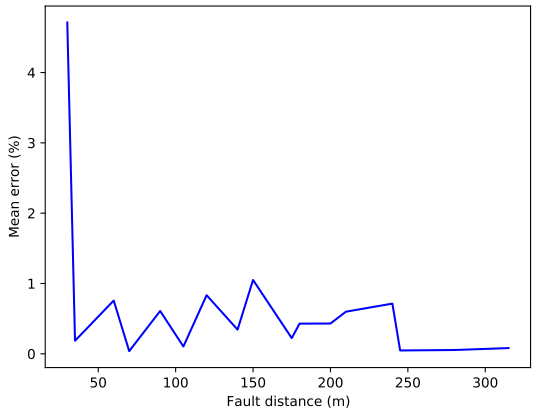

Figure 7.27: PME in relation to the fault distance from the main feeder [4].

Fault resistance

As stressed in section 7.1.7 the effect of the fault resistance on fault location algorithms can be significant. It is an unpredictable parameter whose values can vary notably, therefore, a margin of [0-40] Ω is studied here. In order to analyze the algorithm's behavior in relation to the changes in the fault resistance values, the latter were divided in subsections. Then, the mean test accuracy of the algorithm was calculated for each one of those. As shown in Fig. 7.28, the drop in the accuracy is small as the fault resistance increases, therefore, the algorithm can be considered robust against the fault resistance effect.

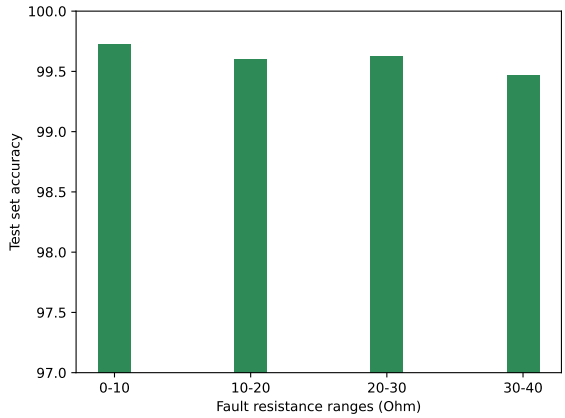

Figure 7.28: Test accuracy of the predictive model in relation to the fault resistance [4].

PV penetration level

In this case study the analysis of the PV penetration level's effect on the algorithm is measured with the use of the MSE instead of the R^2. The reason behind that is that in this case the performed analysis showed that there is a pattern in the relation between the PV penetration level and the MSE. In Fig. 7.29 it can be observed that the higher the PV penetration the bigger the MSE of the model. Nevertheless, the latter

is still exceptionally low and verifies the algorithm's robustness.

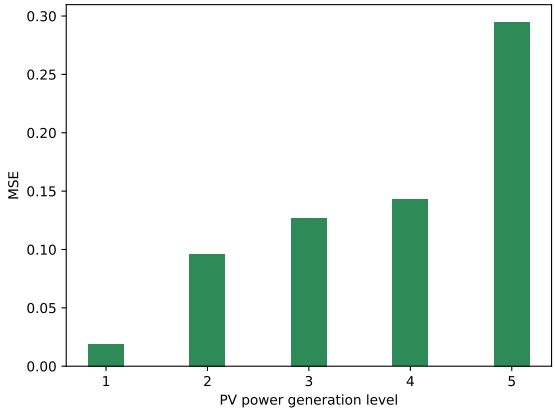

Figure 7.29: MSE of the predictive model in relation to the PV power generation [4].

Load change

Another parameter that could lead to inaccuracies in the location of a fault is the load changes. These directly affect the grid's measured variables and can alter the data patterns learned by the model. In order to test the proposed algorithm's robustness against the effect of the load changes, three different load levels were simulated. These correspond to the 30%, 60% and 80% of the loads' nominal values. The algorithm's accuracy in each case as well as the mean value of the tested cases are presented in Fig. 7.30. As expected, the lower the load level the lower the accuracy of the algorithm, since the measured values diverge more from the originally simulated ones. Nevertheless, the mean accuracy remains high and above 99%.

Measurement noise

Finally, the distortion of the input data due to the noise introduced by the measuring devices is an inevitable part of the data collection. In

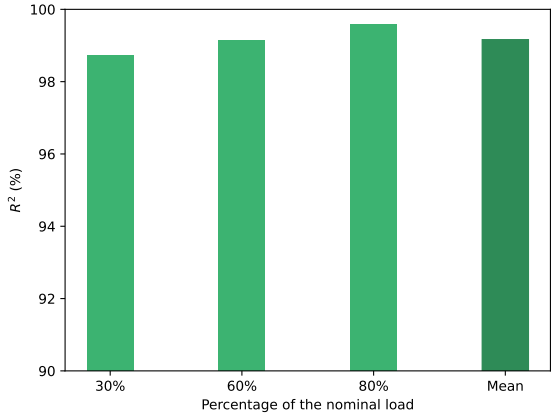

Figure 7.30: Model's test accuracy in relation to the nominal value percentage the grid's loads are operating at.

chapter 6 the incorporation of the measurement noise in the model's training data in case of generated datasets was discussed. Here, the effect of this noise in the algorithm's accuracy is analyzed by contaminating the dataset with 10%, 20%, 40% and 60% noisy data. In order to test the algorithm in the most extreme of cases a $\pm$ 5% noise was added to the measurements. It can be observed that the additional noise in the data is affecting the accuracy of the algorithm, however, it is not causing an error of more than 5.5%.

7.3 Comparison of the possible data minimization and predictive model combinations

The presented fault location method proposes the combination of a data minimization approach with a tree–based predictive model. For this purpose a dimensionality reduction method and a feature selection method were combined with an XGBoost model and a RF respectively. These combinations were made taking into consideration the particular characteristics of the data minimization techniques and the predictive models. The pairings represent the middle cases that are not expected

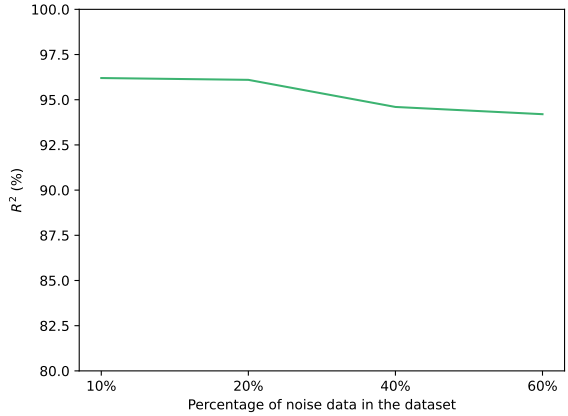

Figure 7.31: Model's test accuracy in relation to the percentage of noisy data included in the dataset.

to lead to either a very high or a very low method performance overall. The goal of this decision was to present a more realistic and generic depiction of the method's performance as well as two viable alternatives.

Nevertheless, the method's accuracy and CT for all the possible combinations was tested and it is presented in Table 7.12. The results verify the hypothesis that the dimensionality reduction–XGBoost and feature selection–RF combinations are both solid solutions that lead to satisfactory results. The best pairing overall, however, is that of the XGBoost model with the SelectFromModel algorithm which combines high speed and accuracy. On the other hand, the worst pairing is that of the dimensionality reduction and the RF. The test results confirm the theoretical comparison presented in chapter 3; the XGBoost model is faster and leads to less overfitting than the RF when used in a fault location application, regrardless of the data minimization technique employed. Furthermore, the SelectFromModel algorithm suits better both models than the T–SVD–based strategy.

Overall, both the XGBoost and the RF models proved to be reliable, flexible and accurate when combined with the appropriate data management techniques, however, the XGBoost is deemed more suitable for

Table 7.12: Comparison of feature selection methods

	RF - SelectFromModel	**XGBoost - SelectFromModel**	**RF - T–SVD**	**XGBoost - T–SVD**
MSE (m^2)	11.65	1.35	206.97	13.26
MAE (m)	0.83	0.49	6.76	1.69
CT with hyperp. tuning (min)	21.01	7.72	29.8	12.27

this application. Moreover, both the T–SVD and the SelectFromModel are capable to decrease the size of the problem without a significant information loss. However, the compression of information by the T–SVD leads to lower predictive accuracy which cannot be counterbalanced by the proposed data management technique that combines original and compressed features. The SelectFromModel algorithm does not require any additional features and leads to more accurate results, thus it is more effective in a fault location algorithm.

