Supervised and Unsupervised Deep Learning Models for Partial Discharge Source Detection and Classification in Electrical Insulation

by

Sara Mantach

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB, Canada

Examining Committee

This thesis was examined and approved by the following examining committee on August 11, 2023:

- **Dr. Behzad Kordi** (Advisor)
  Department of Electrical & Computer Engineering
  University of Manitoba, Winnipeg, MB, Canada
- **Dr. Ahmed Ashraf** (Co-advisor)
  Department of Electrical & Computer Engineering
  University of Manitoba, Winnipeg, MB, Canada
- **Dr. Pooneh Maghoul**
  Department of Civil Engineering
  University of Manitoba, Winnipeg, MB, Canada
- **Dr. Namal Fernando**
  Department of Electrical & Computer Engineering
  University of Manitoba, Winnipeg, MB, Canada
- **Dr. Andrea Cavallini** (External Examiner)
  Department of Electrical Energy and Information "Guglielmo Marconi"
  University of Bologna, Bologna, Italy

The Ph.D. oral examination was chaired by:

- **Dr. Babak Mehran**
  Department of Civil Engineering
  University of Manitoba, Winnipeg, MB, Canada

Abstract

Condition monitoring of electrical insulation of high voltage apparatus is of much importance for the reliable operation of electric power systems. An effective way to monitor the health of such systems is the measurement of partial discharges (PDs) in the insulation material. In order to prevent the PD from progressing in the insulation material, the source of PD should be known. The PD classification problem in high voltage systems changes from a single-class to a multi-class and multi-label classification problem when PD sources take place simultaneously in real-life systems. Regardless of being single-label or multi-label, supervised learning has been employed for PD source classification by academics and researchers where labelled data and the number of PD sources are two prerequisites for the training process. What makes PD classification more complicated in real life scenarios is that the number of PD sources in an insulation system is unknown. This makes the problem of PD classification an unsupervised one, where there is no prior knowledge of the number of PD sources.

Machine learning techniques offer a solution for PD classification by allowing to train models based on extracted features. The performance of such algorithms heavily depends on the choice of features. This can be overcome by using deep learning where feature extraction is done automatically by the algorithm, and the input to such an algorithm is the raw input data.

This research is focused on developing deep learning models for the classification of PD sources in the insulation of high voltage systems. The developed models include: model for classifying multi-source PDs and single-source PDs without introducing multi-source PDs in the training stage, interpretable attention based model that propose non linear filters that are capable of differentiating between PD signals that look alike, and an unsupervised deep learning model for predicting the number of partial discharge sources. These models present the base for smarter next generation PD monitoring software that can be used by researchers and experts in industry in order to overcome the challenges and limitations that are present in current practices.

Acknowledgements

I would like to express my appreciation to my advisor and co-advisor, Dr. Behzad Kordi and Dr. Ahmed Ashraf, for their great support and supervision during this research. I deeply appreciate all their valuable comments and guidance. Many thanks to my thesis examining committee, Dr. Pooneh Maghoul, Dr. Namal Fernando, and Dr. Andrea Cavallini, for their valuable comments and suggestions to enhance this thesis. Financial support from Natural Sciences and Engineering Research Council of Canada (NSERC), University of Manitoba (Faculty of Graduate Studies, Price Graduate Scholarships for Women in Engineering, and Edward R. Toporeck Graduate Fellowship in Engineering) is gratefully acknowledged. Many thanks to my friends and colleagues for their support and help during my PhD research program. Lastly, sincere thanks to my family for their kindness and support while I was pursuing my studies away from them.

Dedications

To my loving family.

Table of Contents

Examining Committee ii
Abstract iii
Acknowledgements v
Dedications vi
List of Figures xi
List of Tables xvii
Nomenclature xix

1 Introduction **1**
1.1 Motivation 1
1.2 Research Objectives 3
1.3 Research Contributions 4
1.3.1 List of Publications 4
1.4 Thesis outline 6

2 Background **9**
2.1 Introduction 9
2.2 Deep Learning 10

2.2.1 Convolutional Neural Networks 12
2.2.2 Autoencoder 14
2.2.3 Recurrent Neural Networks 16
2.3 PD Classification 18
2.3.1 PD Classification Using Convolutional Neural Networks 18
2.3.2 PD Classification Using Autoencoders 29
2.3.3 PD Classification Using Recurrent Neural Networks 32
2.3.4 PD Classification Using Hybrid DL Algorithms 33
2.4 Summary 37

3 Multi-Class Classification of PD Sources Using PRPD Patterns 38
3.1 Experimental Setup 39
3.2 Method 43
3.3 Results 49
3.3.1 Performance Metrics 49
3.3.2 Independent Classifiers 51
3.3.3 Proposed Model 52
3.4 Summary 56

4 Limitations of Off-the Shelf CNN for PD Signal Classification 58
4.1 Presence of Noise and Interference 61
4.1.1 Datasets 61
4.1.2 Implementation of 1D- CNN 63
4.1.3 Findings 65
4.2 Multi-Label PD Classification 68

4.2.1 Datasets 68
4.2.2 Findings 74
4.3 Summary 77

5 An Interpretable CNN Model for Classification of Partial Discharge Waveforms 79

5.1 Materials and Methods 81
5.1.1 3D-Printed PLA Samples 81
5.1.2 Dataset 83
5.1.3 Time-Frequency Map 85
5.1.4 Principal Component Analysis (PCA) of Statistical Features Extracted from PRPD 88
5.1.5 1D-Convolutional Neural Network 90
5.1.6 Adding Interpretability to the CNN 91
5.2 Findings 93
5.3 Summary 98

6 Partial Discharge Adaptive Clustering Method 100

6.1 Experimental Setup and Data Collection 102
6.2 Method 103
6.2.1 Time-Frequency Map 103
6.2.2 Convolutional Autoencoder 106
6.2.3 Proposed Method 108
6.3 Findings 111
6.3.1 Performance Evaluation of the Proposed Technique 111

6.3.2 Sensitivity of the Proposed Model to Noise 114
6.3.3 Comparison with Traditional Machine Learning 115
6.4 Summary 117

7 Concluding Remarks **121**
7.1 Conclusions 121
7.2 Future Work 122

References **124**

List of Figures

2.1 Different deep learning branches: supervised, unsupervised, and reinforcement learning. 12

2.2 A simple CNN architecture: a convolutional layer, pooling layer, fully connected layers followed by the classification layer. 14

2.3 Autoencoder architecture: the output $\tilde{\mathbf{x}}$ is the reconstructed input where a bottleneck enables to compute a latent representation of the original input $\mathbf{x}$. $\mathbf{x}$ is a vector consisting of n elements from x_1 to x_n. 15

2.4 RNN architecture with no output: the network has feedback connections which can be unfolded in time and trained using back-propagation. The input X is processed by incorporating it into the state S that is passed forward through time. 16

3.1 PRPD patterns and their binary representation of various single defects: a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6. 41

3.2 PRPD patterns and their binary representation of various multiple defects: a) class 14; b) class 16; c) class 46; d) class 146. 42

3.3 a) baseline model with independent model for each class; b) proposed model with a common convolutional backbone shared across all classes. 44
3.4 CNN architecture of an independent classifier. 46
3.5 Proposed deep learning model architecture: common convolutional backbone shared across all classes. 48
3.6 Training and validation log losses vs. number of iterations for independent classifiers a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6; g) class 7. 51
3.7 Decision on stopping criteria: the percentage difference between the validation loss and the training loss is minimal in the marked region. 54
3.8 Training and validation log losses vs. number of iterations for the proposed model a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6; g) class 7. 56

4.1 Class 1 with different SNR for AWGN. 62
4.2 Class 1 with different amplitudes of the interference signals. 63
4.3 Classes corresponding to different PD sources in the experimental dataset. . 64
4.4 Class 1 of the experimental dataset with and without added noise. 64
4.5 Class 1 of the experimental dataset with and without interference signal. . . 65
4.6 Sensitivity of 1D- CNN to different interference amplitudes and frequencies for the synthetic dataset. 68
4.7 Synthetic data: different rise times corresponding to different classes. 69
4.8 A sample synthetic waveform with two classes partially overlapping. 70
4.9 (a) Schematic of the experimental setup, (b) Photo of the PLA sample between brass electrodes. For the experiments, the sample is immersed in transformer oil. 71

4.10 Samples of signals collected in the experimental setup. Classes 1 and 2 correspond to the 3D PLA samples with a 0.5 mm and a 2 mm void, respectively. 72

4.11 The average classification accuracy and the false negative rate of the multiclass C12 as a function of increasing the number of classes to be classified. 75

5.1 (a) A schematic of the 3D-printed PLA cylindrical sample showing the location of a cylindrical void at the centre (the common axis of rotational symmetry for the void and the sample is shown as a dashed vertical line through the void), (b) X-ray microtomography scan of a 3D-printed PLA sample, shown looking along the common axis of symmetry (i.e. rotated towards the reader when compared to the schematic), that shows the void in the centre of the sample as well as imperfections generated during the printing process (most of which occur at the base of the sample - farthest from the reader as viewed in this image). 82

5.2 (a) 3D-Printed PLA sample (12mm long and 12mm diameter) held between brass electrodes (external diameter 10mm) (b) Sample and electrodes immersed in Voltesso 35 oil for partial discharge measurement. 84

5.3 Two examples time series waveforms from class 1: the left one shows a positive pulse and the right one shows a negative pulse. Both pulses belong to void size class 1. 85

5.4 PRPD pattern of class 1 sample. 86

5.5 Two examples time series waveforms from class 2: the left one shows a positive pulse and the right one shows a negative pulse. Both pulses belong to void size class 2. 86

5.6 PRPD pattern of class 2 sample. 87

5.7 T-F map of the normalized PD waveforms acquired from 6,000 samples of class 1 and 6,000 of class 2. Some overlap is visible between the two clusters. 88
5.8 First two principal components of data resulting from applying PCA on the statistical features of the PRPD patterns. 89
5.9 1D-CNN with first convolutional layer of filter size 4 and max pooling of 4 followed by a convolutional layer of filter size 3 and max pooling of 4 followed by two fully connected layers of 512 and 64 neurons respectively followed by the classification layer with one neuron. 90
5.10 Flow chart of the proposed framework: mapping the time-series waveforms to a T-F map, training a 1D-CNN using time-series waveforms followed by interpreting the decision making via attention model, and applying PCA on 300 statistical features extracted from PRPD patterns, where within unseen data, there could be more probability of confusion along the second principle component axis. 94
5.11 Intersection of samples resulting from mapping them into the T-F map. . . . 95
5.12 Attention mechanism implementation for class 1 : (a) Positive pulse (b) Negative pulse. In (a) and (b), the model concentrates at the start of the waveform in order to decide on the label of the sample. 96
5.13 Attention mechanism implementation for class 2: (a) Positive pulse (b) Negative pulse. In (a) and (b), the model concentrates at the start of the waveform in order to decide on the label of the sample. 96
5.14 Heat map of the 16 kernels weights learned in first convolutional layer with kernal size of four. 97

6.1 Experimental setup for PD measurement in a generator stator bar: (a) schematic of the laboratory setup; (b) a photo of the lab setup; (c) close-up view of the stator bar in the dummy slot. 104

6.2 (a) A wire is added to the end of the stator bar to simulate corona discharge; (b) Endwinding discharge is simulated by adding metallic particles on the stress control coating. 105

6.3 Flowchart of the proposed method: After training the convolutional autoencoder, the learned bottleneck coefficients are used in order to map the time-series signal into a 16-dimensional space. Applying cosine similarity based clustering technique allows us to predict the number of PD sources. 107

6.4 Proposed adaptive clustering algorithm based on the cosine similarity criteria. (a) The overall flowchart; (b) Detailed implementation of the proposed algorithm. 110

6.5 (a) T-F map of the PD signals captured when 11 kV is applied to the stator bar. The cluster on the left corresponds to PD in microvoids that also exist under 8 kV of applied voltage. The cluster in the middle that corresponds to the surface discharge appears at 11 kV; (b) Representation of clusters based on the proposed cosine similarity criteria for microvoid and surface discharge PD sources where the number of time-series waveforms is shown for each cluster. The proposed method identifies 2 PD sources that is confirmed by the T-F map and the PRPD pattern. 112

6.6 (a) T-F map of the PD signals captured at 11kV of the applied voltage when floating particles are introduced to the setup; (b) Presentation of the clusters based on the cosine similarity criteria the shows the ability to identify the three sources of PD. The number of time-series waveforms is shown for each cluster. 113

6.7 (a) T-F map of the PD signals captured at 11kV of the applied voltage when both floating particles and corona source of PD are introduced to the setup; (b) Presentation of the clusters based on the cosine similarity criteria the shows the ability to identify the four sources of PD. The number of time-series waveforms is shown for each cluster. 115

6.8 A sample of a PD time-series waveform at different levels of signal-to-noise ratio (SNR) of additive white Gaussian noise (AWGN). 116

6.9 T-F map for a dataset including four PD sources at 11 kV for various signal to noise ratios (SNR) of additive white Gaussian noise. (a) original dataset; (b) SNR= 20 dB; (c) SNR= 10 dB; (d) SNR= 5 dB. 119

6.10 Presentation of clusters based on the cosine similarity criteria for void, surface, floating particle and corona PD sources in the presence of AWGN of SNR=5 dB where the number of time-series waveforms is shown for each cluster. 120

List of Tables

2.1 Summary of deep learning algorithms used for classification of PD in various HV applications: specification of characteristics of collected data used and whether multiple-labeled sources of PDs are mentioned. 19

3.1 Different levels of charge magnitude scale setting on the Omicron Software. . 41
3.2 Design specification of an independent classifier. 47
3.3 Accuracy of single and multiple source PRPD patterns. 52
3.4 PCR and PCP for independent classifiers. 53
3.5 Hybrid confusion matrix of independent classifiers. 53
3.6 Accuracy of single and multiple source PRPD patterns. 55
3.7 PCR and PCP for the proposed model. 55
3.8 Hybrid confusion matrix of the proposed model. 57

4.1 Datasets architecture design. 66
4.2 Results of measurement dataset 67
4.3 AUC for the synthetic tested data 67
4.4 Multiple Classes Considered for each Scenario 70
4.5 Datasets Architecture Design 73

4.6 Accuracy of Multiple Classes for Synthetic Data Case Studies 75

4.7 Hybrid Confusion Matrix for the Experimental Data 77

5.1 Confusion matrix of independent classifiers. 94

6.1 Number of clusters reported by traditional machine learning techniques of k-means and PCA that shows they are not always succesful in correct determination of the number of PD clusters. 117

Nomenclature

PD	Partial Discharge
PRPD	Phase Resolved Partial Discharge
T-F Map	Time- Frequency Map
PLA	Polylactic Acid
ML	Machine Learning
PCA	Principle Component Analysis
DL	Deep Learning
CNN	Convolutional Neural Network
AE	Autoencoder
PCR	Arithmetic Mean of Recall
PCP	Arithmetic Mean of Precision

Chapter 1

Introduction

1.1 Motivation

Many essential services in our society necessitates maintaining consistent supply of electrical power with reliable infrastructure that ensures the system is free of faulty conditions [1]. The infrastructure's insulation material is responsible for isolating the components of the electrical equipments from each other and from the ground [2]. On other side, partial discharges (PD) are localized electrical discharge that only partially bridges the insulation between conductors in electrical equipments [3]. The poor design, improper installation, or aging of a high voltage electrical equipment, and the presence of cavities or impurities within the insulation system can trigger PD to take place. Hence, an early detection of PD in the insulation system of any electrical equipment reduces the risk of a total breakdown. Different techniques have been employed to detect these PDs. These techniques are based on acoustic, optical, chemical, ultra high frequency and electrical measurements. The electrical measurement of these PDs in high voltage systems is the focus of the current research. In a perfect scenario,

sensors can be used to monitor different insulation in a high voltage system. Nevertheless, this is highly expensive and time consuming. Therefore, the performance of the insulation is evaluated indirectly by studying the electrical current on the terminals of the system. Each source of partial discharges is accompanied with a unique discharge mechanism which yields unique features. Hence, the goal of the classification of the PD sources is to be aware of the defect which is causing the discharges. Phase related and time resolved recognition are the two basic means to recognize discharges [4]. Time series waveforms resulting from different sources of partial discharges have been used in deciding what PD sources are present. On the other hand, phase resolved partial discharge (PRPD) patterns have been widely used in research and industry as a diagnostic tool to classify different sources of partial discharges. PRPD patterns include the discharge magnitude and the discharge rate with respect to the phase of the applied AC voltage. Engineers and domain experts have been using signal processing and pattern recognition algorithms in order to inspect time series signals and PRPD patterns [5, 6]. These algorithms include data processing and feature extraction. Data processing is important to get informative information of the data. Feature extraction includes dimension reduction by using different linear, non linear, and statistical operators. The above automation process necessitates human expertise for the pre-processing of the input data, specially for the feature extraction stage. These engineered features constitute a characterization vector in a multidimensional space where thresholds have to be specified in order to distinguish different classes where a class, for example, can include PDs from a given source of defect. In order to specify these thresholds, different traditional machine learning classifiers are used such as Support Vector machine (SVM), Fuzzy SVM [7], Kernel SVM, Radial Bases Function Network [8], Probabilistic Neural Networks [9], Naïve Bayes and AdaBoost [10]. A more detailed review on traditional machine learning techniques for PD

classification can be found in [11]. The main restriction of the classical approach, *i.e.* manual feature extraction followed by applying classifiers, is that human expertise is needed for extracting meaningful engineered features. With the availability of computational resources and big data, deep learning, which is an evolving branch in machine learning, started to gain more attention. Deep learning allows the feature selection stage to be integrated with the learning process; thus, making the process all automated. The motivation of this PhD work is studying the integration of deep learning algorithms (supervised and unsupervised) in detecting and classifying partial discharges in the insulation of high voltage systems.

1.2 Research Objectives

The objectives of this research include:

- Developing a deep learning (DL) model for classifying multi-source PDs and single-source PDs without introducing multi-source PDs in the training stage.
- Creation of an interpratable DL model that is capable of differentiating between different time series waveforms that look alike. With the emergence of new technologies, wherein the interference pulses become more similar to PD pulses, established tools that are used in industry fail to differentiate between these pulses.
- The development of an unsupervised system that would predict the number of PD sources taking place in a high voltage (HV) system that is more robust to noise and interference compared to other well established clustering tools that are used in industry.

1.3 Research Contributions

The contributions of this research consist of the following:

- A novel convolutional neural network (CNN) -based architecture for classifying multi-source PDs and single-source PDs without introducing multi-source PDs in the training stage.
- A CNN based model followed by a post-hoc attention mechanism for classifying PD signals that look alike, where established tools (e.g. T-F map) used in industry fail to classify these signals.
- A novel unsupervised deep learning system based on convolutional autoencoder and adaptive clustering technique for predicting the number of partial discharge sources taking place in stator hydro-generator bars.

1.3.1 List of Publications

Journal Papers

1. **S. Mantach**, A. Ashraf, H. Janani, and B. Kordi, "A convolutional neural network-based model for multi-source and single-source partial discharge pattern classification using only single-source training set," *Energies*, 14(5), pp. 1355, March 2021.

   This paper proposed a novel convolutional architecture for single-source PD and multi-source PD classification using training data with ground-truth available only at the level of single-source PDs. The proposed architecture consists of a convolutional backbone feeding into multiple fully connected neural networks (FCNs).

2. **S. Mantach**, P. Gill., D. Oliver, A. Ashraf, and B. Kordi, "An interpretable CNN model for classification of partial discharge waveforms in 3D-printed dielectric samples with different void sizes." *Neural Computing and Applications*, 34(14), pp. 11739-11750, March 2022.

   This paper proposed a 1D-CNN architecture which was designed, implemented, and tested to investigate the PD pulses generated in a void inside a solid dielectric. In addition, an attention mechanism was added to the learned CNN model to introduce interpretability to the decisions made by the deep neural network.

3. **S. Mantach**, A. Lutfi, H. Moradi Tavasani, A. Ashraf, A., A. El-Hag, and B. Kordi, "Deep learning in high voltage engineering: A literature review." *Energies*, 15(14), pp. 5005, July 2022.

   This paper presented a review of the recent literature on the application of deep learning techniques in monitoring high voltage apparatus such as GIS, transformers, cables, rotating machines, and outdoor insulators.

4. **S. Mantach**, M. Partyka, V. Pevtsov, A. Ashraf, and B. Kordi, "Partial Discharge Adaptive Clustering Method for Stator Bars Based on Unsupervised Deep Learning." Accepted for publication in *IEEE Transactions on Dielectrics and Electrical Insulation*.

   In this paper, a robust unsupervised deep learning model which is based on convolutional autoencoder and adaptive clustering technique is proposed for unsupervised clustering.

Conference Papers

1. **S. Mantach**, H. Janani, A. Ashraf and B. Kordi, "Classification of partial discharge signals using 1D convolutional neural networks." *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 12-17 September, 2021, Toronto, ON (online), Canada.

   In this paper, a study was conducted to evaluate the performance of a one-dimensional CNN model for PD detection and to assess how prone it is to noise and interference.

2. **S. Mantach**, P. Gill., D. Oliver, A. Ashraf, and B. Kordi, "Assessing one-vs-all 1D-CNN classifiers for multi-label classification of partial discharge waveforms in 3D-printed dielectric samples with different void sizes," *IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP)*, 31 October-2 November, 2022, Denver, CO, USA.

   In this paper, a study presented on the limitations of multiple one-versus-all, one-dimensional (1D), CNNs for multi-label classification of PD.

3. Contributions were made to the following as a co-author:

   H. Moradi Tavasani, **S. Mantach**, M. Gunawardana, B. Tabei, and B. Kordi, "Localization of partial discharge in power transformer winding using sparse autoencoder," *CIGRÉ Canada Conference and Expo*, 31 October-3 November 2022, Calgary, AB, Canada.

1.4 Thesis outline

This thesis is organized into seven chapters as detailed below:

Chapter 1: Motivation, research objectives, and contributions are presented.

Chapter 2: Background and literature review of previous research on using time series waveforms and PRPD patterns for the classification of PD sources. Some deep learning algorithms that are used for PD applications in literature are presented as well.

Chapter 3: In this chapter, an enhanced convolutional neural network is proposed that is capable of classifying single sources as well as multiple sources of partial discharges without introducing multiple sources in the training phase. The training is done by using only single-source phase-resolved partial discharge (PRPD) patterns, while testing is performed on both single and multi-source PRPD patterns. The proposed model is compared with a single-branch CNN architecture.

Chapter 4: In this chapter, the limitations of traditional one-dimensional convolutional neural network (1D-CNN) is explored. A one dimensional convolutional neural network is designed that takes a set of time series waveforms as the input and is capable of classifying PD sources in the presence of additive Gaussian noise and discrete spectral interference. In addition, the performance of a collection of one-versus-all 1D- CNN is investigated to classify single and multiple PD sources using time-series waveforms. Training of the CNN classification algorithm is done on single classes and testing is done on single and multiple classes. In addition, the effect of the number of the classified classes on the performance of the proposed system is considered.

Chapter 5: In this chapter, a convolutional neural network (CNN) attention based model is proposed and show superior capability over a traditional classification technique to classify partial discharge (PD) waveforms resulting from different voids in polylactic acid (PLA) 3D-printed samples. Extensive investigation of the learned model is conducted in order to interpret the decisions made by the proposed neural network. In particular, adding

an interpretable attention model such as GRAD-CAM to the CNN, shows that while making the decision, the neural network learns to focus more on the regions of the waveform corresponding to the rise of the pulse.

Chapter 6: In this chapter, a robust unsupervised deep learning model which is based on convolutional autoencoder and an adaptive clustering technique is proposed for unsupervised clustering. The input to the system is the unlabelled, time-series PD waveforms and the output is the predicted number of partial discharge sources that are taking place in the high voltage system. The proposed system shows superior results in terms of immunity to noise compared to a traditional classification method.

Chapter 7: This chapter provides the conclusions and suggests future work directions.

Chapter 2

Background

2.1 Introduction

Diagnosis of electrical insulation degradation is essential for monitoring the integrity of an electric power system. A well-known diagnostic method, which has been employed for a number of decades, is the measurement of localized discharges known as partial discharge (PD) [2]. Detecting fault or PD in electric apparatus, such as transformers, rotating machines, cables, gas insulated switchgear (GIS) and outdoor insulators has always required the knowledge of expertise who are able to characterise and differentiate the different sources of fault, PD, defect, or degradation. Throughout the years, different parameters had to be extracted manually from recorded patterns or signals. The aim has been to use the manually-extracted parameters in order to implement a classifier that would be able to perform the task of differentiation and characterization of fault, PD, defect, or degradation. Though the process is partially automated, the fact that experts have to select the features presented

a problem since different features might result in different outcomes. This influences the performance of the classifier due to its dependence on the manually-selected features.

Deep learning allows the feature selection stage to be integrated with the learning process; thus, making the process all automated. In high voltage (HV) applications, the aim has mostly been to classify or localize fault, defect, or PD that occur in HV apparatus or determine the degradation of insulating material. The abundance of computational capabilities and the existence of big data has allowed researchers in different fields to take advantage of deep learning algorithms. Other than the main purpose of classifying and localizing the PD or fault in HV apparatus, a deep learning algorithm, namely the Generative Adversarial Network (GAN), allows researchers to generate more input data from a limited amount of experimental/simulation results (e.g. see [12]).

Classification refers to the process of differentiating between different sources of fault, defect, PD, or levels of degradation. Given that in real life scenarios, fault or PD can happen due to various sources, it is necessary to identify the source. When the source is identified, one can investigate techniques to eliminate that source from the high voltage system. Different sources or causes of fault, defect, PD, or degradation exhibit different characteristics that are unique to each source, making their classification (differentiation) possibly feasible. On the other hand, localization refers to the process of identifying the position of the fault or PD taking place in high voltage apparatus [13].

2.2 Deep Learning

Deep learning is a branch of machine learning that enables data-driven learning of feature representations for input data originating in diverse application domains [14–16]. Unlike tra-

ditional machine learning algorithms, where features need to be extracted explicitly through pre-defined hand-crafted rules, deep learning has the advantage of using raw data, and learn to extract features depending on the task [17]. This is appreciated especially in complex systems where such features are not necessarily known for a given dataset. As a result, deep neural networks subsume the feature extraction step within the learning phase, thereby computing intrinsic representations of the raw input data in an automatic manner.

Similar to traditional machine learning, deep learning also has the following three key paradigms: supervised, unsupervised, and reinforcement learning. For the supervised setting, a labeled dataset is required. The type of output can either be continuous (used in a regression problem) or discrete/categorical (used for classification). For unsupervised systems, data with no labels are given and the objective is either to cluster the data according to their intrinsic characteristics or learn representations which can be later used for downstream supervised or unsupervised settings [18–20] . In scenarios involving agent based learning, exhaustive collection of supervised data is often prohibitively difficult. In such situations, reinforcement learning is a powerful paradigm which allows data collection through interaction with the environment [21]. The agent's goal is to learn policies based on the environment in order to maximize long term expected rewards [21]. In recent years, research in deep reinforcement learning has gained significant traction wherein the agent policies are learnt through deep neural networks [22] (see Fig. 2.1). Depending on the inputs and the desired outputs for most of the high voltage application, a handful of mainly supervised deep learning algorithms have been of interest in this area of research. A brief introduction on major supervised deep learning algorithms is presented in the following.

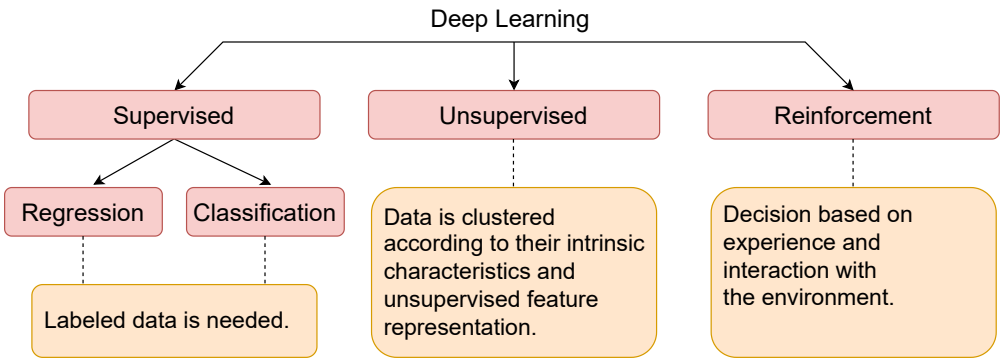


Fig. 2.1: Different deep learning branches: supervised, unsupervised, and reinforcement learning.

2.2.1 Convolutional Neural Networks

Convolutional neural networks (CNNs) represent a class of deep learning architectures which were originally designed for processing data represented in a grid-like topology, e.g. images [17]. A CNN has four main components: convolutional layer, activation function layer, pooling layer, and fully connected layers. Typically, the output of the convolutional layer is passed to an activation function layer where the output of the latter is passed to the pooling layer. In a deep network, this set of the three components are often cascaded multiple times thereby constituting multiple layers and making the network progressively deeper [14,23,24]. While the initial layers usually end up learning low-level features, the deeper layers tend to learn more complex features. The cascade of these layers constitute the automatic feature extraction stage, and the fully connected layers constitute the classification stage [25]. More details on each of these components is presented below:

Convolutional layer: This layer consists of a bank of learnable linear 1D, 2D, or 3D filters which are also called kernels [26]. In the high voltage applications, usually 1D and 2D CNNs are used. The 1D-CNN, for example, is used with time-series waveforms, whereas in problems

involving phase resolved partial discharge (PRPD) patterns or spectrograms, a 2D-CNN is used. Some of the researchers have employed a 2D-CNN for time series waveforms as well, where they considered an image of the signal as an input rather than the 1D data. These filters are convolved with the input data or the output from a previous layer. The output is a set of feature maps, where the number of feature maps is equal to the number of the filters.

Activation function layer: The purpose of adding activation layers is to introduce non-linearity in the input-to-output mapping being learned by the neural network. This is desired because complex data include nonlinear features that need to be detected. Most frequently employed activation functions include sigmoid, ReLU and tanh.

Pooling layer: The aim of pooling layer is to subsample the output feature maps so that wider receptive fields can be spanned during convolution without increasing the size of the filter kernel. Another advantage of this layer is to provide positional invariance or shift-invariance to the network [27]. Commonly-used pooling operations are maximum pooling and average pooling.

