MACHINE LEARNING BASED RESOLUTIONS FOR PARTIAL DISCHARGE DETECTION

XU NING

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Master of Engineering

2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

25-04-2022

. . . . . .
Date

. . . . . .
Xu Ning

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

30-04-2022

. . . . . .
Date

. . . . . .
Assoc. Prof. H. B. Gooi

Authorship Attribution Statement

This thesis contains material from one journal paper under revision, one published conference paper, and one journal paper to be submitted, where I am the first author.

The material contained in this thesis is published as Ning Xu, Hoay Beng Gooi, Lipo Wang, Yuanjin Zheng, Jiawei Yang, “Partial Discharge Detection Based on Long Short-Term Memory Neural Network Classifier with Efficient Feature Extraction Methods,” *The 2021 IEEE 12th Energy Conversion Congress and Exposition – Asia* (*ECCE-ASIA*), 24-27 May 2021, Singapore.

The contributions of the co-authors are as follows:

- I have proposed the partial discharge detection method; conducted verification experiments; and analyzed the results.
- I wrote the drafts of the manuscript. The manuscript was revised together with Mr. Jiawei Yang.
- Prof. Hoay Beng Gooi, Prof. Yuanjin Zheng, and Prof. Lipo Wang closely supervised the research work. They gave suggestions for the grammar and technical concept of the manuscript and reviewed the manuscript for the final submission.

25-04-2022

. . . . . .
Date

. . . . . .
Xu Ning

Acknowledgement

First and foremost, I would like to express my greatest gratitude to my supervisor Assoc. Prof. Gooi Hoay Beng and co-supervisors Assoc. Prof. Zheng Yuanjin and Assoc. Prof. Wang Lipo, whose guidance and endless patience helped me to complete my master studies.

I would like to express my sincere thanks to Prof. Mišák Stanislav and Dr. Fulneček Jan who generously shared the valuable measurement data on the internet. I learned a lot from the collaboration with them. Moreover, I would like to thank all my colleagues and friends for their patient assistance and precious friendship.

I would like to dedicate my deepest love to my parents. I wish to thank all my families.

Finally, this thesis is dedicated to the memory of my grandparents. I still see their benign smiles in my sweet dreams.

Table of Contents

Abstract

Partial discharge (PD) is an important inducement of power failures such as insulation degradation. Effective, timely, and economical detection of PD is the basis for maintaining power system stability. This thesis presents novel resolutions for PD detection of medium voltage (MV) overhead power lines. The research is based on the large PD-related data set shared by Technical University of Ostrava (VSB). The data-driven machine learning-based approaches are employed for the complicated, nonlinear PD detection tasks. The first area of interest in this thesis is to propose a long short-term memory (LSTM) neural network based classifier with optimal feature extraction schemes for PD detection. It establishes a loop optimization process to greatly coordinate the data processing and the data-driven based machine learning algorithm. The combination of the different methods has combined advantages and can overcome each individual's drawback. The second area of interest in this thesis is to propose a novel transformer-based multilevel filtering framework for PD detection. The proposed framework demonstrates superior performance with high robustness and less manual intervention.

List of Figures

List of Tables

Abbreviations and Symbols

Abbreviations

AI	artificial intelligence
CC	covered conductors
CD	capacitive divider
CNN	convolutional neural network
CTN	channel transformer network
CWT	continuous wavelet transform
DAQ	data acquisition
DB1	Daubechies 1
DB4	Daubechies 4
DSI	discrete spectral interference
DWPT	discrete wavelet packet transform
DWT	discrete wavelet transform
ENET	Energy and Environmental Technologies
FBG	Fiber Bragg Grating
FC	fuzzy classifier
FNN	feed-forward neural network
GIS	gas-insulated switch
IEC	International Electrotechnical Commission
LSTM	long short-term memory
LTE	long-term evolution

MMRVM	multi-kernel multi-class relevance vector machine
MV	Medium voltage
NLP	natural language processing
PD	Partial discharge
PE	positional encoding
RNN	recurrent neural network
RPI	random pulse interference
RTU	remote terminal unit
SITS	satellite image time series
SLI	single layer inductor
STL	seasonal and trend decomposition using Loess
Sym8	Symlets 8
UV	ultraviolet
VSB	Technical University of Ostrava
WT	wavelet transform
XLPE	cross-linked polyethylene

Symbols

Acc	Ratio value of the sum of TP_{PD} and TP_{NPD} to the sum of T_{PD} and T_{NPD}
Ave_{RM}	Average value of RM_{PD} and RM_{NPD}
F_{NPD}	Harmonic mean of Pr_{NPD} and R_{NPD}
F_{PD}	Harmonic mean of Pr_{PD} and R_{PD}
NPM_{NPD}	Average number of the possible PD measurements contained in one NonPD Target
NPM_{PD}	Average number of the possible PD measurements contained in one PD Target
P_{NPD}	Number of the NonPD Prediction
P_{PD}	Number of the PD Prediction
PM_{NPD}	Nmber of the possible PD measurement segments belonging to NonPD Targets
PM_{PD}	Nmber of the possible PD measurement segments belonging to PD Targets
PPS_{NPD}	Ratio value of PS_{NPD} and T_{NPD}
PPS_{PD}	Ratio value of PS_{PD} and T_{PD}
Pr_{NPD}	Ratio value of TP_{NPD} and P_{NPD}
Pr_{PD}	Ratio value of TP_{PD} and P_{PD}
PS_{NPD}	Number of the NonPD Targets that contain more than 100 possible PD measurements
PS_{PD}	Number of the PD Targets that contain more than 100 possible PD measurements
R_{NPD}	Ratio value of TP_{NPD} and T_{NPD}
R_{PD}	Ratio value of TP_{PD} and T_{PD}
RM_{NPD}	Ratio value of TPM_{NPD} and PM_{NPD}
RM_{PD}	Ratio value of TPM_{PD} and PM_{PD}
T_{NPD}	Number of the NonPD Target

T_{PD}	Number of the PD Target
TH_A	Threshold of the transformer-based classifier
TH_B	Threshold of the transformer-based multilevel filtering framework
TP_{NPD}	Number of the NonPD Targets that are correctly classified as the NonPD Prediction
TP_{PD}	Number of the PD Targets that are correctly classified as the PD Prediction
TPM_{NPD}	Number of the possible PD measurement segments correctly classified as the predicted NonPD segments
TPM_{PD}	Number of the possible PD measurement segments correctly classified as the predicted PD segments
Tr	Training time of the LSTM neural network classifier

Chapter 1

Introduction

1.1 Motivation

Partial discharge (PD) is a localized dielectric breakdown without full connection of the electrodes [1]. PD signals are known as one of the most conclusive indicators of the deterioration of the electric insulation [2]. An online PD monitoring system is a valuable tool to detect abnormal events and anomalies in operation and set up preventive maintenance plans before the failure of electrical components. Online PD monitoring systems provide the initial measurement data for the inspection and identification of typical failure modes of electrical equipment.

Medium voltage (MV) overhead power lines can be hundreds of kilometers long. They are tasked with supplying electricity to the cities. The MV overhead power lines are usually equipped with covered conductors (CC) that consist of the aluminum core with the cross-linked polyethylene (XLPE) insulation on the surface. The XLPE insulation of CC prevents inter-phase short circuits when faults occur on the MV overhead power lines [3]. However, the PD usually happens after the occurrence of the faults. The PD is an indication of dielectric breakdown and also a cause of insulation degradation for the MV overhead power lines [4]. The economic and effective long-term online PD detection resolution is an important part

of the insulation monitoring and maintenance of the MV overhead power lines.