7.4 Conclusions

In this chapter two fault location case studies were presented. Their goal was not only to verify the proposed algorithm's accuracy but also to present an insight to the design choices made during the development of a ML–based solution. Thus, each case study tested different sides of the fault location algorithm presented in chapter 6 and proposed solutions that ensure the successful practical implementation of the method under the most common operational scenarios. The first case study evaluated the use of dimensionality reduction techniques for the minimization of the training data and an XGBoost regression model as the predictor. Additionally, it included an all–important data analysis that presented the correlation and importance of the data in different versions of the original dataset. Then, the second case study evaluated the efficiency of feature selection methods in minimizing the training data and the predictive capabilities of a RF regressor. More-

over, it analyzed the algorithm's performance when synchronized measurements are not available and it validated the effectiveness of the model's retraining scheme in case of topology changes. Finally, both case studies included a sensitivity analysis examining the effect of the fault's location, the fault's resistance and the PV's penetration level on the algorithm's accuracy. The first case study also analyzed the effect of the dataset's size and potential data loss on the method. The second case study complemented those results with the evaluation of the load changes' and the measurement noise's impact on the model's performance.

Many conclusions can be drawn from these case studies. First of all, the combined normal and faulty operation dataset generated from such an active LV grid is not normally distributed and has a loose correlation with the location of the fault. The most compact dataset that can be formed by these data is the one containing the difference between the before and after the fault values. Moreover, even though the use of phasors leads to higher location accuracy, the method's error in case only magnitude values are available is not much higher. Additionally, the meters providing the most important measurements are those place in the beginning and the middle of each feeder.

Furthermore, the proposed data management strategies not only improve the model's performance but also facilitate its re–training. Thus, the presented smart re–training strategy is fast and efficient for all switching events with a minimum number of additional examples. Finally, the best predictor–data minimization technique combination is that of the XGBoost and the SelectFromModel algorithm. The overall results verify that the method is highly effective in locating faults and can be easily applied to any active LV grid since it is not dependent to any topology–specific parameters.

8

Conclusions

This thesis presented a complete fault diagnosis solution for distribution grids. The term fault refers to all ten types of shunt faults. Each part of the fault diagnosis process targeted a specific research gap related to it, as those arose from the literature review presented in chapter 2. The steps of the fault diagnosis are three: the detection of a fault, the classification of the type of fault and the location of the faulted point. Most research gaps were identified in the research related to the LV grid, therefore, the study was mainly focus on that part of the grid and, in particular, on active LV grids. The ongoing smart grid transition has posed many challenges to the traditional fault diagnosis techniques, however, it has also provided opportunites for the development of innovative solutions. Thus, taking into consideration the increased observability over the grid that is enabled by the continuous technological advancements, the proposed methods were all data–driven. More specifically, emphasis was placed on the use of ML, whose advanced pattern identification abilities have been verified in multiple applications.

Regarding the first step of the fault diagnosis process, the fault detection method presented in this research focuses on the detection of faults in LV grids with increased fast charging and ultra–fast charging penetration. The problem constitutes a binary classification problem, therefore, the CatBoostClassifier is proposed as the predictive model. It

constitutes a state–of–the–art tree–based boosting model that is highly efficient and easy to train. Moreover, in order to develop a ML–based fault detection algorithm with high accuracy and robustness against the EV charging stochasticity, the algorithm's training is performed with static simulation data. For the generation of the training data two static simulation scenarios were tested and compared, one that ignored all EV charging and one that considered that all EV chargers installed in the grid were simultaneously occupied and operating at nominal power. The algorithm was trained with each dataset and tested on out–of–sample data generated from the simulation of possible intermediate charging and loading states. The tests showed that the algorithm achieved a high detection accuracy in both cases, however, it yielded the best results when trained with the first dataset, i.e. the one generated with no EV charging simulation. Thus, the proposed algorithm can be effectively trained with static simulation data and without requiring the prior knowledge of the number or position of the EV chargers. Hence, it is easily applicable to any active LV grid and combines generalizability with speed and accuracy. Its performance was also tested against the effect of the fault resistance, potential data loss, the PV penetration's level, the dataset's size and the charging stations' occupancy rate.

The second step of the fault diagnosis process, the fault classification, constitutes the least studied part of the process. Therefore, the research was extended to both the MV and the LV grid, with emphasis on the MV. The proposed approach in this case is more traditional and aims to present an updated and automatized version of the threshold–based techniques, which account for the majority of the available fault classification techniques. Thus, after the establishment of some generic criteria to be used for the identification of each type of fault, an algorithm automatizing the criteria's threshold–setting process was presented. The criteria use only three–phase current values before and after the fault. The developed algorithm analyzes the data generated by the grid at hand and proposes suitable thresholds for each criterion. The process is mainly defined by the influence of the fault resistance on the fault current. Thus, it constitutes the first method that is not limited to testing its robustness against this very important parameter but also incorporates its effect on the threshold–setting process. This

is possible by simulating some representative fault scenarios with various fault resistance values. In this way the method's accuracy and adaptability to each specific grid are ensured and facilitated, rendering the proposed fault classification method universal, effective and easily applicable to any conventional grid.

The final step of the fault diagnosis process is the location of the faulted point. The literature review revealed that even though there is an abundance of related papers developed for MV grids, there is a lack of research for LV grids. Hence, the proposed fault location method was developed for and tested on an active LV grid. In order to avoid phenomena such as the multiple location estimation the fault location process has been split into two parts: i) the identification of the faulted branch and ii) the calculation of the distance between the main feeder and the faulted point. Both of these functions are performed by ML models, rendering this an ensemble ML–based method. The faulted branch identification constitutes a multiclass classification problem since the target value is the last node of each branch. Therefore, a RF classifier was used as the predictive model. This is a reliable and easy–to–implement model that has been vastly used and tested, thus it was selected as a fitting model for this simpler part of the method. The tests confirmed its suitability by showcasing its high accuracy.

The calculation of the fault's distance from the main feeder constitutes a regression problem, since the target value is a continuous number representing the distance to the fault. The distance calculation is a complex problem whose result can be easily influenced by multiple parameters. Therefore, two regression tree–based models were tested and compared, a RF and an XGBoost. These represent the traditional and the advanced generation of tree–based models respectively. The performance of ML models though is not only determined by their inherent characteristics but also by the tuning of their hyperparameters and the quality and management of the input data. Thus, in order to boost the models' performance, apart from the necessary hyperparameter tuning, the proposed algorithm includes: i) a smart data storage scheme, which groups the data based on the switches' states during the data collection, ii) a three–stage data pre–processing which includes the scaling, shuffling and splitting of the data into training and testing datasets and iii) a data minimization strategy which reduces the problem's size, and

the model's training time and overfitting. The employed techniques not only increase the algorithm's efficiency but also facilitate the implementation of a fast and effective retraining scheme in the case of topology changes due to switching events. Furthermore, in order to ensure the algorithm's applicability, its measurement requirements were also addressed with the different possibilities being analyzed. To the author's knowledge, this is the first research that provides an all–round ML–based fault location algorithm comprising smart data storage and minimization schemes and an efficient re–training strategy. Finally, it is the first algorithm to present a thorough data analysis that evaluates the correlation between the data that are usually available on the grid such as the voltage and current phasors before and after the fault and the fault distance as well as the importance of these data to the model's predictive process.

A crucial part of the aforementioned data management process is that of the data minimization for which multiple solutions have been developed over the years. In this research two common approaches were utilized and compared, the feature selection and the dimensionality reduction. For each of the two approaches the most popular and/or sophisticated techniques were applied and compared. These are:

- for the feature selection the:
    1. SelectFromModel
    2. Boruta
- for the dimensionality reduction the:
    1. PCA
    2. KPCA
    3. FastICA
    4. T–SVD
    5. ISOMAP

Each data minimization approach was combined with a ML model in order to test the algorithm's performance. The feature selection techniques were paired with the RF regressor and the dimensionality

reduction techniques were paired with the XGBoost regressor. These pairings were made based on the algorithms' expected behavior and aimed to form combinations with similar, satisfactory performances that support the part that is considered less efficient; in the case of the data minimization techniques this is the dimensionality reduction, and in the case of the predictive models this is the RF. In this way two viable solutions are presented. The accuracy and computational speed of the other two possible combinations, however, were also computed in order to identify the best pair overall for this application.