Fully connected layers: In a fully connected (FC) layer, every neuron in one layer is connected to every neuron in the next layer. FC layers are also referred to as dense layers in the literature [17]. In a CNN, the input to the first fully connected layer is the output of the last set of the first three components mentioned above, where the corresponding features maps are flattened into 1D vectors. For classification problems, the architecture is appended by FC layers and ends with a classification layer where the number of neurons is equal to the number of classes. A typical CNN architecture is shown in Fig. 2.2. The main advantage

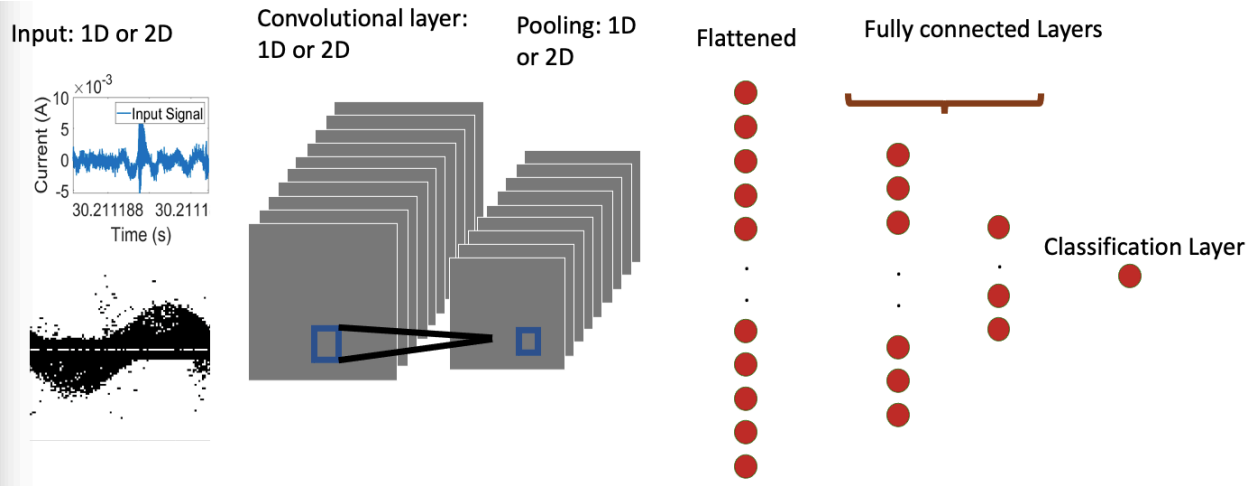


Fig. 2.2: A simple CNN architecture: a convolutional layer, pooling layer, fully connected layers followed by the classification layer.

of CNNs compared to traditional neural networks is the weight sharing when training the learnable kernels, which reduces the learnable parameters in the network [28].

2.2.2 Autoencoder

Autoencoders (AE) were introduced in 1980s [29] in order to learn useful representations in an unsupervised fashion by the use of the input data on its own [30]. They were then reintroduced in 2006 with the booming of the deep learning architectures [31]. The idea behind an autoencoder is to train a neural network such that the model learns a latent intrinsic representation of the original input. An autoencoder consists of an encoder-decoder architecture, wherein the role of the encoder is to transform the input to a latent representation, while the decoder is responsible for transforming the latent representation back to the original data. The two parts (encoder and decoder) are learned jointly so as to minimize

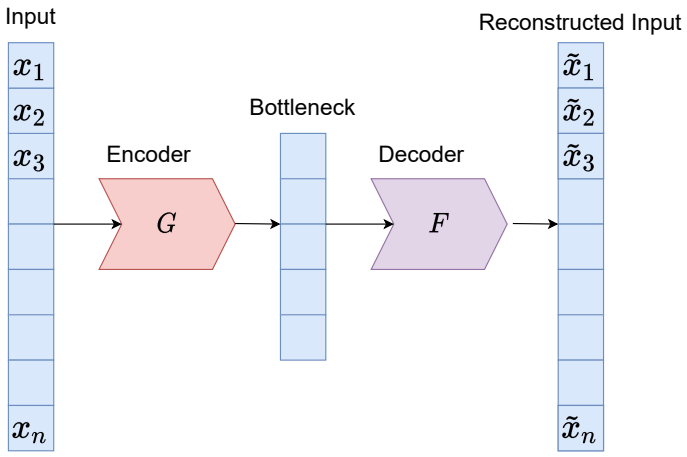


Fig. 2.3: Autoencoder architecture: the output $\tilde{\mathbf{x}}$ is the reconstructed input where a bottleneck enables to compute a latent representation of the original input $\mathbf{x}$. $\mathbf{x}$ is a vector consisting of n elements from x_1 to x_n.

the reconstruction error between the decoder's output and the network input. A simple illustration of an autoencoder model is shown in Fig. 2.3.

Assuming that the input is $\mathbf{x}$ and the reconstructed output is $\tilde{\mathbf{x}}$, the model is trained to minimize the reconstruction error $\mathcal{L}(\mathbf{x}, \tilde{\mathbf{x}})$. The encoder and decoder can be fully connected layer networks or any deep learning architecture. The encoder is expressed as a function G such that

$$\mathbf{b}_i = G(\mathbf{x}_i) \tag{2.1}$$

where $\mathbf{b}_i$ represents the latent feature representation (bottleneck) of a single observation sample $\mathbf{x}_i$. The decoder F accepts $\mathbf{b}_i$ as input and produces $\tilde{\mathbf{x}}_i$ is. This is shown in

$$\tilde{\mathbf{x}}_i = F(\mathbf{b}_i) = F(G(\mathbf{x}_i)). \tag{2.2}$$

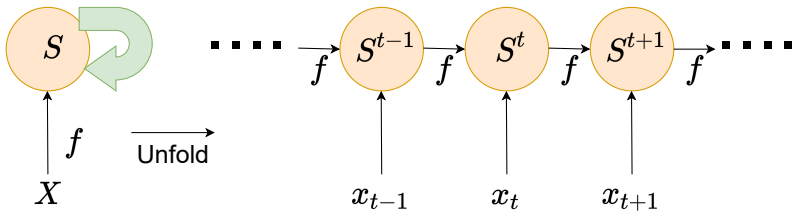


Fig. 2.4: RNN architecture with no output: the network has feedback connections which can be unfolded in time and trained using back-propagation. The input X is processed by incorporating it into the state S that is passed forward through time.

The goal is then to find F and G that would minimize

$$\arg\min_{F,G} \sum_i \left[\mathcal{L}\left(\mathbf{x}_i, F\left(G\left(\mathbf{x}_i\right)\right)\right)\right] \tag{2.3}$$

where the summation is over all the observations during training.

2.2.3 Recurrent Neural Networks

Recurrent Neural Network (RNN) is another family of deep learning architectures which are intended for the processing of sequential data [32, 33]. A simple illustration of an RNN model is shown in Fig. 2.4. RNNs process data from each time point in a sequential manner. However, the output is not just influenced by data at the current time, but also by the entire history of inputs that have been fed into the RNN previously. This is reflected by the cycles in the architecture, which are maintained in the hidden unit as a state vector including the history of the previous time points. RNN cell has one common set of weights, and when backpropagation runs, data from different time points contribute in updating the same set of weights.

The mathematical representation of the RNN is shown as [34]

$$\mathbf{h}_t = f_{\mathbf{w}}\left(\mathbf{h}_{t-1}, \mathbf{x}_t\right) \tag{2.4}$$

where $\mathbf{h}_t$ represents the updated state vector, $\mathbf{h}_{t-1}$ is the hidden state vector from the previous time step, $\mathbf{x}_t$ represents the input vector at time t, and $f_{\mathbf{w}}$ represents a given function corresponding to learnable weight vector $\mathbf{w}$. The input can either be a vector or a sequence, and the output can either be a vector, sequence, or a value. For example, given a high voltage problem where classification of different PD pulses is required, the input is a vector of PD pulses and the output is a label corresponding to a PD source.

The drawback of a typical RNN is the long-term dependency where the current state depends on all the previous states, which causes the vanishing gradient problem [17]. The vanishing gradient emerges from the fact that as RNN processes more time steps, repeated multiplication of small weights causes the gradients to approach zeros. To overcome this problem, long short-term memory (LSTM) architecture is used [35]. The main difference in an LSTM architecture is that instead of computing the hidden state directly from the previous one, LSTM computes additional states, and this structure allows alternative paths to gradients to flow during the backpropagation avoiding repeated matrix multiplications [36]. An LSTM cell has two hidden states $\mathbf{c}_t$ corresponding to the cell state and $\mathbf{h}_t$ corresponding to the hidden state which are calculated as [34]

$$\begin{aligned} \mathbf{c}_t &= \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{g} \\ \mathbf{h}_t &= \mathbf{o} \odot \tanh\left(\mathbf{c}_t\right) \end{aligned} \tag{2.5}$$

where **i** is the input gate, **f** is the forget gate, **o** is the output gate, and **g** is the gate. The operator $\odot$ is element-wise multiplication operation. The input gate decides what new information will be stored in the cell state, the forget gate decides what information will be removed from the cell state, and the output gate decides what information from the cell state will be used in the output [37].

2.3 PD Classification

A summary of the papers on classification of PD using deep learning is shown in Table 2.1. The following subsections will focus on the application of DL algorithms for different high voltage applications. Classification of PDs has been a standard procedure in the maintenance of high voltage assets. Various distinctive parameters extracted from PD measurements have been introduced for the PD classification application. Starting with current or voltage signals, time series waveforms have been proven to acquire unique behavior for each source of fault or PD [38]. Phase resolved partial discharge (PRPD) patterns have also been used to differentiate between different PD sources [5].

2.3.1 PD Classification Using Convolutional Neural Networks

In this section, the application of convolutional neural networks (CNNs) for PD classification is presented. The literature is divided based on the high voltage apparatus.

GIS

Gas insulated switchgear (GIS) is widely used in industry [74]. GIS has its components close to each other which makes the fault occurrence in one component transfer to other compo-

Table 2.1: Summary of deep learning algorithms used for classification of PD in various HV applications: specification of characteristics of collected data used and whether multiple-labeled sources of PDs are mentioned.

HV Application	Data Collected	Field/ Lab/ simulations	Multiple sources(?)	DL Technique	References
GIS	PRPD	Lab	No	Stacked Sparse AE, LSTM ,Siamese CNN network, multi-head self attention LSTM and self attention based neural network model	[39], [40], [41], [42]
GIS	PRPD	Both	No	CNN(LeNet-5), Variational AE	[43], [44]
Solid Insulation	PRPD	Lab	No	DBN , CNN	[45], [46]
Transformer	PRPD	Lab	No	CNN-LSTM , CNN , lightweight attention mechanism Squeeze-and-Excitation (SE) module on top of CNN, CNN	[47], [48], [49], [50]
Transformer	PRPD	Lab	Yes	Novel architecture based on CNN , LSTM	[51], [52]
Transformer	PRPD	Lab	Yes *	ResNet	[53]
Cables	PRPD	Lab	No	Transfer Learning on CNN	[54]
Rotating Machines	PRPD	Field	Yes	ANNs incorporated in a hierarchical fashion	[55]
Cables	T-S waveforms	Simulation	No	CNN	[56]
Cables	T-S waveforms	Lab	No	CNN, DBN	[57], [58]
Power Cables	T-S waveforms	Field	No	ensemble of deep learning algorithms (CNN, convolutional RNN, LSTM and bidirectional LSTM)	[59]
Hydro-generators	T-S waveforms	Field	No	Variational Autoencoder	[60]
GIS	T-S waveforms	Lab/ simulation	No	Conditional Variation Autoencoder and CNN, Convolutional Autoencoder, CNN(AlexNet), depth-wise CNN, 1D-CNN model where a multiple scale convolution kernel, CNN-LSTM	[61], [62], [63], [64] [65], [66]
GIS	T-S waveforms	Lab/ simulation/ Field	No	domain adaptive deep transfer learning (DADTL) CNN	[67]
GIS	T-S waveforms	Simulation	No	CNN-LSTM	[68]
Power Lines	T-S waveforms	Field	No	1D-CNN with Global Average Pooling layer, Dual Cycle-consistency network	[69], [70]
Conductors	T-S waveforms	Field	No	time-series decomposition and LSTM, CNN-LSTM with attention layer	[71], [72]
Insulators	T-S waveforms	Lab	No	CNN with Bayesian optimization for hyperparameters tuning	[73]

* Treating multiple PDs as a new class

nents easily. The components of the GIS are power conducting components and the control system. The power conducting components are responsible of ensuring the flow of the electric current in the system, and the control systems work on monitoring the behavior of the conducting components. Thermal, mechanical, and electrical faults comprise the main faults that take place in a GIS platform. The authors of [43] published in 2018 collected PRPD data from experimental setup and more than 30 live GIS substations. The five defects studied in this work are floating electrode discharge, surface discharge, corona discharge, insulation void discharge, and free metal particle discharge. A known CNN architecture, LeNet-5, was employed and compared with back propagation neural network (BPNN) and SVM, where statistical features were extracted for the later two and raw data was fed to the CNN. The statistical features extracted from the PRPD patterns include skewness, steepness, asymmetry and cross correlation coefficient of the PD amplitude and rate in both positive and negative half cycles of the applied voltage. In order to optimize the weights of the CNN, the authors trained an autoencoder and used the weights as an initialization for the CNN training. The authors reported an improved average accuracy of the CNN compared to that of SVM and BPNN. Four sources of PD that usually occur in GIS (protruding electrodes, floating electrodes, void defects, and free particles) were simulated in [41]. The authors used a Siamese network where the raw input data are pairs of PRPDs. The motivation behind using Siamese network is that PDs usually result in small datasets. Two identical independent CNN models are trained and the distance between the embedded features resulting from the two CNN models is calculated. As a result, a decision is made whether the pair belong to the same or a different class. The authors compared their proposed architecture with SVM and a CNN model. They reported that the proposed method performed better compared to the latter two. In 2019, the authors of [61] investigated GIS PD data, where

time-series waveforms were collected from laboratory experiments and finite difference time domain simulations. A conditional variational autoencoder (CVAE) was used to generate more training data. A seven-layer CNN model was used for the classification of four different sources of PD, free metal particle, metal tip defects, floating electrode defects, and insulation void. The authors also reported a visualization of the feature maps from the first two convolutional layers. They compared their results with support vector machine (SVM), decision trees (DT), back propagation neural network (BPNN), and a few CNN architecture (LeNet5, AlexNet and VGG16). The proposed CNN model outperformed the above mentioned models. In 2021, four types of PDs that take place in a GIS platform (free metal particle defects, metal tip defects, floating electrode defects and insulation void defects) were experimentally simulated and time-series waveforms corresponding to each type of PD were recorded using UHF sensors (butterfly antennas) [64]. The variability in the dataset was introduced by randomly changing the position of the defect. The deep learning model proposed by the authors was based on depth-wise CNN model where the convolution is divided into two parts: the first part is composed of convolving one channel at a time with the convolution kernel (i.e depth-wise convolution) and the second part is to mix the feature map using a 1×1 convolution kernel (i.e point-wise convolution). A generative adversarial network (GAN) was also used in order to generate more data. The proposed model wass compared with other CNN-based models such as MobileNetV1, MobileNetV2, Xception, ResNet, and LeNet models. The proposed model reported enhanced classification accuracy compared to the other models. Visualization of the feature maps of some layers is presented as well, which highlight what each layer was capable of learning. In the same year, a research group simulated 4 different sources of PDs in a GIS controlled lab environment [65]. Varying the defect location for each of the artificial defects ensured the variability in the collected dataset. The

authors proposed a 1D-CNN model where a multiple scale convolution kernel is used instead of a single scale convolution kernel. Channel shuffling was used on the outputs of the two feature maps produced by the multiple scale convolutional kernel in order to have a unified feature map. Since having labelled data is time consuming and needs expertise knowledge, the authors proposed domain adversarial transfer strategy (DATS) which is inspired by the GAN. Four different unbalanced datasets were acquired from an actual GIS in order to test the performance of the proposed model. The 1D-CNN that was trained on the experimental data is used to classify the on-site GIS PD data where some data had no labels. For the proposed 1D-CNN model, the authors compared the results of the proposed model with traditional 1D- and 2d-CNN models. In regards to the performance of DATS, the authors compared the results with other transfer learning (TL) techniques such as fine-tuning TL and domain adaptation TL. The study reports enhanced results using the proposed framework. The authors suggested that future work will focus on the automatic optimization of hyperparameters and on trying the platform in an online monitoring system. In [67], the aim of the research work was to investigate transfer learning, especially domain adaptive deep transfer learning (DADTL) CNN, for GIS PD diagnosis. The authors used four different datasets for the training of their model. The datasets consisted of measured and simulated data. Data set A included field data from three types of fault (rolling element, inner ring, and outer ring). Dataset B corresponded to the GIS PD simulation data using the finite difference time domain (FDTD) technique, where four sources of PDs (metal particle, tip, floating electrode and insulator air gap defect) were simulated. Dataset C corresponded to a 252kV GIS experimental platform, where signals where captured corresponding to the four defects mentioned in Dataset B. Finally, dataset D corresponded to the PD samples collected from a provincial power company's GIS failure. By the use of maximum mean

discrepancy to minimize the sliced Wasserstein distance (SWD), the authors aimed to ensure that transferable features have minimal discrepancy. The proposed model starts with data pre-processing, where samples from the larger datasets are classified as source domain samples, and samples from smaller dataset are classified as target domain samples. The aim is to minimize the differences between the features learned from both source and target datasets. The authors used residuals units in the CNN based architecture. The authors compared the results with traditional CNNs (LeNet and AlexNet) with the same number of layers. The proposed model reported improved results compared to the other deep learning models especially when the dataset is small.

Transmission Line Networks

Transmission line networks are used to enable the long distance transmission of power. A few research groups have developed deep learning models to identify PD from non-PD signals collected from a publicly-available dataset. ENET Centre in Czech Republic developed a meter to measure the voltage signal induced by the stray electric field along covered conductors, that contained PD or fault signals. The dataset contains noisy real world measurements from high-frequency voltage sensors, where the objective is to identify damaged three-phase, medium-voltage overhead power lines [75]. In 2020, the authors of [70] performed pre-processing of the raw data in order to remove noise and low-frequency component of the signals . The output of this process was the time and frequency representation of the signal by applying short term Fourier transform. The time and frequency domain positive and negative half cycle signals are the input to the proposed deep learning algorithm. The proposed model is a Dual Cycle-Consistency network. Both time and frequency domain branches consists of three blocks. Each block contains a 2D convolutional layer, a Rectified Linear

Unit (ReLU), and a batch normalization layer. The output from block-3 of the time-domain and frequency-domain branches is passed through a global average pooling layer, a shared fully connected layer, and a sigmoid layer. In order to calculate the cycle-consistency loss, the outputs are fed to the dual-domain attention module block (DDAM) for joint learning. The prediction is then based on the weighted average of the output from the fully connected layers and the output from the DDAM block. The results are compared with other models such as Random Forest, Resnet18 + VggNet11, and LSTM. The performance metric used is the Matthews Correlation Coefficient (MCC), in addition to precision, recall, and F1-score. The authors reported better results compared to the other approaches. In 2021, the authors of [69] aimed to classify PD versus no-PD signals using the same publicly-available dataset of damaged power lines [75]. The proposed model was a traditional 1D CNN model where Global Average Pooling (GAP) layer is employed before the fully connected layer. Each sample in the dataset is compromised of voltage of the three phases over one period. For each phase, a highpass filter is used to remove the power frequency, after which a maximum filter is used to extract a set of pulses. Each set of pulses from each phase is the input to the trained 1D CNN. Finally, the decision on the label of the power line is based on the three outputs of each phase. In order to visualize what the model is looking at, in order to decide on the label, pulse activation map (PAM) was used. The evaluation metrics used are Matthews Correlation Coefficient, precision, recall, and accuracy. The authors compared their results with other publicly reported results where models such as LSTM were used. The proposed model showed enhanced results, and the authors suggested that a larger dataset will be more compatible for hyperparameters tuning.

Cables and Solid Insulation

Electric power can be transmitted by underground cables or by overhead transmission lines. The main advantage of underground cables compared to overhead lines is the low maintenance cost. This is linked to the fact that overhead lines are exposed to environmental factors such as storms or lightning. An underground cable consists of one or more conductors which are covered with suitable insulation and the external component is the protecting cover [76]. The major disadvantage of using underground cables though is the problem of degradation and failure of the insulation under high voltage stress. Hence, detecting PDs is crucial for assessing the health of the system. This section reviews the application of deep learning to solve the problem of classification of PDs in cable insulation and solid dielectric using PRPD patterns or time-series waveforms as the input to train the classification model. The performance of a CNN model was evaluated on the prediction of the ageing stage of high voltage insulation material using PRPD data [46]. Three classes of start, middle, and end as well as noise/disturbance were defined for the electrical insulation degradation process representing the ageing that occur in an insulation specimen under electrical stress. Precision, recall, and F1 score were the metrics used for the evaluation of the CNN model. The author reported that the performance is consistent even with changes in the CNN hyper-parameters values. The effect of noise in PRPD patterns on the classification accuracy of different artificial defects in a 11 kV cross linked polyethylene (XLPE) cable joints has been investigated in [54]. There are a total number of five PD sources considered in this study. After training a CNN architecture using noise-free PRPD patterns, transfer learning was performed where the authors used this model to start training another CNN architecture but this time with noisy PRPDs. The results were compared with those obtained using traditional machine learning classifiers where hand crafted features were extracted. The authors reported that the CNN-

based model was able to outperform the models that use manual feature extraction with an increase of 16.9% in the classification accuracy. In 2019, a traditional CNN model was used to differentiate between synthetic PD pulses in power cables [56]. The variability in the synthetic dataset was introduced by the signal to noise ratio (SNR) and the position at which the PD initiated. The model was compared with support vector machine (SVM), where the study reported enhanced results using the proposed algorithm. In the same year, five types of artificial defects in ethylene-propylene-rubber cables in a high voltage laboratory were collected to generate signals containing PD data [57]. Seventeen features were extracted from the time-series waveforms corresponding to characteristics such as pulse width, rise time, fall time, peak voltage, pulse polarity, mean voltage, and root mean square (RMS) voltage. In addition, 16 wavelet features were extracted from the transient signals using Wavelet Transform. In total, 33 features constituted the input corresponding to each signal to the proposed CNN model. Analysis was performed on the effect of the change in the hyperparameters of the CNN architecture such as the number of layers and the convolution kernel sizes. The results were compared with those obtained using SVM and back propagation neural network (BPNN) models. The study reported better classification accuracy when compared with the other two models.

Transformers

Power transformers play a significant role in power systems, so any failure in this apparatus may interrupt the power supply and cause outages and loss of profit. One of the beneficial methods for preventing the failure in the power transformers and raising the reliability of these systems is detecting faults in power transformers accurately and promptly. The following section summarizes the literature on the use of deep learning for the PD sources

classification in transformers. In 2020, the authors of [48] simulated six types of PDs that take place in power transformers using artificial cells in a laboratory setup. They collected the PRPDs of the six PDs which include protruding electrode, moving particle, floating object, surface discharges, bad contact between windings, and void. In order to reduce the input size of the PRPD, the authors used the phase-amplitude (PA) response that is extracted from PRPDs. The authors proposed a CNN model for classification of PD sources. Comparing the classification accuracy of the proposed architecture versus other machine learning classifiers, such as linear and nonlinear SVM, the authors reported a better performance. They also reported that using the PA response as an input increases the accuracy by 1.46% compared to using the raw PRPDs as the input to the CNN model. The PRPD data of different PD sources in a transformer were collected in a laboratory-controlled setup and reported in [49]. The PD sources included tip discharge, surface discharge, air gap discharge, and suspended discharge. The squeeze-and-excitation (SE) module, that is a lightweight attention mechanism, and the nonlinear function hard-swish (h-swish) were used in addition to a CNN model in order to decrease the accuracy loss of the model further. The authors performed image pre-processing such as segmentation, binarization, and enhancement of the data before feeding it to the training model. They compared the results of their model versus other models such as AlexNet, ResNet-18, and VGG16. They reported enhanced average accuracy versus the other models, in addition to less weight storage and reduction of parameters. In the same year, an investigation of a transformer bushing insulation quality, which was affected by poor drying and impregnation, was reported in [50]. The authors used a simple CNN (i.e. 3,300 parameters) for the identification of four types of dry impregnation defects using PRPDs as the input to the proposed CNN. The performance metrics used for the evaluation of the model are the precision rate, recall rate, and F1 score. The authors reported 97.1%

average accuracy rate and indicated that their model can be used for online monitoring as it is a small model. A novel convolutional architecture for single and multiple source PD classification, where the model is trained on single-source PDs, was proposed in [51]. The dataset included PRPDs of single and multiple sources of PD taking place in air, oil, and SF_6 which mimic common sources of PD. The six single PD sources of floating electrode in SF_6, moving particle in SF_6, fixed protrusion in SF_6, free particle in transformer oil, needle electrode in transformer oil, and corona in air were simulated in a laboratory setup. The proposed architecture has a convolutional backbone feeding into multiple fully connected neural networks (FCNs). The performance metrics used are the arithmetic mean of recall and precision in addition to the classification accuracy and false negative rate. The authors compared their results with one-versus-all CNN and reported that their model has better results than the traditional single-branch CNN architecture. Single and multiple sources of corona discharge in a controlled lab environment were simulated in [53]. The four single sources were: sphere–plane, sphere–sphere, needle–plane, and needle–needle. The three multiple sources were: needle–needle and sphere–sphere, needle–plane and sphere–sphere, and sphere–plane and needle–needle. The PRPD patterns were collected and pre-processing was performed by filtering discharges that have small magnitude. The input to the deep learning models were greyscale images of 75 by 75 pixels. The classes were labeled from 0 for the first single class to 6 for the double-sourced configuration, that is considering the multi-source classes as a new class. The authors proposed an optimized ResNet model which they compared with other DL models such as AlexNet, Inception-V3, residual network (ResNet), and DenseNet. The study reported enhanced classification accuracy and least computational cost.

2.3.2 PD Classification Using Autoencoders

In 2017, the authors of [39] simulated four sources of PDs that take place in a GIS platform,where PRPD patterns were collected. The four sources of PDs are: protrusion, contamination, gap, and particle defects. For each of the sources of PDs, four severity states of the PD were collected: normal state, attention state, serious state, and dangerous state. The authors proposed a stacked sparse autoencoder (AE) model, where the output of the middle layer (bottleneck) of the preceding AE is the input to the next AE. The output of middle layer of the final AE is the input to a softmax layer which decided on the assigned severity level label for each sample. The effect of changing different hyperparameters, such as the number of stacked AE or number of nodes in the middle layer, were examined. The proposed model was compared with support vector machine (SVM), where nine statistical characteristics were extracted from the PRPD patternss.The study reports enhanced average classification accuracy of the PD severity compared to SVM. In 2019, the authors of [44] simulated PDs using a laboratory setup and collected PRPD data from a live substation. The four PD sources considered in this work that take place in a GIS platform are: floating electrode defects, metallic protrusion defects, insulation void discharge defects, and free metal particle discharge defects. A variational autoencoder (VAE) was trained to extract the eigenvalues corresponding to the PRPD data. The training set included a mix of both the laboratory and substation data. For the test dataset, a matching algorithm based on cosine distance was used in order to decide to what class the test PRPD belongs to. The proposed method was compared with statistical features, deep belief networks (DBN) [77] and CNNs. The authors reported that the eigenvalues extracted from the VAE feature vector have improved results over the other methods used. Using a laboratory setup, four defects that take place in medium voltage switchgear were replicated [62],where time-series waveforms were

collected. The four sources of PDs included: cable termination floating earth, earth cable in contact with cable termination insulation, voltage presence indicating systems (VPIS) bushing screen disconnected, and earth grounding spring missing on bus bar connector. The spectrogram of the PD signals collected using a coupling capacitor is generated by applying the continuous wavelet transform (CWT). In addition, spectrograms from noises and other HF signals are generated. The authors proposed a convolutional autoencoder (CAE) that is able to reconstruct the spectrograms of the different sources of PDs and noise. After the CAE is trained, the decoder part of the autoencoder is removed and substituted by a fully connected layer followed by the classification layer. This model is trained using a labeled dataset, where the model is able to output the percentage of belonging of a tested spectrogram to each of the 4 PD sources and the noise/ HF signals classes. The study reported high performance ability of the proposed model. In another work [63], four sources of PDs in a gas-insulated switchgear were simulated in a lab setup. An existing CNN architecture, AlexNet, was the method used in this work where the inputs are the time series waveforms treated as images. The results of the proposed method are compared with the fractal method and mean discharge method. The time series waveforms were transformed into a PRPD plot for the sake of applying the fractal method and mean discharge method. Both the fractal and mean discharge methods provide features which are considered as input to two fully connected neural networks. The study included reporting the average classification accuracy of the three models with different percentage of noise added to the signals. The proposed method showed improved results compared with the other two methods especially with a high noise percentage. In addition, the author reported that the time consumed for PD classification was the least using the proposed CNN-based method.

In rotating machines, voltages are generated due to time-varying magnetic fields, which is the result of the the change in the flux [78]. The change in the flux results from the mechanical motion of the rotating machine. Rotating machines consist of stator and rotor structures which are made of thin lamination of electrical steel, insulated from each other in order to reduce losses and prevent discharges and faults to take place. Various types of stress, such as thermal, electrical, ambient, or mechanical stress, can affect the insulation system of rotating machines. Statistical data show PD activities have preceded a large number of stator failures [79] and as such, PD detection in rotating machines has been attracting attention. Extensive research has been done on using traditional machine learning techniques, such as Naïve Bayes-, SVM-, and kNN-based techniques [80], for rotating machine electrical insulation diagnosis. A framework was proposed using visual data analysis for PD source classification in hydrogenerators with a minimum of labeled data [60]. A convolutional encoder was used to project the PD signals acquired from the generator stators to a 2D-visualization latent space. This serves as a visual aid for the expert to analyze the distribution of the training dataset. After being labeled by the experts, the labeled data is trained by a neural network classifier. Other unlabeled data are tested using the already trained classifier, and if any conflict area appeared on the 2D latent space, the human experts will have to label by conflict area sample data. The new labeled data is then added to the dataset and this procedure is done in an iterative manner until the area of conflicted data is minimized. This study reported a base that integrate both expert knowledge and the advantages of deep learning in order to have a correctly-labeled dataset of PD sources.