Efficient and effective PD monitoring and diagnosis depend on the performance of the sensors, measurement techniques, and data interpretation techniques. In conventional methods, the data interpretation usually depends on statistical methods, expert experience, specific calculation formulas, or thresholds. Such data interpretation modalities are usually specific to the fixed electrical model and highly dependent of the artificial choice. Unintelligent data interpretation costs a lot of manpower and financial resources. And its accuracy and effectiveness cannot be effectively guaranteed for complex and dynamic power systems.

With the development of artificial intelligence (AI) technologies, machine learning-based data interpretation methods are applied to wide application fields. There have been many research studies on machine learning-based PD detection methods. However, there is still a lack of machine learning-based resolutions that balance the artificial intelligent level of the data interpretation and the practical engineering application requirements and limitations. For the real-world online monitoring task for the power system, e.g., the MV overhead power lines, the PD detection resolutions are more likely required to interpret the dynamic large-scale online measurement data disturbed by complex background noises. It is a meaningful and urgent research topic with wide application prospects requesting optimal resolutions for the conflict of the lack of the ground truth and the performance of the machine learning algorithm; the conflict of the manual intervention and the robustness and adaptability of the PD detection method; and the conflict of the large-scale measurement data and the limited computation ability.

1.2 Objectives

The main objective of this thesis is to propose effective and novel machine learning-based resolutions for PD detection for MV overhead power lines. The resolutions

are on the basis of the PD measurements shared by Smart Grid Laboratory, Centre for Energy and Environmental Technologies (ENET) of Technical University of Ostrava (VSB). The PD measurements are online voltage signals obtained via cost-effective long-term data acquisition (DAQ) platforms from 23 MV overhead power lines in the Czech Republic. The machine learning-based resolutions aim to efficiently interpret the voltage signals and correctly detect whether there are any PD activities occurring in the detected voltage signals. Specifically,

1. The PD measurements are large-scale online data with millions of measurements. The first objective is to propose PD detection resolutions with the ability to effectively and efficiently process large-scale online data with limited hardware and computation requirements.

2. The second objective is to improve the noise reduction scheme of the PD detection resolutions to reduce the interference of the complex background noises in the online PD measurements of the real-world MV overhead power lines.

3. The third objective is to improve the robustness and adaptability of the PD detection resolutions and reduce the manual intervention.

To obtain optimal PD detection resolutions for the MV overhead power lines, different machine learning algorithms and models are adopted for systematic analysis and comparison. Extensive experiments are conducted to improve the resolutions of the PD measurements interpretation including noise reduction, feature extraction, and data classification. The final objective is to propose intelligent PD detection resolutions with superior technical and economic performance, that can be widely extended to practical engineering applications.

1.3 Major Contribution of the Thesis

The major contributions of this thesis are:

1. This thesis systematically reviews the history and development status of the PD detection. The PD measurement modalities and machine learning-based data interpretation methods are reviewed. The PD-related data set that is analyzed in this thesis is introduced in detail. The construction of the DAQ platform and the PD measurement procedure for the MV overhead power lines are presented.

2. A novel long short-term memory (LSTM) based classifier with optimal feature extraction schemes for the PD detection of the MV overhead power lines is proposed in this thesis. Its effectiveness and practicality are verified by the large PD-related data set when compared with a number of different signal processing methods and machine learning models.

3. The influence of the data processing scheme such as the noise reduction and feature extraction to the PD detection resolutions is systematically analyzed via extensive experiments. The optimal data processing scheme for the MV overhead power lines is determined. The loop optimization rules for PD detection solutions have been summarized based on extensive experiments and analysis.

4. A novel transformer-based multilevel filtering framework for the PD detection of the MV overhead power lines is proposed in this thesis. Extensive experiments are conducted to verify its effectiveness, robustness, and flexibility. It outperforms the previous LSTM-based classifier and other techniques in the literature based on the experimental results.

1.4 Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 firstly introduces the definition of the PD and its models. Various PD measurement methods are reviewed. Then the PD-related data set shared

by VSB is presented. The PD measurements for MV overhead power lines are described. The description of the DAQ platform is also presented. Then the machine learning-based data interpretation methods are reviewed. A review of the history and applications of the LSTM neural network and the transformer algorithm are presented in detail.

Chapter 3 presents a novel and effective LSTM-based PD detection resolution for the MV overhead power lines. It includes an optimal signal processing scheme and an LSTM neural network classifier. The proposed PD detection resolution is verified by the VSB PD-related data set. It is optimized by systematic comparison and analysis of the experimental results. A variety of feature extraction methods, noise reduction methods, and AI algorithms are analyzed and compared. The loop optimization rules for PD detection solutions have been summarized based on extensive experiments and analysis.

Chapter 4 proposes a novel transformer-based multilevel filtering framework that is competent for intelligent PD detection for large-scale online data of complex power systems. It combines the advanced and intelligent computation ability of the transformer algorithm and the multilevel filtering mechanism. They make the whole framework more applicable and robust to complicated and dynamic power systems and various practical engineering applications. The transformer-based multilevel filtering framework is also verified by the VSB PD-related data set. It provides an economic and optimal resolution for PD monitoring of MV overhead power lines with superior accuracy, high availability, and less manual intervention. Extensive experiments demonstrate its high robustness and great efficiency for PD detection.

Chapter 5 concludes the whole thesis. The future works of the machine learning applications for PD detection and monitoring are introduced.

Chapter 2

Literature Review

2.1 Partial Discharge

2.1.1 Partial Discharge Model

The PD is defined by the International Electrotechnical Commission (IEC) 60270 as a kind of localized electrical discharge that only partially bridges the insulation between conductors and can or cannot occur adjacent to a conductor [5]. PD is usually the result of localized electrical stress concentrations in an insulator or on the surface of an insulator. Typically, the partial discharges appear as pulses of much less than 1-microsecond duration. The history of PD research can be traced back to the 1940s [6]. The IEC adopted the nomenclature "partial discharges" in 1967 [7]. So far, numerous researches on PD have been carried out, including physical model research, measurement, pattern recognition, detection, and prevention.

PD occurs when the electric field in the specific spot is higher than the inception field and an initial electron is present [8]. There are three types of the PD: corona discharge, surface discharge, and internal discharge [9]. Various electric equipment from gas insulation to liquid and solid insulation suffers from the PD [10–12]. The

PD is detrimental to the insulation with a cumulative effect. In this thesis, the PD along MV overhead power lines is the focus of the research.

MV overhead power lines can be hundreds of kilometers long. They are tasked with supplying electricity to the cities. The CC is widely used in MV overhead power lines due to their streamlined construction space, optimal operational reliability, and robust protection [13]. The surface discharge and internal discharge are directly related to the CC of the MV overhead power lines. Surface discharge occurs along the surface of a dielectric material, especially the interface between the different dielectric layers of the CC [14]. Internal discharge happens within dielectrics with low strength. When such discharges occur, the insulation of CC prevents inter-phases short circuits. As the value of the abnormal leakage current is very small, the standard digital relay protection is unable to react with such a low current [15]. However, the PD slowly damages the power line, and even leads to the outage of the power network [16].

Various models have been proposed for the different PD pulses, such as the three-capacitance model (abc-model), induced charge theory, particle in cell (PIC) method, fluid model, and etc [8, 17, 18]. The PD pulse waveform can be directly used for PD detection of microseconds duration. While for the PD detection of milliseconds duration, there is a greater demand for balancing detection costs, noise environment interference, and detection accuracy.

2.1.2 Partial Discharge Measurement

The PD measurement is based on the energy exchange that takes place at the localized part of the insulation defect. The energy exchange can be in the form of electromagnetic radiation, sound, electromagnetic impulses, thermal radiation, and chemical formation [19–22]. These energy exchanges indicate a possible PD activity. In [19], the authors design the fluorescent optical fiber sensor for the PD measurement of the gas-insulated switch (GIS) and transformer. The fluorescent

optical fiber, as a probe, is connected with the transmission optical fiber through a coupler. The photomultiplier tube converts the optical signal into an electrical signal for oscilloscope display. In [23], the authors employ the Fiber Bragg Grating (FBG) sensor to detect free metal particle discharge and suspended metal electrode discharge in oil. It detects the equipment vibration caused by PD-stimulated ultrasonic waves. In [24], the authors employ an acoustic emission array sensor for the PD detection and direction-finding research. In [25], the authors design the optimal size parameters of the Rogowski coil for PD and current measurement.