The method's testing was performed with the use of simulation data generated by properly modified versions of the CIGRE European LV grid. The results showed that for the specific application the best–performing feature selection technique was the SelectFromModel algorithm and the best–performing dimensionality reduction technique was the T–SVD. The latter was combined with the use of original features in order to form a more efficient data minimization strategy. Moreover, the best overall data minimization technique–predictive model pairing was that of the SelectFromModel and the XGBoost algorithms. The XGBoost outperformed the RF in both cases leading to lower errors and CTs. Similarly, the SelectFromModel algorithm outperformed the T–SVD in both cases. Finally, the conducted sensitivity analysis showed that the proposed algorithm is robust against the fault's location, the fault resistance, the PV penetration, the dataset's size and potential data loss, the changes in the load and the noise in the measurements.

8.1 Future work

This research has presented solutions for some of the more pressing research gaps in the field of fault diagnosis. Nevertheless, there are more sides of the problem to be explored and as the grid transformation and the technological advancements continue new challenges and opportunities arise. Potential future work topics in the field of fault diagnosis are the following:

1. With regard to the fault detection:
    - Development of a generalizable solution for the detection of high impedance faults in active LV grids.

- Testing of the algorithm with different input data, i.e. different collected measurements.

2. With regard to the fault classification:
   - Testing of the technique on an active grid.
   - Testing of the technique exclusively on a LV grid.
   - Threshold establishment with the use of ML and comparison with the conventional computational techniques.
3. With regard to the fault location:
   - Utilization of a physics–based ML model for the location of faults.
   - Expansion of the method to other types of faults, such as series faults.
   - Testing of the algorithm on high impedance faults.
4. With regard to all the proposed algorithms:
   - Testing of the techniques on weaker grids.
   - Testing on larger grids.
   - Testing on an experimental setting or with the use of real data, if the testings proposed above are performed with simulation data, as is the case in this research.

Bibliography

[1] P. Stefanidou-Voziki, N. Sapountzoglou, B. Raison, and J.L. Dominguez-Garcia. A review of fault location and classification methods in distribution grids. *Electric Power Systems Research*, 209:108031, August 2022. XVII, 4, 8, 9, 10, 11, 17, 34

[2] P. Stefanidou-Voziki, D. Cardoner-Valbuena, R. Villafafila-Robles, and J.L. Dominguez-Garcia. Data analysis and management for optimal application of an advanced ML-based fault location algorithm for low voltage grids. *International Journal of Electrical Power & Energy Systems*, 142:108303, November 2022. XVIII, XIX, XX, 4, 48, 51, 117, 118, 119, 120, 121, 122, 123, 124, 125, 128, 129, 131, 132, 133

[3] Paschalia Stefanidou-Voziki, Cristina Corchero, and Jose Luis Dominguez-Garcia. A Practical Algorithm for Fault Classification in Distribution Grids. In *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*, pages 1–6, June 2020. XIX, 4, 90, 91

[4] Paschalia Stefanidou-Voziki, David Cardoner-Valbuena, Roberto Villafafila-Robles, and Jose Luis Dominguez-Garcia. Feature Selection and Optimization of a ML Fault Location Algorithm for Low Voltage Grids. In *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*, pages 1–6, September 2021. XX, 4, 140, 143, 144, 145

[5] IEEE Guide for Electric Power Distribution Reliability Indices. *IEEE Std 1366-2012 (Revision of IEEE Std 1366-2003)*, pages 1–43, May 2012. 1

[6] A.A. Girgis, C.M. Fallon, and D.L. Lubkeman. A fault location technique for rural distribution feeders. *IEEE Transactions on Industry Applications*, 29(6):1170–1175, December 1993. 2, 14, 19

[7] Jun Zhu, D.L. Lubkeman, and A.A. Girgis. Automated fault location and diagnosis on electric power distribution feeders. *IEEE Transactions on Power Delivery*, 12(2):801–809, April 1997. 2, 19

[8] Y. Liao. Generalized Fault-Location Methods for Overhead Electric Distribution Systems. *IEEE Transactions on Power Delivery*, 26(1):53–64, January 2011. 2, 20

[9] R.H. Salim, M. Resener, A.D. Filomena, K. Rezende Caino de Oliveira, and A.S. Bretas. Extended Fault-Location Formulation for Power Distribution Systems. *IEEE Transactions on Power Delivery*, 24(2):508–516, April 2009. 2, 18, 20

[10] Rajarshi Dutta and S.R. Samantaray. Assessment of impedance based fault locator for AC micro-grid. *Renewable Energy Focus*, 26:1–10, September 2018. 2, 20

[11] A. Borghetti, M. Bosetti, C. A. Nucci, M. Paolone, and A. Abur. Integrated Use of Time-Frequency Wavelet Decompositions for Fault Location in Distribution Networks: Theory and Experimental Validation. *IEEE Transactions on Power Delivery*, 25(4):3139–3146, October 2010. 2, 23

[12] A.M. El-Zonkoly. Fault diagnosis in distribution networks with distributed generation. *Electric Power Systems Research*, 81(7):1482–1490, July 2011. 2, 13, 14, 24

[13] M. Goudarzi, B. Vahidi, R.A. Naghizadeh, and S.H. Hosseinian. Improved fault location algorithm for radial distribution systems with discrete and continuous wavelet analysis. *International*

Journal of Electrical Power & Energy Systems, 67:423–430, May 2015. 2, 14, 24

[14] D. Thukaram, H. P. Khincha, and H. P. Vijaynarasimha. Artificial neural network and support vector Machine approach for locating faults in radial distribution systems. *IEEE Transactions on Power Delivery*, 20(2):710–721, April 2005. 2, 11, 26, 28, 106

[15] Hadi Zayandehroodi, Azah Mohamed, Hussain Shareef, and Masoud Farhoodnea. A Novel Neural Network and Backtracking Based Protection Coordination Scheme for Distribution System with Distributed Generation. *International Journal of Electrical Power & Energy Systems*, 43(1):868–879, December 2012. 2, 26, 27

[16] M. Majidi, M. Etezadi-Amoli, and M. Sami Fadali. A Novel Method for Single and Simultaneous Fault Location in Distribution Networks. *IEEE Transactions on Power Systems*, 30(6):3368–3376, November 2015. 2

[17] M. U. Usman, J. Ospina, and M. O. Faruque. Fault Classification and Location Identification in a Smart Distribution Network Using ANN. In *2018 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–6, August 2018. 2, 11, 26, 27

[18] Fernanda C. L. Trindade, Walmir Freitas, and Jose C. M. Vieira. Fault Location in Distribution Systems Based on Smart Feeder Meters. *IEEE Transactions on Power Delivery*, 29(1):251–260, February 2014. 2, 30

[19] M. Majidi, A. Arabali, and M. Etezadi-Amoli. Fault Location in Distribution Networks by Compressive Sensing. *IEEE Transactions on Power Delivery*, 30(4):1761–1769, August 2015. 2, 26, 28, 31

[20] A. Bahmanyar, A. Estebsari, E. Pons, E. Patti, S. Jamali, E. Bompard, and A. Acquaviva. Fast fault location for fast restoration of smart electrical distribution grids. In *2016 IEEE International Smart Cities Conference (ISC2)*, pages 1–6, September 2016. 2, 30

[21] Angel Silos-Sanchez, Roberto Villafafila-Robles, and Pau Lloret-Gallego. Novel fault location algorithm for meshed distribution networks with DERs. *Electric Power Systems Research*, 181:106182, April 2020. 2, 31, 32

[22] Lei Ye, Dahai You, Xianggen Yin, Ke Wang, and Junchun Wu. An improved fault-location method for distribution system using wavelets and support vector regression. *International Journal of Electrical Power & Energy Systems*, 55:467–472, February 2014. 2, 33

[23] Fernanda C. L. Trindade and Walmir Freitas. Low Voltage Zones to Support Fault Location in Distribution Systems With Smart Meters. *IEEE Transactions on Smart Grid*, 8(6):2765–2774, November 2017. 2, 11, 33, 37

[24] Ramón Perez, Carmen Vásquez, and Amelec Viloria. An intelligent strategy for faults location in distribution networks with distributed generation. *Journal of Intelligent & Fuzzy Systems*, 36(2):1627–1637, March 2019. 2, 11, 33

[25] Turan Gönen. *Electric Power Distribution Engineering*. 2014. 2

[26] Geng Niu, Long Zhou, Wei Pei, and Zhiping Qi. A novel fault location and recognition method for low voltage active distribution network. In *2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT)*, pages 876–881, Changsha, China, November 2015. IEEE. 3, 36, 37

[27] Kongming Sun, Qing Chen, and Zhanjun Gao. An Automatic Faulted Line Section Location Method for Electric Power Distribution Systems Based on Multisource Information. *IEEE Transactions on Power Delivery*, 31(4):1542–1551, August 2016. 3, 37