2.3.3 PD Classification Using Recurrent Neural Networks

A long short-term memory (LSTM) recurrent neural network (RNN) has been used to classify PRPDs in a GIS in [40]. In this work, four different types of defects have been simulated in a controlled lab environment, where the PRPD patterns have been collected. The PD sources simulated are: protruding electrodes, floating electrodes, free particles, and void defects. In addition to that, the authors simulated noise by using an air purifier and the noise signals were obtained using the external UHF sensor. The authors compared the classification accuracy with SVM and fully dense artificial neural network. The study reported that although the proposed model takes more training time, the classification accuracy is superior to the other two machine learning models. Reference [42] presents the simulation of artificial defects that take place in a GIS platform, where the PRPD patterns corresponding to each source of PD were recorded. The four PD sources in this work are: corona, floating electrode, particle, and void. The authors proposed a multi-head self attention LSTM based model for PD (LSANPD) and a self attention based neural network model for PD (SANPD). They compared the classification of the two models with their previous published work which used an LSTM-RNN based model. They reported that SANPD and LSANPD are better in terms of classification accuracy and that SANPD is better than the LSANPD and LSTM-RNN model in terms of complexity.

The authors of [71] developed a model based on time-series decomposition and LSTM for to classify PD from non-PD signals from the same public dataset that was used in [70] . Seasonal-Trend decomposition using Loess (STL) was used to decompose each raw signal into three parts: trend, seasonal, and residual. PD is mostly reflected in the residual part. Four different STL modules with different seasonal window lengths were used to generate four different residual components. Feature engineering was then applied on the residual parts

where a sequential feature vector is extracted. As a result, many-to-one sequential data are generated and are considered as the input to the long short-term memory network (LSTM) classifier. The proposed model was compared with other classifiers such as fully connected layers, SVM, XGBoost, and Multivariate Logistic Regression (MLR). The proposed model showed enhanced classification accuracy compared to the other models.

Adam et. al [52] simulated six artificial PD sources in a controlled lab environment which mimic PDs in power transformer. The PD sources include two discharge sources in air and four discharge sources in mineral oil. The time at which the discharge takes place, the apparent discharge in pC, and the phase angle are recorded for each PD event. 100 PD events constituted a sample. Superimposed patterns were created by using the single sources patterns, where 30 different combinations of samples with two class labels are formed. The authors proposed an LSTM model which is able to classify multiple and single sources of PDs, where the training was done just on single sources of PDs. The study reported the multi-label accuracy in addition to the single-label accuracy. The multi-label accuracy is defined as the proportion of the correctly predicted labels to the total number of labels for each sample. The model showed a 99% average accuracy for single PD sources and 43% for the average multi-label classification problem.

2.3.4 PD Classification Using Hybrid DL Algorithms

In 2018, five sources of PD were generated in a GIS tank model in a laboratory setup that include a floating electrode, a metal protrusion on the conductor and the tank, surface contamination, and free metal particles [66]. Four planar spiral antennas were installed at different locations on the tank. For each time-series signal collected, the authors calculated three different short time Fourier transform (STFT) by changing the window lengths. The

different window lengths correspond to high time resolution, high frequency resolution and medium resolution. The proposed model was a CNN-LSTM based model. The three different STFTs calculated from each signal were used to train three different CNN models, where the three outputs of the CNN models are combined by a fully connected layer. The output of the fully connected layer is the input to the LSTM. Since there are four sensors, the model is comprised of four fully-connected layers which are the input to four separate LSTMs. The outputs of the LSTMs collectively decide on the label of each input sample. The authors compared the model performance with other baseline models and with the case where a single window length was used for the STFT. The model showed improved results compared to the other models. Another work which used simulated data is presented in [68], where a CNN-LSTM network is proposed for the classification of PD sources in a GIS system. Time-series waveforms were collected. The model consists of two blocks of convolutional layers followed by pooling layers. The output of the second block is fed to an LSTM layer which is followed by a fully connected layer and ending with the classification layer. In order to generate the dataset, simulation software XFDTD is used. The four sources of PDs are metal tip defect, insulator air gap defect, floating electrode defect, and free metal particle defect. The authors reported the precision, recall, and F1-score of the four sources of PDs. They compared the performance of the proposed model with other models like SVM, LSTM and CNN. The proposed model reported high average classification accuracy compared to the other models.

In [72], the authors aimed to classify PD versus no PD using the publicly available dataset that was introduced earlier [75], which included the three phase voltage signals. FFT noise reduction algorithms were used on the raw data. The proposed model was a CNN-LSTM model with an attention layer before the classification layer. Starting with two blocks of

convolutional and max-pooling layers, the output is fed to a fully connected layer which is considered as the input to the LSTM layer. The output of the LSTM is the input to an attention layer, where multiplication of the feature vector obtained from the LSTM is done with learnable weight coefficients. The output of the attention layer is fed to to a sigmoid which decides on PD versus no PD label. The performance metrics used are precision, recall, and F1-score. The proposed model is compared with other traditional models such as SVM, CNN, and bidirectional LSTM. The study reported higher average accuracy compared to the other models.

An ensemble of deep learning algorithms was used to differentiate between PD signals and noise signals in medium voltage power cables [59]. The samples were collected from offline in-service cables. The idea behind using ensemble learning is to allow more than one neural network to make the classification decision. CNN, convolutional RNN, LSTM, and bidirectional LSTM (BILSTM) were used in the ensemble frame but two of these models were used at a time. In this paper, two scenarios were considered. In one scenario where there is difference in the prediction of a sample between two deep learning (DL) models, a human expert will have to decide on the label of that sample, and in the other, the output from the activation function of two different models were added together. Five different cables were tested on the trained models. Adaption training was done for each of the five cables where the classifier layer is re-trained with the measured calibration pulse specific to each cable. The authors reported the results of the five cables for the two ensemble scenarios and with different binary selection of the above-mentioned DL algorithms. They also reported the results when each DL model is used alone. It was reported that the CNN paired with the BILSTM gave the best results.

Four typical transformer insulation defects were simulated in [47] that include metal protrusion, oil paper void, surface discharge, and floating potential defects. The authors developed a CNN-LSTM based model where the input is the PRPD data. They compared the results with a CNN-only model and an LSTM-only model. The evaluation metric used was the classification accuracy where the authors reported that CNN-LSTM has better overall recognition accuracy than CNN and LSTM alone.

Research has been focused on using different DL techniques for the purpose of identifying PD sources varying from autoencoders, CNN, RNNs , or a combination of the above techniques. Authors have compared their proposed models with other DL models or traditional machine learning models. Despite the advantages of using different DL techniques, future work should focus on the quality of the input data in regards to the interference, noise, and the fact that multiple PDs can take place at the same time. The integration of the developed models in real-life systems would present its own challenges especially when it comes to developing industrial standards or regulations, and implementing condition-based maintenance asset management policies depending on the severity of the situation. Although the investigation of different DL techniques is essential for the sake of completeness of any work, it is observed that having common, publicly-available datasets can help in focusing on the generalization of any developed algorithm. In addition,similar to the application of DL in any area, the prerequisite of developing any DL algorithm to assess the health of an asset is the availability of reliable training data. Most of papers in the literature, either employ simulated lab data or even numerically-generated data. Training a DL algorithm with such data will make the performance of the algorithm in the field require proper validation.

2.4 Summary

Monitoring of electrical insulation of high voltage apparatus is crucial for the reliable operation of power systems. Extensive research has been done on the classification of sources of partial discharge (PD). Modern techniques have been based on machine learning methods, where the input to such methods is composed of manually-extracted features, i.e. feature extraction has required the intervention of human experts. Deep learning, which is a branch of machine learning, has been used to enhance the performance of PD classification. This enhancement is attributed to the capability of deep learning techniques to use raw data as the input to the classification model. In other words, instead of using manually-extracted features, raw data such as PRPD patterns, time-series waveforms, or images are used as the input to the deep learning systems. This allows the classification model to be fully automated where the feature extracting stage is integrated in the learning stage. Gaps have been identified in regarding to the classification of multi-sources of PDs and the identification of number of PD sources.

Chapter 3

Multi-Class Classification of PD Sources Using PRPD Patterns

All of the prior methods reported in literature to classify multi-source PDs depend on the availability of training data from multi-source PD inputs [81]. There are a number of drawbacks associated with this choice. Such a training data is difficult to collect in practice, is time consuming, and, by its very combinatorial nature, precludes the collection of examples for all possible combinations of concurrently occurring defects. In this chapter, to address these drawbacks, a novel convolutional architecture is proposed for single-source PD and multi-source PD classification using training data with ground-truth available only at the level of single-source PDs. The proposed architecture consists of a convolutional backbone feeding into multiple fully connected neural networks (FCNs). The input to the convolutional part of the network is the PRPD pattern matrix. The output of this CNN stage is a common feature representation which is broadcast to different FCNs, wherein each FCN is trained to output the probability of occurrence of a specific PD. Thus the proposed hybrid

architecture moves from extracting general representations to more fine-tuned representation in a hierarchical fashion. The overall loss of the network is the combination of individual binary cross entropy losses from each of the FCNs. This loss is jointly optimized with respect to the parameters of the CNN stage and the FCNs. At testing time, our network produces a multi-label output vector signifying the probability of the presence of respective PDs.

3.1 Experimental Setup

Several experiments were performed by different groups to classify partial discharges by the use of their phase resolved partial discharge patterns. PD classification and identification using laboratory data has been used to establish proof of concept for a number of techniques available in the literature *(*e.g. [39,82]). Lab experiments were done by Janani *et al.* [83,84] to simulate artificial defects. The experimental setup consisted of a high voltage transformer, a capacitive divider to measure the AC voltage, the test cell, and the PD measurement system. The lab setups simulate common sources of PD in air, oil, and SF_6. PD data collection was conducted in accordance with IEC 60270 standard [85]. The test cells include three sources (floating electrode, moving particle, and fixed protrusion) of partial discharge in SF_6, two sources of PD (free particle and needle electrode) in transformer oil, and corona in air which has the same setup as floating electrode but filled with air. For the floating electrode, the distance between the gap between the two electrodes is 1 mm. For the free particle, a small bearing with a diameter of 3.17 mm was placed on a concave dish ground electrode. For the point plane electrode, the needle has a diameter of of 20 μm. More details on the experimental setups are explained in [84]. In total, there are six different PD patterns generated. In addition, four different combinations of multiple partial discharges were simulated by using

simultaneously two or three test cells. A commercial PD measurement system (Omicron MPD 600) is used to acquire the PRPD of each test cell.

The output data from the Omicron software is exported as binary files. This data includes information about partial discharges taking place relative to the applied phase voltage. The discharge magnitude and phase are divided into 400 and 500 bins respectively. This results in a 400×500 matrix $M(x, y)$, where the number in each bin represents the number of discharges occurring at a specific phase angle (x) and a specific discharge magnitude (y). Figure 3.1 shows a visual representation of the six single-source PRPD patterns. The 400×500 matrix is reduced to 100×100 by summing up the counts in each 40×50 sub-matrix. In addition, background noise is unavoidable even with perfect measurement conditions. Background noise is reflected by having an offset charge over all the phase windows. In this work, noise is removed for all the PRPD patterns. In addition to the six classes, an additional no-pattern class corresponding to the cases not involving the presence of any PD is added. In order to encourage the model to learn features related to the shape of PRPDs, the samples are converted into binary samples, where zero threshold is considered for binarization. A visual representation of the binary matrices are shown on the side of the Omicron representation of each of the single PD source- six classes.Fig. 3.2 shows the PRPD patterns of the four multi-source PD classes. To make the systems insensitive to changes in charge magnitude settings on Omicron software and to different applied voltages, samples are extracted using multiple magnitude settings; hence, introducing variability in the dataset. Particularly, for each class, the following charge magnitude settings for the Omicron software were employed that are summarized in Table 3.1.

The numbering of the six single-source classes is done as follows:

- *Class 1* (corona in air)

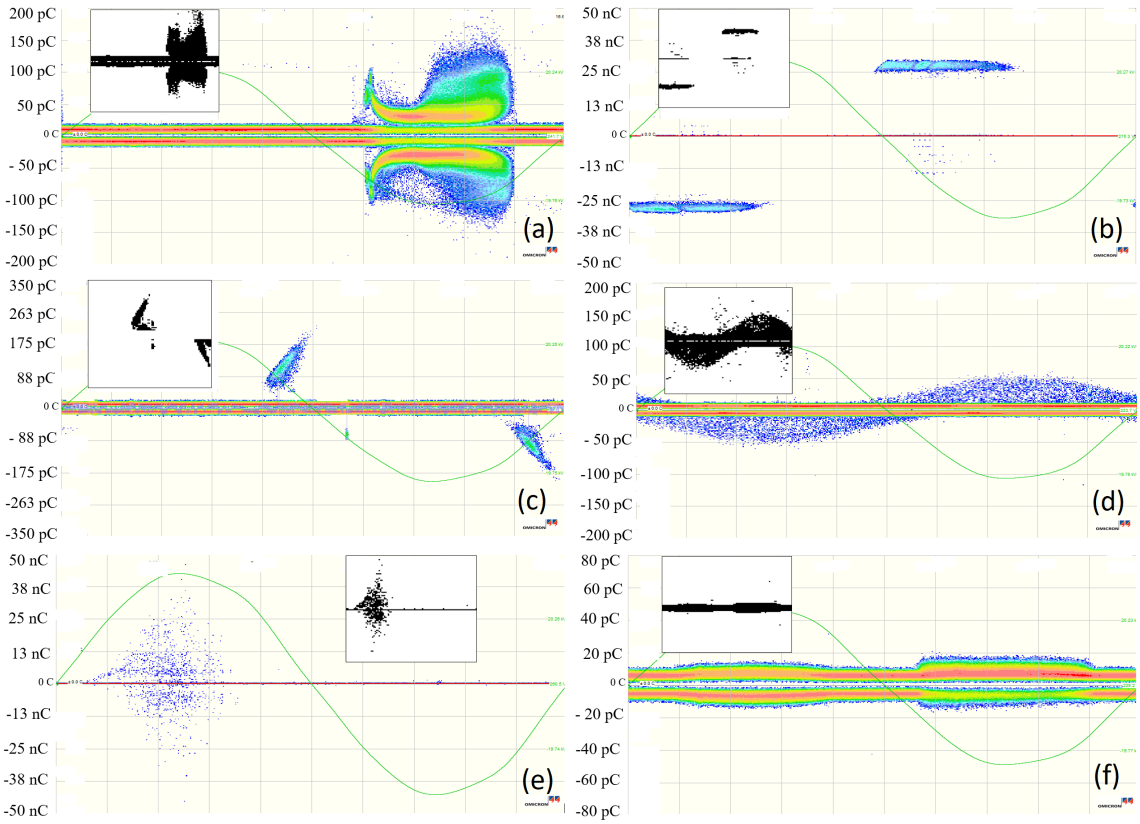


Fig. 3.1: PRPD patterns and their binary representation of various single defects: a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6.

Table 3.1: Different levels of charge magnitude scale setting on the Omicron Software.

Classes	Charge Magnitude Scale Setting
Class 1	100, 200, 250, 500, and 1000 pC
Class 2	70, 100, and 200 nC
Class 3	200, 300, and 350 pC
Class 4	70, 150, and 250 pC
Class 5	10, 50, and 100 nC
Class 6	20, 50, 100, and 200 pC

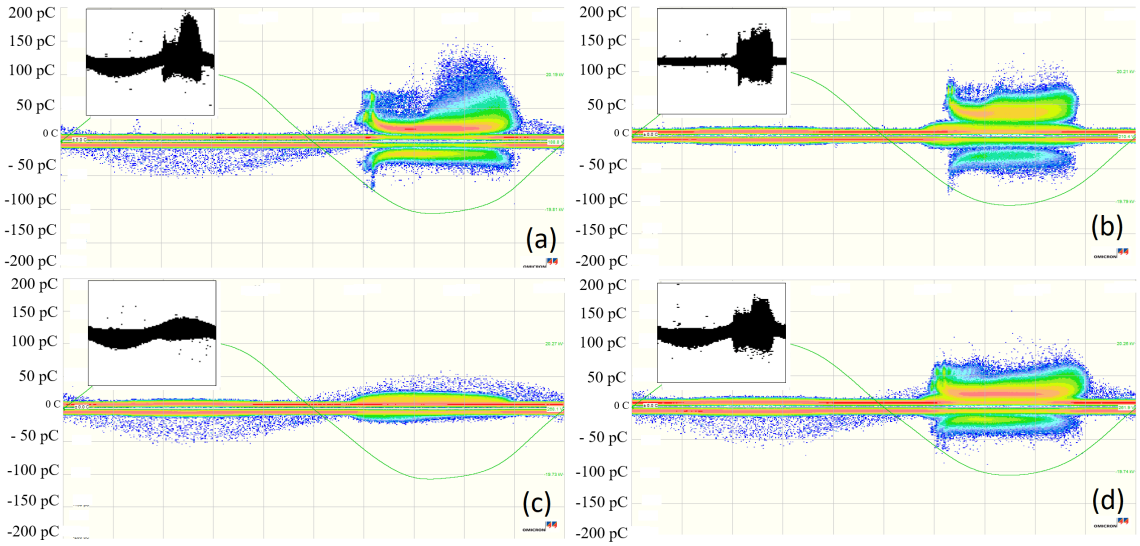


Fig. 3.2: PRPD patterns and their binary representation of various multiple defects: a) class 14; b) class 16; c) class 46; d) class 146.

- *Class 2* (floating electrode in SF_6)
- *Class 3* (free particle in oil)
- *Class 4* (free particle in SF_6)
- *Class 5* (point plane electrode in oil)
- *Class 6* (point plane electrode in SF_6)

The numbering of the four multiple-source classes is done as follows:

- *Class 14* (corona in air and free particle in SF_6)
- *Class 16* (corona in air and point plane electrode in SF_6)
- *Class 46* (free particle in SF_6 and point plane electrode in SF_6)
- *Class 146* (corona in air, free particle in SF_6 and point plane electrode in SF_6)

3.2 Method

In classification problems, a data-point can belong to a single class (mutually exclusive membership) or it could belong to multiple categories at the same time. The latter is usually referred to as multi-label classification. Since PRPD patterns from multiple sources can occur concurrently, PD detection is essentially a multi-label classification problem. In the presence of training data with various combinations of co-occurring multi-source PD labels, building a multi-label classification model is tenable. However, as mentioned in Section 1, collection of such a dataset is expensive, time consuming, and may not allow to span all possible combinations of PDs. On the other hand, it is practically more feasible to collect single-source PD data in large quantities. We therefore focus on methods to capitalize single-source training data for solving multi-label classification problem.

Let K be the number of PD sources. To enable explicit detection of cases with no PDs, we define a separate category representing the absence of all the PDs. Let the training data consisting of N examples be represented as $\{\mathbf{X}_i, \mathbf{y}_i\}_{i=1}^{N}$, where $\mathbf{X}_i \in \mathbb{R}^{H \times W}$ is the i^{th} PRPD pattern image, and $\mathbf{y}_i \in \{0,1\}^{K+1}$ is the corresponding $(K+1)$-dimensional* binary label vector, signifying the presence or absence of each PD. Since only single-source examples are considered during training, each $\mathbf{y}_i$ is a one-hot vector. At testing time, the label vector for a test-case can contain multiple 1s.

Multiple single-source classifiers (baseline): To achieve multi-label classification, a traditional way is to learn multiple $(K+1)$ independent binary classifiers, each trained to detect an individual PD defect. The loss for the k^{th} model, given the training dataset, is

*The label vector is $(K+1)$-dimensional because, as described above, we have defined an additional class for cases with no PDs.

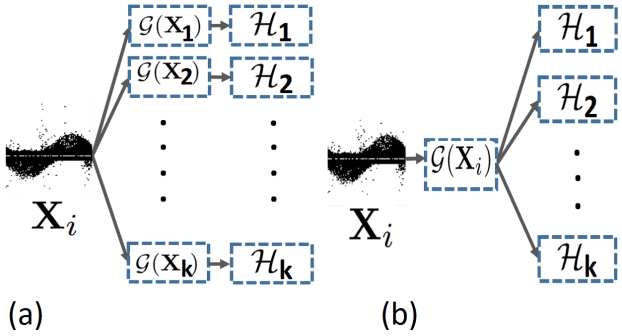


Fig. 3.3: a) baseline model with independent model for each class; b) proposed model with a common convolutional backbone shared across all classes.

given by

$$\mathcal{L}_k(\boldsymbol{\theta}_k) = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_{ik}\log(\mathcal{F}_{\boldsymbol{\theta}_k}(\mathbf{X}_i)) + (1-y_{ik})\log(1-\mathcal{F}_{\boldsymbol{\theta}_k}(\mathbf{X}_i))\Big] \tag{3.1}$$

where $\mathcal{F}_{\boldsymbol{\theta}_k}$ is the function encoded by the k^{th} model, depending on parameters $\boldsymbol{\theta}_k$. y_{ik} is the k^{th} element of $\mathbf{y}_i$. After the training phase, given a test case $\mathbf{X}^{(\text{test})}$, one then needs to invoke $K+1$ models to build a multi-label output, $\hat{\mathbf{y}}^{(\text{test})} = \{\mathcal{F}_{\boldsymbol{\theta}_k}(\mathbf{X}^{(\text{test})})\}_{k=1}^{K+1}$. An example convolutional architecture that accepts a PRPD pattern image and performs a binary classification for the presence of a specific single-source PD is shown in Fig. 3.3 (a).

Joint model with shared CNN parameters (Proposed): While the baseline approach described above may learn excellent single-source classifiers, it is not expected to generalize for multi-label classification task. This is due primarily to over-tuned class specific parameters $\{\boldsymbol{\theta}_k\}_{k=1}^{K+1}$ learned independently for each single-source PD. To address these issues, decomposition of the network parameters into two sets has been proposed: a shared set of common parameters, $\boldsymbol{\rho}_{\text{CNN}}$ (for the convolutional part), and class specific parameters, $\{\boldsymbol{\phi}_{\text{FCN}_k}\}_{k=1}^{K+1}$ (for fully connected networks). In particular, the proposed architecture has a

shared convolutional stage for feature extraction. These features are then distributed to multiple FCNs. The motivation is to encourage the CNN to learn to extract more general feature representations which are useful for all classes. The FCNs accept these general features to learn class specific models in a joint fashion. The proposed architecture is shown in Fig. 3.3 (b). Let the CNN part be represented by the network $\mathcal{G}$, and each of the fully connected networks be represented by $\mathcal{H}_k$. The joint loss function is then given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\rho}_{\text{CNN}}, \{\boldsymbol{\phi}_{\text{FCN}_k}\}_{k=1}^{K+1}) = -\frac{1}{N(K+1)} \sum_{k=1}^{K+1} \sum_{i=1}^{N} \Big[y_{ik} \log \mathcal{H}_k(\mathcal{G}(\mathbf{X}_i)) \\ + (1 - y_{ik}) \log(1 - \mathcal{H}_k(\mathcal{G}(\mathbf{X}_i))) \Big].\end{aligned} \tag{3.2}$$

Design details for network layers: Two CNN layers consisting of 36 filters, a kernel of size 3×3 followed by two dense layers with 128 and 64 filters respectively, and ending with a classification layer of seven nodes constitute the network used in this study. Batch normalization is used in order to decrease the effect of over-fitting. The schematic for one of the classifiers in Fig. 3.3(a) is shown in Fig. 3.4. The design details for the implemented classifiers are shown in Table 3.2. The hyper parameters of the neural network such as the number of layers, number of nodes per layer, kernel size are chosen by running experiments for different values of the parameters and plotting the training and validation accuracy curves as a function of epochs. The design of the layers is kept the same for both the baseline model and the proposed model so that the difference in performance due to the proposed parameter-sharing based architecture can be investigated. The proposed model architecture is shown in Fig. 3.5.

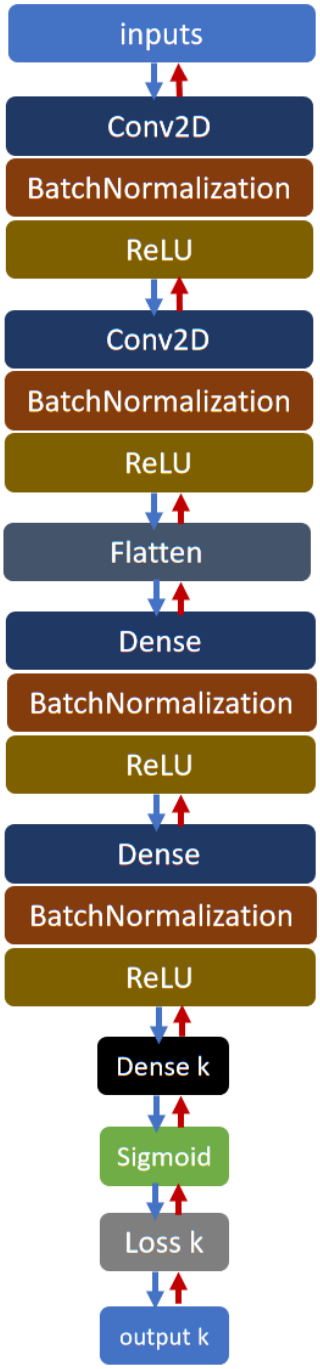


Fig. 3.4: CNN architecture of an independent classifier.

Table 3.2: Design specification of an independent classifier.

Layer Type	Output Shape
Input Layer	(None,100,100,1)
Batch Normalization	(None,100,100,1)
Convolution1 2D	(None,100,100,36)
Max-pooling1 2D	(None,50,50,36)
Batch Normalization1	(None,50,50,36)
Activation1	(None,50,50,36)
Convolution2 2D	(None,50,50,36)
Max-pooling2 2D	(None,25,25,36)
Batch Normalization2	(None,25,25,36)
Activation2	(None,25,25,36)
Flatten	(None,22500)
Dense1	(None,128)
Batch Normalization3	(None,128)
Activation2	(None,128)
Dense2	(None,64)
Batch Normalization4	(None,64)
Activation4	(None,64)
Dense3	(None,1)
Activation4	(None,1)

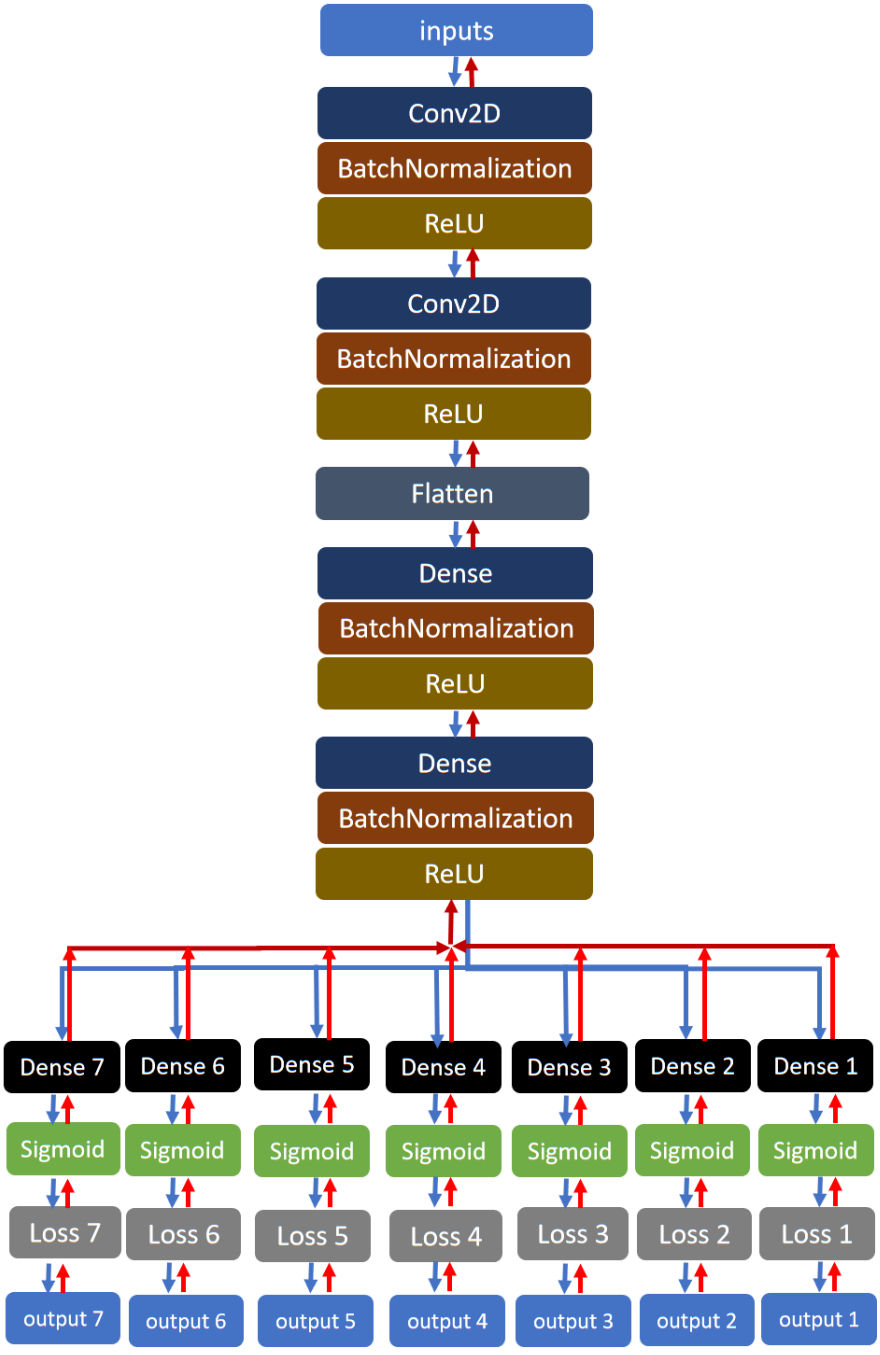


Fig. 3.5: Proposed deep learning model architecture: common convolutional backbone shared across all classes.