Smart Grid Laboratory of ENET Centre of VSB shares a large PD-related data set of MV overhead power lines on the internet. This thesis analyzes this VSB PD-related data set to propose novel and effective PD detection resolutions for the MV overhead power lines. The PD measurement is conducted by the DAQ platform cheaper than the Rogowski coil to measure the high-frequency voltage signals from the MV overhead power lines [15, 26, 27]. The DAQ platforms are set up on 23 MV overhead power lines in the Czech Republic. Fig. 2.1 shows the diagram of the DAQ platform. The DAQ platform consists of several parts: the single layer inductor (SLI) sensor, high voltage capacitive divider (CD), DAQ oscilloscope card, and control unit.

The SLI sensor is used for MV electric field strength measurement. It is a coil wound tightly on the surface of the CC. Fig. 2.2a shows the SLI sensor during installation. The number of turns of wire in an SLI sensor varies depending on the outside diameter of the CC. The standard configuration is 75 coils of wire. The SLI sensor is covered by a layer of semi-conductive insulation, as shown in Fig. 2.2b. The semi-conductive insulation layer provides necessary ultraviolet (UV) protection and seals the outdoor sensor. The inner surface of the SLI sensor creates capacitance coupling to the surface of the CC. The output of the SLI sensor is connected to the high voltage CD to lower the voltage. Fig. 2.3 presents the on-location photo of the high voltage CD.

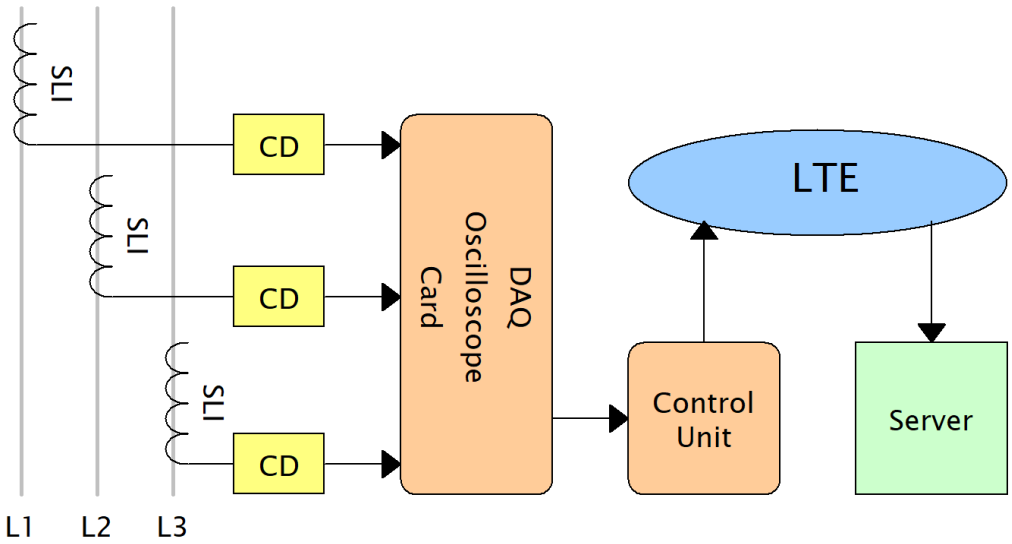


Figure 2.1: Diagram of the DAQ platform.

The output of the high voltage CD is connected to the DAQ platform. The whole DAQ platform, as shown in Fig. 2.4, is based on the modular remote terminal unit (RTU) developed by the ELVAC company [15]. The RTU units are used widely in the electrical distribution network of the Czech Republic for control and measurement. This module was designed especially for PD detection. Unlike the others, it has low impedance inputs (75 Ohms) with simultaneous sampling. The input impedance is combined with the high voltage CD to form a high pass filter which suppresses the low-frequency components of the acquired signals, such as the 50-Hz power grid carrier. Such a low impedance input is also more resistant to interference than the high impedance inputs.

All signals of the three phases of the MV overhead power line, i.e., L1, L2, and L3 in Fig. 2.1, are acquired simultaneously. Therefore, there are three signals acquired during each measurement. Acquired data are transmitted to the server via the long-term evolution (LTE) network. The acquired data represent the high-frequency component of the line-to-ground voltage of the MV overhead power lines. The data are acquired once an hour with a 40 MS/s sample rate. Each acquired signal, hereinafter called the voltage signal, consists of 800,000 measurements in the period of 20 ms, i.e., one period of the power grid cycle. The examples of the

(a) An SLI sensor (red) during installation.

(b) An SLI sensor with semi-conductive insulation.

Figure 2.2: The SLI sensors shot on location.

Figure 2.3: A high voltage CD shot on location.

Figure 2.4: A DAQ platform shot on location.

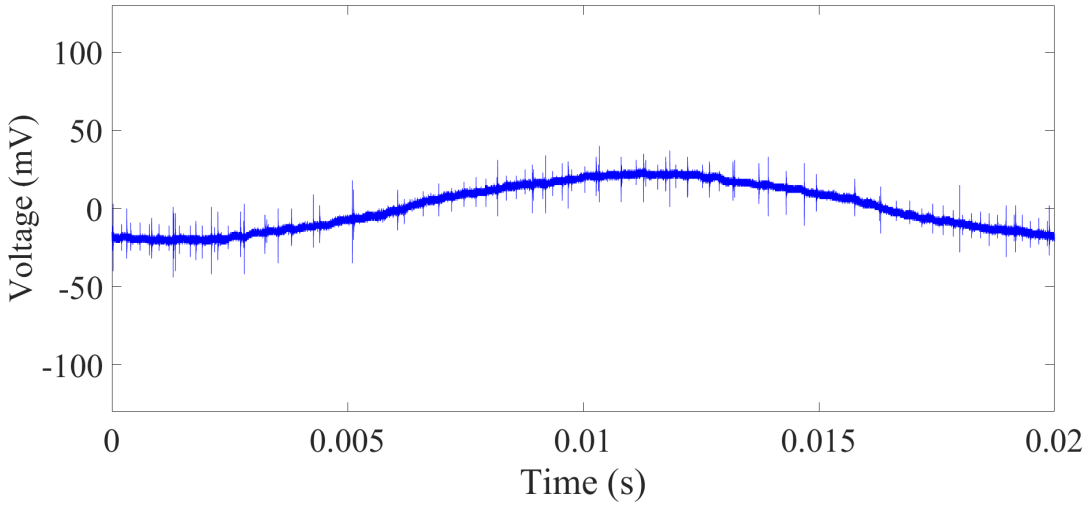
voltage signal are shown in Fig. 2.5.

There are 29,049 voltage signals in the VSB PD-related data set. The experts mark 2,104 of them as PD and the other 26,945 signals as NonPD. The PD and NonPD labels are distinguished according to the impulse component generated by PD activities. The relative amplitude and phase of the impulse components are also the focus of this thesis. Meanwhile, all the voltage signals contain a high amount of background noises. The four main sources of background noises are discrete spectral interference (DSI) caused by radio emissions; random pulse interference (RPI) caused by switching operations, lightning, and corona; repetitive pulses interference due to power electronics; and ambient and amplifier noise.