[28] Nikolaos Sapountzoglou, Jesus Lago, and Bertrand Raison. Fault diagnosis in low voltage smart distribution grids using gradient boosting trees. *Electric Power Systems Research*, 182:106254, May 2020. 3, 35, 36, 38

[29] Jen-Hao Teng. A Direct Approach for Distribution System Load Flow Solutions. *IEEE Transactions on Power Delivery*, 18(3):882–887, July 2003. 7

[30] Keith Malmedal, Ben Kroposki, and P. K. Sen. Distributed Energy Resources and Renewable Energy in Distribution Systems: Protection Considerations and Penetration Levels. In *2008 IEEE Industry Applications Society Annual Meeting*, pages 1–8, October 2008. ISSN: 0197-2618. 7

[31] Abouzar Estebsari, Edoardo Patti, and Luca Barbierato. Fault Detection, Isolation and Restoration Test Platform Based on Smart Grid Architecture Model Using Intenet-of-Things Approaches. In *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I CPS Europe)*, pages 1–5, June 2018. 8

[32] Christian Andre Andresen, Bendik Nybakk Torsæter, Hallvar Haugdal, and Kjetil Uhlen. Fault Detection and Prediction in Smart Grids. In *2018 IEEE 9th International Workshop on Applied Measurements for Power Systems (AMPS)*, pages 1–6, September 2018. ISSN: 2475-2304. 8

[33] Fadi M. AlAlamat, Eyad A. Feilat, and Mohammed A. Hajahmed. New Distance Protection Scheme for PV Microgrids. In *2020 6th IEEE International Energy Conference (ENERGYCon)*, pages 668–673, September 2020. 9

[34] Whei-Min Lin, Chin-Der Yang, Jia-Hong Lin, and Ming-Tong Tsay. A fault classification method by RBF neural network with OLS learning procedure. *IEEE Transactions on Power Delivery*, 16(4):473–477, October 2001. 9

[35] Joe-Air Jiang, Cheng-Long Chuang, Yung-Chung Wang, Chih-Hung Hung, Jiing-Yi Wang, Chien-Hsing Lee, and Ying-Tung Hsiao. A Hybrid Framework for Fault Detection, Classification, and Location—Part I: Concept, Structure, and Method-

ology. *IEEE Transactions on Power Delivery*, 26(3):1988–1998, July 2011. 9

[36] K. M. Silva, K.M.C. Dantas, B. A. Souza, N.S.D. Brito, F. B. Costa, and J.A.C.B. Silva. Haar Wavelet-Based Method for Fast Fault Classification in Transmission Lines. In *2006 IEEE/PES Transmission Distribution Conference and Exposition: Latin America*, pages 1–5, August 2006. 9

[37] P. Rajaraman, N.A. Sundaravaradan, Rounak Meyur, M. Jaya Bharatha Reddy, and D.K. Mohanta. Fault Classification in Transmission Lines Using Wavelet Multiresolution Analysis. *IEEE Potentials*, 35(1):38–44, January 2016. 9

[38] Zhengyou He, Sheng Lin, Yujia Deng, Xiaopeng Li, and Qingquan Qian. A rough membership neural network approach for fault classification in transmission lines. *International Journal of Electrical Power & Energy Systems*, 61:429–439, October 2014. 9

[39] Yilmaz Aslan. An alternative approach to fault location on power distribution feeders with embedded remote-end power generation using artificial neural networks. *Electrical Engineering*, 94(3):125–134, September 2012. 11, 26, 27

[40] F. Dehghani, F. Khodnia, and E. Dehghan. Fault location of unbalanced power distribution feeder with distributed generation using neural networks. *CIRED - Open Access Proceedings Journal*, 2017(1):1134–1137, 2017. 11, 26, 27

[41] Dabit Sonoda, A.C. Zambroni de Souza, and Paulo Márcio da Silveira. Fault identification based on artificial immunological systems. *Electric Power Systems Research*, 156:24–34, March 2018. 11, 29

[42] P. Janik and T. Lobos. Automated Classification of Power-Quality Disturbances Using SVM and RBF Networks. *IEEE Transactions on Power Delivery*, 21(3):1663–1669, July 2006. 11

[43] Ali Ghaemi, Amin Safari, Hadi Afsharirad, and Hossein Shayeghi. Accuracy enhance of fault classification and location in a smart

distribution network based on stacked ensemble learning. *Electric Power Systems Research*, 205:107766, April 2022. 12, 29

[44] B. Das. Fuzzy logic-based fault-type identification in unbalanced radial power distribution system. *IEEE Transactions on Power Delivery*, 21(1):278–285, January 2006. 12

[45] Wen-Hui Chen, Chih-Wen Liu, and Men-Shen Tsai. On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method. *IEEE Transactions on Power Delivery*, 15(2):710–717, April 2000. 12

[46] J. Mora-Florez, V. Barrera-Nuez, and G. Carrillo-Caicedo. Fault Location in Power Distribution Systems Using a Learning Algorithm for Multivariable Data Analysis. *IEEE Transactions on Power Delivery*, 22(3):1715–1721, July 2007. 12

[47] Mostafa Gilanifar, Hui Wang, Jose Cordova, Eren Erman Ozguven, Thomas I. Strasser, and Reza Arghandeh. Fault classification in power distribution systems based on limited labeled data using multi-task latent structure learning. *Sustainable Cities and Society*, 73:103094, October 2021. 12

[48] Zhongjian Kang, Aina Tian, and Yanyan Feng. A New Method for Fault Type Identification Based on HHT and Neural Network in Distribution Network. In Wenjiang Du, editor, *Informatics and Management Science IV*, volume 207, pages 187–194. Springer London, London, 2013. 12

[49] A. Ngaopitakkul, C. Pothisarn, S. Bunjongjit, and B. Suechoey. DWT and RBF Neural Networks Algorithm for Identifying the Fault Types in Underground Cable. In *TENCON 2011 - 2011 IEEE Region 10 Conference*, pages 1379–1382, November 2011. 12

[50] Majid Jamil, Rajveer Singh, and Sanjeev Kumar Sharma. Fault identification in electrical power distribution system using combined discrete wavelet transform and fuzzy logic. *Journal of Electrical Systems and Information Technology*, 2(2):257–267, September 2015. 12

[51] Ali Rafinia and Jamal Moshtagh. A new approach to fault location in three-phase underground distribution system using combination of wavelet analysis with ANN and FLS. *International Journal of Electrical Power & Energy Systems*, 55:261–274, February 2014. 12, 33

[52] J. Zhang, Z.Y. He, S. Lin, Y.B. Zhang, and Q.Q. Qian. An ANFIS-based fault classification approach in power distribution system. *International Journal of Electrical Power & Energy Systems*, 49:243–252, July 2013. 12, 13

[53] L. Sousa Martins, J. F. Martins, V. Fernao Pires, and C. M. Alegria. The application of neural networks and Clarke-Concordia transformation in fault location on distribution power systems. In *IEEE/PES Transmission and Distribution Conference and Exhibition*, volume 3, pages 2091–2095 vol.3, October 2002. 12, 13, 33

[54] L. S. Martins, J. F. Martins, C. M. Alegria, and V. F. Pires. A network distribution power system fault location based on neural eigenvalue algorithm. In *2003 IEEE Bologna Power Tech Conference Proceedings,*, volume 2, pages 6 pp. Vol.2–, June 2003. 12, 13, 33

[55] S. M. Torabi. Fault location and classification in distribution systems using clark transformation and neural network. In *16th Electrical Power Distribution Conference*, pages 1–8, April 2011. 12, 13, 33

[56] Eyada A. Alanzi, Mahmoud A. Younis, and Azrul Mohd Ariffin. Detection of faulted phase type in distribution systems based on one end voltage measurement. *International Journal of Electrical Power & Energy Systems*, 54:288–292, January 2014. 12, 13

[57] R.H. Salim, K. de Oliveira, A.D. Filomena, M. Resener, and A.S. Bretas. Hybrid Fault Diagnosis Scheme Implementation for Power Distribution Systems Automation. *IEEE Transactions on Power Delivery*, 23(4):1846–1856, October 2008. 13, 14, 33

[58] U. D. Dwivedi, S. N. Singh, and S. C. Srivastava. A wavelet based approach for classification and location of faults in distribution systems. In *2008 Annual IEEE India Conference*, volume 2, pages 488–493, December 2008. 13, 22, 23