3.3 Results

3.3.1 Performance Metrics

Since the model is trained using single PRPD patterns only, the generalization of the model is tested by evaluating the performance on a new hybrid dataset that includes PRPD patterns from single as well as from multiple partial discharge sources. In addition to that, samples with different charge magnitude specification on the Omicron software are tested. Different standard multi-label classification metrics have been used in the literature to evaluate the performance of trained models. Some of these metrics include mean average precision, 0-1 exact match, macro and micro F1, per class precision, per class recall, overall precision and overall recall [86]. The individual recall (Recall(k)) and the individual precision (Precision(k)) are calculated for each of the classes 1 to 7 by taking into account both single-source PDs and multiple-source PDs. The recall reflects the proportion of the positive examples that is correctly classified, and the precision reflects the proportion of the examples predicted to be positive that are actually positive. PCR and PCP represent the arithmetic mean of recall and precision respectively,

$$\text{PCR} = \frac{1}{K}\sum_{k=1}^{K}\text{Recall(k)} \tag{3.3}$$

$$\text{PCP} = \frac{1}{K}\sum_{k=1}^{K}\text{Precision(k)}. \tag{3.4}$$

In addition, classification accuracy and false negative accuracy are evaluated for single as well as for multiple classes. The classification accuracy is calculated considering equal weights for all classes, while the false negative accuracy is calculated taking into consideration only the

true class or classes that the sample truly belongs to. The importance of calculating the false negative accuracy metric in this context comes from the fact that it is of high importance to detect the correct source of PD in high voltage systems. Consistent false identification of a PD source will put the high voltage apparatus in failing condition, in addition to safety risk for employees working near this apparatus. The false negative accuracy reflects on the performance of the model by quantitatively evaluating single classes and multiple classes separately in comparison to the individual recall and precision. If a PRPD pattern belongs to classes one, four and six, then the ground truth is [1001010]. The classification accuracy is then calculated by checking the matching elements in each of the seven-element vector [1001010]. The classification accuracy for a single sample is calculated as

$$P_{\text{classification}} = \sum_{k=1}^{7} \frac{M_k}{7} \times 100 \tag{3.5}$$

where M_k is equal to one when the element k in the ground truth vector agrees with the prediction of the model for the corresponding class k, and zero otherwise. The ideal classification accuracy is 100% and the worst is 0%.

The false negative accuracy for a single sample is calculated as:

$$P_{\text{false negative}} = \sum_{j=1}^{T} \frac{N_j}{T} \times 100 \tag{3.6}$$

where N_j equals one when the element j in the ground truth vector which is equal to one does not agree with the prediction of the model for the corresponding class j, and zero otherwise. In this metric, checking the matching prediction is performed only on the class or classes that the sample truly belongs to. T is the number of classes that the sample truly belongs to. In our tested dataset, T can be 1 for single-sourced PDs, 2 or 3 for multiple-sourced

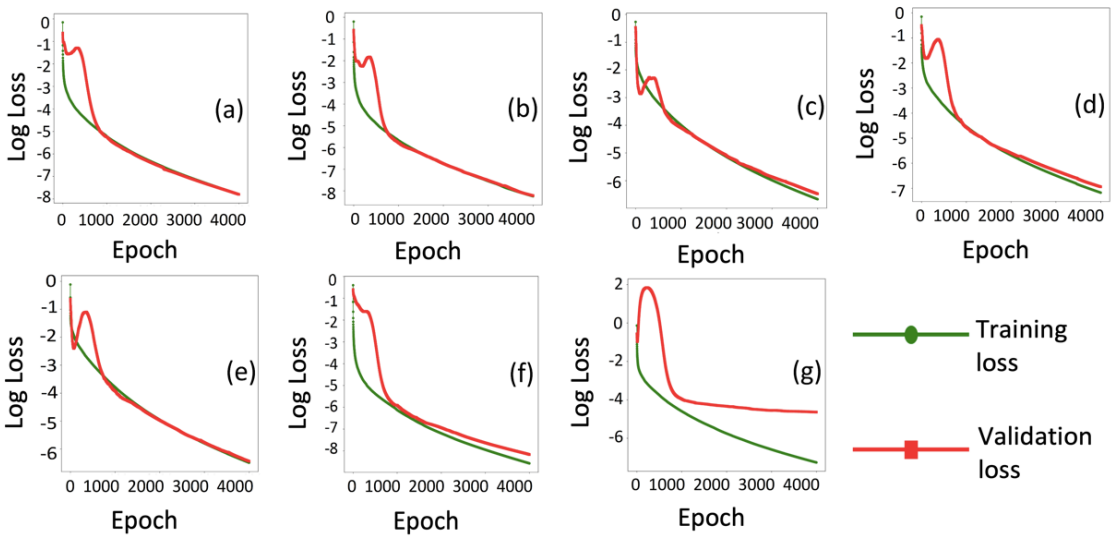


Fig. 3.6: Training and validation log losses vs. number of iterations for independent classifiers a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6; g) class 7.

PDs. Hence, the ideal false negative accuracy is 0% and the worst is 100%. Calculating the classification and false negative accuracies over a number of samples is done by averaging (3.5) and (3.6) over the number of samples.

3.3.2 Independent Classifiers

Fig. 3.6 shows calculated loss of the trained model, using (3.1), as a function of epochs (or iterations) for both the training and the validation dataset in log scale. An epoch is when the entire dataset is passed forward and backward through the neural network. Underfitting is clearly seen for classes six and seven where the gap between the training loss and validation loss increases at epoch 1000. Table 3.3 shows the classification and the false negative accuracies.

As seen in Table 3.3, the model does not generalize well to the multiple classes especially Class 16 and Class 146. The arithmetic mean of both precision and recall are shown in

Table 3.3: Accuracy of single and multiple source PRPD patterns.

Classes	Classification Accuracy	False Negative Accuracy
Class 1	87.71%	0%
Class 2	97.7%	0%
Class 3	100%	0%
Class 4	99.82%	0%
Class 5	100%	0%
Class 6	88.44%	0%
Class 7	100%	0%
Class 16	79.42 %	44 %
Class 46	78.28 %	0 %
Class 14	85.71 %	0 %
Class 146	65.71 %	50 %

Table 3.4. The precision for each of class three and class four is low similar to the recall of class six. In Table 3.5 we show the hybrid confusion matrix in which the rows and columns represent the input and predicted classes respectively. The true positives are highlighted for better visibility.

3.3.3 Proposed Model

As we proceed with training the model, a trade off takes place between generalization and learning deeper features about single partial discharges. Generalization comes in the context of correct classification of multiple source-PRPD patterns. The training of the model is terminated when the validation accuracy is observed to start shifting from the training accuracy. During the training phase, a portion of the dataset is used for validation purposes where this portion is used to calculate the loss for back propagation in each epoch. The decision is collectively made by analyzing the average validation and training loss of the seven

Table 3.4: PCR and PCP for independent classifiers.

Classes	Recall(i)	Precision(i)
Class 1	1	0.83
Class 2	1	1
Class 3	1	0.38
Class 4	0.89	0.6
Class 5	1	0.89
Class 6	0.65	0.99
Class 7	1	0.95
Arithmetic Mean	PCR: 0.93	PCP: 0.8

Table 3.5: Hybrid confusion matrix of independent classifiers.

		Predicted Classes						
		1	2	3	4	5	6	7
Input Classes	1	100	0	49	37	0	0	0
	2	0	140	10	0	1	0	10
	3	0	0	130	0	0	0	0
	4	0	0	10	97	0	1	0
	5	0	0	0	0	130	0	0
	6	0	0	5	90	10	120	0
	7	0	0	0	0	0	0	200
	14	50	0	50	50	0	0	0
	16	50	0	14	14	0	6	0
	46	50	0	26	50	0	50	0
	146	50	0	45	25	0	0	0

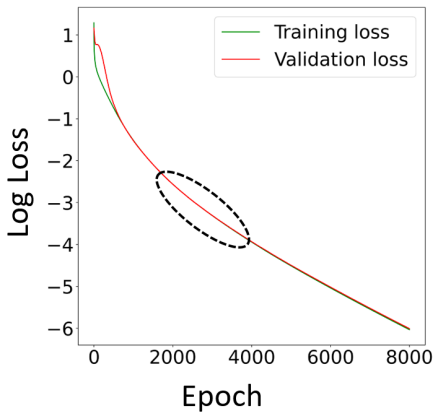


Fig. 3.7: Decision on stopping criteria: the percentage difference between the validation loss and the training loss is minimal in the marked region.

classes. For epoch 4000, the percentage difference between the validation and the training loss is 0.8% compared to 2.7% for epoch 8000 as shown in Fig. 3.7, and consequently, the training is stopped at iteration (epoch) 4000.

The calculated loss of the trained model as a function of epochs or iterations for both the training and the validation dataset in log scale, using (3.1), is shown in Fig. 3.8. The classification accuracy and false negative accuracy are shown in Table 3.6. In comparison with Table 3.3, better performance is recorded where the average classification accuracy for single classes increased from 96.2% to 99.6%. The average false negative accuracy for the multiple classes dropped from 23.5% to 8.7%. On the other hand, the arithmetic mean of both precision and recall are calculated in Table 3.7. Comparing Table 3.7 with Table 3.4, ideal recall is recorded for class 6 and ideal precision is recorded for all classes. This indicates that our proposed model enhanced the prediction of true positives. The hybrid confusion matrix for the proposed model is shown in Table 3.8. As seen in these tables, compared to Table 3.5, our proposed model has enhanced classification ability not only for single-source PDs, but also for multi-source PDs. This is shown in the last four rows of Table 3.8

Table 3.6: Accuracy of single and multiple source PRPD patterns.

Class	Classification Accuracy	False Negative Accuracy
Class 1	100%	0%
Class 2	100%	0%
Class 3	100%	0%
Class 4	97.17%	0%
Class 5	100%	0%
Class 6	100%	0%
Class 7	100%	0%
Class 16	100%	0%
Class 46	100%	0%
Class 14	96.28%	13%
Class 146	90.57%	22%

Table 3.7: PCR and PCP for the proposed model.

Class	Recall(i)	Precision(i)
Class 1	1	1
Class 2	1	1
Class 3	1	1
Class 4	0.86	0.89
Class 5	1	1
Class 6	1	1
Class 7	1	1
Arithmetic Mean	PCR: 0.98	PCP: 0.99

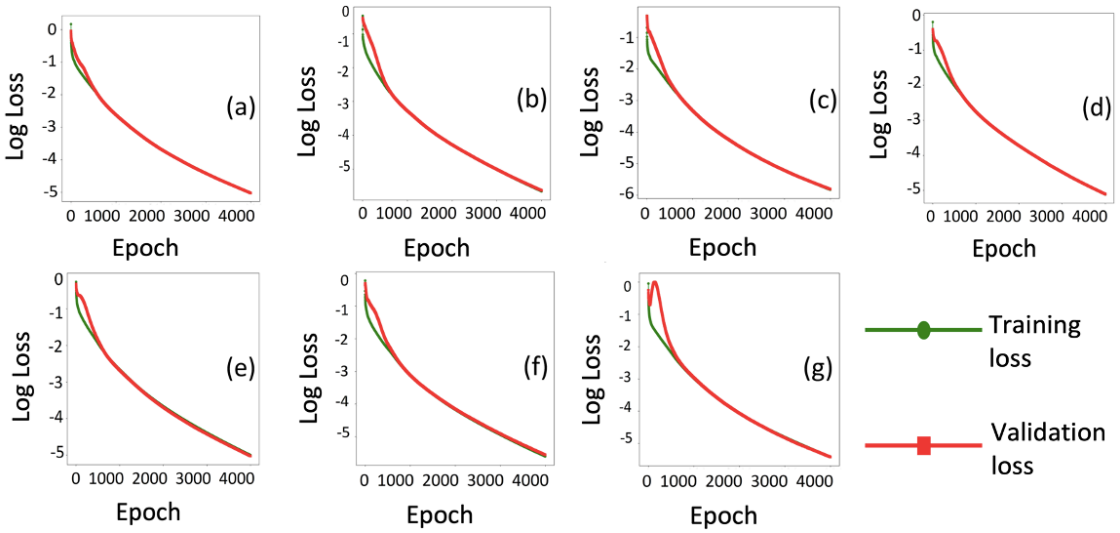


Fig. 3.8: Training and validation log losses vs. number of iterations for the proposed model a) class 1; b) class 2; c) class 3; d) class 4; e) class 5; f) class 6; g) class 7.

corresponding to the multiple classes and comparing them with that of Table 3.5. This indicates that our proposed model decreased false negatives predictions.

3.4 Summary

A novel architecture based on deep learning algorithm particularly CNN has been developed in order to identify single and multiple sources PDs which can occur in high voltage insulation systems. The difficulty of identifying multiple sources PDs using training set of single sources PDs results from the fact that the PRPD patterns are partially overlapping. As a result, traditional machine learning techniques which are based on the manual extraction of features get confused when multiple source PRPD patterns are set to be classified. Different algorithms should be deployed in order to decide on the separation criteria between these overlapped PRPDs. A new architecture based on CNN has been shown to be useful for

Table 3.8: Hybrid confusion matrix of the proposed model.

Input Classes \ Predicted Classes	1	2	3	4	5	6	7
1	100	0	0	0	0	0	0
2	0	140	0	0	0	0	0
3	0	0	130	0	0	0	0
4	0	0	0	97	0	12	0
5	0	0	0	0	130	0	0
6	0	0	0	0	0	120	0
7	0	0	0	0	0	0	200
14	50	0	0	37	0	0	0
16	50	0	0	0	0	50	0
46	0	0	0	50	0	50	0
146	50	0	0	17	0	50	0

this problem through the proposed enhanced version based on sharing the weights among different classes. The essence of the proposed model is that the training is done on single sources of PDs only. This is appreciated in the industry where additional financial resources and time are needed to acquire data from simultaneous sources of PDs. The model is robust to electric interference as well as to applied phase voltage. The average percentage improvements of the proposed architecture for single-source PDs and multi-source PDs are 99.6% and 96.7%, respectively compared to 96.2% and 77.3% for that of the independent classifiers architecture.

Chapter 4

Limitations of Off-the Shelf CNN for PD Signal Classification

A CNN - based model for multi-source PD classification using PRPD patterns as an input was proposed in the previous chapter. In this chapter, the attention of the research is continued towards the multi-source PD classification using time-series waveforms as an input. The first step is detecting the limitations of off-the-shelf CNNs for the PD time-series waveforms classification ; whether it is for immunity to noise and interference ,or classification of multi-source PDs using training of single source time-series waveforms. This is the focus of this chapter.

Diagnosis of the insulation degradation in any high voltage electrical system is very important for monitoring its performance. Measurement of partial discharges (PD) has been a well-known tool for this purpose. With the expansions in the field of pattern recognition, automating partial discharge recognition and classification got more attention. One of the pioneers for developing methods to measure partial discharges were Okamoto and

Tanaka [38]. It has been shown that defects in the insulation system can be detected by analyzing the magnitude-phase representation of the PD signals by using phase resolved partial discharge (PRPD) patterns. PRPD patterns represent the amplitude and the rate of the pulses generated versus the time reference of the test voltage signal. Using PRPD as a classification tool assumes each PD source has a unique discriminating PRPD pattern. However, using this technique can be problematic for a number of reasons; first, capturing and recording PRPD patterns require extra hardware equipment. Second, in the presence of increased noise during online measurements, the patterns often get distorted to the extent that the identification of the PD source becomes impossible. Thus, more research has been focused on interpreting the time-series waveforms of the pulses corresponding to the active sources of PDs [87, 88]. Although the PD signals differ from one measurement system to another, the classification of PD sources within a particular measurement system can be done by analysing the PD signals. The characteristics of each of the captured PD signals are determined not only by the measurement system, but also by their physics and the travelling path to the PD acquisition system [89]. This is accompanied by the necessity of applying effective techniques to filter out the background noise [90,91]. With the advancement of power electronic devices, not only is background noise of concern, the internal noise and external interferences also make the problem difficult. All these disturbances are factors responsible for degrading the functionality of a PD detection system. Thus, an extensive amount of research has been carried out to suppress noise and interference in a PD detection system. Researchers have been using FFT-based methods and digital filtering methods to eliminate interference since the 1990's [92] . Many approaches have been implemented in order to differentiate PD pulses from interference and noise pulses since then. Some of the approaches are, for example, based on signal energy ratios analysis [93], wavelet techniques [94], and

pulse clustering by means of time-frequency (T-F) maps [95]. Over the years, statistical time-domain and frequency-domain methods have been used to extract features from the PD waveforms. Selecting discriminate features (from the pool of all the extracted features) is typically performed that will be used to train a classifier. As a result, the performance of the PD detection system is dependent on the features used for training the classifier. Recent advances in the application of deep learning algorithms on raw data has enabled the integration of feature selection stage within the classifier. Convolutional neural networks (CNN) and long short-term memory (LSTM) models have been used to classify different sources of PDs [56,61]. However, the raw input data to these algorithms mostly consist of high quality noise-free signals which are not representative of real-life noisy signals.

Moreover, in real life, different sources of PD might exist in an electrical insulation system simultaneously. This means that the PD classification problem changes from a single-class to a multi-class and multi-label classification problem, where a PD pulse sample can be assigned multiple labels at the same time. However, most of the existing research on PD source classification has focused on the classification of single-class sources of PD. To tackle the multi-class PD classification problem, most approaches consider the multi-source scenario as a new class to be trained rather than a combination of multiple single classes [96,97]. The authors in [98] used long short-term memory (LSTM) in order to classify different sources of PD using time-series waveforms. The above work depends on the assumption that PDs do not happen at the same time i.e., while PRPD patterns can be superimposed, time-series waveforms of different sources do not overlap. The same authors expanded their work in [52] to train a sequence of given number of PD pulses instead of a single PD pulse using LSTM. Using the training of single sources of PDs in a sequence, the aforementioned model was reported to be able to detect multiple sources in each sequence.

In this chapter, the performance of one-dimensional CNN model is explored when noise is added to the PD pulses. In addition, a study is presented to examine the limitations of multiple one-versus-all, one-dimensional (1D), CNNs for multi-label classification of PD.

4.1 Presence of Noise and Interference

This study is conducted to evaluate the performance of a one-dimensional CNN model and to evaluate how prone it is to noise and interference. Two case studies have been explored where synthetic and experimental datasets are used. Independent binary classifiers are implemented for each dataset, where the evaluation of the model is performed on samples with different interference frequencies and different levels of signal to noise ration (SNR).

4.1.1 Datasets

In order to test the classification capability of the 1D-CNN, a controlled dataset is generated to establish a proof of concept. Two classes corresponding to two types of PD sources have been represented as Gaussian waveforms of unity magnitude with two different rise times (5ns and 10ns). This synthetic dataset consists of 1000 samples per class, where the pulse moves from time $t = 0$ ns and ending at $t = 450$ ns with a shift of 0.4 ns. In order to have a simplified presentation of real signals, a sinusoidal interference is introduced to the dataset with variation in the frequency, amplitude, and random shift angles. Interference signals corresponding to the following frequencies were synthesized: $10, 30, 50, 60, 80$ and 100 MHz. The corresponding rise time for each of the interference signals are: $30, 10, 5, 3.75$, and 3 ns. For each of the interference frequencies, different interference amplitudes are considered: $0.2, 0.5, 0.6, 0.8$,and 1. In addition, additive white Gaussian noise (AWGN)

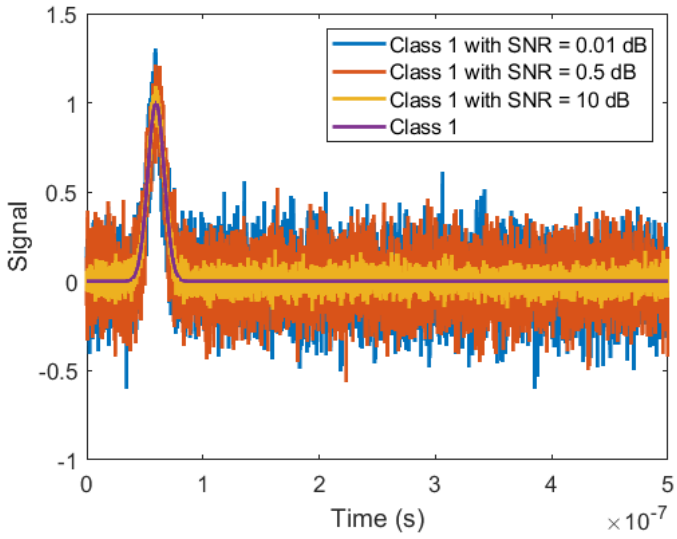


Fig. 4.1: Class 1 with different SNR for AWGN.

with different levels of signal-to-noise ratio (SNR=$0.01, 0.5$ and 10dB) are added to the two classes. Examples of the AWGN and interference signals are shown in Fig. 4.1 and Fig. 4.2 respectively. To avoid redundancy, only waveforms corresponding to class 1 (i.e. a risetime of 10ns) are shown in these figures. The training dataset consisted of the two classes with an interference frequency of 100 MHz and an amplitude of 0.6. Eighty percent of the samples are used for training and 20% are used for validation. The model is evaluated on the testing dataset which includes all the signals with different frequencies and amplitudes of the interference signals in addition to the AWGN signals.

The second dataset is an experimental dataset that is adopted from lab measurements where artificial defects were employed to generate PD [99]. The lab measurement setup consisted of a high voltage transformer, a capacitive divider to measure the AC voltage, the test cell, and the PD measurement system [99]. The lab setups simulate common sources of PD in air and in SF6. The test cells include three sources: floating electrode in SF6 (class 1),

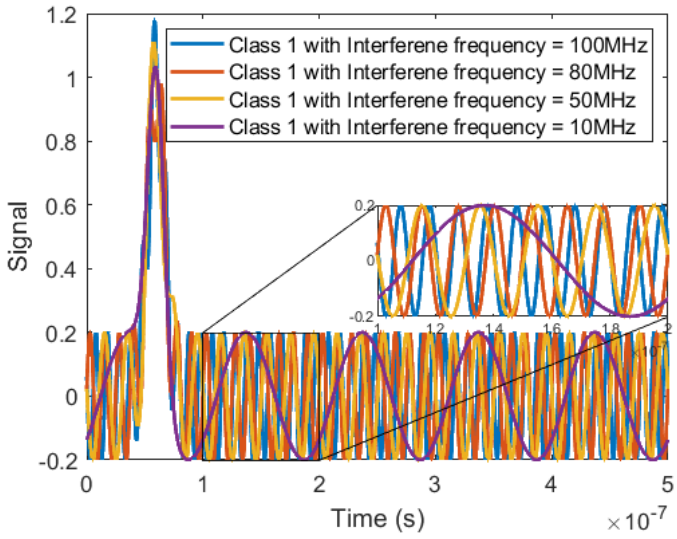


Fig. 4.2: Class 1 with different amplitudes of the interference signals.

moving particle in SF6 (class 2), fixed protrusion in SF6 (class 3), and corona in air (class 4). In total, there are four different PD waveforms generated. A 2.5 GHz digital oscilloscope was used to acquire the waveforms. Samples of the input waveforms of the four PD sources are shown in Fig. 4.3. 1000 samples per class resulting in 4000 samples are used for the training stage. Eighty percent of the samples are used for training and 20% are used for validation. For the testing dataset, AWGN and interference signals are added to the measured signals. An example of class 1 is shown in Fig. 4.4 and Fig. 4.5 corresponding to the AWGN and the interference signals, respectively.

4.1.2 Implementation of 1D- CNN

The simplest way to start with training multiple K classes (in this study, $K = 2$ for the synthetic dataset and $K = 4$ for the experimental dataset) is to start with K one-versus-all binary classifiers for which each classifier is trained to learn an individual PD source. The

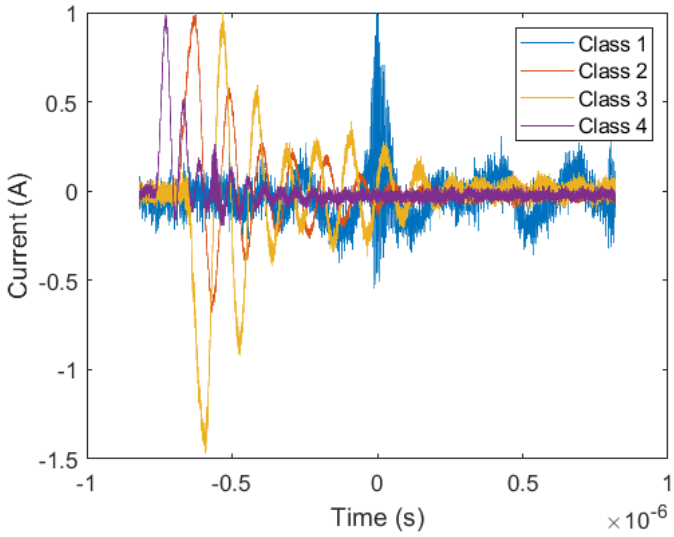


Fig. 4.3: Classes corresponding to different PD sources in the experimental dataset.

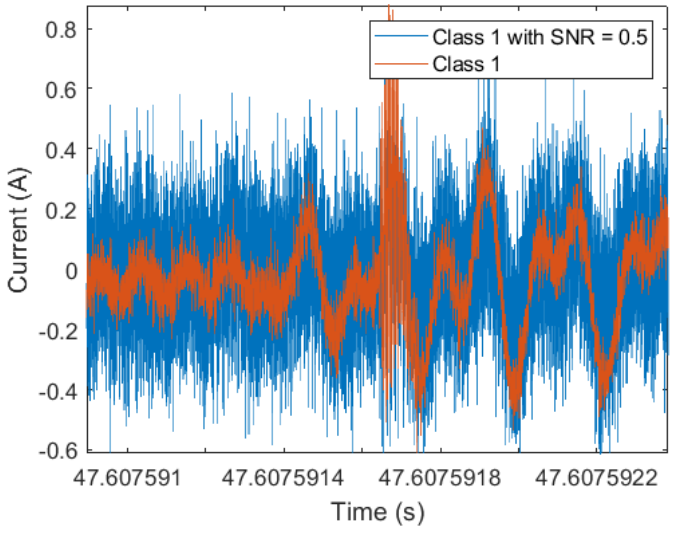


Fig. 4.4: Class 1 of the experimental dataset with and without added noise.

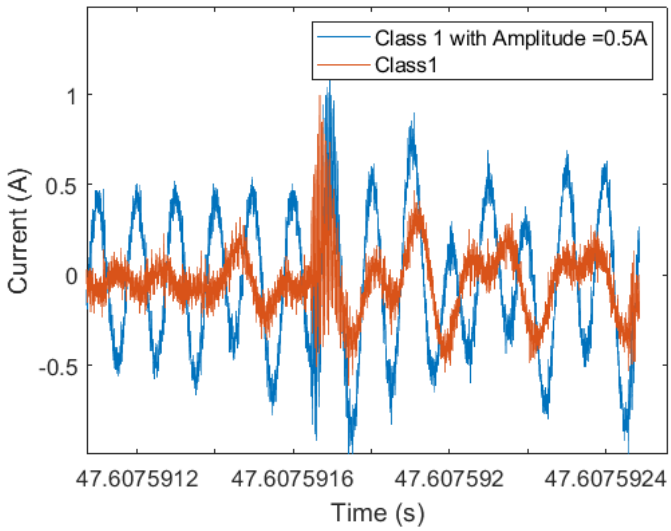


Fig. 4.5: Class 1 of the experimental dataset with and without interference signal.

decision on the hyper parameters such as the number of layers, number of kernels per layer, the size of the kernels is specific to each dataset.

The architecture details for each of the two datasets is shown in Table 4.1. Number of kernels specifies the number of filters the model learns for each layer. The kernel size specifies the size of time window that will be used for the filter learning. Max-pooling reflects the extent of temporal sub-sampling while moving from layer to layer. Batch normalization is used in both models, and the loss used for back propagation is the binary cross entropy. The models were coded using Keras and Tensorflow in Python, and Adam optimizer was used [100].

4.1.3 Findings

The performance metric that is used in this study is the average accuracy percentage of the classifiers, where a sample is considered correctly classified if the probability of correct

Table 4.1: Datasets architecture design.

Layers	Number of Kernels		Maxpooling
	Synthetic Dataset	Experimental Dataset	
Conv 1	64	16	3
Conv 2	64	16	8
Conv 3	32		
Conv 4	32		
FC 1	64		
FC 2	16		
Output	1		

classifier is larger than 0.5 and the probability of the other classifiers is less than 0.5. The accuracy is then calculated as

$$\text{Accuracy} = \frac{\text{Correctly classified samples}}{N} \times 100 \tag{4.1}$$

where N is the number of tested samples. In addition to the accuracy, for the synthetic dataset, receiver operating characteristic (ROC) analysis is performed by computing the area under the curve (AUC) for the ROC. The ROC-AUC is an estimate for the probability of distinguishing the two classes over all possible thresholds [101].

Table 4.2 shows the results of the measurement dataset. The model was able to classify correctly the four classes in the presence of different levels of AWGN. For the interference, the model starts to show a drop in performance as the amplitude of the interference signal increases above 0.5. For the synthetic dataset, the percentage accuracy for the AWGN is 100% for all SNR levels. Fig. 4.6 shows the classification accuracy of the model versus different frequencies of the interference signals and different amplitudes. The figure shows that as the interference frequency shifts from the trained frequency 100 MHz, the performance

Table 4.2: Results of measurement dataset

Measurement Dataset			
AWGN		Interference	
SNR Level (dB)	Accuracy	Amplitude Level	Accuracy
10,0.5,0.01	89%	0.1, 0.2	100%
		0.5	89.5%

Table 4.3: AUC for the synthetic tested data

Frequency \ Amplitude		0.2	0.5	0.6	0.8	1
100 MHz	C1	1	1	1	1	1
	C2	1	1	1	1	1
80 MHz	C1	1	1	1	1	1
	C2	1	1	1	1	1
50 MHz	C1	1	0.673	0.655	0.98	1
	C2	1	1	1	1	1
30 MHz	C1	1	0.577	0.556	0.666	0.991
	C2	1	1	1	0.964	0.391
10 MHz	C1	0.674	0.673	1	1	1
	C2	1	0.485	0	0	0

of the model degrades. This behavior is evident for the frequency of 50 MHz. For 30 MHz interference signals, the model performance drops. This is because of the fact that the rise time of the sinusoidal interference is equal to that for one of the classes. Table 4.3 shows the values for ROC-AUC. The AUC is a number between 0 and 1, where 1 reflects a perfect classifier and 0.5 reflects a purely random classifier.