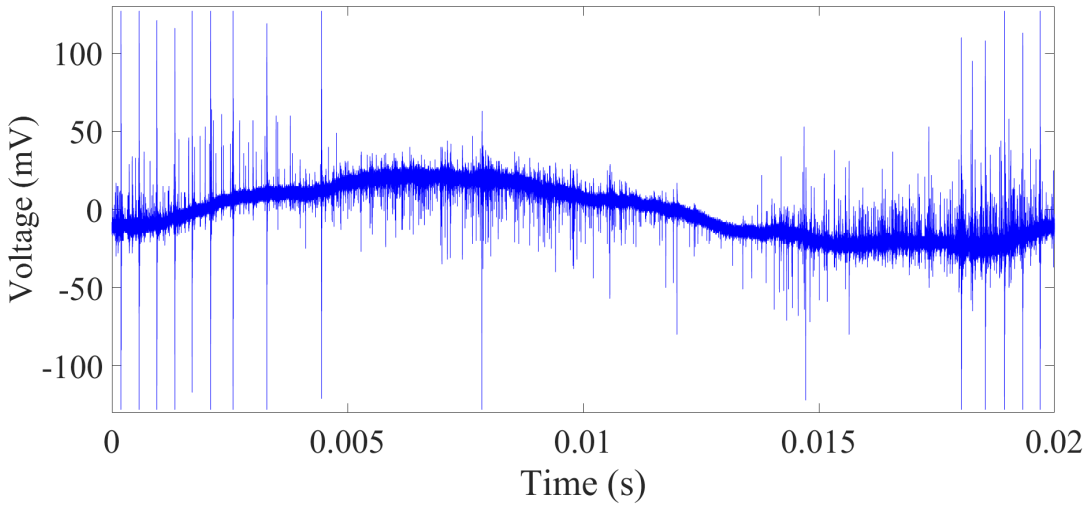
2.2 Machine Learning-Based Data Interpretation

The data can be interpreted according to model-based approaches or data-driven approaches [28]. The former is based on the understanding of the physical mechanisms within the system and their accurate modeling. The latter uses empirical operational data without references to accurate model representations. With the development of AI technology, machines are provided with more advanced capabilities to use data and knowledge [29, 30].

Machine learning approaches have been widely applied to PD detection [31, 32]. The authors in [33] use a modified 10-fold cross-validation procedure with a convolutional neural network (CNN)-based classifier to detect PD. In [34], the authors combine the Cepstrum analysis with a multi-layer feed-forward neural network (FNN) to recognize the PD types of different sources simultaneously under noisy conditions. Reference [35] proposes a novel method on the basis of the multi-kernel multi-class relevance vector machine (MMRVM) for fault detection. The authors in [36] introduce the fuzzy techniques to PD classification. They analyze the signal by predicting the conditional probability of its measurement vector be-



(a) A voltage signal labelled by the experts as NonPD.



(b) A voltage signal labelled by the experts as PD.

Figure 2.5: The examples of the detected signals.

longing to a particular PD class. In [37], the authors propose a PD classification method using a fuzzy classifier (FC).

2.2.1 LSTM Neural Network

The LSTM neural network was first introduced in 1997 [38]. It is a kind of recurrent neural network (RNN). The RNN is named after its typical cyclic connection architecture that enables RNN to update the current state in accordance with the past state and current input data [39]. The LSTM neural network is improved

from the RNN with the ability to handle the vanishing and exploding gradient problems [40].

The LSTM neural network is suitable for processing and predicting important events with very long intervals and delays in time series. It is widely applied to the fields such as natural language processing (NLP) [41, 42], transportation system [43], financial market [44], energy consumption [45,46], and medical science [47, 48] for the classification, prediction, and forecast tasks. There is also much work for the LSTM applications for PD detection. The authors in [49] propose a PD detection method that combines discrete wavelet transform (DWT) and the LSTM neural network. The DWT decomposition features are extracted to train the many-to-one LSTM model. Reference [50] introduces the seasonal and trend decomposition using Loess (STL) as a preprocessing method for the LSTM-based classifier.

2.2.2 Transformer

Recently, a machine learning algorithm named transformer has set off a wave of research and applications. The transformer model is first described in [51] for addressing machine translation. The classic way to do machine translation in a neural network is encoder/decoder. The dominant sequence transduction models are based on the complex RNN or CNN. The best-performing models also connect the encoder and decoder through an attention mechanism. The new simple network architecture, the transformer, is proposed based solely on attention mechanisms, dispensing with recurrence and convolutions entirely [52].

The transformer-series algorithms are known for achieving fully parallel computing. The transformer mechanism has derived lots of excellent models and algorithms suitable for various fields, such as BERT [53], Electra [54], GPT-2 [55], Roberta [56], ViT [57], DETR [58], and wav2vec 2.0 [59]. They develop fast to perform tasks on different fields such as the NLP, computer vision, and digital

media [60–62].

Besides, the transformer mechanism is applied to various practical applications. Reference [63] proposes a CNN-transformer approach to perform the crop classification. The multi-layer encoder from transformer architecture is used for pattern recognition of multi-temporal sequences. Reference [64] uses the transformer model for image captioning. The transformer model not only establishes a relationship between the natural language and the image but also is able to focus on the most contributing image regions. Reference [65] proposes a novel parameter-free method named channel transformer network (CTN) to decrease or increase channels for transformer modules in CNN whilst keeping most information with lower computation complexity. Reference [66] proposes a traffic sign recognition system using a deep neural network that comprises convolutional and spatial transformer layers. Reference [67] applies BERT to the hyperspectral image classification. Reference [68] uses the pretrained and fine-tuned BERT model to deal with the satellite image time series (SITS) classification.

Chapter 3

LSTM-Based Resolution with Optimal Signal Processing Schemes for Partial Discharge Detection

3.1 Introduction

MV overhead power lines are tasked with supplying electricity to the cities. The PD detection for MV overhead power lines is an important part of the insulation protection, power system monitoring, and power lines maintenance. As introduced in Chapter 2, VSB shares a large PD-related data set of MV overhead power lines. The data set consists of PD-related voltage signals that are already marked with a PD or NonPD label by experts by evaluating the impulse component of the voltage signals [15, 27]. Such manual detection is extremely labor-intensive and time-consuming. An automatic PD detection method for the voltage signals leads to a continuous and economical monitoring method to detect faults from MV overhead power lines. The machine learning models create more possibilities

for various automatic fault detection and pattern recognition tasks. They are combined with data processing such as noise reduction and feature extraction to perform the various PD detection tasks.

As for the interpretation of microsecond-level data, the data volume is usually less than several thousand. Such lengths are suitable for the AI algorithms to display their intelligence functions [69,70]. While for the measurement data at the level of milliseconds, the number of measurements in one signal can be a hundred million. Such a large scale of data cannot be analyzed directly due to the limitation of computation. Moreover, the PD detection of large-scale data confronts more challenges of noise corruption and irrelevant information interference. The noise reduction and data refining are more important for such PD detection tasks [71,72].

In the literature, various noise reduction methods, features, and AI techniques are employed to different PD detection tasks [33, 49, 50, 71, 72]. However, the procedure of noise reduction, feature extraction, and AI processing are selected and optimized independently. There is a lack of the internal coordination and linkage optimization. And there are often deficiencies in the analysis for the rules of determining scheme details, which are also very important for practical engineering applications.

This thesis dedicates to the PD detection method that coordinates the data processing and data-driven based machine learning algorithm with internal linkages. It aims to explore an optimal resolution for the automatic online PD detection of MV overhead power lines. An LSTM-based classifier with optimal feature extraction schemes is proposed based on large real trial data measured from the MV overhead power lines. This proposed resolution employs the loop optimization to greatly coordinate the noise reduction, feature extraction, and the LSTM neural network as a whole. The individual components in the whole are selected based on the most essential characteristics of PD. And they have great practicality with low requirements for computation and hardware. The proposed LSTM-based res-

olution maximizes their effectiveness through the optimal combination to achieve a better solution for the PD detection, while other methods in the literature are in general single ones, and not dedicated designed for the resolution. This comprehensive resolution is with advantage than the literature ones. Moreover, this chapter proposes the loop optimization rules of the adaptive framework based on the systematic analysis. The loop optimization rules have great potential for the promotion of wider engineering applications.