[59] Damir Novosel, David Hart, Yi Hu, and Jorma Myllymaki. System for locating faults and estimating fault resistance in distribution networks with tapped loads, November 1998. 14, 19, 77, 78, 86, 89

[60] H. Nezami and F. Dehghani. A new fault location technique on radial distribution systems using artificial neural network. In *22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013)*, pages 0375–0375, Stockholm, Sweden, 2013. Institution of Engineering and Technology. 14, 26, 27

[61] J. Coser, D. T. do Vale, and J. G. Rolim. Design and Training of Artificial Neural Networks for Locating Low Current Faults in Distribution Systems. In *2007 International Conference on Intelligent Systems Applications to Power Systems*, pages 1–6, November 2007. 14, 26, 27

[62] Yuan Liao. A novel method for locating faults on distribution systems. *Electric Power Systems Research*, 117:21–26, December 2014. 14, 20

[63] S. A. M. Javadian, A. M. Nasrabadi, M. Haghifam, and J. Rezvantalab. Determining fault's type and accurate location in distribution systems with DG using MLP Neural networks. In *2009 International Conference on Clean Electrical Power*, pages 284–289, June 2009. 14, 26, 27

[64] Teke Gush, Syed Basit Ali Bukhari, Raza Haider, Samuel Admasie, Yun-Sik Oh, Gyu-Jung Cho, and Chul-Hwan Kim. Fault detection and location in a microgrid using mathematical morphology and recursive least square methods. *International Journal of Electrical Power & Energy Systems*, 102:324–331, November 2018. 14, 31

[65] S. Mallat and W.L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, March 1992. 15

[66] Mohamed A. Gabr, Doaa K. Ibrahim, Eman S. Ahmed, and Mahmoud I. Gilany. A new impedance-based fault location scheme for overhead unbalanced radial distribution networks. *Electric Power Systems Research*, 142:153–162, January 2017. 19, 21

[67] A.A. Girgis, D.G. Hart, and W.L. Peterson. A new fault location technique for two- and three-terminal lines. *IEEE Transactions on Power Delivery*, 7(1):98–107, January 1992. 19

[68] Myeon-Song Choi, Seung-Jae Lee, Duck-Su Lee, and Bo-Gun Jin. A new fault location algorithm using direct circuit analysis for distribution systems. *IEEE Transactions on Power Delivery*, 19(1):35–41, January 2004. 19

[69] M. Choi, S. Lee, S. Lim, D. Lee, and X. Yang. A Direct Three-Phase Circuit Analysis-Based Fault Location for Line-to-Line Fault. *IEEE Transactions on Power Delivery*, 22(4):2541–2547, October 2007. 19

[70] R. Das, M.S. Sachdev, and T.S. Sidhu. A technique for estimating locations of shunt faults on distribution lines. In *IEEE WESCANEX 95. Communications, Power, and Computing. Conference Proceedings*, volume 1, pages 6–11, Winnipeg, Man., Canada, 1995. IEEE. 20

[71] S.-J. Lee, M.-S. Choi, S.-H. Kang, B.-G. Jin, D.-S. Lee, B.-S. Ahn, N.-S. Yoon, H.-Y. Kim, and S.-B. Wee. An Intelligent and Efficient Fault Location and Diagnosis Scheme for Radial Distribution Systems. *IEEE Transactions on Power Delivery*, 19(2):524–532, April 2004. 20

[72] E.C. Senger, G. Manassero, C. Goldemberg, and E.L. Pellini. Automated Fault Location System for Primary Distribution Networks. *IEEE Transactions on Power Delivery*, 20(2):1332–1340, April 2005. 20

[73] Wanjing Xiu and Yuan Liao. Novel fault location methods for ungrounded radial distribution systems using measurements at substation. *Electric Power Systems Research*, 106:95–100, January 2014. 20

[74] A.S. Bretas, C. Orozco-Henao, J. Marín-Quintero, O.D. Montoya, W. Gil-González, and N.G. Bretas. Microgrids physics model-based fault location formulation: Analytic-based distributed energy resources effect compensation. *Electric Power Systems Research*, 195:107178, June 2021. 20

[75] Shamam Fadhil Alwash, Vigna K. Ramachandaramurthy, and Nadarajah Mithulananthan. Fault-Location Scheme for Power Distribution System with Distributed Generation. *IEEE Transactions on Power Delivery*, 30(3):1187–1195, June 2015. 20

[76] Alireza Bahmanyar and Sadegh Jamali. Fault location in active distribution networks using non-synchronized measurements. *International Journal of Electrical Power & Energy Systems*, 93:451–458, December 2017. 20

[77] F.M. Aboshady, D.W.P. Thomas, and Mark Sumner. A new single end wideband impedance based fault location scheme for distribution systems. *Electric Power Systems Research*, 173:263–270, August 2019. 20

[78] Rahman Dashti and Javad Sadeh. Accuracy improvement of impedance-based fault location method for power distribution network using distributed-parameter line model. *International Transactions on Electrical Energy Systems*, 24(3):318–334, 2014. 20

[79] R.H. Salim, K.C.O. Salim, and A.S. Bretas. Further improvements on impedance-based fault location for power distribution systems. *IET Generation, Transmission & Distribution*, 5(4):467, 2011. 20

[80] Hassan Nouri and Mohsen Mohammadi Alamuti. Comprehensive Distribution Network Fault Location Using the Dis-

tributed Parameter Model. *IEEE Transactions on Power Delivery*, 26(4):2154–2162, October 2011. 20

[81] Haizhu Yang, Xiangyang Liu, Yiming Guo, and Peng Zhang. Fault Location of Active Distribution Networks Based on the Golden Section Method. *Mathematical Problems in Engineering*, 2020:1–9, February 2020. 20

[82] S. Das, N. Karnik, and S. Santoso. Distribution Fault-Locating Algorithms Using Current Only. *IEEE Transactions on Power Delivery*, 27(3):1144–1153, July 2012. 21

[83] Charalampos G Arsoniadis. A voltage-based fault location algorithm for medium voltage active distribution systems. *Electric Power Systems Research*, page 11, 2021. 21

[84] F.H. Magnago and A. Abur. A new fault location technique for radial distribution systems based on high frequency signals. In *199 IEEE Power Engineering Society Summer Meeting. Conference Proceedings (Cat. No.99CH36364)*, volume 1, pages 426–431, Edmonton, Alta., Canada, 1999. IEEE. 22, 23

[85] Javad Sadeh, Ehsan Bakhshizadeh, and Rasoul Kazemzadeh. A new fault location algorithm for radial distribution systems using modal analysis. *International Journal of Electrical Power & Energy Systems*, 45(1):271–278, February 2013. 22, 23

[86] H. Hizman, P.A. Crossley, P.F. Gale, and G. Bryson. Fault section identification and location on a distribution feeder using travelling waves. In *IEEE Power Engineering Society Summer Meeting,*, pages 1107–1112, Chicago, IL, USA, 2002. IEEE. 22, 23

[87] A. Borghetti, S. Corsi, C.A. Nucci, M. Paolone, L. Peretto, and R. Tinarelli. On the use of continuous-wavelet transform for fault location in distribution power systems. *International Journal of Electrical Power & Energy Systems*, 28(9):608–617, November 2006. 23, 25

[88] A. Borghetti, M. Bosetti, M. Di Silvestro, C.A. Nucci, and M. Paolone. Continuous-Wavelet Transform for Fault Location

in Distribution Power Networks: Definition of Mother Wavelets Inferred From Fault Originated Transients. *IEEE Transactions on Power Systems*, 23(2):380–388, May 2008. 23

[89] Rui Liang, Guoqing Fu, Xueyuan Zhu, and Xue Xue. Fault location based on single terminal travelling wave analysis in radial distribution network. *International Journal of Electrical Power & Energy Systems*, 66:160–165, March 2015. 24

[90] N. I. Elkalashy, N. A. Sabiha, and M. Lehtonen. Earth Fault Distance Estimation Using Active Traveling Waves in Energized-Compensated MV Networks. *IEEE Transactions on Power Delivery*, 30(2):836–843, April 2015. 24

[91] D.W.P. Thomas, R.J.O. Carvalho, and E.T. Pereira. Fault location in distribution systems based on traveling waves. In *2003 IEEE Bologna Power Tech Conference Proceedings,*, volume 2, pages 468–472, Bologna, Italy, 2003. IEEE. 24