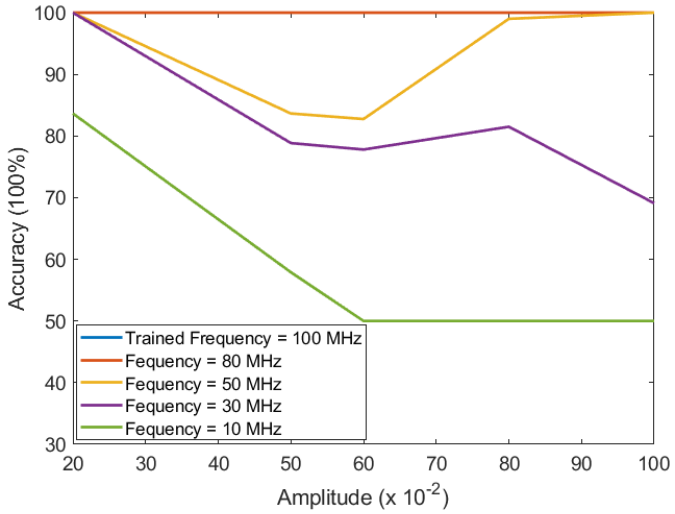


Fig. 4.6: Sensitivity of 1D- CNN to different interference amplitudes and frequencies for the synthetic dataset.

4.2 Multi-Label PD Classification

4.2.1 Datasets

Since gathering experimental data is resource-intensive, it is advantageous to start with synthetic data. This brings an additional benefit of complete control over the initial data sets used for training the CNN. In this study, Gaussian waveforms of unity magnitude with different rise times ranging from 1 to 17 μs were generated with a 0.1 μs time step. A sample of the waveform corresponding to the different classes is shown in Fig. 4.7. A set of waveforms for each corresponding rise time was generated, each set of waveforms consists of 1000 samples per class, where the pulse moves along the time axis with a shift of 0.3 μs. As commonly used, 80% of the samples were used for training and 20% were used for testing. For the testing dataset, multiple classes in one sample were taken into consideration.

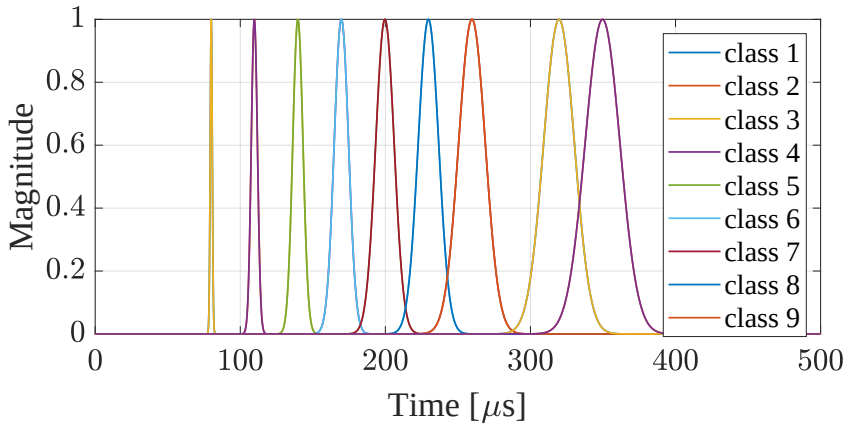


Fig. 4.7: Synthetic data: different rise times corresponding to different classes.

Since this is a time-series waveform application, the timing of the pulses is important. In one scenario, pulse corresponding to two different classes may occur simultaneously (i.e. completely overlap), and in the other, the the pulses for two classes could be far apart in time (i.e. no overlapping takes place). The third scenario is when pulses partially overlap, shown by example in Fig. 4.8.

A total of 4 different case studies were considered: 3-, 4-, 7-, and 10-class classification problems. For the testing datasets, different classes were randomly combined. By way of an example, for the 7-class classification scenario, the tested dataset included: classes 1 and 6, classes 2 and 5, classes 5 and 6, classes 1 and 2. These combinations were, respectively, labeled as C16, C25, C56, and C12. A summary of the random multiple classes for the different scenarios is shown in Table 4.4. Since, in this study, the purpose was to investigate the effect of increasing the number of classes to be classified on the performance of a one-vs-all 1D-CNN, no interference or noise is added to the waveforms.

The second dataset is the experimental dataset. 3D-printed polylactic acid (PLA) cylindrical samples that contained a void of specific size inside them were designed and fabricated

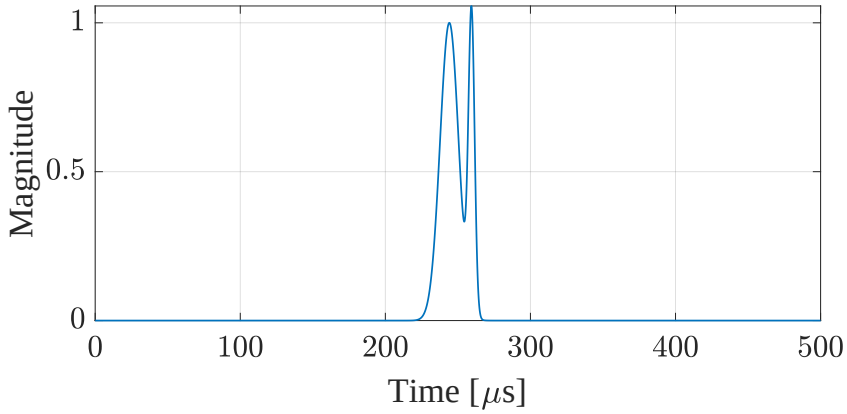
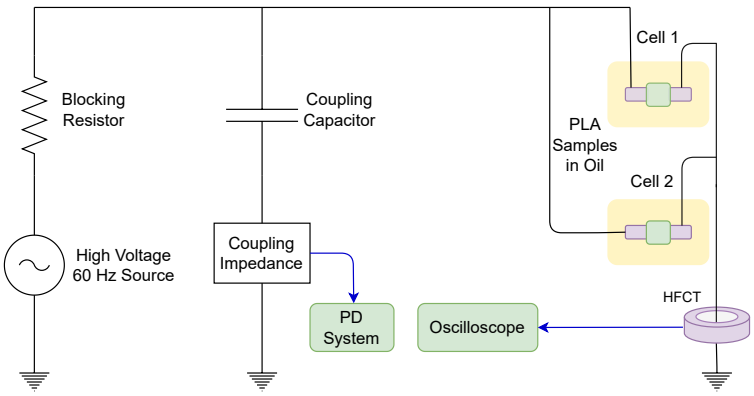


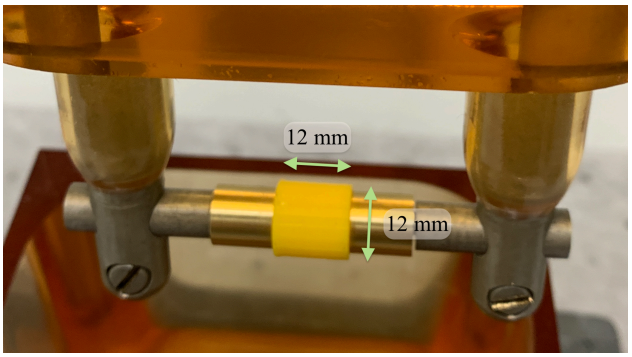
Fig. 4.8: A sample synthetic waveform with two classes partially overlapping.

Table 4.4: Multiple Classes Considered for each Scenario

Case Study	Test Dataset
Three class classification	C12
Four class classification	C12, C13, C23
Seven class classification	C12, C25, C56, and C16
Ten class classification	C12, C13, C19, and C27



(a)



(b)

Fig. 4.9: (a) Schematic of the experimental setup, (b) Photo of the PLA sample between brass electrodes. For the experiments, the sample is immersed in transformer oil.

to generate PD in a controlled laboratory environment. A schematic of the experimental setup, and a sample of the 3D PLA sample are shown in Fig. 4.9a and Fig. 4.9b respectively. In this experiment, classes 1 and 2 refer to PD resulting from samples with, respectively, a 0.5 mm and 2 mm void. More information on the PLA samples and the experimental setup can be found in [102]. The laboratory measurement setup consisted of a high voltage transformer, a coupling capacitor, test cells, and a commercial PD measurement system. A

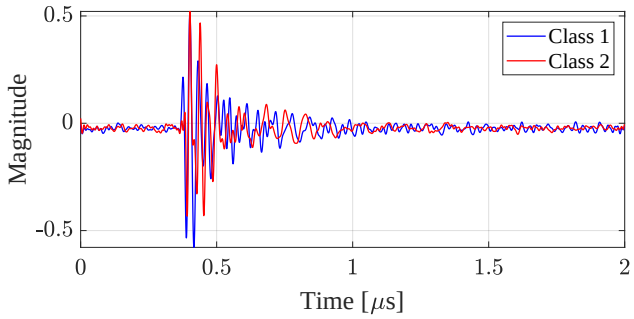


Fig. 4.10: Samples of signals collected in the experimental setup. Classes 1 and 2 correspond to the 3D PLA samples with a 0.5 mm and a 2 mm void, respectively.

high-frequency current transformer (HFCT) connected to a 2.5-GHz digital oscilloscope with segmented memory (Agilent DSO9254A) was used to acquire the PD time-series waveforms. The experimental setup does not permit more than two cells to be connected in parallel to record the signals, limiting the investigation of the experimental data to two classes. Adding class 0, representing the class of no PD, makes this problem a 3-class classification case. A sample of the signals corresponding to the two classes acquired by the HFCT is shown in Fig. 4.10. The magnitude of all signals was normalized to unity to eliminate the magnitude factor from the learning criteria of the deep learning algorithm. For each class, 2048 signals were recorded. As commonly used, 80% of the samples were used for training and 20% were used for testing. When only one cell is connected in the setup, the ground truth is known (i.e. either class 1 or class 2). To investigate the multi-class scenario where both class 1 and class 2 were present, two cells were connected in parallel in the experimental setup. However, when both cells were connected, the ground truth of the recorded signal is not intuitively known as the PD pulse could arise from cell 1, or cell 2, or both. To label correctly the data samples resulting from connecting both cells in parallel, an investigation of the time-series waveforms frequency spectrum was conducted. Out of the 2048 signals recorded, 845 sig-

Table 4.5: Datasets Architecture Design

Layers	Number of kernels	Maxpooling
Conv 1	32	5
Conv 2	16	3
Conv 3	16	2
FC 1	512	
FC 2	128	
FC 3	64	
Output	1	

nals were labeled as class 12 since the frequency spectrum investigation yielded that the PD resulted from both cells. However, without manual inspection on each of these signals, it was not possible to determine whether the PD pulses resulting from the two cells overlapped completely, partially, or did not overlap at all.

The details of the neural network architecture for the synthetic and experimental datasets are shown in Table 4.5. The learned models were optimized in regard to the hyperparameter values such as the learning rate, number of layers, number of neurons in each layer, and the size of the kernels. Since the synthetic and experimental datasets are different from each other, the architecture of each dataset is unique with respect to the size of kernels used in each convolutional layer. The binary cross entropy loss function and batch normalization were used in both algorithms. The algorithms were implemented using Keras and TensorFlow in Python.

4.2.2 Findings

The performance metrics used in this study are the average accuracy percentage and the false negative rate of the classifiers. For the average accuracy percentage, a sample is considered correctly-classified if the probability of the correct classifier is larger than 0.5 and the probability of the other classifiers is less than 0.5. For the signals which correspond to two classes, the average classification accuracy is considered by taking the mean of correct classification for each class. The false negative rate is an important metric in high voltage applications since the failure to detect PD might eventually result in equipment failure. While the classification accuracy gives equal weight to all classes, the false negative rate focuses on the labels that the sample belongs to. Details on the calculation of the average classification accuracy and false negative rate is found in Chapter 3. Table 4.6 shows the average classification accuracy and the false negative rate for each of the considered multi-labeled classes corresponding to each of the four case studies using the synthetic data. For every case study, it was noticed that the performance was better when the two classes have distinct rise times. For example, in the ten-class classification problem, the average accuracy was higher for the multi-labeled class C19 compared to that of C12. In addition, the effect of increasing the number of classified classes on the average classification accuracy of the signals which result from two classes was investigated. Fig. 4.11 shows the average classification accuracy and the false negative rate of the multiclass C12 as a function of increasing the number of classes to be classified. In this figure, the percentage accuracy values for the multiclass C12 for all the different classification case studies (i.e. 3, 4, 7, and 10 classes) were compared. As the number of classes to be classified is increased, the performance of the one-vs-all 1D CNN classifiers decays, and this is reflected by investigating the average classification accuracy and the false negative rate. This may be attributed to the fact that with more classes, the

Table 4.6: Accuracy of Multiple Classes for Synthetic Data Case Studies

		Average Accuracy %	False Negative Rate %
Three classes	C12	90.2	12.0
Four classes	C12	85.4	13.0
	C13	94.6	8.4
	C23	81.4	15.3
Seven classes	C12	74.8	20.1
	C16	87.0	13.1
	C25	85.3	16.3
	C56	82.0	17.4
Ten classes	C12	60.2	30.2
	C13	75.5	22.3
	C19	80.1	13.1
	C27	85.0	12.3

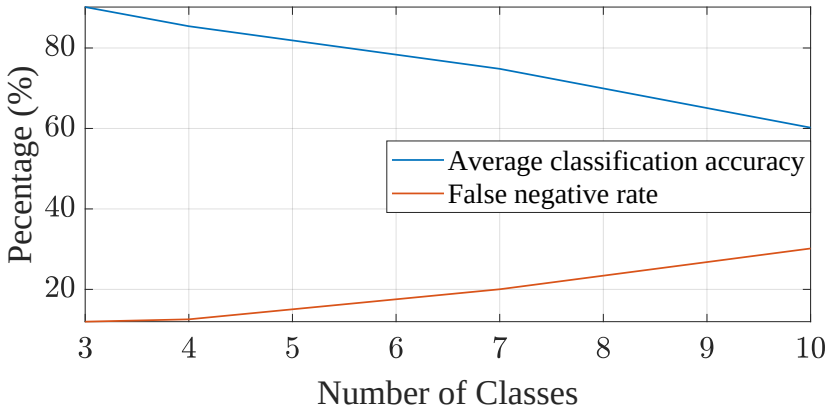


Fig. 4.11: The average classification accuracy and the false negative rate of the multiclass C12 as a function of increasing the number of classes to be classified.

'all' category in a one-vs-all setting tends to contain datapoints from increasingly diverse classes while the algorithm is still being forced to learn representations that treat examples from the 'all' category similarly, thus making the classification problem harder. Another aspect to investigate was the effect of overlapping signals on the classification accuracy for multiclass labels. When the PD signals resulting from two classes do not overlap at all, the classification accuracy is 99%. However, when the two PD signals start overlapping, the accuracy decreases to 84% until it becomes 67% when the signals overlap completely. This is expected since, with no overlapping, the two signals preserve the intrinsic characteristics of the single classes of class 1 and class 2 separately. However, when the signals start to overlap, the model struggles to incorporate the learned intrinsic characteristics of the single classes in order to determine the label of the overlapping signal.

For the experimental data, the investigation of the performance of the model used the confusion matrix as the laboratory conditions do not completely control the time of occurrence of both signals resulting from the cells. The confusion matrix is a decision-making tool which shows how the classification model is confused when trying to label the data and making predictions. As the training phase includes the single classes only, a hybrid confusion matrix is shown in Table 4.7, where the rows and columns represent the input and predicted classes, respectively. In the confusion matrix (Table 4.7), the true positives are highlighted for better visibility. The model successfully predicted the signals resulting from single classes. This is expected since the training of the model was a result of learning the intrinsic characteristics of the single classes. When it comes to having a signal that has a label of C12, the model performance drops substantially to an average classification accuracy of 67.1% and a negative false rate of 49.4%. This is consistent with the results of the synthetic data.

Table 4.7: Hybrid Confusion Matrix for the Experimental Data

		Predicted Class		
		C0	C1	C2
Input Class	C0	432/432	0	0
	C1	0	389/389	0
	C2	0	1	408/408
	C12	0	266/845	589/845

4.3 Summary

One dimensional convolutional neural networks were used for the classification of PD pulses and showed promising results for distinguishing PD pulses from random noise and interference. The random noise is represented as Gaussian noise and the interference pulses are represented as sinusoidal signals. The proposed method does not require any pre-processing of the signals. A synthetic dataset was used to have complete control on the variability of the training and testing dataset. Random noise with different SNR levels, in addition to different rise time and amplitude in the interference signals, were tested. A measurement dataset of four sources of PDs was tested as well. The designed model starts degrading in performance when the rise time of the interference signal is the same as the rise time of the PD. In addition, a one-vs-all one-dimensional convolutional neural network classifier was used for the classification of PD pulses. Single classes corresponding to single sources of PDs were trained, and testing was performed using single and multiclass signals corresponding to single and multiple sources of PDs taking place. A synthetic dataset was generated to have complete control over the dynamics of the signals especially for multiclass labeled data. An experimental dataset using 3D PLA samples with voids inside them was generated and

tested as well. While the performance is promising in regarding to the single classes, the performance of the one-vs-all 1D-CNN started dropping when multiclass labeled signals were tested. This is shown by investigating the average classification accuracy and the false negative rate . In addition, as the number of the classes to be classified increased, the one-vs-all 1D-CNN showed lower levels of performance. After noting the limitations of off-the-shelf CNN architectures, the continuing work of the research is focused on developing CNN based models for PD classification where time-series waveforms are the input the learning models.

Chapter 5

An Interpretable CNN Model for Classification of Partial Discharge Waveforms

In this chapter, a 1D-CNN is designed, implemented, and tested to investigate the PD pulses generated in a void inside a solid dielectric. In addition, we add attention mechanism to the learned CNN model to introduce interpretability to the decisions made by the deep neural network. A number of 3D-printed polylactic acid (PLA) cylindrical samples which have voids inside them are designed and fabricated. Each sample had one 0.5 or 2 mm void. The PD pulses collected from these samples are used to evaluate the performance of the proposed 1D-CNN. Studying the PD waveforms resulting from either void sizes makes the study a void characterization problem, where the network is trained in order to output the void-size catergory. As such, the input to the CNN model is the time series PD waveforms, and the output is the label corresponding to the identified void class. Since the amplitude

of the PD waveforms is specified by the measurement system, the time-series waveforms are normalized to make the filters of the CNN invariant to the amplitude of the waveforms. One of the main contributions of this work is the establishment of a connection between the void size and the corresponding bandwidth of the associated PD waveform, which in turn is related to the corresponding rise time. In fact, based on our observations, only human visual inspection - based decisions will be very difficult because the pulses resulting from samples of two void sizes look similar to each other. The fact that the pulses look similar across the classes actually makes the problem difficult rather than easy. This visual observation is further validated by the time-frequency (T-F) map, where two partially overlapping clusters appear. Therefore, it is required to use a non-linear mapping (such as CNN) on the two classes where the difference between the two classes corresponding to pulses of different void sizes is amplified. Hence, the 1D-CNN findings are compared with those obtained from T-F map, which is an established tool used in industry to classify different PD pulses. Moreover, the findings are also compared with extracted statistical features from the PRPD patterns. While the T-F map is based on the application of one filter, *i.e.* Fast Fourier transform, the 1D-CNN learns different adaptive filters automatically which allow it to learn more compact and general characteristics of the time series waveforms. The 1D-CNN model shows improved classification results compared to the T-F map, where the metric used is the classification accuracy. The attention mechanism along with interpreting the filter's coefficients of the first convolutional layer verify that the filters of the CNN model automatically learn to home in on those sub-sections of the waveform which are associated with the rise time of the PD pulses corresponding to the two classes.

5.1 Materials and Methods

5.1.1 3D-Printed PLA Samples

The samples considered were fabricated from polylactic acid (PLA) filament using a commercially-available 3D printer (AXIOM Dual Extruder) controlled with APEX software. The choice of the particular filament type is not critical to the outcome of this study. PLA is commonly used for 3D printing applications and, therefore, readily available. Several samples with a void were printed for the partial discharge measurement as well as control samples which were solid cylinder samples with no void at the center [103]. Each sample is a cylinder with a height of 12 mm, and a diameter of 12 mm. The CAD models of our 3-D printed samples include a cylindrical void located at the mid-point of the cylindrical axis as shown in Fig. 5.1a. The first group of the samples had a void with a height and diameter of 2 mm, whereas in the second group of the samples both the height and the diameter of the void were reduced to 0.5 mm. * Before running the partial discharge tests on the samples, X-ray microtomography imaging using a commercial system (Sky-scan 1275) was used to confirm the presence/absence of the void. The images obtained from X-ray microtomography provide insight into deviations from the desired void geometry/dimension that may arise from the 3D-printing process as shown in Fig. 5.1b. The imperfections shown in Fig. 5.1b are also present in the solid (control) samples and partial discharge signatures were not obtained from these control samples. We conclude, therefore, that the most significant contribution to the partial discharge signature does arise from the void at the center of the samples studied. Moreover, to ensure that the PD is generated by the void, we had a setup with the electrodes with no sample up to 26 kV, where no PD was recorded. Next, we added

*The 3D-printed samples were fabricated by Puneet Gill. Courtesy of Puneet Gill for the x-ray image in Fig. 5.1b [104].

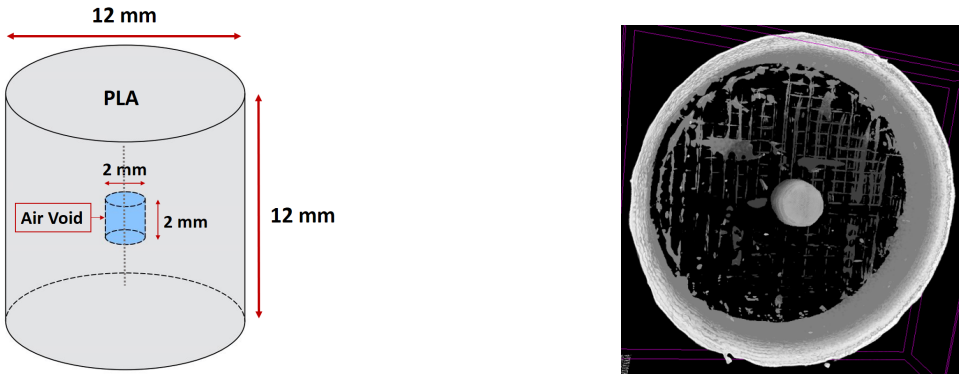


Fig. 5.1: (a) A schematic of the 3D-printed PLA cylindrical sample showing the location of a cylindrical void at the centre (the common axis of rotational symmetry for the void and the sample is shown as a dashed vertical line through the void), (b) X-ray microtomography scan of a 3D-printed PLA sample, shown looking along the common axis of symmetry (i.e. rotated towards the reader when compared to the schematic), that shows the void in the centre of the sample as well as imperfections generated during the printing process (most of which occur at the base of the sample - farthest from the reader as viewed in this image).

the control samples (solid cylinder samples with no void at the center) which had the same quality with regards to the same imperfections and still no PD was recorded. And finally PD was recorded when the sample with a void was introduced. This confirms that PD was generated by the void. The accuracy of the 3D-printing process is heavily dependent on the geometry and sequencing of the raster pattern that the print head follows as well as a number of parameters including: the local temperature at the print head, the speed at which the print head moves across the sample during printing, and the rate at which the filament is inserted into the print head.

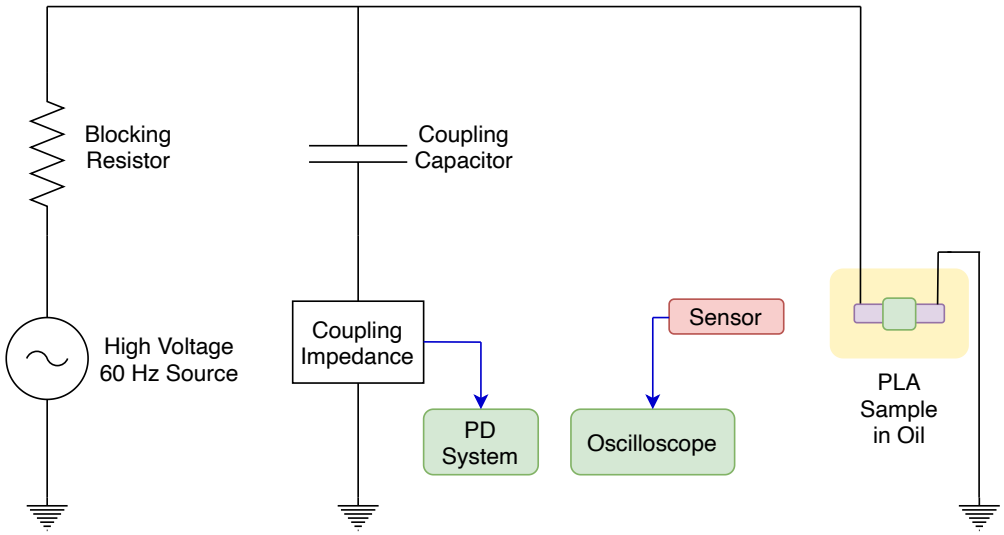
A schematic of the experimental setup is shown in Fig. 5.2a. A commercial instrument (including a coupling capacitor, a coupling impedance, and a PD system) is used to generate the PRPD patterns. A capacitive sensor (by HVPD) connected to a 2.5-GHz digital oscil-

loscope with segmented memory (Agilent DSO9254A) is used to acquire the PD waveforms. The 3D-printed PLA sample is held between two brass electrodes of an external diameter of 10 mm. The sample and electrodes are immersed in transformer oil (Voltesso 35) as shown in Fig. 5.2b. In order to confirm whether or not there is any oil absorption by the samples, a number of 3D-printed PLA samples are immersed in oil for two weeks. The mass of each sample is determined before and after immersion using a microbalance (Perkin Elmer - AD 6 Autobalance with Controller). In all cases, the variation in the recorded mass within instrumental uncertainty confirms the hypothesis that the 3D-printed PLA samples do not absorb oil in sufficient quantities to impact the outcome of the partial discharge measurements.

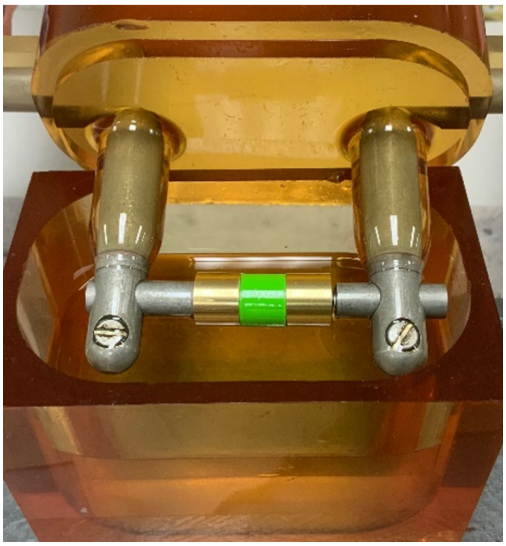
5.1.2 Dataset

The dataset investigated consists of PD waveforms from the 2-mm void (class 1) and 0.5-mm void (class 2). In order to generate the PDs, the applied voltage for samples of class 1 is 14 kV, and the applied voltage for the samples of class 2 is 22 kV. The PRPD patterns are recorded for samples with both the 0.5 and 2 mm void sizes. Typical acquired waveforms for class 1 are shown in Fig. 5.3. The PRPD pattern of class 1 samples (*i.e.* with a 2-mm void) is shown in Fig. 5.4. A sample of the PRPD pattern and PD waveforms corresponding to class 2 are shown as well (Figs. 5.5 and 5.6). As seen for each of the classes, two different waveforms are present in the time series waveforms recorded. This corresponds to the clusters formed in the PRPD pattern, where each cluster is at a different phase angle of the applied voltage. The number of waveforms corresponding to each of the clusters is roughly the same.

Six thousand time-series waveforms are recorded for each of the two classes. The length of the time series is 10,000 for each waveform. Depending on the triggering setup on the oscilloscope, a 4 μs window is considered as the input to the 1D-CNN model for both classes.



(a)



(b)

Fig. 5.2: (a) 3D-Printed PLA sample (12mm long and 12mm diameter) held between brass electrodes (external diameter 10mm) (b) Sample and electrodes immersed in Voltesso 35 oil for partial discharge measurement.

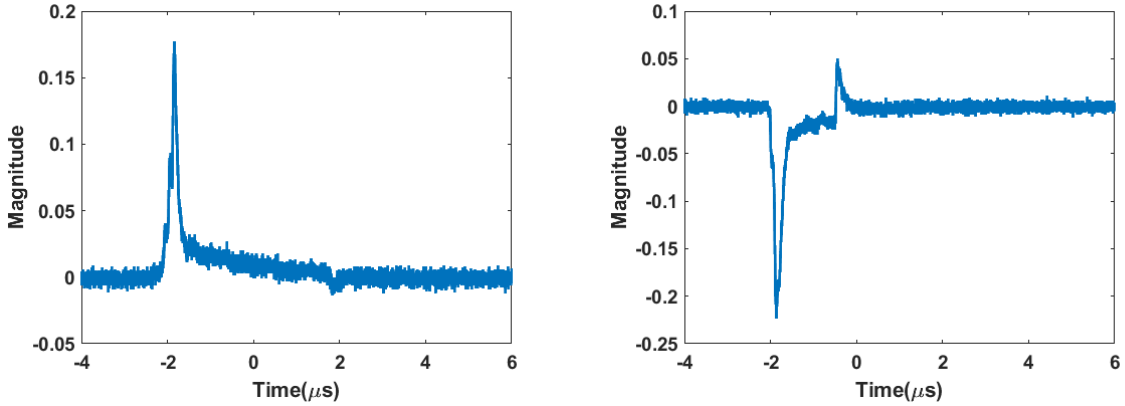


Fig. 5.3: Two examples time series waveforms from class 1: the left one shows a positive pulse and the right one shows a negative pulse. Both pulses belong to void size class 1.

Hence, the length of the input time series waveforms considered is 4000. Since the internal resonance frequency of the capacitive sensor and the connecting cable is high enough, the acquired waveforms are not oscillatory. As a result, no pre-processing of the waveforms is required other than normalizing the magnitude. Even with damped oscillation present in the waveforms, that may be caused by the measurement system, machine learning algorithms will learn representations that ignore the similarity across classes. In this case, since the damped oscillations would be similar across the two classes, the non-oscillatory waveforms used in this study are suitable to evaluate the developed CNN model.