The structure of the remaining parts of this chapter is as follows: Section 3.2 introduces the architecture and methodologies of the proposed LSTM-based resolution. Section 3.3 demonstrates its novelty. The results of more than 200 different models have been analyzed. The details of parameters are determined. The common loop optimization rules are also established. Section 3.4 summarizes the chapter.

3.2 Architecture and Methodologies

The overall flowchart of the proposed LSTM-based resolution is shown in Fig. 3.1. To effectively process such large-scale voltage signals, series of noise reductions are conducted to reduce the complicated background noises in the complex real environment. Multiple sequence features are employed to capture the PD characteristics that are closely related to phase-dependent variation rather than independent values. The LSTM neural network classifier classifies the input features into the PD or NonPD class. Being labelled as PD means that the voltage signal contains the impulse components of PD. Being labelled as NonPD means that there is no PD activity. The proposed LSTM-based resolution aims to achieve the effective distinction of PD components from other ambient background noises through the comprehensive contribution of the optimal noise reduction methods, feature extraction methods, and the LSTM neural network classifier. This section introduces the methodologies for each step.

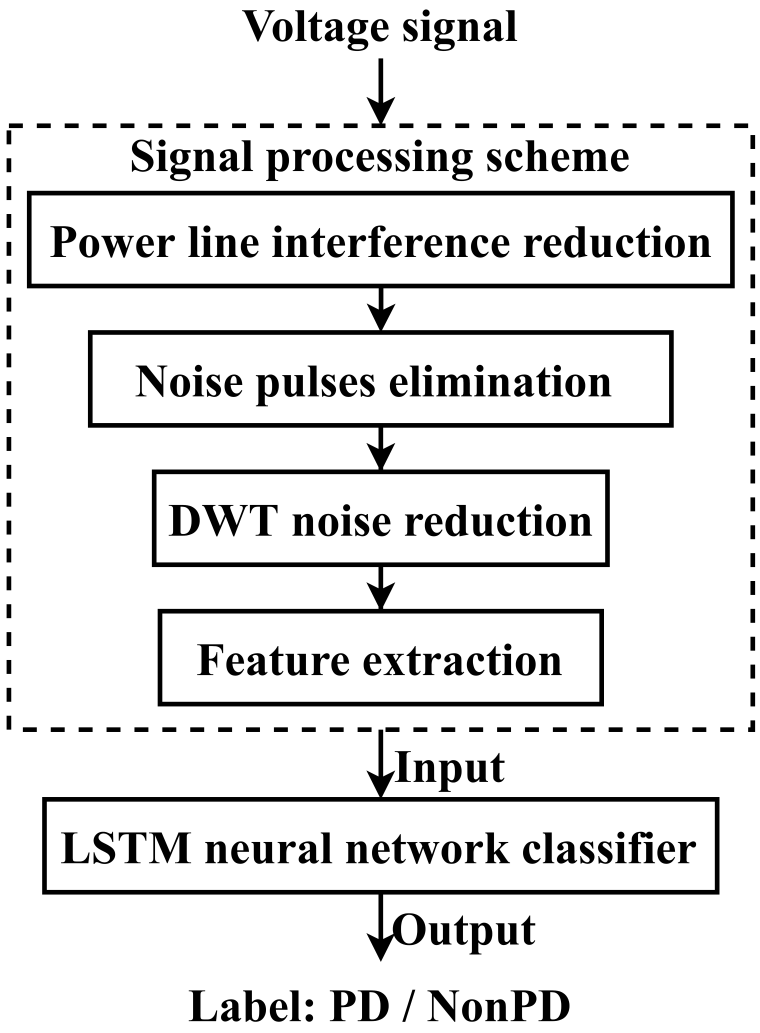


Figure 3.1: The flowchart of the proposed LSTM-based resolution for PD detection.

3.2.1 Noise Reduction

Firstly, the digital trap filter is used to reduce the power line interference. The transfer function of the digital trap filter is as follows:

$$H(z) = \frac{z^2 - 2z + 1}{z^2 - 1.998z + 0.998} \tag{3.1}$$

As an example, Fig. 3.2 shows a voltage signal before and after the digital trap filter. Fig. 3.2a is a voltage signal labelled by experts as PD. Fig. 3.2b shows the outcome of the digital trap filter. It can be seen from the comparison of Fig. 3.2a and Fig. 3.2b that the digital trap filter (3.1) effectively reduces the 50-Hz power

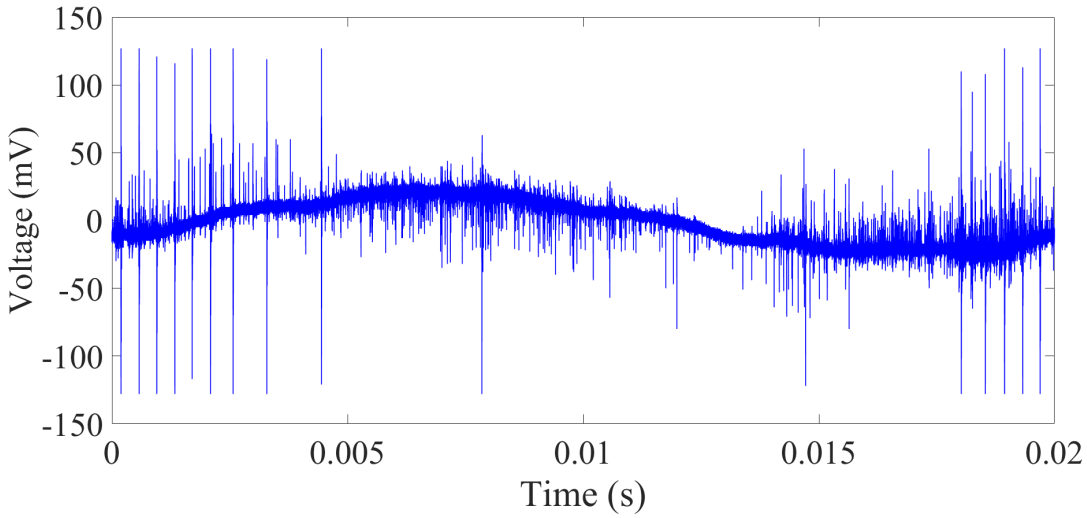
line interference. The pulse components of PD and other noises are more obvious and intuitive.

Secondly, the random pulse interference (RPI) is reduced by eliminating its high amplitude pulse components. The VSB research team describes one of the most significant background noises of the PD-related data as the false hit peaks caused by RPI (annotated in Fig. 3.2b). The false hit peaks usually have higher amplitude than PD pulses. They are often shown as symmetrical pulses with damped oscillations [15]. And the noise is more likely to affect negative measurements. According to the knowledge, this thesis focuses on the measurements that have abnormally large negative values, or abnormally large negative derivatives. Local signals near such measurements are considered to have been disturbed by noise pulses. They will be removed from the voltage signals.