[92] D. W. P. Thomas, Ricardo J. O. Carvalho, Elisete T. Pereira, and Christos Christopoulos. Field trial of fault location on a distribution system using high frequency transients. In *2005 IEEE Russia Power Tech*, pages 1–7, St. Petersburg, Russia, June 2005. IEEE. 24

[93] H. Nouri, Chun Wang, and T. Davies. An accurate fault location technique for distribution lines with tapped loads using wavelet transform. In *2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502)*, volume vol.3, page 4, Porto, Portugal, 2001. IEEE. 24

[94] S. Robson, A. Haddad, and H. Griffiths. Fault Location on Branched Networks Using a Multiended Approach. *IEEE Transactions on Power Delivery*, 29(4):1955–1963, August 2014. 24

[95] R. Razzaghi, G. Lugrin, H. M. Manesh, C. Romero, M. Paolone, and F. Rachidi. An Efficient Method Based on the Electromagnetic Time Reversal to Locate Faults in Power Networks. *IEEE Transactions on Power Delivery*, 28(3):1663–1673, July 2013. 25

[96] J. C. S. Souza, M. A. P. Rodrigues, M. T. Schilling, and M. B. Do Coutto Filho. Fault location in electrical power systems using intelligent systems techniques. *IEEE Transactions on Power Delivery*, 16(1):59–67, January 2001. 26, 27

[97] Meshal A Al-shaher, Manar M Sabry, and Ahmad S Saleh. Fault location in multi-ring distribution network using artificial neural network. *Electric Power Systems Research*, page 6, 2003. 26, 27

[98] Patrick E. Farias, Adriano Peres de Morais, Jean Pereira Rossini, and Ghendy Cardoso. Non-linear high impedance fault distance estimation in power distribution systems: A continually online-trained neural network approach. *Electric Power Systems Research*, 157:20–28, April 2018. 26

[99] Z. Galijasevic and A. Abur. Fault location using voltage measurements. *IEEE Transactions on Power Delivery*, 17(2):441–445, April 2002. 26, 28, 30

[100] J. J. Mora, G. Carrillo, and L. Perez. Fault Location in Power Distribution Systems Using ANFIS Nets and Current Patterns. In *2006 IEEE/PES Transmission Distribution Conference and Exposition: Latin America*, pages 1–6, August 2006. 26, 28

[101] Prashant P. Bedekar, Sudhir R. Bhide, and Vijay S. Kale. Fault section estimation in power system using Hebb's rule and continuous genetic algorithm. *International Journal of Electrical Power & Energy Systems*, 33(3):457–465, March 2011. 26, 28

[102] Q. Jin and R. Ju. Fault Location for Distribution Network Based on Genetic Algorithm and Stage Treatment. In *2012 Spring Congress on Engineering and Technology*, pages 1–4, May 2012. 26, 28

[103] Hatice Okumus and Fatih M. Nuroglu. A random forest-based approach for fault location detection in distribution systems. *Electrical Engineering*, August 2020. 26, 28, 47

[104] Zakaria El Mrabet, Niroop Sugunaraj, Prakash Ranganathan, and Shrirang Abhyankar. Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems. *Sensors*, 22(2):458, January 2022. 26, 28

[105] Kunjin Chen, Jun Hu, Yu Zhang, Zhanqing Yu, and Jinliang He. Fault Location in Power Distribution Systems via Deep Graph Convolutional Networks. *IEEE Journal on Selected Areas in Communications*, 38(1):119–131, January 2020. 28

[106] Guomin Luo, Yingjie Tan, Meng Li, Mengxiao Cheng, Yanmei Liu, and Jinghan He. Stacked Auto-Encoder-Based Fault Location in Distribution Network. *IEEE Access*, 8:28043–28053, 2020. 28

[107] Xin Shi, Robert Qiu, Zenan Ling, Fan Yang, Haosen Yang, and Xing He. Spatio-Temporal Correlation Analysis of Online Monitoring Data for Anomaly Detection and Location in Distribution Networks. *IEEE Transactions on Smart Grid*, 11(2):995–1006, March 2020. 29

[108] Mladen Kezunovic. Smart Fault Location for Smart Grids. *IEEE Transactions on Smart Grid*, 2(1):11–22, March 2011. 30

[109] R. A. F. Pereira, L. G. W. da Silva, M. Kezunovic, and J. R. S. Mantovani. Improved Fault Location on Distribution Feeders Based on Matching During-Fault Voltage Sags. *IEEE Transactions on Power Delivery*, 24(2):852–862, April 2009. 30

[110] Hazlie Mokhlis and Haiyu Li. Non-linear representation of voltage sag profiles for fault location in distribution networks. *International Journal of Electrical Power & Energy Systems*, 33(1):124–130, January 2011. 30

[111] S. Lotfifard, M. Kezunovic, and M. J. Mousavi. Voltage Sag Data Utilization for Distribution Fault Location. *IEEE Transactions on Power Delivery*, 26(2):1239–1246, April 2011. 30

[112] A. Teninge, C. Pajot, B. Raison, and D. Picault. Voltage profile analysis for fault distance estimation in distribution network. In *2015 IEEE Eindhoven PowerTech*, pages 1–5, June 2015. 30

[113] S. M. Brahma. Fault Location in Power Distribution System With Penetration of Distributed Generation. *IEEE Transactions on Power Delivery*, 26(3):1545–1553, July 2011. 31

[114] Cristian Grajales-Espinal, Juan Mora-Flórez, and Sandra Pérez-Londoño. Advanced fault location strategy for modern power distribution systems based on phase and sequence components and the minimum fault reactance concept. *Electric Power Systems Research*, 140:933–941, November 2016. 31

[115] Aleksandar Janjic and Lazar Velimirovic. Integrated fault location and isolation strategy in distribution networks using Markov decision process. *Electric Power Systems Research*, 180:106172, March 2020. 31

[116] Liang Rui, Peng Nan, Yang Zhi, and Firuz Zare. A novel single-phase-to-earth fault location method for distribution network based on zero-sequence components distribution characteristics. *International Journal of Electrical Power & Energy Systems*, 102:11–22, November 2018. 31, 32

[117] Penghui Liu and Chun Huang. Detecting Single-Phase-to-Ground Fault Event and Identifying Faulty Feeder in Neutral Ineffectively Grounded Distribution System. *IEEE Transactions on Power Delivery*, 33(5):2265–2273, October 2018. 31, 32

[118] M. Pignati, L. Zanni, P. Romano, R. Cherkaoui, and M. Paolone. Fault Detection and Faulted Line Identification in Active Distribution Networks Using Synchrophasors-Based Real-Time State Estimation. *IEEE Transactions on Power Delivery*, 32(1):381–392, February 2017. 31, 32

[119] Fan Chunju, K. K. Li, W. L. Chan, Yu Weiyong, and Zhang Zhaoning. Application of Wavelet Fuzzy Neural Network in Locating Single Line to Ground Fault (SLG) in Distribution Lines.

International Journal of Electrical Power & Energy Systems, 29(6):497–503, July 2007. 33

[120] S. Hongchun, W. Xu, X. Qi, W. Qinjin, and T. Xincui. A fault location method of traveling wave for distribution network with only two-phase current transformer using artificial neutral network. In *2010 3rd International Congress on Image and Signal Processing*, volume 6, pages 2942–2945, October 2010. 33

[121] M. Pourahmadi-Nakhli and A. A. Safavi. Path Characteristic Frequency-Based Fault Locating in Radial Distribution Systems Using Wavelets and Neural Networks. *IEEE Transactions on Power Delivery*, 26(2):772–781, April 2011. 33

[122] Sadegh Jamali, Siavash Ranjbar, and Alireza Bahmanyar. Identification of faulted line section in microgrids using data mining method based on feature discretisation. *International Transactions on Electrical Energy Systems*, 30(6):e12353, 2020. 33

[123] J A Momoh, L G Dias, and N Laird. An Implementation of a Hybrid Intelligent Tool for Distribution System Fault Diagnosis. page 6. 33

[124] Nan Peng, Rui Liang, Guanhua Wang, Peng Sun, Chunyu Chen, and Tianyu Hou. Edge Computing-Based Fault Location in Distribution Networks by Using Asynchronous Transient Amplitudes at Limited Nodes. *IEEE Transactions on Smart Grid*, 12(1):574–588, January 2021. 33

[125] J. Ren, S. S. Venkata, and E. Sortomme. An Accurate Synchrophasor Based Fault Location Method for Emerging Distribution Systems. *IEEE Transactions on Power Delivery*, 29(1):297–298, February 2014. 33