5.1.3 Time-Frequency Map

A typical visualization and classification tool that is used by researchers to classify PDs from different sources is a clustering technique based on analysing the time behavior and the frequency content of the PD time-series waveforms [87, 105]. The clustering technique is based on calculating equivalent time (T_{eq}) and equivalent bandwidth (W_{eq}) of the PD

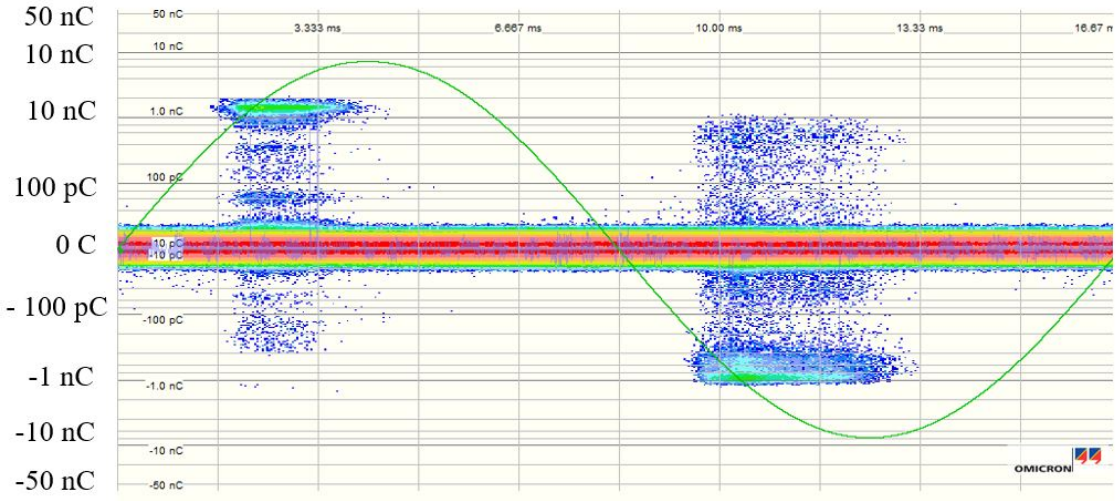


Fig. 5.4: PRPD pattern of class 1 sample.

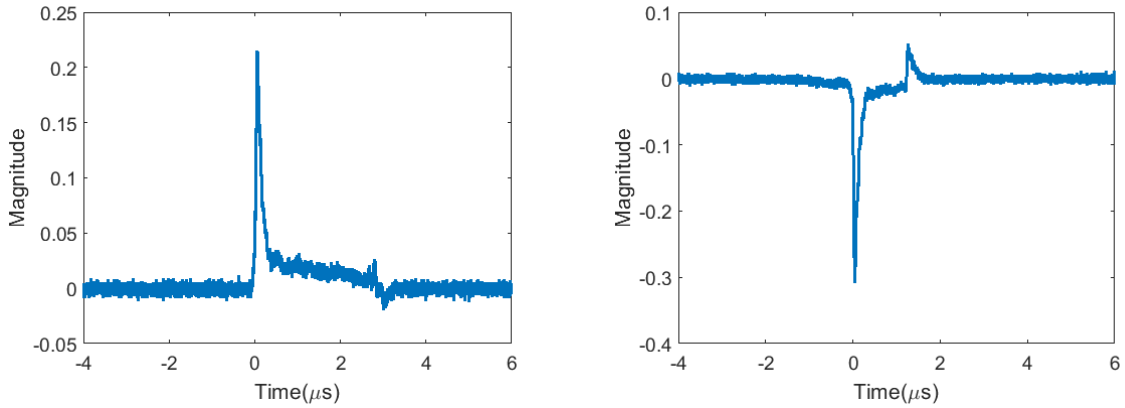


Fig. 5.5: Two examples time series waveforms from class 2: the left one shows a positive pulse and the right one shows a negative pulse. Both pulses belong to void size class 2.

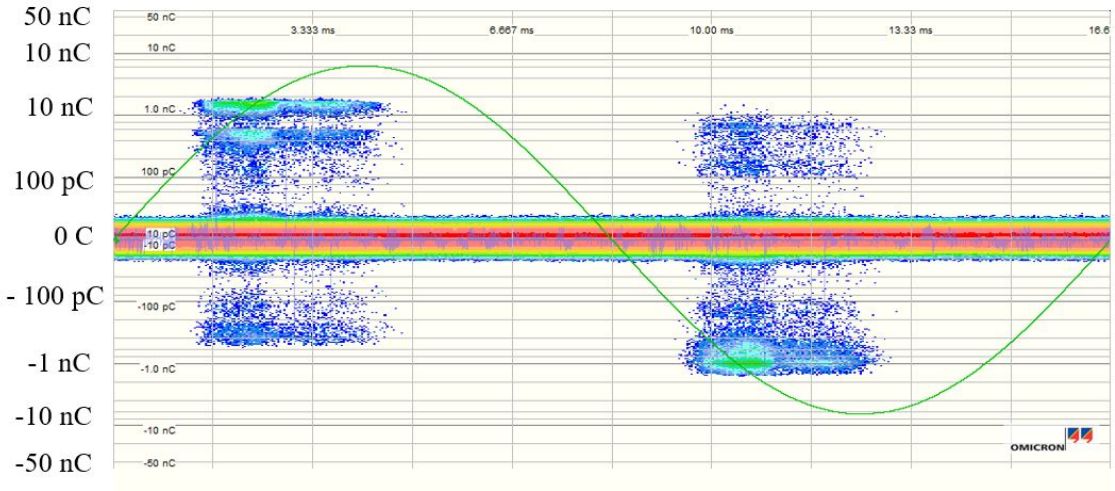


Fig. 5.6: PRPD pattern of class 2 sample.

waveform (denoted as $s(t_i)$ in the time domain and $S(f_i)$ in the frequency domain) that are defined as [105]

$$T_{eq} = \sqrt{\frac{\sum_{i=1}^{N}(t_i - t_0)^2 s^2(t_i)}{\sum_{i=1}^{N} s^2(t_i)}} \tag{5.1}$$

where

$$t_0 = \frac{\sum_{i=1}^{N} t_i s^2(t_i)}{\sum_{i=1}^{N} s^2(t_i)}$$

and

$$W_{eq} = \sqrt{\frac{\sum_{i=1}^{N} f_i^2 |S(f_i)|^2}{\sum_{i=1}^{N} |S(f_i)|^2}} \tag{5.2}$$

In (5.1) and (5.2), N represents the number of samples of each waveform. $S(f_\mathrm{i})$ is obtained by applying the Fast Fourier Transform (FFT) to the time-domain waveform $s(t_i)$.

T_{eq} and W_{eq} of each PD waveform maps it to a point on the T-F map. The PD waveforms with a similar shape are mapped into a cluster on the T-F map. Using the T-F map to

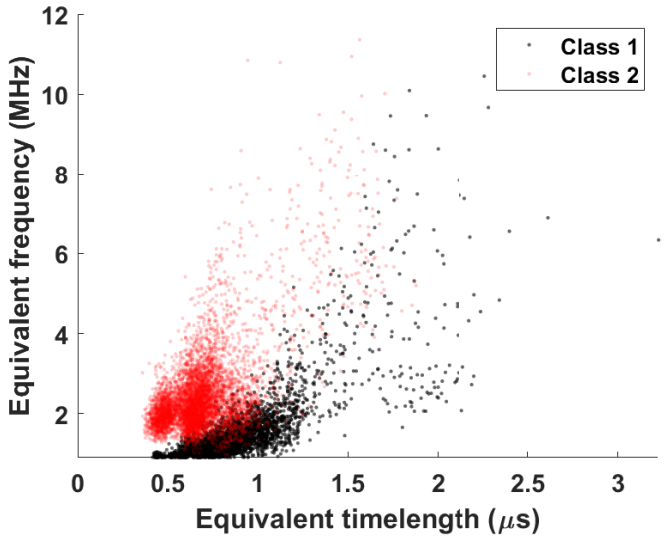


Fig. 5.7: T-F map of the normalized PD waveforms acquired from $6,000$ samples of class 1 and $6,000$ of class 2. Some overlap is visible between the two clusters.

classify different PD sources and mapping them into different clusters has been useful in many applications [106]. However, loss of signal information is one disadvantage of this technique. This is expected since this mapping is essentially a compression technique, where a signal is represented in a compacted form on the T-F map. Figure 5.7 shows the T-F map of PD waveforms acquired from both classes.

5.1.4 Principal Component Analysis (PCA) of Statistical Features Extracted from PRPD

Since the source of partial discharge is the same for all the samples, that is a void in solid dielectric, the PRPD patterns visually appear to be the same for both the 2 and 0.5 mm sizes of the void. However, one can extract additional statistical features from a PRPD recording.

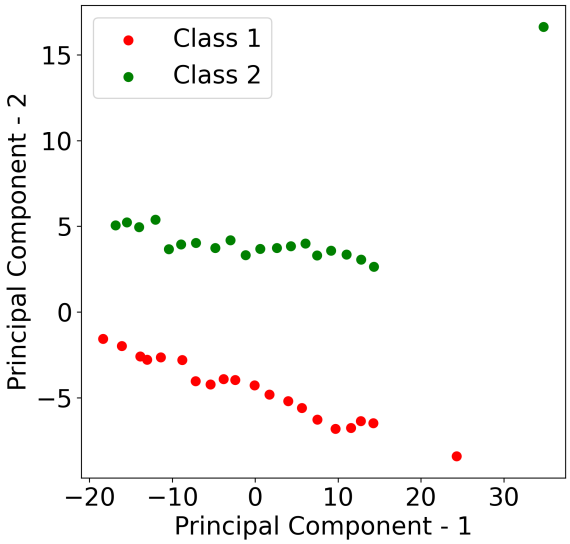


Fig. 5.8: First two principal components of data resulting from applying PCA on the statistical features of the PRPD patterns.

To do this, the 2π phase angle is divided into a number of windows (in this work 100 windows of $\pi/50$ width). The number of discharges, the maximum value of the discharge magnitude, and the average of discharge magnitude is computed for each phase window [107]. As a result, a feature vector of 300 elements is extracted for each sample. Principal component analysis (PCA) is applied on the feature vectors. The first two principal components are shown in Fig. 5.8. The variance preserved from the first two principle components is 68%. As shown in Fig. 5.8, the values along the first principle component, which has the highest variance, overlap for the two classes. However, along the second principle component, which has less variance, we observe some separability. Hence, the performance of a PCA-based classifier is expected to be limited.

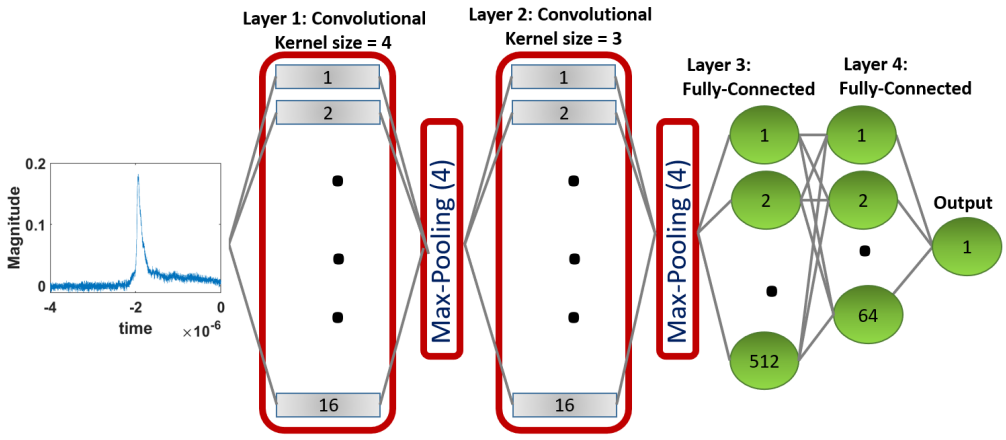


Fig. 5.9: 1D-CNN with first convolutional layer of filter size 4 and max pooling of 4 followed by a convolutional layer of filter size 3 and max pooling of 4 followed by two fully connected layers of 512 and 64 neurons respectively followed by the classification layer with one neuron.

5.1.5 1D-Convolutional Neural Network

CNNs belong to a class of deep learning models which apply a series of successive convolutional and pooling layers to learn to derive feature representations useful for a particular task at hand e.g. classification [14]. The application of CNNs was originally intended for images, but recently their use has been extended to a range of applications involving one dimensional waveforms [56, 108, 109]. Owing to the complexity of PD pulses, we have implemented a 1D-CNN model for time series waveform classification depending on the different void sizes inside the samples. The architecture used for this study is shown in in Fig. 5.9. Each layer in the CNN architecture has a specific role. The proposed architecture consists of four learnable layers: two convolutional layers, and two fully-connected layers. The first convolutional layer accepts the waveform as input and learns multiple time-domain filters to process the waveform. The response of each of the filters suppresses or enhances certain aspects of the input signal e.g., filtering noise or amplifying peaks. It must be noted that

the coefficients of the filters are not pre-specified, rather are learned so as to aid in the final classification of the waveform. To allow the CNN to model non-linear transformations, a non-linear activation function, rectified linear unit (ReLU) [110], was added after each layer. The output of the convolutional layer was subsampled (using a max-pooling operation [111]) to allow the subsequent layer to learn filters that model long-range interactions in time. The filter-responses from the second convolutional layer were concatenated as a vector and fed into a cascade of two fully-connected neural networks (layers 3 and 4) [112]. The output of layer 4 is a scalar which is passed through a sigmoid activation (producing a number between 0 and 1), such that the final output of the overall neural network can be interpreted as the probability of the waveform belonging to a particular void category. Since we are dealing with a two-class problem, as a loss, we minimize the binary cross entropy (BCE) [113] between the network output and the actual ground-truth labels for the void category. To minimize the loss, the Adam optimizer was used [100]. The model is coded using Keras and Tensorflow in Python. Four-fold cross validation [114] was employed to assess the robustness of the traning process to the choice of the training and testing sets. Eighty percent of the dataset is used for training and 20% is used for testing.

5.1.6 Adding Interpretability to the CNN

CNNs extract features from input data such that the convolutional layers retain detailed spatial information, while the fully connected layers retain high level information [115]. Capitalizing on this property, techniques have been developed for introducing class-specific interpretability within the CNN framework [116–118]. Chief among them is the Gradient Class Activation Map (Grad - CAM) method [118]. Grad-CAM is a post-hoc interpretability method that takes in a pre-trained CNN, and analyzes what part of the last convolutional

layer contributes to the final decision of the network. In particular, the method computes the gradient of the predicted class with respect to the last convolutional layer's feature maps. This gradient reflects the part of the input which contributes mostly to the classification output. Grad-CAM has been used widely for a range of applications [118–121]. In [122], a two CNN-based multitask model that classifies simultaneously the fault types and the working conditions of rotating machinery was proposed. Grad-CAM method was used to visualize the weight vectors of multiple convolutional layers and accordingly, the part of the signal which is of interest was localized.

A Grad-CAM approach is implemented in this chapter to interpret and localize parts of the input PD waveform responsible for an output decision corresponding to a particular void size class. To achieve this goal, after training the model, first, the model output and the last convolutional layer output for a given tested sample are computed. The gradient ($Grad$) of the winning void size class y^c is computed with respect to the feature map activations A^k of the last convolutional layer that is defined as

$$Grad = \frac{\partial y^c}{\partial A^k}. \tag{5.3}$$

The value of the gradient depends on the input PD waveform and the void size class to which it belongs. This is due to the fact that the feature maps A^k are determined by the input PD waveform. In our model, there are $k = 16$ feature maps. The gradient with respect to the k^{th} feature map is then averaged to compute its overall effect on the c^{th} class output. The averaged gradient α_k^c is calculated as

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^{Z} \frac{\partial y^c}{\partial A_i^k} \tag{5.4}$$

where Z (that is equal to 30 in this study) is the length of one feature map, and A_i^k is the i^{th} element of the k^{th} feature map. Finally, we use the averaged gradients, α_k^c, as the weight of the corresponding feature map and calculate a weighted sum of feature maps as the final Grad-CAM heat map. In the last step, we apply a rectified liner unit (ReLU) operation to emphasize only the positive values and turn all the negative values into 0. This is defined as

$$L^{\text{c}}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{5.5}$$

Since the feature maps in the last convolutional layer in most CNNs are small compared to the input, up-sampling operation is required before visualizing the heat map.

5.2 Findings

In order to summarize the methodology, Fig. 5.10 shows a flow chart of the different techniques used in order to emphasise the importance of the proposed CNN model. The performance metric that is used in this study is the accuracy percentage. The accuracy is calculated as

$$\text{Accuracy} = \frac{\text{Correctly Classified Samples}}{N} \times 100 \tag{5.6}$$

where N is the number of tested samples. Table 5.1 presents the confusion matrix in which the rows and columns represent the input and predicted classes, respectively. The per-class accuracy in addition to the average accuracy are calculated. The classification accuracy for class 1 is 99.16% and the classification accuracy for class 2 is 98.83%. The f1- score is 0.9899.

As already shown in Fig. 5.7, the two clusters intersect in a small portion. This is illustrated more clearly in Fig. 5.11. One thousand one hundred eighty data points of class

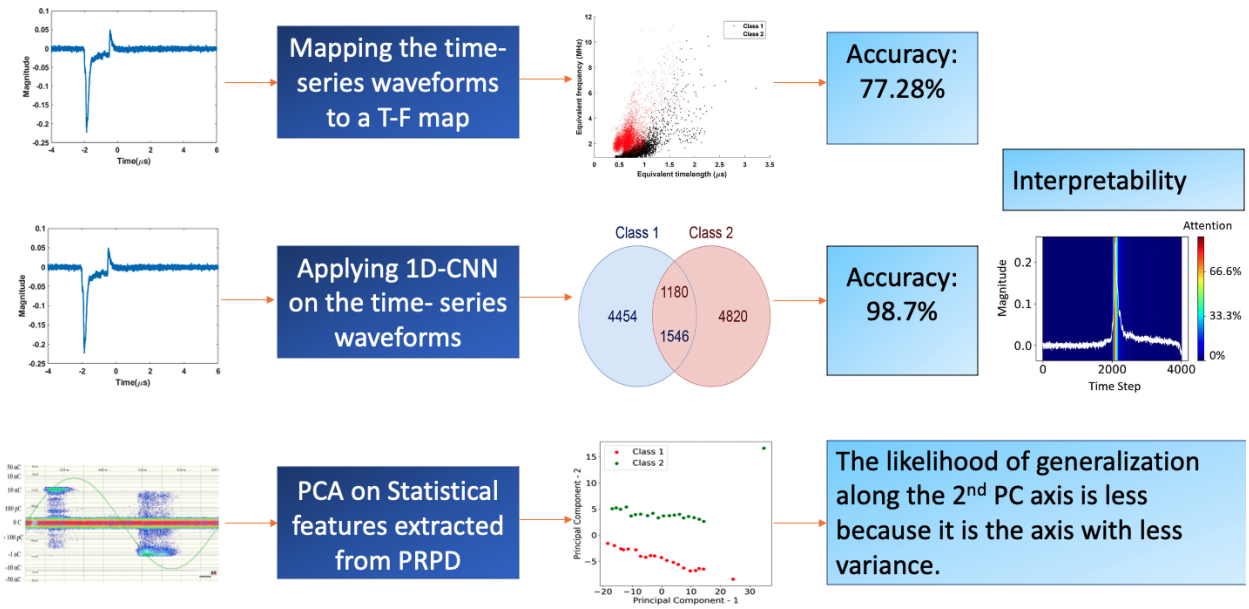


Fig. 5.10: Flow chart of the proposed framework: mapping the time-series waveforms to a T-F map, training a 1D-CNN using time-series waveforms followed by interpreting the decision making via attention model, and applying PCA on 300 statistical features extracted from PRPD patterns, where within unseen data, there could be more probability of confusion along the second principle component axis.

Table 5.1: Confusion matrix of independent classifiers.

		Predicted Class	
		1	2
Input Class	1	1190	10
	2	14	1186

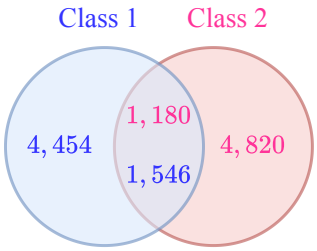


Fig. 5.11: Intersection of samples resulting from mapping them into the T-F map.

2 overlap with those from class 1 and 1, 546 data points of class 1 overlap with those from class 2. Of the 12, 000 data points that correspond to the time series waveforms of class 1 and class 2, 2, 726 are not assigned to either clusters. Therefore, 77.28% of the data points can be assigned to their true label. Moreover, although applying the PCA on the statistical features of the PRPD is able to differentiate between the two clusters, a restriction is that this approach is not applicable for DC applications where no phase reference exists.

On the other hand, as mentioned above, a 4-fold cross validation is adopted for the 1D-CNN model. The average scores for all folds has an accuracy of 98.7%. In addition, the loss of the model is 0.003. Therefore, the 1D-CNN shows improved results with 99% classification accuracy compared with that of the T-F map with 77.28%.

Class specific attention maps as computed by the Grad-CAM approach are shown in Figs. 5.12 and 5.13, where it is seen that the model concentrates at the start of the signal in order to decide on the label of each sample. Since CNN classification problems tend to find common features for same-class samples, it is expected that the model focuses on the positive pulse rather the negative pulse.

In order to have a better understanding of what the model is learning, we further analyse the learnable kernels of the convolutional layers. Figure 5.14 shows the heat map of the 16

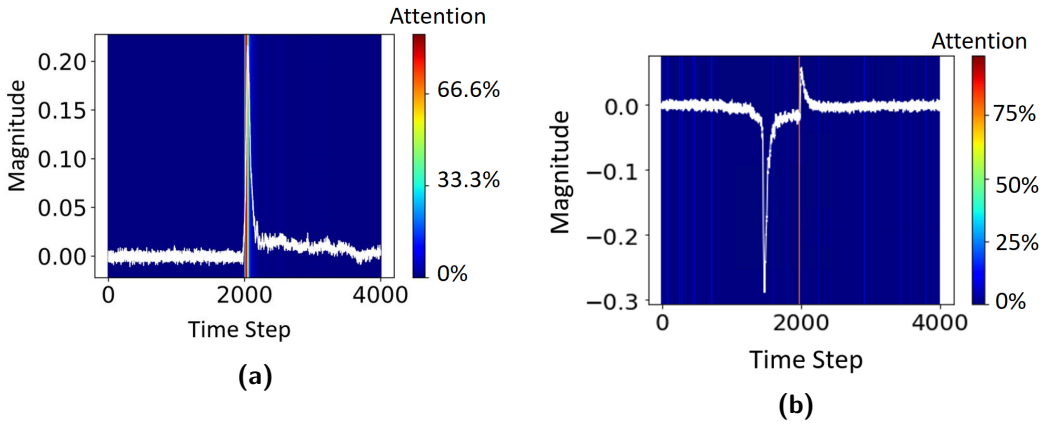


Fig. 5.12: Attention mechanism implementation for class 1 : (a) Positive pulse (b) Negative pulse. In (a) and (b), the model concentrates at the start of the waveform in order to decide on the label of the sample.

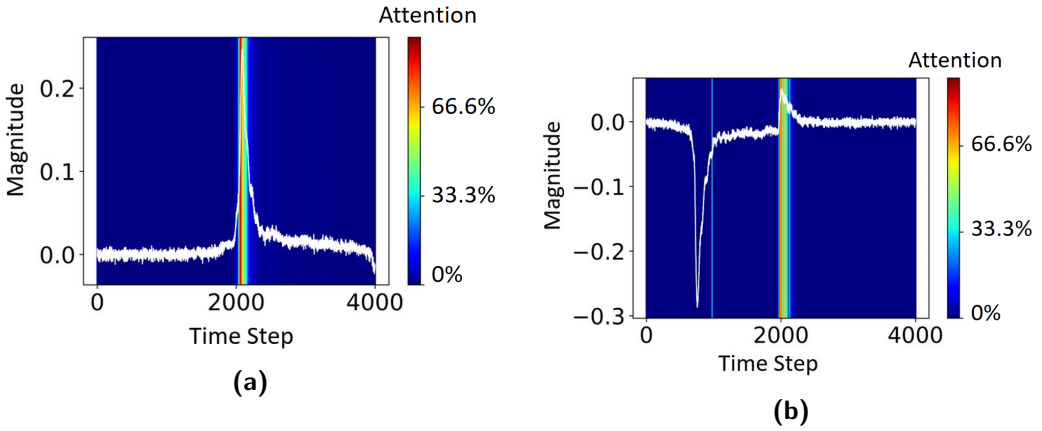


Fig. 5.13: Attention mechanism implementation for class 2: (a) Positive pulse (b) Negative pulse. In (a) and (b), the model concentrates at the start of the waveform in order to decide on the label of the sample.

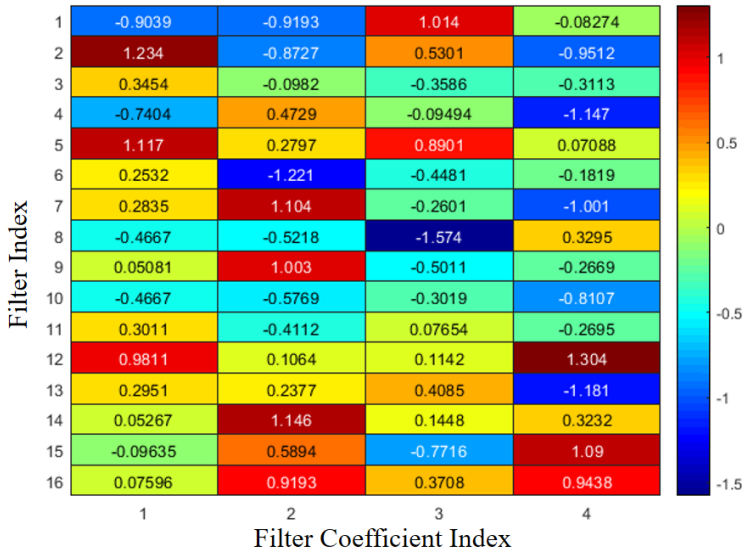


Fig. 5.14: Heat map of the 16 kernels weights learned in first convolutional layer with kernal size of four.

filters that are learned in the first convolutional layer. As shown, the filters are a mix of third order difference-filters, where variations of positive and negative filter coefficients show that the model is learning different aspects of one particular signal including: detection of steps (both up and down), along with peaks and dips.

When interpreting the PD pulse resulting from any source, one should be able to localize what part of the signal is important. When automating this interpretation, experts should make sure that the automated system is looking at the right part of the signal. As such, interpretability methods such as Grad-CAM are of crucial importance as they enable validating the learned outcome of the CNN model. As mentioned before, the T-F map makes use of the Fast Fourier Transform (FFT) in order to map a signal to a single point. The usage of FFT indirectly implies that one type of filter is used for the mapping of the waveforms into the T-F map. On the other side, the CNN learns different sets of filters where no constraints

are forced on these learnable filters. As a consequence, comparing the T-F map to the CNN, it is possible that CNN is allowing to learn a mapping that is more diverse and more general than T-F mapping. This is reflected in the classification accuracy of the T-F map versus the CNN.

Further, the filter visualization findings in addition to the class attention map findings (see Figs. 5.12, 5.13, 5.14) show that the CNN model learns to focus on the rise time of the waveforms in order to decide on the label assignment of a tested sample. This shows that, in a given PD pulse, the CNN model is not only able to learn where to look at, but also it can classify waveforms according to their bandwidth which is associated with the rise time and the frequency content of the PD waveform.

5.3 Summary

A CNN attention based model with improved capability over traditional classification techniques (*e.g.* T-F map) to classify different void sizes resulting from a cavity in PLA 3D-printed samples was presented. Two classes corresponding to 2 and 0.5 mm sizes of void were studied. Mapping the waveforms corresponding to the two classes on the T-F map showed a classification accuracy of 77.28%. Also, three statistical features were extracted from recorded PRPD patterns, which then by applying PCA, revealed two clusters corresponding to the two classes. Although the latter is a successful tool, it is limited to AC applications in addition to the fact that additional hardware and software are needed to record the PRPD patterns. The 1D-CNN showed an improved classification accuracy of 98.7% with an increase of 21.42% compared to the T-F map. In order to have a deeper understanding of what the model is learning, a CNN interpretability method, Grad-CAM,

was implemented, where it is shown that the CNN learns to home in on the beginning of the waveform in order to assign the sample to a given class. The filter visualization findings showed that the CNN is learning finite-difference filters of various orders. This is in line with the inference that the CNN model is looking at the rise time of the waveforms in order to decide on the label assignment of a tested sample. In other words, the 1D-CNN was able to differentiate between the bandwidths of different PD pulses. After establishing the work done in Chapter 4 and Chapter 5, the last part of this research is directed toward considering more complicated scenario where the number of PD sources is not known. An unsupervised system is developed where convolutional autoencoder based model is able to detect the number of PD sources taking place in the insulation of a HV apparatus. The input of the unsupervised model is the time-series waveforms.

Chapter 6

Partial Discharge Adaptive Clustering Method

The focus of this chapter is on the identification of the number of PD sources that occur simultaneously in an insulation system. The proposed approach is implemented to identify the number of PD sources that take place in the stator bars in hydro-generators. Considerable research on classification of PD sources in rotating machine electrical insulation diagnosis has been done using supervised traditional machine learning techniques [11,79,80] and supervised deep learning techniques for classification purposes [55, 123–126]. For unsupervised tasks, unsupervised deep learning has been used in integration with experts knowledge to have a correctly labeled PD sources dataset in [60]. On the other hand, in order to design a PD system that would detect the number of PD sources, unsupervised traditional machine learning has been explored in literature. Other than the visual inspection of phase resolved partial discharge (PRPD) pattern, research has focused on extracting distinctive features from PD pulses. Separation maps (2D or 3D) would then be used after applying dimension

reduction techniques in order to decide on the number of clusters which would correspond to the number of PD sources. Different traditional clustering techniques include K- means, density based clustering algorithms, and Gaussian mixed models [127]. However, critical threshold choices should be made for such algorithms. For example, number of clusters and minimum distance between data points must be assumed in some algorithms. The authors in [128] proposed a new clustering technique to discriminate multiple PD sources from electrical noise. The proposed method considers each data point as an object with two criteria: a mass and a resultant local force generated by its neighbors. After extracting the features, clustering is applied. The authors reported better performance compared to traditional clustering techniques such as K-means. Accordingly, a gap has been identified in literature regarding identifying the number of PD sources with the use of raw data (PD pulses) and no prior critical thresholds to be assumed.