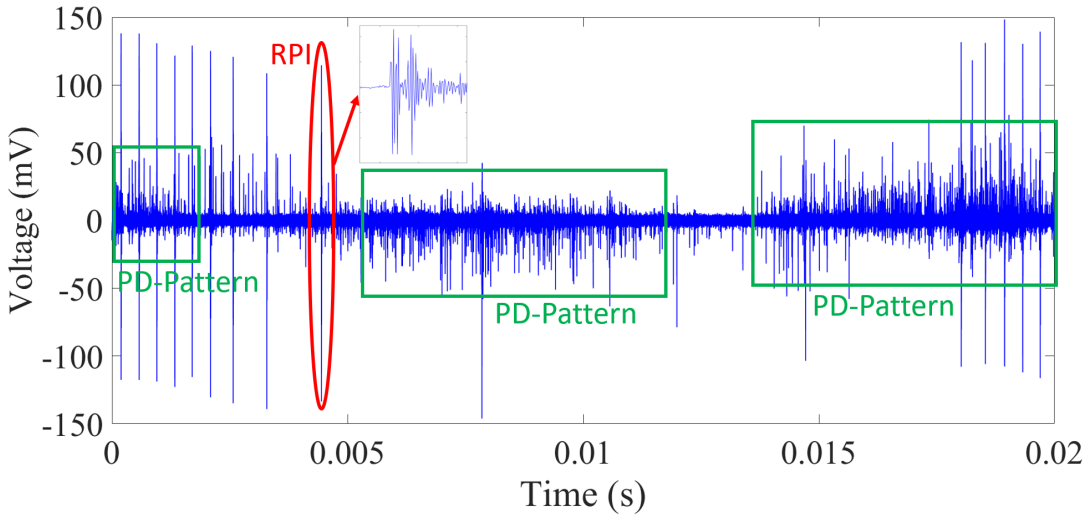
In addition to the aforementioned targeted noise pulse elimination, the wavelet transform (WT) noise reduction is carried out on the voltage signal. Wavelets are a systematic way to represent functions on unbounded domains by linear combinations of orthogonal basis functions that are compactly supported and overlapped [73]. The WT is known to have excellent energy compaction characteristics [74]. And it can provide perfect reconstruction. It is commonly known to have the following three categories: discrete wavelet transform (DWT); discrete wavelet packet transform (DWPT); and continuous wavelet transform (CWT). The DWT is a computationally efficient WT. It introduces the repetitive application of low-pass and high-pass filters to calculate the wavelet expansion of a given sequence of discrete numbers [75]. An orthonormal compactly supported wavelet basis of $L^2(R)$ is formed by the dilation and translation of the single real-valued function $\psi(t)$, called the mother wavelet given by

$$\psi_{j,k}(t) = 2^{-j/2}\psi\frac{t - k2^j}{2^j} \tag{3.2}$$

where j and k belong to the set of integers. In (3.2), the function ψ satisfies the



(a) An example of a voltage signal labelled by experts as PD.



(b) The signal shown in Fig. 3.2a after the digital trap filter. The positions of the RPI and PD patterns are annotated with red circles and green boxes, respectively.

Figure 3.2: The example of a voltage signal (PD) before and after the digital trap filter.

following two-scale difference equation,

$$\psi_{j,k}(t) = \sqrt{2}\sum_{k=0}^{L-1} h_k.\varphi(2t-k) \tag{3.3}$$

where $\varphi(t)$ is a companion function of the wavelet function called the scaling function.

The WT computation requires a pair of filters. The high-pass filter calculates the wavelet coefficients, whereas the low-pass filter applies the scaling function. Recursively, the low-pass outputs are used to compute the wavelet coefficients at the next level of the resolution [76,77]. With the DWT, a signal $y(t)$ is decomposed as a series of approximation and detail coefficients. As a reverse process, the signal $y(t)$ can be reconstructed with the scaling function $\varphi(t)$ and the wavelet function $\psi(t)$ as

$$y(t) = \sum_{i=1}^{N/2^j} a_j(i)\varphi_{j,i}(t) + \sum_{j=1}^{J}\sum_{i=1}^{N/2^j} d_j(i)\psi_{j,i}(t) \tag{3.4}$$

where J is the series of wavelet decomposition; N is the total number of coefficients of wavelet decomposition; $a_j(i)$ is the approximate coefficient; and $d_j(i)$ is the detailed coefficient. $a_j(i)$ and $d_j(i)$ are expressed as follows:

$$a_j(i) = \langle y(t), \varphi_{j,i}(t)\rangle \tag{3.5}$$

$$d_j(i) = \langle y(t), \psi_{j,i}(t)\rangle \tag{3.6}$$

The proposed resolution adopts the DWT noise reduction. The noise is reduced by applying the soft threshold and heuristic variant of Stein's unbiased risk threshold selection rules to the wavelet coefficients. And the denoised signal is reconstructed with manipulated detail signals.

Fig. 3.3 shows the voltage signal in Fig. 3.2b after the DWT noise reduction. As an example, the 11-level DWT noise reduction with Symlets 8 (Sym8) wavelet is performed on the signal. Fig. 3.3a and Fig. 3.3b compare the DWT noise reduction

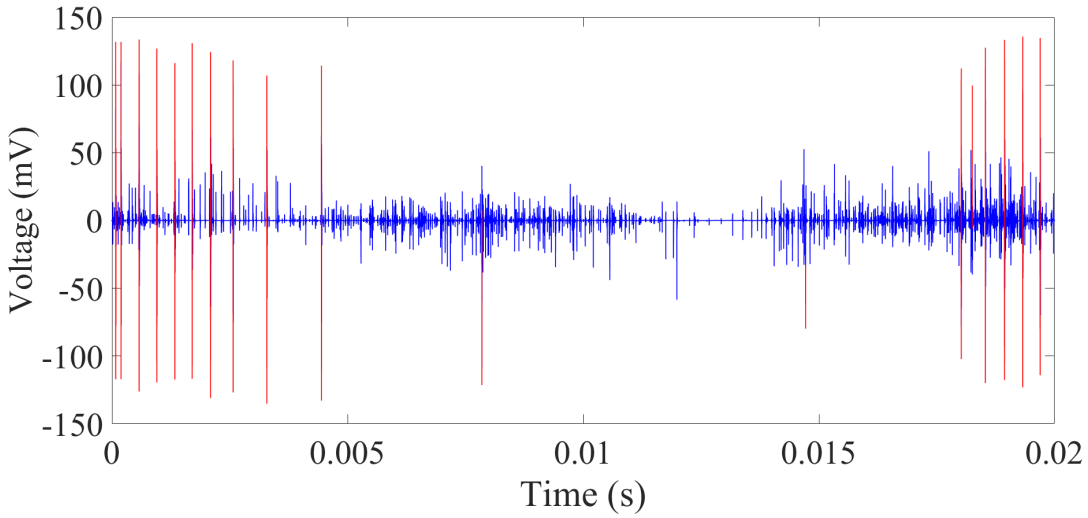
performance with and without the aforementioned RPI noise pulse elimination. It can be seen that the DWT noise reduction cannot eliminate the RPI pulses with very high amplitude, which is highly likely to interfere with PD detection. The RPI noise pulse elimination is proved effective.

In summary, Fig. 3.4 shows the voltage signals before and after all three noise reduction processes. Specifically, Fig. 3.4a presents a NonPD signal before and after the noise reduction processes. Fig. 3.4b presents a PD signal before and after the noise reduction processes. It can be seen that most irrelevant information is filtered out, which also leads to the reduction of signal amplitude. The intensive pulses caused by PD are much clearer than before. The options of the DWT mother wavelets and the specific performances of the various noise reduction schemes will be analyzed and discussed in Section 3.3.