[126] A. Estebsari, E. Pons, E. Bompard, A. Bahmanyar, and S. Jamali. An improved fault location method for distribution networks exploiting emerging LV smart meters. In *2016 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)*, pages 1–6, June 2016. 33

[127] Hamid Mirshekali, Rahman Dashti, Ahmad Keshavarz, Amin J. Torabi, and Hamid Reza Shaker. A Novel Fault Location Methodology for Smart Distribution Networks. *IEEE Transactions on Smart Grid*, 12(2):1277–1288, March 2021. 33, 71

[128] D.S. Gazzana, G.D. Ferreira, A.S. Bretas, A.L. Bettiol, A. Carniato, L.F.N. Passos, A.H. Ferreira, and J.E.M. Silva. An integrated technique for fault location and section identification in distribution systems. *Electric Power Systems Research*, 115:65–73, October 2014. 33

[129] Yimai Dong, Ce Zheng, and Mladen Kezunovic. Enhancing Accuracy While Reducing Computation Complexity for Voltage-Sag-Based Distribution Fault Location. *IEEE Transactions on Power Delivery*, 28(2):1202–1212, April 2013. 33

[130] Chaoqun Zhu, Qing Chen, Zhanjun Gao, Tingting Bo, Pu Zhao, and Yi Zhu. A new fault location method for distribution networks using multi-source information. In *2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 1–5, Brisbane, Australia, November 2015. IEEE. 33

[131] S. Jamali and A. Bahmanyar. A new fault location method for distribution networks using sparse measurements. *International Journal of Electrical Power & Energy Systems*, 81:459–468, October 2016. 33

[132] Felix Glinka, Nicolas Schulte, R. Bertram, Armin Schnettler, and Mitja Koprivšek. Solutions for blinding of protection in today's and future German LV grids with high inverter penetration – simulative and experimental analysis. *The Journal of Engineering*, 2018(15):1256–1260, 2018. 35

[133] Nikolaos Sapountzoglou, Bertrand Raison, and Nuno Silva. Fault Detection and Localization in LV Smart Grids. In *2019 IEEE Milan PowerTech*, pages 1–6, Milan, Italy, June 2019. IEEE. 35, 36, 131

[134] Mehdi Shafiei, Houman Pezeshki, Gerard Ledwich, and Ghavameddin Nourbakhsh. Fault Detection in LV Distribution

Networks Based on Augmented Complex Kalman Filter. In *2019 29th Australasian Universities Power Engineering Conference (AUPEC)*, pages 1–5, November 2019. ISSN: 2474-1507. 35

[135] Nuno Silva, Francisco Basadre, Paulo Rodrigues, Mario Serafim Nunes, Antonio Grilo, Augusto Casaca, Francisco Melo, and Luis Gaspar. Fault detection and location in Low Voltage grids based on distributed monitoring. In *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6, Leuven, Belgium, April 2016. IEEE. 35, 37

[136] M. Nunes, A. Grilo, A. Casaca, N. Silva, F. Basadre, P. Rodrigues, F. Melo, and L. Gaspar. Fault Detection and Location in Low Voltage Grids Based on RF-Mesh Sensor Networks. In *CIRED Workshop 2016*, pages 143 (4 .)–143 (4 .), Helsinki, Finland, 2016. Institution of Engineering and Technology. 35, 37

[137] Nikolaos Sapountzoglou, Jesus Lago, Bart De Schutter, and Bertrand Raison. A generalizable and sensor-independent deep learning method for fault detection and location in low-voltage distribution grids. *Applied Energy*, 276:115299, October 2020. 35, 38, 93

[138] Neelesh Yadav and Narsa Reddy Tummuru. A Real-Time Resistance Based Fault Detection Technique For Zonal Type Low-Voltage DC Microgrid Applications. *IEEE Transactions on Industry Applications*, 56(6):6815–6824, November 2020. 35

[139] Iman Sadeghkhani, Mohamad Esmail Hamedani Golshan, Ali Mehrizi-Sani, Josep M. Guerrero, and Abbas Ketabi. Transient Monitoring Function–Based Fault Detection for Inverter-Interfaced Microgrids. *IEEE Transactions on Smart Grid*, 9(3):2097–2107, May 2018. 35

[140] Bhaskar Patnaik, Manohar Mishra, Ramesh C. Bansal, and Ranjan K. Jena. MODWT-XGBoost based smart energy solution for fault detection and classification in a smart microgrid. *Applied Energy*, 285:116457, March 2021. 35, 36

[141] James J. Q. Yu, Yunhe Hou, Albert Y. S. Lam, and Victor O. K. Li. Intelligent Fault Detection Scheme for Microgrids With Wavelet-Based Deep Neural Networks. *IEEE Transactions on Smart Grid*, 10(2):1694–1703, March 2019. 35, 36

[142] G. A. Orcajo, J. M. Cano, M. G. Melero, M. F. Cabanas, C. H. Rojas, J. F. Pedrayes, and J. G. Norniella. Diagnosis of Electrical Distribution Network Short Circuits Based on Voltage Park's Vector. *IEEE Transactions on Power Delivery*, 27(4):1964–1972, October 2012. 36, 37

[143] S. Navaneethan, J. J. Soraghan, W. H. Siew, F. McPherson, and P. F. Gale. Automatic fault location for underground low voltage distribution networks. *IEEE Transactions on Power Delivery*, 16(2):346–351, April 2001. 37

[144] W Siew, John Soraghan, Martin Stewart, David Fisher, David Fraser, PSI-Electronics Ltd, and Muhammad Asif. INTELLIGENT FAULT LOCATION FOR LOW VOLTAGE DISTRIBUTION NETWORKS. (0327):4, 2007. 37

[145] A. M. Pasdar, Y. Sozer, and I. Husain. Detecting and Locating Faulty Nodes in Smart Grids Based on High Frequency Signal Injection. *IEEE Transactions on Smart Grid*, 4(2):1067–1075, June 2013. 37

[146] K. Jia, Z. Ren, T. Bi, and Q. Yang. Ground Fault Location Using the Low-Voltage-Side Recorded Data in Distribution Systems. *IEEE Transactions on Industry Applications*, 51(6):4994–5001, November 2015. 37

[147] Mohsen Mohammadi Alamuti, Hassan Nouri, Rade M. Ciric, and Vladimir Terzija. Intermittent Fault Location in Distribution Feeders. *IEEE Transactions on Power Delivery*, 27(1):96–103, January 2012. 37

[148] Lefeng Cheng and Tao Yu. A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems. *International Journal of Energy Research*, 43(6):1928–1973, 2019. 37

[149] Muhammad Sohail Ibrahim, Wei Dong, and Qiang Yang. Machine learning driven smart electric power systems: Current trends and new perspectives. *Applied Energy*, 272:115237, August 2020. 37

[150] L. Souto, J. Meléndez, and S. Herraiz. Fault Location in Low Voltage Smart Grids Based on Similarity Criteria in the Principal Component Subspace. In *2020 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, February 2020. 38

[151] Eduardo F. Ferreira and J. Dionísio Barros. Faults Monitoring System in the Electric Power Grid of Medium Voltage. *Procedia Computer Science*, 130:696–703, 2018. 41

[152] Dávid Raisz and János Gönczi. Fault location methods at compensated MV networks. In *2014 49th International Universities Power Engineering Conference (UPEC)*, pages 1–5, September 2014. 41

[153] Juan José Mora Flórez. Localización de faltas en sistemas de distribución de energía eléctrica usando métodos basados en el modelo y métodos basados en el conocimiento. page 151, 2006. 41

[154] Shahram Kazemi. Reliability Evaluation of Smart Distribution Grids. page 150. 41

[155] Hahn Tram. Technical and operation considerations in using Smart Metering for outage management. In *2008 IEEE/PES Transmission and Distribution Conference and Exposition*, pages 1–3, April 2008. ISSN: 2160-8563. 41

[156] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Routledge, Boca Raton, October 2017. 47

[157] Muhammad Waseem Ahmad, Monjur Mourshed, and Yacine Rezgui. Trees vs Neurons: Comparison between random forest

and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147:77–89, July 2017. 49

[158] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Ensemble Learning. In *The Elements of Statistical Learning*, pages 605–624. Springer New York, New York, NY, 2009. 50

[159] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. *arXiv:1706.09516 [cs]*, January 2019. arXiv: 1706.09516 version: 5. 50, 59

[160] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM. 50

[161] Abdullahi Ibrahim, Ridwan Raheem, Muhammed Muhammed, Rabiat Abdulaziz, and Saheed Ganiyu. Comparison of the CatBoost Classifier with other Machine Learning Methods. *International Journal of Advanced Computer Science and Applications*, 11:11, December 2020. 50, 59