In this paper, a novel unsupervised deep learning system based on a convolutional autoencoder and an adaptive clustering technique is proposed to predict the number of partial discharge sources present in hydro-generator stator bars. Four common PD sources are simulated in a laboratory experimental setup in a stator bar. The inputs to the unsupervised system are the unlabeled PD pulses, and the output is the number of the PD sources. The findings are compared with the visual inspection of time-frequency (T-F) maps, which are used in commercial PD instruments. In addition, the proposed method is compared with traditional clustering techniques (K-means clustering and principal component analysis) that have been used for PD clustering in the literature. The visual inspection of T-F maps becomes challenging when two clusters overlap partially or completely. In addition, the equivalent time (T_{eq}) and equivalent bandwidth (W_{eq}) are affected by the signal to noise ratio (SNR), sampling rate, and other acquisition characteristics that affect the shape and

magnitude of a cluster in the T-F map [129]. Unlike the T-F map that uses two manually defined filters, i.e. the equivalent time (T_{eq}) and equivalent bandwidth (W_{eq}), in order to represent a PD pulse in a latent space, the proposed model automatically learns filters that perform better when the PD pulse is mapped into a latent space.

6.1 Experimental Setup and Data Collection

The winding and the core are the main components that comprise the stator of a generator. Stator winding insulation failure is a major cause of generators failure. The stator windings are made up of multiple bars/coils which are composed of insulated copper, are connected in series, and are held in the stator core slots [130]. Groundwall insulation prevents short circuits between the grounded stator core and the conductors. This insulation is usually formed of layers of mica tape saturated with resin.

Typical sources of PD that take place in stator bars have been studied in [131, 132]. The common sources of PDs include microvoids, surface discharge, corona discharge, and endwinding discharges that are simulated in this thesis. Microvoid PDs result from the manufacturing process of stator windings, where small air pockets in the groundwall insulation are formed [133]. Corona discharge occurs in the surrounding air of a high voltage conductor. Surface discharge occurs when the stress control coating around the winding is compromised, which leads to discharges between the winding surface and the grounded stator core. Surface discharges start appearing after increasing the applied voltage above a certain threshold.

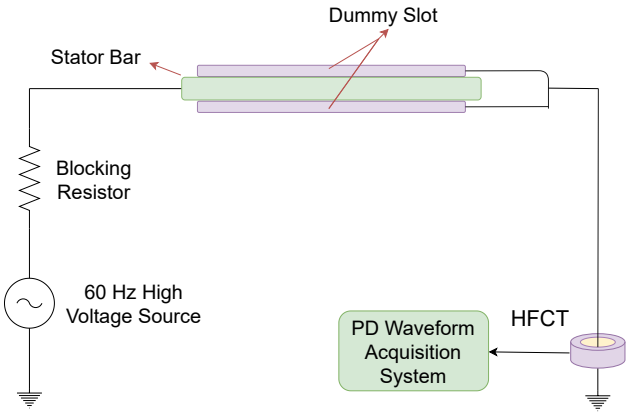
Fig. 6.1a shows the schematic of the experimental setup employed in this work. The setup consists of a high voltage transformer, a coupling capacitor, a generator stator bar, and a commercial PD measurement system. The stator bar is held in a grounded dummy slot,

which mimics a real stator slot. A high-frequency current transformer (HFCT) is connected to the ground wire to collect the PD time-series waveforms. A metal sphere is placed on the end of the bar in order to avoid the generation of unwanted corona. Figs. 6.1b and 6.1c show a photo of the laboratory setup and a close-up of the stator bar in the dummy slot. Corona discharge is simulated by removing the metal sphere from the end of the bar, and connecting a wire as shown in Fig. 6.2a. Endwinding discharges occur due to contamination, which allow the formation of conductive paths. This is simulated by spreading metallic particles on the stress control coating (see Fig. 6.2b). A commercial PD measurement system is used to acquire the PD time series waveform. A total of $57,696$ time-series waveforms are captured with a 10 ns time step. The recording time is 2 μs, so the length of the time series is 200 for each waveform. The PD sources are introduced sequentially into the experimental setup. The magnitude of all signals was normalized to unity to eliminate the magnitude factor from the learning criteria of the deep learning algorithm. $10,362$ samples (i.e. time-series waveforms) are captured when 11 kV is applied to the stator bar, $15,034$ samples are captured when floating particles are introduced to the setup, and $32,300$ samples are captured when corona source is added to the setup.

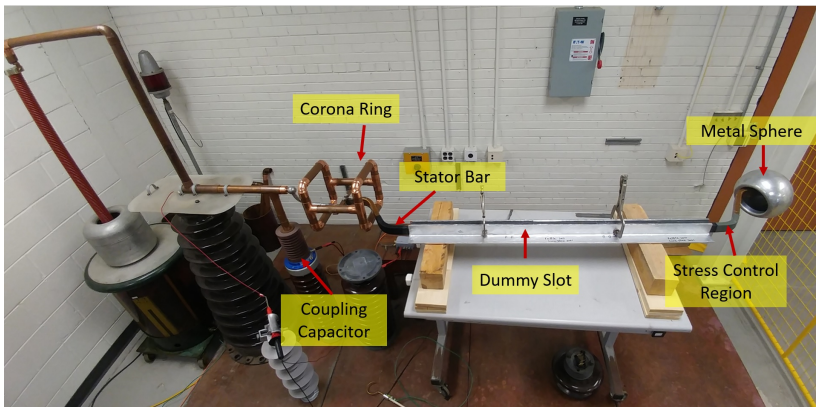
6.2 Method

6.2.1 Time-Frequency Map

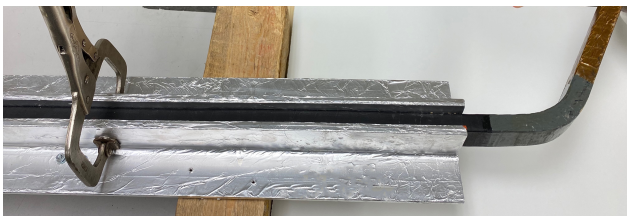
Analysing the time and frequency content of PD time-series waveforms allows researchers to classify different PD sources [87, 105]. This is done by computing the T-F map, which is a clustering technique that calculates the equivalent time, T_{eq}, and equivalent bandwidth, W_{eq} (see section 5.1.3 for details). A PD waveform is represented as a point on the T-F map



(a)



(b)



(c)

Fig. 6.1: Experimental setup for PD measurement in a generator stator bar: (a) schematic of the laboratory setup; (b) a photo of the lab setup; (c) close-up view of the stator bar in the dummy slot.

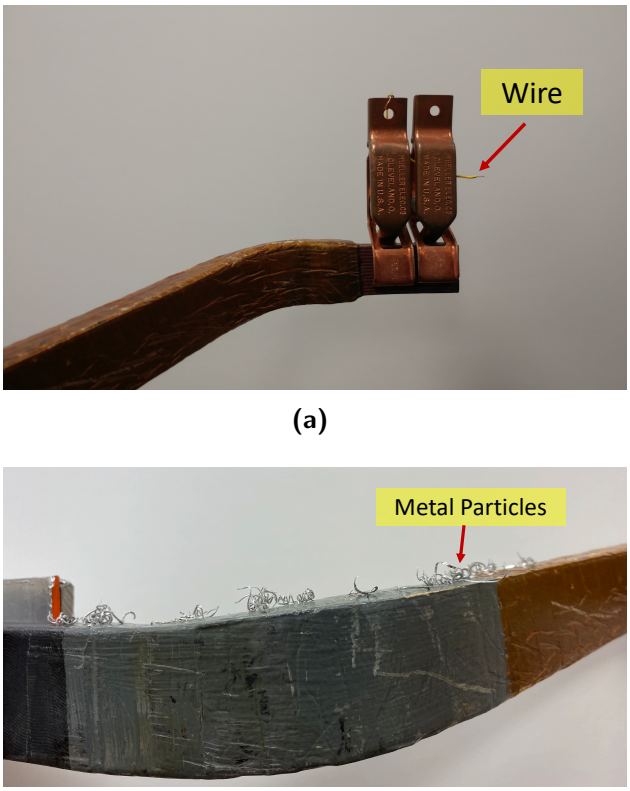


Fig. 6.2: (a) A wire is added to the end of the stator bar to simulate corona discharge; (b) Endwinding discharge is simulated by adding metallic particles on the stress control coating.

using the two values T_{eq} and W_{eq}. It is assumed that the PD waveforms due to a specific PD source are represented as a cluster on the T-F map; hence, it has been a common practice to use T-F map for clustering purposes and identifying new PD sources that are introduced to a high voltage asset [134]. However, using the T-F map becomes challenging when the clusters on the T-F map overlap, which suggests that the application of T_{eq} and W_{eq} as the discriminating features is not sufficient in some cases.

6.2.2 Convolutional Autoencoder

Autoencoders (AE) were introduced in the 1980s [29,30] in order to learn compact representations of unlabeled input data. With the emerging capabilities of deep learning architectures in 2006, they regained attention again [31]. Autoencoders consist of two main blocks: an encoder and a decoder. While training, the encoder transforms the input data into a latent space by representing it with a compacted representation. On the other hand, the decoder transforms the compacted representation into the original form. Performing this simultaneously by the encoder and the decoder, the neural network is trained by minimizing the reconstruction error between the input data and the decoder's output. The encoder and decoder can be represented by different neural network architectures, where the simplest one includes fully connected layers. Convolutional neural networks (CNNs) on the other hand represent a class of deep neural networks. Although CNNs were originally intended to be used for images, they have shown advantageous performance for different applications [25, 135, 136]. CNNs consist of a series of successive convolutional and pooling layers. A convolutional layer is made up of a collection of linear 1D, 2D or 3D filters which are convolved with an input to produce feature maps. In order to introduce non-linearity to the learning process, the output is usually passed through a non-linear activation function e.g.

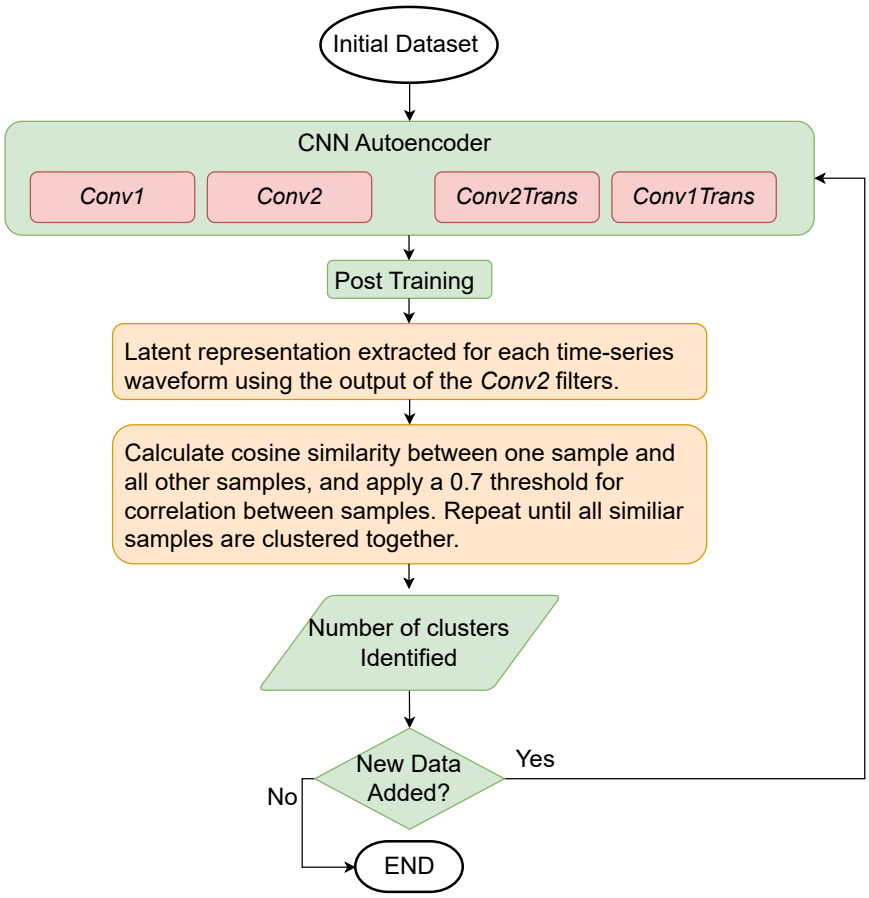


Fig. 6.3: Flowchart of the proposed method: After training the convolutional autoencoder, the learned bottleneck coefficients are used in order to map the time-series signal into a 16-dimensional space. Applying cosine similarity based clustering technique allows us to predict the number of PD sources.

a rectified linear unit (ReLU). A pooling layer is used to subsample the input, where usually maximum pooling is used which preserves the maximum value in a specific window. In this paper, due to the complexity of PD pulses, we have implemented a 1D-CNN based encoder and decoder for the PD time series waveforms; hence, having a convolutional autoencoder architecture.

6.2.3 Proposed Method

The flowchart of the proposed method is shown in Fig. 6.3. A convolutional autoencoder consisting of two convolutional layers is trained using an existing dataset. The hyper-parameters such as the filter size, number of filters, and learning rate are optimized. The length of the time series is $m = 200$ for each waveform. This is the input to the convolutional autoencoder. The first and second convolutional layers consist of 16 learnable filters of size 10×1 and 3×1 respectively. In order to sub-sample the input, a 2×1 max-pooling was performed after each convolutional layer. This constituted the encoder part. Since we are interested in the learnable filters after the second convolutional layers, we opted out of the flatten layer and proceeded to the decoder part. The inference speed measurements were performed on the NVIDIA Tesla T4 under CUDA 12.0, the number of floating point operations is $834,923$, and the number of the trainable parameters is $4,465$. The model is coded using Keras and Tensorflow in Python. Batch normalization is used, and the loss used for back propagation is the mean square error (MSE). Eighty percent of the samples are used for training and 20% are used for validation. After training the convolutional autoencoder, the compact representation from the second convolutional layer in the encoder is extracted. The compact representation that is learned by the convolutional autoencoder for each waveform is 50×16. Averaging of the 16 filters is done which allows each waveform to be represented by a 16-dimensional sample. The end representation vector of the N waveforms is $N \times 16$.

There are a number of similarity metrics in the literature to assess the closeness of data points in a feature space, such as Jaccard similarity, cosine similarity and Pearson's similarity. Jaccard similarity [137] computes the similarity between two samples by measuring the intersection divided by the union of the sample sets. Cosine similarity [138] computes the similarity between two samples by taking the dot product into consideration. Pearson's

correlation computes the correlation between two jointly distributed random variables [139]. Since we are trying to get the similarity between 16-dimensional samples with no prior knowledge if the samples are from the same distribution or different distributions, cosine similarity is adopted in the proposed method. The cosine similarity of vectors **a** and **b** is defined as [138]

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{6.1}$$

Two proportional vectors have a cosine similarity of 1, two opposite vectors have a cosine similarity of -1, and two orthogonal vectors have a cosine similarity of 0. In our proposed model, the threshold where two samples are considered similar is a hyper-parameter. The hyper-parameter is set to 0.7 which reflects an angle of 45° between two samples. A visual representation of the proposed adaptive clustering algorithm is shown in Fig. 6.4b. As shown in Fig. 6.4a, the algorithm starts off by drawing a random sample from N waveforms and assigns it to be the reference waveform. The absolute values of the cosine similarity between the reference sample and each of the rest of $N-1$ samples are computed. All waveforms with similarity equal to or larger than 0.7 are clustered with the reference waveform as Cluster 1, and all corresponding waveforms of the factors that are less than 0.7 are clustered together into an Auxiliary Cluster. The same procedure is repeated again on the Auxiliary Cluster. This is repeated until Auxiliary Cluster is null. As such, if this procedure is repeated K times, this means that the proposed model predicted $K-1$ clusters. The algorithm thus does not need a priori knowledge about the number of clusters and adaptively estimates the number of clusters depending on the sample distribution. The final check is done on the size of

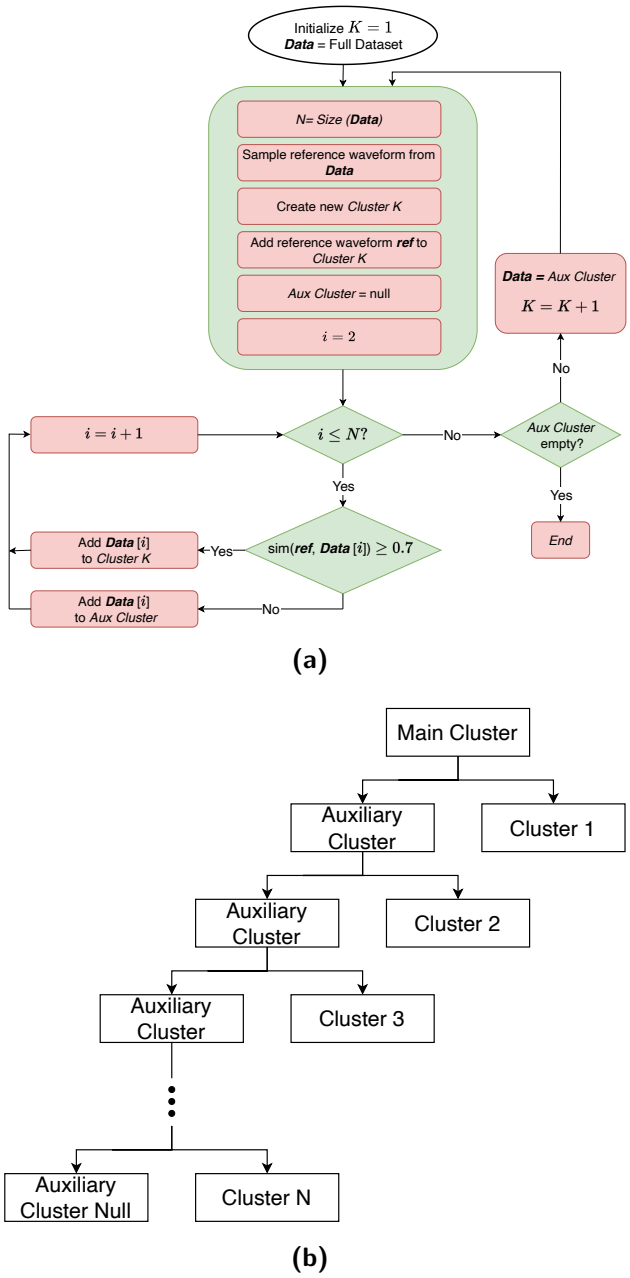


Fig. 6.4: Proposed adaptive clustering algorithm based on the cosine similarity criteria. (a) The overall flowchart; (b) Detailed implementation of the proposed algorithm.

the clusters. If any cluster is less than 1% of the total size of the evaluated dataset, it is considered anomaly and it is disregarded.

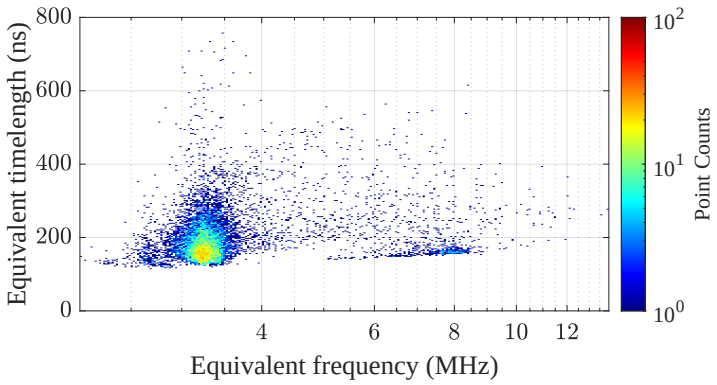
6.3 Findings

6.3.1 Performance Evaluation of the Proposed Technique

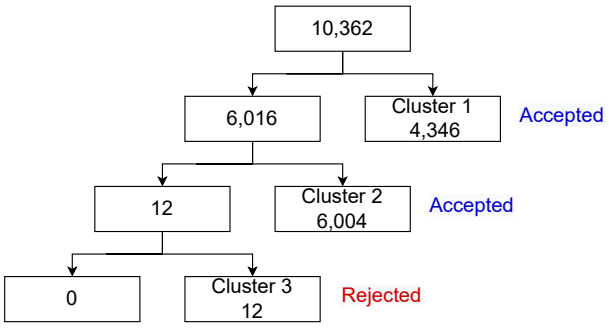
The performance of the proposed technique is evaluated for three different case studies, where PD sources are introduced sequentially to the lab setup.

Case #1

A voltage of 8 kV is applied that only excites PD in the microvoids of the stator bar insulation, and time-series PD waveforms are measured. These PD waveforms correspond to the left-hand side cluster of the T-F map shown in Fig. 6.5a. As the applied voltage is increased to 11 kV, another cluster starts forming (the right hand side cluster in Fig. 6.5a) that correspond to surface discharges. In the proposed method, 10,362 time-series PD waveforms are used to train the convolutional autoencoder. The results of the proposed model output are shown in Fig. 6.5b. The model suggests that there are two PD sources, which agrees with the T-F map and the known human experts knowledge. The proposed method predicts two accepted clusters (4,346 and 6,004 waveforms each) and a rejected cluster (with 12 time series waveforms). The rejected cluster is disregarded and considered as an anomaly as it compromises of less than 1% of the PD dataset.



(a)



(b)

Fig. 6.5: (a) T-F map of the PD signals captured when 11 kV is applied to the stator bar. The cluster on the left corresponds to PD in microvoids that also exist under 8 kV of applied voltage. The cluster in the middle that corresponds to the surface discharge appears at 11 kV; (b) Representation of clusters based on the proposed cosine similarity criteria for microvoid and surface discharge PD sources where the number of time-series waveforms is shown for each cluster. The proposed method identifies 2 PD sources that is confirmed by the T-F map and the PRPD pattern.

Case #2

Metallic particles are then spread on the stress control coating, and measurements are repeated at 11 kV. A new cluster (right-hand side cluster in Fig. 6.6a) appears in addition to the anticipated two clusters which correspond to internal and surface discharges (see Fig.

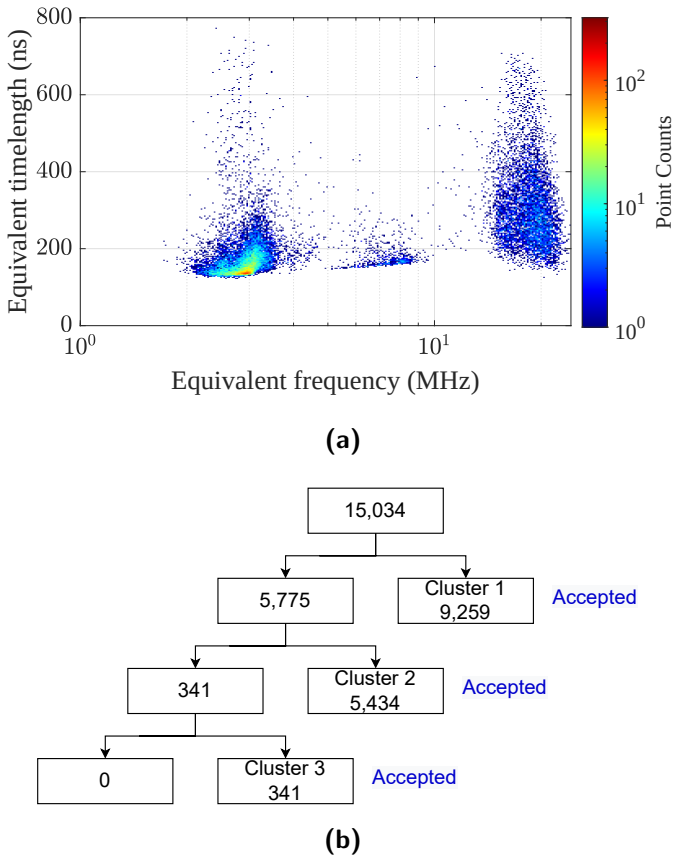


Fig. 6.6: (a) T-F map of the PD signals captured at 11kV of the applied voltage when floating particles are introduced to the setup; (b) Presentation of the clusters based on the cosine similarity criteria the shows the ability to identify the three sources of PD. The number of time-series waveforms is shown for each cluster.

6.5a). The performance of the proposed technique is demonstrated in Fig. 6.6b where three clusters are identified which correspond to microvoid, surface discharge, and floating particle PD sources. Since the three predicted clusters have 61.58% ,36.14%, and 2.26% of the PD dataset, respectively, none of the clusters is rejected.

Case #3

The fourth PD source is then introduced to the stator bar by creating a source of corona discharge. Since the floating particles PD source resembles a corona discharge, it is expected to see the cluster of the anticipated corona discharge close to the floating particle PD on the T-F map. This is shown in Fig.6.7 (a). The proposed model suggests five clusters of 61.13%, 28.81%, 8.26%, 1.64% and 0.14% of the total PD dataset. The fifth cluster with a density of 0.14% is disregarded, resulting is an identification of 4 PD sources.

6.3.2 Sensitivity of the Proposed Model to Noise

To evaluate the immunity of the proposed model to noise, additive white Gaussian noise (AWGN) with different signal to noise ratio (SNR) levels is added to the dataset [87]. An example of the AWGN with different SNR levels added to a measured PD waveform when the four PD sources are present is shown in Fig. 6.8. As shown in Fig. 6.9, the T-F map for each level of SNR shows a different behaviour in regards to the clusters position and shape. Although the number of PD sources is the same for the four levels of SNR, relying on the T-F map solely will predict four PD sources in the original dataset and two PD sources in the PD dataset with AWGN of SNR= 5 dB which is incorrect. Applying the proposed model, however, on the noisy dataset on the other hand gives consistent results regardless of the level of SNR for added AWGN. For illustration purposes, the presentation of the proposed model for the PD dataset with AWGN of SNR=5 dB is shown in Fig. 6.10 where four clusters/PD sources are still identified.

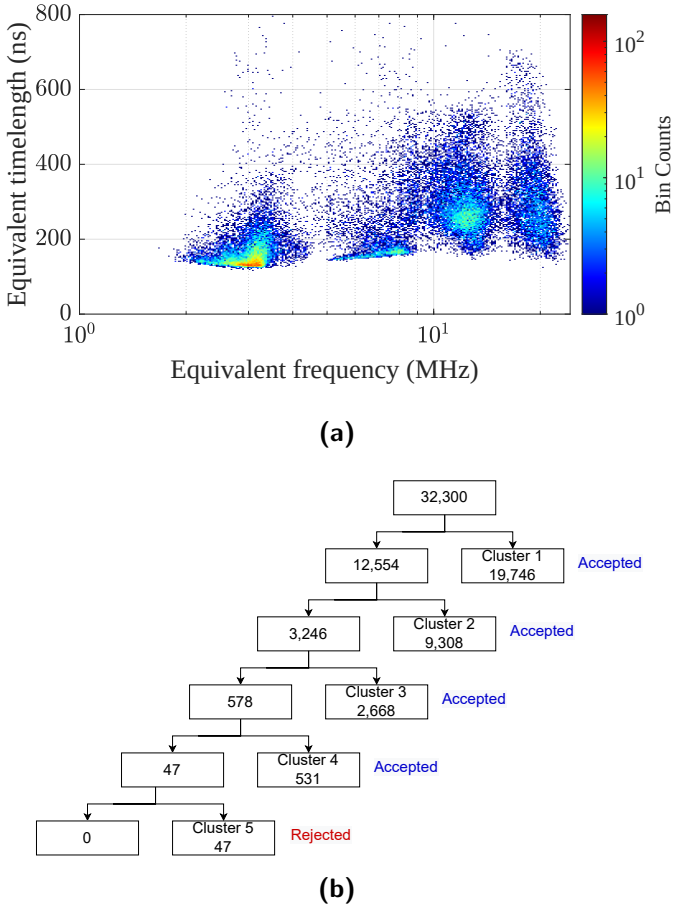


Fig. 6.7: (a) T-F map of the PD signals captured at 11kV of the applied voltage when both floating particles and corona source of PD are introduced to the setup; (b) Presentation of the clusters based on the cosine similarity criteria the shows the ability to identify the four sources of PD. The number of time-series waveforms is shown for each cluster.

6.3.3 Comparison with Traditional Machine Learning

Unsupervised learning seeks to find an underlying pattern in an unlabelled dataset and group the latter according to what similarities the learning process concludes [140]. Hence, one result of unsupervised learning is intended to be used for clustering purposes. Some of the traditional unsupervised machine learning algorithms are: hierarchical clustering, k-means

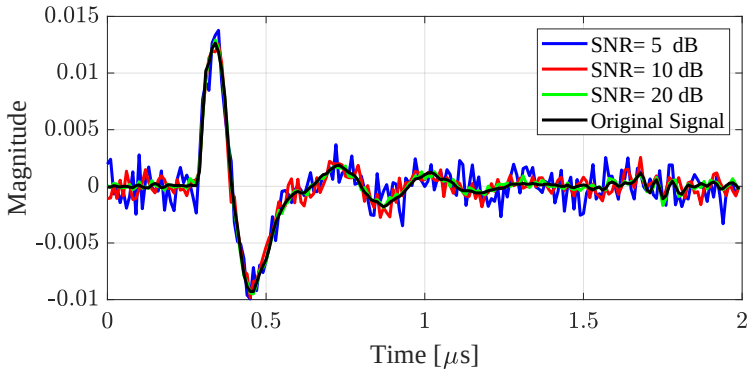


Fig. 6.8: A sample of a PD time-series waveform at different levels of signal-to-noise ratio (SNR) of additive white Gaussian noise (AWGN).

clustering, anomaly detection, principle component analysis, and independent component analysis. More information on these algorithms can be found in [141–145].

In this work, to demonstrate the performance of traditional methods, k-means and principle component analysis (PCA) are used on the raw data (i.e. time-series PD waveforms). K-means has a prerequisite of knowing the distribution of the clusters or the number of clusters [105]. For the sake of predicting the number of clusters, the elbow method is used [146] where the sum of squared error is used after running k-means for variable number of assumed clusters. The PCA is also applied on the raw time-series waveforms and the first three principle components are visualized in a 3D-plot. Using visual inspection, it is decided how many clusters results in the 3D-plot.

The four cases where PD waveforms generated by each of the PD sources are tested using the k-means and PCA methods and the number of clusters determined by these methods is shown in Table 6.1. In this table, the expected number of clusters is the true value. In addition, the energy preserved from the first three principle components is also reported, which shows that most of the variance in the time-series waveforms is captured by the first

Table 6.1: Number of clusters reported by traditional machine learning techniques of k-means and PCA that shows they are not always succesful in correct determination of the number of PD clusters.

	Number of clusters			
	Expected	Elbow and k-means	PCA	Energy preserved in the first 3 PCs
2 sources	2	2	3	98.4%
3 sources	3	3	1	96.6%
4 sources	4	3	2	93.0%

three principle components. The number of clusters predicted by either PCA or k-means doesn't reflect the true values accurately.