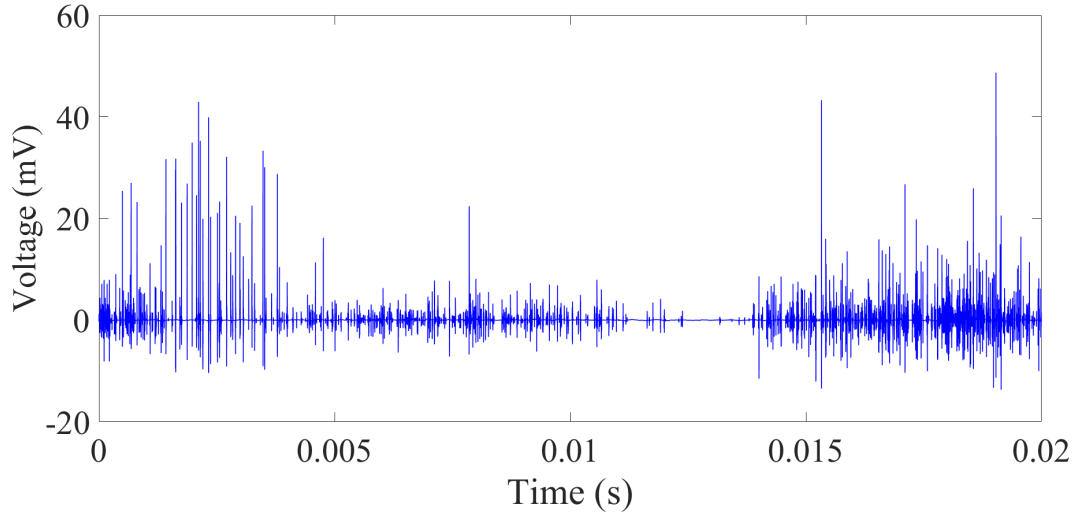
3.2.2 Feature Extraction

According to the limitation of the practical conditions, it is impossible to directly establish the mapping relationship between nearly millions of voltage measurements and the PD. The amount of the data must be reduced prior to the classification. Hence, it is necessary to extract features from the voltage signal. The feature can be a value or a sequence. This thesis adopts sequences as the features. Compared with a single-element feature, the sequence feature makes the variation and correlation between voltage amplitude and phase more involved in the analysis process of the PD detection, while reducing the degree of manual intervention of the proposed LSTM-based resolution.

The feature length, which means the length of each sequence, must be consistent. To construct a sequence feature with N elements, the voltage signal is divided into N segments. The length of each segment is $800,000/N$. One value is extracted from each segment. There are various extraction methods, such as taking the maximum value; taking the standard deviation; taking the average value; and so

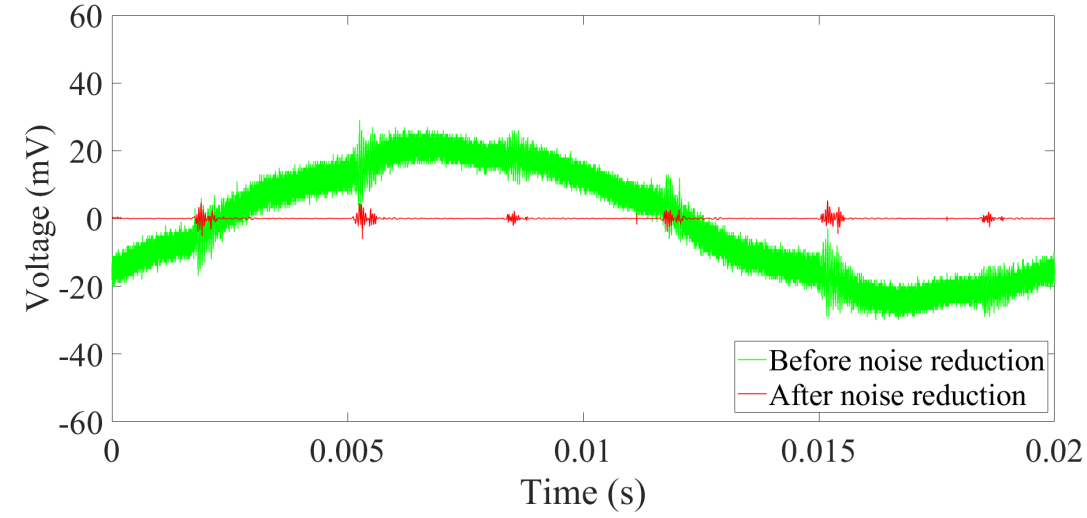


(a) The signal shown in Fig. 3.2b after the DWT noise reduction without the RPI noise pulse elimination. The RPI noise pulses are highlighted in red.

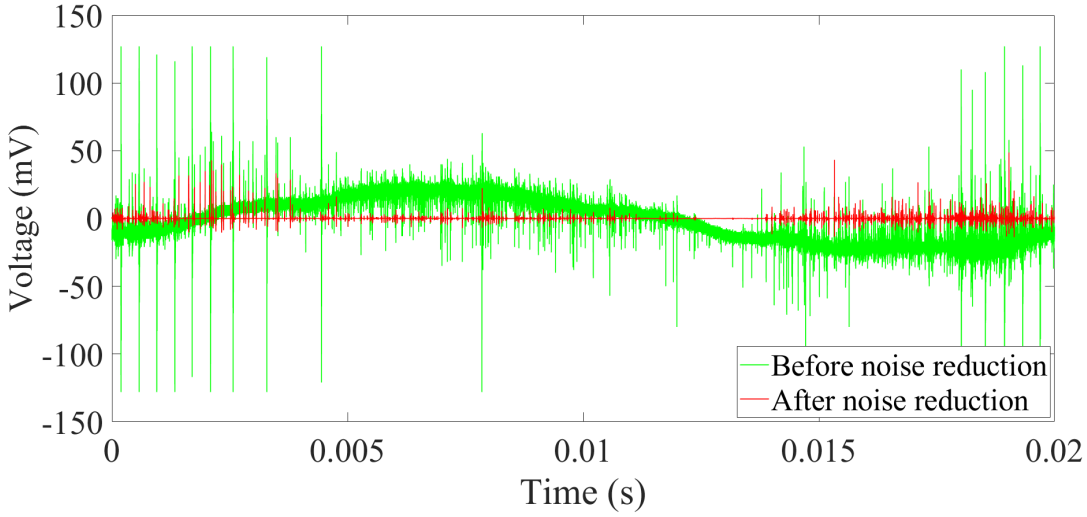


(b) The signal shown in Fig. 3.2b after the RPI noise pulse elimination and DWT noise reduction.

Figure 3.3: The examples of the signal shown in Fig. 3.2b after the DWT noise reduction with and without the RPI noise pulse elimination.



(a)



(b)

Figure 3.4: The examples of the voltage signals before and after all three noise reduction processes. (a) is labelled as NonPD while (b) as PD.

on. When a method is selected, all the N segments use this method to extract values. All the N values form a sequence with length of N as an N-element feature for the voltage signal. To reduce the confusion, the sequences are always rearranged in phase order of 0°- 360°. All the features in this thesis are extracted in this manner.

Thirteen different features have been analyzed. Feature 1 is the maximum value, i.e., the 1^{st} largest value. Feature 2 is the 3^{rd} largest value. Feature 3 is the 5^{th} largest value. Feature 4 is the 10^{th} largest value. Feature 5 is the 25^{th} largest value. By introducing Features 2 to 5, the effect of the individual large amplitude noise pulses is reduced. The numbers 3^{rd}, 5^{th}, 10^{th}, and 25^{th} are randomly selected. Although the PD pulses contain both positive and negative peaks, this thesis focuses on the positive extreme values, which are intended to reduce the interference of RPI pulses.

Besides the amplitude, PD also changes the data distribution of the voltage signal. The data distribution is also considered as a feature. The extraction method is explained as follows: In order to describe the distribution of data belonging to interval M, the number of measurements belonging to interval M is firstly counted. The ratio of this number to the total number of measurements is the desired feature. Feature 6 is the distribution of data less than 1 mV. Feature 7 is the distribution of data less than 0.5 mV. Feature 8 is the distribution of data belonging to interval of [-1 mV, 1 mV]. Feature 9 is the distribution of data belonging to interval of [-0.5 mV, 0.5 mV]. Feature 10 is the distribution of data less than 0.1 mV. Feature 11 is the distribution of data belonging to interval of [-0.1 mV, 0.1 mV]. The distribution of the large amplitude measurements fluctuates greatly due to different measuring environments. Therefore, the above small values are selected to reflect the commonness of PD. The specific values of these intervals are randomly selected.