[162] Weizeng Wang, Yuliang Shi, Gaofan Lyu, and Wanghua Deng. Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform. *DEStech Transactions on Computer Science and Engineering*, 0(aiea), 2017. 51

[163] Raza Abid Abbasi, Nadeem Javaid, Muhammad Nauman Javid Ghuman, Zahoor Ali Khan, Shujat Ur Rehman, and Amanullah. Short Term Load Forecasting Using XGBoost. In Leonard Barolli, Makoto Takizawa, Fatos Xhafa, and Tomoya Enokido, editors, *Web, Artificial Intelligence and Network Applications*, Advances in Intelligent Systems and Computing, pages 1120–1131, Cham, 2019. Springer International Publishing. 51

[164] Wu Yucong and Wang Bo. Research on EA-Xgboost Hybrid Model for Building Energy Prediction. *Journal of Physics: Conference Series*, 1518:012082, April 2020. 51

[165] Global EV Outlook 2021. Technical report, International Energy Agency, April 2021. 55

[166] J. A. P. Lopes, F. J. Soares, and P. M. R. Almeida. Integration of Electric Vehicles in the Electric Power System. *Proceedings of the IEEE*, 99(1):168–183, January 2011. 56

[167] Sulabh Sachan and Nadia Adnan. Stochastic charging of electric vehicles in smart power distribution grids. *Sustainable Cities and Society*, 40:91–100, July 2018. 56

[168] J. R. Pillai, P. Thøgersen, J. Møller, and B. Bak-Jensen. Integration of Electric Vehicles in low voltage Danish distribution grids. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–8, July 2012. 56

[169] Linni Jian, Yanchong Zheng, Xinping Xiao, and C.C. Chan. Optimal scheduling for vehicle-to-grid operation with stochastic connection of plug-in electric vehicles to smart grid. *Applied Energy*, 146:150–161, May 2015. 56

[170] Niels Leemput, Frederik Geth, Juan Van Roy, Jeroen Büscher, and Johan Driesen. Reactive power support in residential LV distribution grids through electric vehicle charging. *Sustainable Energy, Grids and Networks*, 3:24–35, September 2015. 56

[171] V. A. Katić, A. M, Stanisavljević, B. P. Dumnić, and B. P. Popadić. Impact of V2G operation of electric vehicle chargers on distribution grid during voltage dips. In *IEEE EUROCON 2019 - 18th International Conference on Smart Technologies*, pages 1–6, July 2019. 56

[172] Antonio Zecchino and Mattia Marinelli. Analytical assessment of voltage support via reactive power from new electric vehicles supply equipment in radial distribution grids with voltage-dependent loads. *International Journal of Electrical Power & Energy Systems*, 97:17–27, April 2018. 56

[173] E. Akhavan-Rezai, M. F. Shaaban, E. F. El-Saadany, and A. Zidan. Uncoordinated charging impacts of electric vehicles on electric distribution grids: Normal and fast charging comparison. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–7, July 2012. 56

[174] Niels Leemput, Frederik Geth, Juan Van Roy, Pol Olivella-Rosell, Johan Driesen, and Andreas Sumper. MV and LV Residential Grid Impact of Combined Slow and Fast Charging of Electric Vehicles. *Energies*, 8(3):1760–1783, March 2015. 56

[175] Conseil international des grands réseaux électriques Comité d'études C6 and International Council on Large Electric Systems. *Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources: Task Force C6.04*. [Brochures thématiques]. CIGRÉ, 2014. 62, 111

[176] Chen Tsai-Hsiang and Liao Rih-Neng. Analysis of Charging Demand of Electric Vehicles in Residential Area. pages 26–30. Atlantis Press, August 2013. 64

[177] Jairo Quirós-Tortós, Alejandro Navarro Espinosa, Luis F. Ochoa, and Tim Butler. Statistical Representation of EV Charging: Real Data Analysis and Applications. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7, June 2018. 64

[178] Su Su, Yong Hu, Tiantian Yang, Shidan Wang, Ziqi Liu, Xiangxiang Wei, Mingchao Xia, Yutaka Ota, and Koji Yamashita. Research on an Electric Vehicle Owner-Friendly Charging Strategy Using Photovoltaic Generation at Office Sites in Major Chinese Cities. *Energies*, 11(2):421, February 2018. 64

[179] M. Kezunovic, P. Spasojevic, C.W. Fromen, and D.R. Sevcik. An expert system for transmission substation event analysis. *IEEE Transactions on Power Delivery*, 8(4):1942–1949, October 1993. 77, 86, 89

[180] Panagiotis D. Diamantoulakis, Vasileios M. Kapinas, and George K. Karagiannidis. Big Data Analytics for Dynamic En-

ergy Management in Smart Grids. *Big Data Research*, 2(3):94–101, September 2015. 98

[181] S. S. Shapiro and M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4):591–611, 1965. 100

[182] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*, pages 1–4. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 100

[183] Alexei Stepanov. On the Kendall Correlation Coefficient. *arXiv:1507.01427 [math, stat]*, July 2015. arXiv: 1507.01427. 100

[184] Clark Wissler. The Spearman Correlation Formula. *Science*, September 1905. 100

[185] sklearn.feature_selection.SelectFromModel — scikit-learn 0.24.2 documentation. 104

[186] Miron Kursa and Witold Rudnicki. Feature Selection with Boruta Package. *Journal of Statistical Software*, 36:1–13, September 2010. 104

[187] Mahdis Amiri, Hamid Reza Pourghasemi, Gholam Abbas Ghanbarian, and Sayed Fakhreddin Afzali. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. *Geoderma*, 340:55–69, April 2019. 104

[188] Behzad Najafi, Monica Depalo, Fabio Rinaldi, and Reza Arghandeh. Building characterization through smart meter data analytics: Determination of the most influential temporal and importance-in-prediction based features. *Energy and Buildings*, 234:110671, March 2021. 104

[189] Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459, 2010. 106

[190] Joshua B. Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing Systems 10*, pages 682–688. MIT Press, 1998. 106

[191] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999. 106

[192] Tony F. Chan and Per Christian Hansen. Computing Truncated Singular Value Decomposition Least Squares Solutions by Rank Revealing QR-Factorizations. *SIAM Journal on Scientific and Statistical Computing*, 11(3):519–530, May 1990. 106

[193] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, December 2000. 106

A

List of Publications

In this chapter, the list of publications, both journal and conference papers, derived from the development of the thesis are presented.

A.1 Journal articles

[**J1**] P. Stefanidou–Voziki, N. Sapountzoglou, B. Raison, and J.L. Dominguez Garcia. A review of fault location and classification methods in distribution grids. Electric Power Systems Research, 209:108031, August 2022. XIII, 8, 13, 29, 30

[**J2**] P. Stefanidou–Voziki, D. Cardoner Valbuena, R. Villafafila Robles, and J.L. Dominguez Garcia. Data analysis and management for optimal application of an advanced ML-based fault location algorithm for low voltage grids. International Journal of Electrical Power & Energy Systems, 142:108303, November 2022. XIII, XV, 42, 44, 104, 105, 106, 107, 108, 109, 110, 111, 112, 114, 116, 117, 118, 119

[**J3**] P. Stefanidou–Voziki, D. Cardoner Valbuena, R. Villafafila Robles, and J.L. Dominguez Garcia. A Practical Fault Location Algorithm with Increased Adaptability for Active Low Voltage Grids. IEEE Transactions on Industrial Applications. *Under Review*

A.2 Conference articles

[**C1**] Paschalia Stefanidou–Voziki, Cristina Corchero, and Jose Luis Dominguez Garcia. A Practical Algorithm for Fault Classification in Distribution Grids. In 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), pages 1–6, June 2020. XIV, 78, 79

[**C2**] Paschalia Stefanidou–Voziki, David Cardoner Valbuena, Roberto Villafafila Robles, and Jose Luis Dominguez Garcia. Feature Selection and Optimization of a ML Fault Location Algorithm for Low Voltage Grids. In 2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), pages 1–6, September 2021. XV, XVI, 125, 128, 129, 130

[**C3**] P. Stefanidou–Voziki, N. Sapountzoglou, R. Villafafila Robles, and J.L. Dominguez Garcia. A study on the effect of EVs' charging stochasticity on a ML–based fault detection algorithm. In 2022 CIGRE Session, 28 August–02 September 2022, Paris, France.