6.4 Summary

In this chapter, an unsupervised deep learning method based on convolutional autoencoder and cosine similarity based clustering technique was proposed for predicting the number of PD sources given unlabeled time-series waveforms. Lab measurements were performed on generator stator bar to evaluate the performance of the proposed technique. Two common PD sources that take place in generator windings, corona discharge and floating particles, are introduced to the lab setup in addition to the inherent microvoids and surface discharges resulting in four PD sources in total. The proposed model was compared with T-F map, which is a commonly used tool in industry in order to monitor the PD in high voltage insulation systems. The proposed unsupervised deep learning method was able to identify the number of PD sources correctly that is confirmed by using the T-F maps. Further, additive white Gaussian noise is added to the measured PD waveforms acquired when four

sources of PD are present in the setup. The results show that while the performance of the T-F map technique degrades with increasing the noise level and it doesn't reflect the true number of PD sources, the proposed model is immune to AWGN.

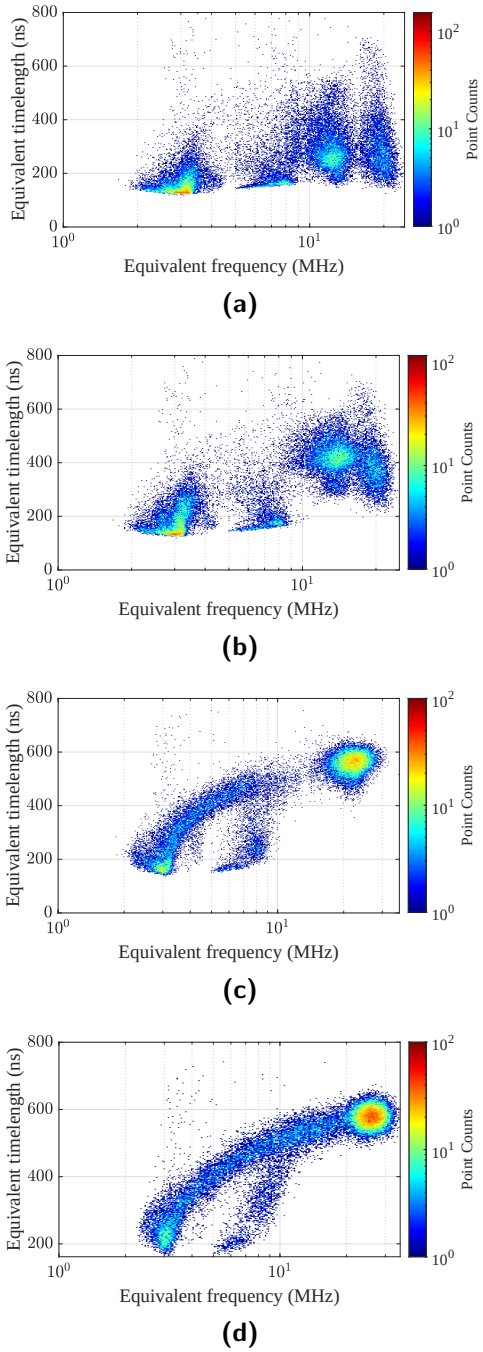


Fig. 6.9: T-F map for a dataset including four PD sources at 11 kV for various signal to noise ratios (SNR) of additive white Gaussian noise. (a) original dataset; (b) SNR= 20 dB; (c) SNR= 10 dB; (d) SNR= 5 dB.

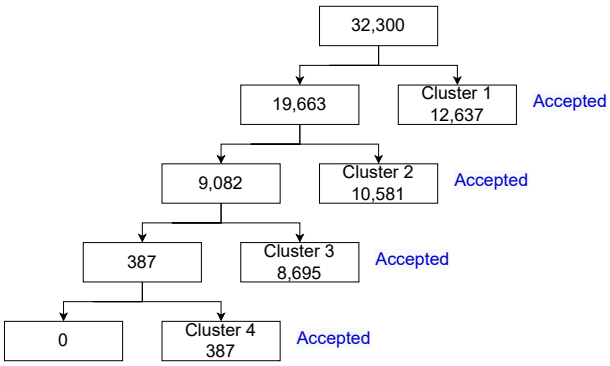


Fig. 6.10: Presentation of clusters based on the cosine similarity criteria for void, surface, floating particle and corona PD sources in the presence of AWGN of SNR=5 dB where the number of time-series waveforms is shown for each cluster.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Electrical insulation systems play a major role in the safe, reliable operation of all electric devices and apparatus and similar to other systems are prone to failure under operational conditions. One factor for estimating such a failure is measuring and analysing partial discharges. Insulation failure could be catastrophic especially when the operating voltage is high. Operating under a high voltage or equivalently a high electric field, a specific insulation system degrades faster, and as a result, small local breakdown/discharges take place that are caused by the electrical field, leading to the deterioration of the insulation system. This breakdown is called partial discharge (PD). Therefore, diagnosing a defective insulation system by detecting the partial discharges and classifying the type of the defect in early stages is important for having a reliable electric system. The thesis's scope is developing models and techniques within deep learning for PD classification application. As the models discussed are data-driven models, it is crucial to highlight the importance of feeding contin-

uous updated PD data to the machine. The models proposed in this research were compared with different traditional machine learning methods, where better performance was recorded for the investigated problem settings.

A limitation of using the classification models without fine tuning is present. In addition, knowing that the models were trained on data that represent particular PD sources, these models will perform poorly when they are tested on data representing different PD sources. A second limitation is linked to the use of the PLA 3D-printed samples. Changing the conditions, whether it is the shape and size of the void or the dielectric material would definitely be of interest to investigate in future work. The evolution of the CNN weights as we vary the experimental conditions could shed light on the role of these changes in classifying the void sizes. In addition, the post-hoc attention would be valuable in this analysis as well where we can interpret as to which changes conditions contribute to the final decision of the CNN. A third limitation is linked to the use of a specific detection PD system. If another detection PD system is to be used, the pre-trained models can't be used; therefore, the need for re-training is necessary. The last limitation is linked to the change in the measurement conditions (e.g. an added impedance). Re-training of the model is crucial as this affects the intrinsic characteristic of the measured pulses/patterns.

7.2 Future Work

With the rapid evolving of deep learning algorithm, future work can focus on comparing new deep learning algorithm with the already existing models. In addition, to overcome the limitation present earlier, a potential future work is to collect data including more PD

sources, where the models will have the ability to generalize over more scenarios when used in a real-life setting.

While the proposed work related to the attention based model prove that that there is a connection between the void size and the corresponding bandwidth of the associated PD waveform, future work can be directed toward integrating the physics behind the PD waveforms into the learning process. In addition, future work can be done in regards to implementing a new visual map, which can take into consideration different learned filters instead of the known equivalence time and equivalence frequency.

A future path for the unsupervised system include involving more PD sources. Transfer learning can be implemented and explored.

References

[1] F. A. Rizk and G. N. Trinh, *High voltage engineering*. CRC Press, 2018.

[2] Y. Luo, Z. Li, and H. Wang, "A review of online partial discharge measurement of large generators," *Energies*, vol. 10, no. 11, p. 1694, 2017.

[3] J. Kuffel and E. Kuffel, *High voltage engineering fundamentals*. Elsevier, 2000.

[4] F. Kreuger, E. Gulski, and A. Krivda, "Classification of partial discharges," *IEEE transactions on Electrical Insulation*, vol. 28, no. 6, pp. 917–931, 1993.

[5] C. Cachin and H. J. Wiesmann, "Pd recognition with knowledge-based preprocessing and neural networks," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 2, no. 4, pp. 578–589, 1995.

[6] A. Krivda, "Automated recognition of partial discharges," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 2, no. 5, pp. 796–821, 1995.

[7] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE transactions on neural networks*, vol. 13, no. 2, pp. 464–471, 2002.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification second edition john wiley & sons," *New York*, vol. 58, p. 16, 2001.

[9] D. F. Specht, "Probabilistic neural networks," *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[11] M. Wu, H. Cao, J. Cao, H.-L. Nguyen, J. B. Gomes, and S. P. Krishnaswamy, "An overview of state-of-the-art partial discharge analysis techniques for condition monitoring," *IEEE electrical insulation magazine*, vol. 31, no. 6, pp. 22–35, 2015.

[12] J. A. Ardila-Rey, J. E. Ortiz, W. Creixell, F. Muhammad-Sukki, and N. A. Bani, "Artificial generation of partial discharge sources through an algorithm based on deep convolutional generative adversarial networks," *IEEE Access*, vol. 8, pp. 24 561–24 575, 2020.

[13] H. Sinaga, B. Phung, and T. Blackburn, "Partial discharge localization in transformers using uhf detection method," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 19, no. 6, pp. 1891–1900, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] J. J. Levy, A. J. Titus, C. L. Petersen, Y. Chen, L. A. Salas, and B. C. Christensen, "Methylnet: an automated and modular deep learning approach for dna methylation analysis," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–15, 2020.

[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT Press Cambridge, 2016, vol. 1, no. 2.

[18] G. Zhang, Y. Liu, and X. Jin, "A survey of autoencoder-based recommender systems," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 430–450, 2020.

[19] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, p. 36, 2018.

[20] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep q-learning," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 486–489.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[26] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[27] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *International conference on rough sets and knowledge technology*. Springer, 2014, pp. 364–375.

[28] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[30] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *ICML Workshop on Unsupervised and Transfer Learning*, Bellevue, WA, USA, July 2012, pp. 37–49.

[31] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[33] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[34] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*. Springer, 2005, pp. 799–804.

[37] J. Michalek and J. Vaněk, "A survey of recent dnn architectures on the timit phone recognition task," in *International Conference on Text, Speech, and Dialogue*. Springer, 2018, pp. 436–444.

[38] T. Okamoto and T. Tanaka, "Novel partial discharge measurement computer-aided measuremnet systems," *IEEE Transactions on Electrical Insulation*, vol. EI-21, no. 6, pp. 1015–1019, 1986.

[39] J. Tang, M. Jin, F. Zeng, X. Zhang, and R. Huang, "Assessment of pd severity in gas-insulated switchgear with an ssae," *IET Science, Measurement & Technology*, vol. 11, no. 4, pp. 423–430, 2017.

[40] M.-T. Nguyen, V.-H. Nguyen, S.-J. Yun, and Y.-H. Kim, "Recurrent neural network for partial discharge diagnosis in gas-insulated switchgear," *Energies*, vol. 11, no. 5, p. 1202, 2018.

[41] V.-N. Tuyet-Doan, T.-D. Do, N.-D. Tran-Thi, Y.-W. Youn, and Y.-H. Kim, "One-shot learning for partial discharge diagnosis using ultra-high-frequency sensor in gas-insulated switchgear," *Sensors*, vol. 20, no. 19, p. 5562, 2020.

[42] V.-N. Tuyet-Doan, T.-T. Nguyen, M.-T. Nguyen, J.-H. Lee, and Y.-H. Kim, "Self-attention network for partial-discharge diagnosis in gas-insulated switchgear," *Energies*, vol. 13, no. 8, p. 2102, 2020.

[43] H. Song, J. Dai, G. Sheng, and X. Jiang, "Gis partial discharge pattern recognition via deep convolutional neural network under complex data source," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 25, no. 2, pp. 678–685, 2018.

[44] J. Dai, Y. Teng, Z. Zhang, Z. Yu, G. Sheng, and X. Jiang, "Partial discharge data matching method for gis case-based reasoning," *Energies*, vol. 12, no. 19, p. 3677, 2019.

[45] M. Karimi, M. Majidi, H. MirSaeedi, M. M. Arefi, and M. Oskuoee, "A novel application of deep belief networks in learning partial discharge patterns for classifying corona, surface, and internal discharges," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 4, pp. 3277–3287, 2019.

[46] M. Florkowski, "Classification of partial discharge images using deep convolutional neural networks," *Energies*, vol. 13, no. 20, p. 5496, 2020.

[47] X. Zhou, X. Wu, P. Ding, X. Li, N. He, G. Zhang, and X. Zhang, "Research on transformer partial discharge uhf pattern recognition based on cnn-lstm," *Energies*, vol. 13, no. 1, p. 61, 2019.

[48] T.-D. Do, V.-N. Tuyet-Doan, Y.-S. Cho, J.-H. Sun, and Y.-H. Kim, "Convolutional-neural-network-based partial discharge diagnosis for power transformer using uhf sensor," *IEEE Access*, vol. 8, pp. 207 377–207 388, 2020.

[49] Y. Sun, S. Ma, S. Sun, P. Liu, L. Zhang, J. Ouyang, and X. Ni, "Partial discharge pattern recognition of transformers based on mobilenets convolutional neural network," *Applied Sciences*, vol. 11, no. 15, p. 6984, 2021.

[50] Y. Liu, M. Hu, Q. Dai, H. Le, and Y. Liu, "Online recognition method of partial discharge pattern for transformer bushings based on small sample ultra-micro-cnn network," *AIP Advances*, vol. 11, no. 4, p. 045221, 2021.

[51] S. Mantach, A. Ashraf, H. Janani, and B. Kordi, "A convolutional neural network-based model for multi-source and single-source partial discharge pattern classification using only single-source training set," *Energies*, vol. 14, no. 5, p. 1355, 2021.

[52] B. Adam and S. Tenbohlen, "Classification of superimposed partial discharge patterns," *Energies*, vol. 14, no. 8, p. 2144, 2021.

[53] M. Borghei and M. Ghassemi, "A deep learning approach for discrimination of single- and multi-source corona discharges," *IEEE Transactions on Plasma Science*, vol. 49, no. 9, pp. 2936–2945, 2021.

[54] W. J. K. Raymond, C. W. Xin, L. W. Kin, and H. A. Illias, "Noise invariant partial discharge classification based on convolutional neural network," *Measurement*, vol. 177, p. 109220, 2021.

[55] R. C. Araújo, R. M. de Oliveira, and F. J. Barros, "Automatic PRPD image recognition of multiple simultaneous partial discharge sources in on-line hydro-generator stator bars," *Energies*, vol. 15, no. 1, p. 326, 2022.

[56] M. A. Khan, J. Choo, and Y.-H. Kim, "End-to-end partial discharge detection in power cables via time-domain convolutional neural networks," *Journal of Electrical Engineering & Technology*, vol. 14, no. 3, pp. 1299–1309, 2019.

[57] X. Peng, F. Yang, G. Wang, Y. Wu, L. Li, Z. Li, A. A. Bhatti, C. Zhou, D. M. Hepburn, A. J. Reid *et al.*, "A convolutional neural network-based deep learning methodology for recognition of partial discharge patterns from high-voltage cables," *IEEE Transactions on Power Delivery*, vol. 34, no. 4, pp. 1460–1469, 2019.

[58] Z. Li, Y. Xu, and X. Jiang, "Pattern recognition of dc partial discharge on xlpe cable based on adam-dbn," *Energies*, vol. 13, no. 17, p. 4566, 2020.

[59] J. Yeo, H. Jin, A. R. Mor, C. Yuen, W. Tushar, T. K. Saha, and C. S. Ng, "Identification of partial discharge through cable-specific adaption and neural network ensemble," *IEEE Transactions on Power Delivery*, 2021.

[60] R. Zemouri, M. Levesque, N. Amyot, C. Hudon, O. Kokoko, and S. A. Tahan, "Deep convolutional variational autoencoder as a 2d-visualization tool for partial discharge source classification in hydrogenerators," *IEEE Access*, vol. 8, pp. 5438–5454, 2019.

[61] Y. Wang, J. Yan, Z. Yang, T. Liu, Y. Zhao, and J. Li, "Partial discharge pattern recognition of gas-insulated switchgear via a light-scale convolutional neural network," *Energies*, vol. 12, no. 24, p. 4674, 2019.

[62] S. Barrios, D. Buldain, M. P. Comech, and I. Gilbert, "Partial discharge identification in mv switchgear using scalogram representations and convolutional autoencoder," *IEEE Transactions on Power Delivery*, vol. 36, no. 6, pp. 3448–3455, 2020.

[63] F.-C. Gu, "Identification of partial discharge defects in gas-insulated switchgears by using a deep learning method," *IEEE Access*, vol. 8, pp. 163 894–163 902, 2020.

[64] Y. Wang, J. Yan, Z. Yang, Y. Zhao, and T. Liu, "Optimizing gis partial discharge pattern recognition in the ubiquitous power internet of things context: A mixnet deep learning model," *International Journal of Electrical Power & Energy Systems*, vol. 125, p. 106484, 2021.

[65] Y. Wang, J. Yan, Z. Yang, J. Wang, and Y. Geng, "A novel 1dcnn and domain adversarial transfer strategy for small sample gis partial discharge pattern recognition," *Measurement Science and Technology*, vol. 32, no. 12, p. 125118, 2021.

[66] G. Li, X. Wang, X. Li, A. Yang, and M. Rong, "Partial discharge recognition with a multi-resolution convolutional neural network," *Sensors*, vol. 18, no. 10, p. 3512, 2018.

[67] Y. Wang, J. Yan, Z. Yang, Q. Jing, Z. Qi, J. Wang, and Y. Geng, "A domain adaptive deep transfer learning method for gas-insulated switchgear partial discharge diagnosis," *IEEE Transactions on Power Delivery*, 2021.

[68] T. Liu, J. Yan, Y. Wang, Y. Xu, and Y. Zhao, "Gis partial discharge pattern recognition based on a novel convolutional neural networks and long short-term memory," *Entropy*, vol. 23, no. 6, p. 774, 2021.

[69] G. Michau, C.-C. Hsu, and O. Fink, "Interpretable detection of partial discharge in power lines with deep learning," *Sensors*, vol. 21, no. 6, p. 2154, 2021.

[70] M. Zunaed, A. Nath, M. Rahman *et al.*, "Dual-cycon net: a cycle consistent dual-domain convolutional neural network framework for detection of partial discharge," *arXiv preprint arXiv:2012.11532*, 2020.

[71] M. Dong and J. Sun, "Partial discharge detection on aerial covered conductors using time-series decomposition and long short-term memory network," *Electric Power Systems Research*, vol. 184, p. 106318, 2020.

[72] Z. Li, N. Qu, X. Li, J. Zuo, and Y. Yin, "Partial discharge detection of insulated conductors based on cnn-lstm of attention mechanisms," *Journal of Power Electronics*, vol. 21, no. 7, pp. 1030–1040, 2021.

[73] B. Vigneshwaran, M. W. Iruthayarajan, and R. Maheswari, "Recognition of shed damage on 11-kv polymer insulator using bayesian optimized convolution neural network," *Soft Computing*, pp. 1–13, 2022.

[74] A. Sahoo, A. Subramaniam, S. Bhandari, and S. K. Panda, "A review on condition monitoring of gis," in *2017 International Symposium on Electrical Insulating Materials (ISEIM)*, vol. 2. IEEE, 2017, pp. 543–546.

[75] ENET-Centre, "VSB power line fault detection," https://www.kaggle.com/c/vsb-power-line-fault-detection,, 2019, last accessed June 9, 2022.

[76] G. Chen, M. Hao, Z. Xu, A. Vaughan, J. Cao, and H. Wang, "Review of high voltage direct current cables," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 2, pp. 9–21, 2015.

[77] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[78] A. E. Fitzgerald, C. Kingsley, S. D. Umans, and B. James, *Electric machinery*. McGraw-Hill New York, 2003, vol. 5.

[79] M. Lévesque, N. Amyot, C. Hudon, M. Bélec, and O. Blancke, "Improvement of a hydrogenerator prognostic model by using partial discharge measurement analysis," in *Annual Conference of the PHM Society*, vol. 9, no. 1, 2017.

[80] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33–47, 2018.

[81] B. Ganguly, S. Chaudhury, S. Biswas, D. Dey, S. Munshi, B. Chatterjee, S. Dalai, and S. Chakravorti, "Wavelet kernel based convolutional neural network for localization of partial discharge sources within a power apparatus," *Transactions on Industrial Informatics*, 2020.

[82] E. Gulski and F. Kreuger, "Computer-aided recognition of discharge sources," *IEEE Transactions on Electrical Insulation*, vol. 27, no. 1, pp. 82–92, 1992.

[83] H. Janani, N. D. Jacob, and B. Kordi, "Automated recognition of partial discharge in oil-immersed insulation," in *IEEE Electrical Insulation Conference (EIC)*, 2015, pp. 467–470.

[84] H. Janani and B. Kordi, "Towards automated statistical partial discharge source classification using pattern recognition techniques," *IET High Voltage*, vol. 3, no. 3, pp. 162–169, 2018.

[85] I. 60270:, "High-voltage test techniques: partial discharge measurements," *Tech. Rep.*, 2000.

[86] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 647–657.

[87] A. Cavallini, G. Montanari, A. Contin, and F. Pulletti, "A new approach to the diagnosis of solid insulation systems based on PD signal inference," *Electrical Insulation Magazine*, vol. 19, no. 2, pp. 23–30, 2003.

[88] H. Okubo and N. Hayakawa, "A novel technique for partial discharge and breakdown investigation based on current pulse waveform analysis," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 4, pp. 736–744, 2005.

[89] F. Alvarez, J. Ortego, F. Garnacho, and M. Sanchez-Uran, "A clustering technique for partial discharge and noise sources identification in power cables by means of waveform parameters," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 23, no. 1, pp. 469–481, 2016.

[90] A. Cavallini, A. Contin, G. C. Montanari, and F. Puletti, "Advanced pd inference in on-field measurements. i. noise rejection," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 10, no. 2, pp. 216–224, 2003.

[91] A. Contin and S. Pastore, "Classification and separation of partial discharge signals by means of their auto-correlation function evaluation," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 16, no. 6, pp. 1609–1622, 2009.

[92] V. Nagesh and B. Gururaj, "Evaluation of digital filters for rejecting discrete spectral interference in on-site pd measurements," *IEEE Transactions on Electrical Insulation*, vol. 28, no. 1, pp. 73–85, 1993.

[93] J. Martínez-Tarifa, G. Robles, M. Rojas-Moreno, and J. Sanz-Feito, "Partial discharge pulse shape recognition using an inductive loop sensor," *Measurement Science and Technology*, vol. 21, no. 10, p. 105706, 2010.

[94] L. Hao, P. Lewin, J. Hunter, D. Swaffield, A. Contin, C. Walton, and M. Michel, "Discrimination of multiple pd sources using wavelet decomposition and principal component analysis," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 18, no. 5, pp. 1702–1711, 2011.

[95] H. J. van Breen, E. Gulski, J. Smit, H. Verhaart, and W. De Leeuw, "Experience with on-line pd measurements on generators in frequency and time domain," in *Conference Record of the the 2002 IEEE International Symposium on Electrical Insulation (Cat. No. 02CH37316)*. IEEE, 2002, pp. 31–35.

[96] M. Cacciari, A. Contin, G. Mazzanti, and G. Montanari, "Identification and separation of two concurrent partial discharge phenomena," in *IEEE Conference on Electrical Insulation and Dielectric Phenomena*, vol. 2, August 1996, pp. 476–479.

[97] E. Lalitha and L. Satish, "Wavelet analysis for classification of multi-source PD patterns," *Transactions on Dielectrics and Electrical Insulation*, vol. 7, no. 1, pp. 40–47, 2000.

[98] B. Adam and S. Tenbohlen, "Classification of multiple pd sources by signal features and lstm networks," in *2018 IEEE International Conference on High Voltage Engineering and Application (ICHVE)*. IEEE, 2018, pp. 1–4.

[99] H. Janani, P. Jayasinghe, M. J. Jozani, and B. Kordi, "Statistical feature extraction and system identification algorithms for partial discharge signal classification using laguerre polynomial expansion," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 27, no. 6, pp. 1924–1932, 2020.

[100] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–2.

[101] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[102] S. Mantach, P. Gill, D. R. Oliver, A. Ashraf, and B. Kordi, "An interpretable cnn model for classification of partial discharge waveforms in 3d-printed dielectric samples with different void sizes," *Neural Computing and Applications*, pp. 1–12, 2022.

[103] M. Borghei, M. Ghassemi, B. Kordi, P. Gill, and D. Oliver, "A finite element analysis model for internal partial discharges in an air-filled cylindrical cavity inside solid dielectric," in *IEEE Electrical Insulation Conference (EIC)*, 2021, pp. 7–21.

[104] P. Gill, "3d printed poly-lactic acid for partial discharge studies," Master's thesis, University of Manitoba, 2022.

[105] A. Contin, A. Cavallini, G. Montanari, G. Pasini, and F. Puletti, "Digital detection and fuzzy classification of partial discharge signals," *Transactions on Dielectrics and Electrical Insulation*, vol. 9, no. 3, pp. 335–348, 2002.

[106] A. Cavallini, G. Montanari, F. Puletti, and A. Contin, "A new methodology for the identification of pd in electrical apparatus: properties and applications," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 2, pp. 203–215, 2005.

[107] H. Janani, "Partial discharge source classification using pattern recognition algorithms," Ph.D. dissertation, University of Manitoba, 2016.

[108] X. Shen, Z. Ni, L. Liu, J. Yang, and K. Ahmed, "Wipass: 1d-cnn-based smartphone keystroke recognition using wifi signals," *Pervasive and Mobile Computing*, vol. 73, p. 101393, 2021.

[109] W. Chen and K. Shi, "A deep learning framework for time series classification using relative position matrix and convolutional neural network," *Neurocomputing*, vol. 359, pp. 384–394, 2019.

[110] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[111] I. Goodfellow, H. Lee, Q. Le, A. Saxe, and A. Ng, "Measuring invariances in deep networks," *Advances in neural information processing systems*, vol. 22, pp. 646–654, 2009.

[112] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[113] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.

[114] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989.

[115] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[116] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391*, 2018.

[117] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[118] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[119] T. He, J. Guo, N. Chen, X. Xu, Z. Wang, K. Fu, L. Liu, and Z. Yi, "Medimlp: using grad-cam to extract crucial variables for lung cancer postoperative complication prediction," *IEEE journal of biomedical and health informatics*, vol. 24, no. 6, pp. 1762–1771, 2019.

[120] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons & Fractals*, vol. 140, p. 110190, 2020.

[121] D. Cian, J. van Gemert, and A. Lengyel, "Evaluating the performance of the lime and grad-cam explanation methods on a lego multi-label image classification task," *arXiv preprint arXiv:2008.01584*, 2020.

[122] F. Feng, C. Wu, J. Zhu, S. Wu, Q. Tian, and P. Jiang, "Research on multitask fault diagnosis and weight visualization of rotating machinery based on convolutional neural network," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 42, no. 11, pp. 1–14, 2020.

[123] R. Oliveira, R. C. Araújo, F. J. Barros, A. P. Segundo, R. F. Zampolo, W. Fonseca, V. Dmitriev, and F. S. Brasil, "A system based on artificial neural networks for automatic classification of hydro-generator stator windings partial discharges," *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, vol. 16, pp. 628–645, 2017.

[124] R. Srivastava, V. Avasthi *et al.*, "Deep convolutional neural network for partial discharge monitoring system," *Advances in Engineering Software*, vol. 180, p. 103407, 2023.

[125] S. Lu, H. Chai, A. Sahoo, and B. Phung, "Condition monitoring based on partial discharge diagnostics using machine learning methods: A comprehensive state-of-the-art review," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 27, no. 6, pp. 1861–1888, 2020.

[126] Y. Wang, J. Yan, Z. Wang, D. Zhao, R. He, J. Wang, and Y. Geng, "Multi-source partial discharge diagnosis in gas-insulated switchgear via zero-shot learning," *Measurement*, vol. 217, p. 113033, 2023.

[127] L. C. C. Heredia and A. R. Mor, "Density-based clustering methods for unsupervised separation of partial discharge sources," *International Journal of Electrical Power & Energy Systems*, vol. 107, pp. 224–230, 2019.

[128] C. Boya-Lara, O. Rivera-Caballero, and J. A. Ardila-Rey, "Clustering by communication with local agents for noise and multiple partial discharges discrimination," *Expert Systems with Applications*, vol. 225, p. 120067, 2023.

[129] A. R. Mor, L. C. Heredia, and F. Muñoz, "Effect of acquisition parameters on equivalent time and equivalent bandwidth algorithms for partial discharge clustering," *International Journal of Electrical Power & Energy Systems*, vol. 88, pp. 141–149, 2017.

[130] N. Dehlinger and G. Stone, "Surface partial discharge in hydrogenerator stator windings: Causes, symptoms, and remedies," *Electrical Insulation Magazine*, vol. 36, no. 3, pp. 7–18, 2020.

[131] X. Jiang, G. Liu, G. Zhang, and G. Wu, "Study on PD characteristics of insulation internal defects of large generator stator bar," in *IEEE International Symposium on Electrical Insulation*, Boston, MA, USA, August 2002, pp. 14–18.

[132] G. C. Stone, E. A. Boulter, I. Culbert, and H. Dhirani, *Electrical insulation for rotating machines: design, evaluation, aging, testing, and repair*, 2nd ed. John Wiley & Sons, 2014.

[133] IEC/TS Standard 60034 Ed 1.0, *Rotating electrical machines - Part 27-2: On-line partial discharge measurements on the stator winding insulation of rotating electrical machines*. IEC, 2012.

[134] J. C. Chan, H. Ma, and T. K. Saha, "Time-frequency sparsity map on automatic partial discharge sources separation for power transformer condition assessment," *Transactions on Dielectrics and Electrical Insulation*, vol. 22, no. 4, pp. 2271–2283, 2015.

[135] L. Duan, J. Hu, G. Zhao, K. Chen, S. X. Wang, and J. He, "Method of inter-turn fault detection for next-generation smart transformers based on deep learning algorithm," *High Voltage*, vol. 4, no. 4, pp. 282–291, 2019.

[136] S. Polisetty, A. El-Hag, and S. Jayram, "Classification of common discharges in outdoor insulation using acoustic signals and artificial neural network," *High Voltage*, vol. 4, no. 4, pp. 333–338, 2019.

[137] A. H. Murphy, "The finley affair: A signal event in the history of forecast verification," *Weather and forecasting*, vol. 11, no. 1, pp. 3–20, 1996.

[138] A. Singhal *et al.*, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[139] M. L. Gaddis and G. M. Gaddis, "Introduction to biostatistics: part 6, correlation and regression," *Annals of emergency medicine*, vol. 19, no. 12, pp. 1462–1468, 1990.

[140] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[141] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[142] T. M. Kodinariya, P. R. Makwana *et al.*, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[143] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[144] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[145] J. V. Stone, "Independent component analysis: an introduction," *Trends in Cognitive Sciences*, vol. 6, no. 2, pp. 59–64, 2002.

[146] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336, no. 1. Surabaya, Indonesia: IOP Publishing, November 2018, p. 012017.