Feature 12 is the average value. It reflects the energy of the voltage signal. Feature

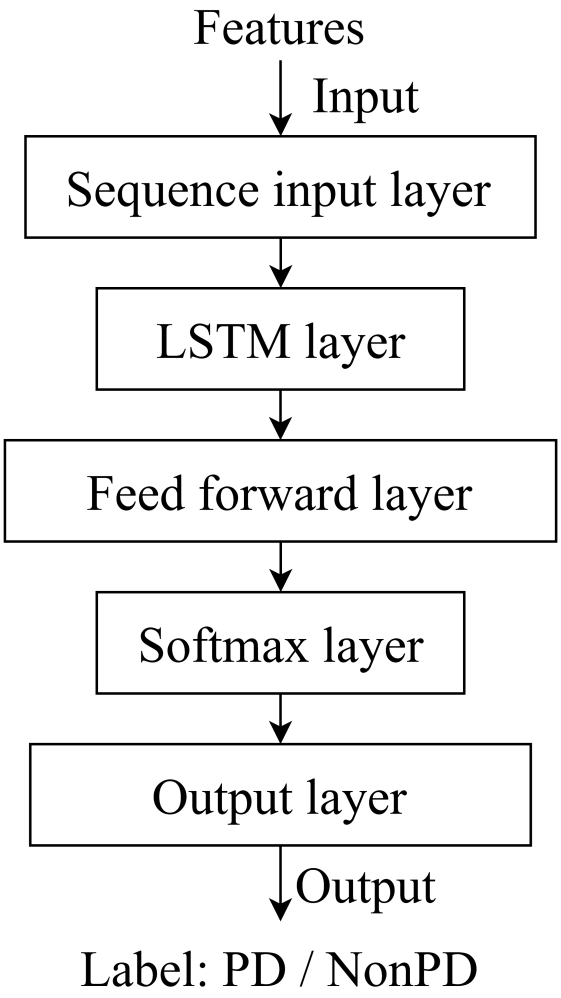


Figure 3.5: The architecture of the LSTM neural network classifier.

13 is the standard deviation value. It reflects the degree of the dispersion of the voltage signal.

3.2.3 LSTM Neural Network Classifier

The proposed LSTM-based resolution utilizes an LSTM neural network classifier to classify the features. The LSTM neural network classifier, as shown in Fig. 3.5, consists of a sequence input layer that receives the input features, an LSTM layer, a feed forward layer with fully connected FNN, a Softmax layer, and an output layer.

The LSTM neural network classifier uses the advantage of the optimal recurrence

mechanism of the LSTM neural network. The LSTM neural network is able to not only store information as an internal state, but also precisely control the time scale variables of the hidden state [38]. It is applicable to sequence analysis, such as the voltage signal classification. The LSTM cell, as shown in Fig. 3.6, introduces the forget gate, input gate, and output gate. Such a structure is able to flexibly control the influence of the previous information on the present state. The LSTM cell also introduces the addition operation between different internal state information. This could effectively reduce the gradient disappearance problem caused by the multiplicative effect of small gradients.

In Fig. 3.6, x_t is the input of the present time step, and h_{t-1} is the output of the previous time step. Specifically, in this thesis, the input of the LSTM neural network classifier is an $M \times N$ matrix, where M is the number of the sequence features, and N is the length of each sequence feature:

$$x_t \in \mathbb{R}^M, \qquad t \in 1, 2, ..., N \tag{3.7}$$

x_t and h_{t-1} are combined into the four internal variables: g, i, f, and k. Among them, g is squashed by the hyperbolic tangent activation function. Variable i of the input gate, f of the forget gate, and k of the output gate utilize the sigmoid activation function. The calculation formulas are as follows:

$$g = tanh(x_t W_x^g + h_{t-1} W_h^g + b^g) \tag{3.8}$$

$$i = \sigma(x_t W_x^i + h_{t-1} W_h^i + b^i) \tag{3.9}$$

$$f = \sigma(x_t W_x^f + h_{t-1} W_h^f + b^f) \tag{3.10}$$

$$k = \sigma(x_t W_x^k + h_{t-1} W_h^k + b^k) \tag{3.11}$$

where W_x^g, W_x^i, W_x^f, and W_x^k are the weights for the current input; W_h^g, W_h^i, W_h^f, and W_h^k are the weights for the previous output; and b^g, b^i, b^f, and b^k are the

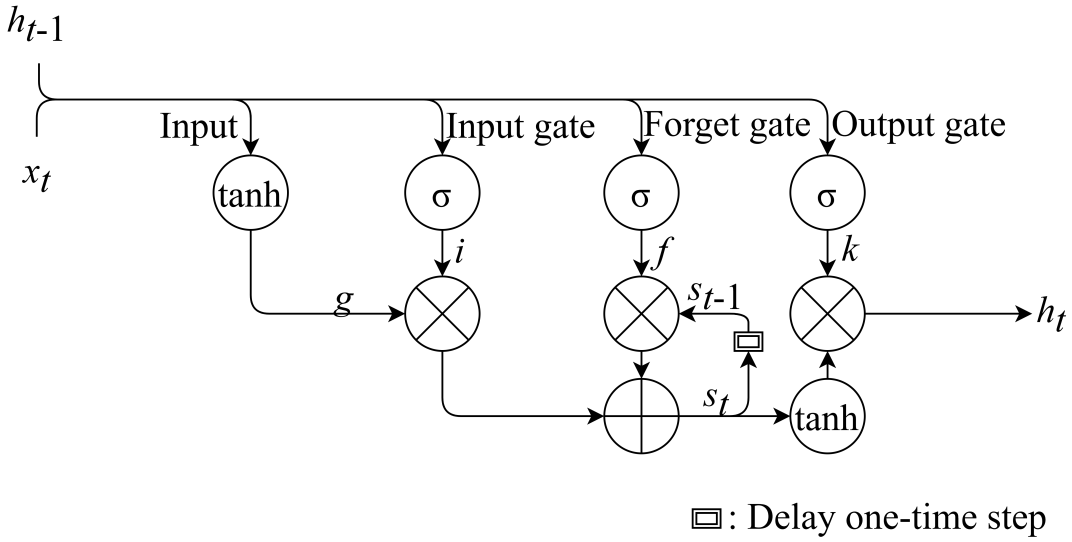


Figure 3.6: Diagram of the LSTM cell.

input biases. s_t is the inner state of the LSTM cell. The current inner state s_t is related to s_{t-1}, g, f, and i:

$$s_t = s_{t-1}f + gi \tag{3.12}$$

Therefore, the cell constructs an internal recurrence loop, which establishes a connection between inputs at different times. The inner state is through the hyperbolic tangent activation function. The result element-wise multiplies the output of the output gate to yield the final output of the LSTM cell:

$$h_t = tanh(s_t)k \tag{3.13}$$

With the above logic unit, the LSTM layer learns long-term dependencies between time steps in sequence features of the voltage signal. Its output is used as the input of the fully connected FNN. The Softmax function is used as the last activation function to normalize the output of the FNN to a probability distribution over the PD class and the NonPD class. Finally, the output classification layer computes the cross-entropy loss for classification tasks with the PD and NonPD classes and decides the predicted result, PD/NonPD, to which the sequence features belong.

The training-validation data and testing data for the LSTM neural network classi-

fier are the sequence features extracted from the voltage signals randomly selected from the VSB PD-related data set. In the training step, both the sequence features and the PD/NonPD labels are given to the LSTM neural network classifier. The classifier is improved by establishing the mapping relationship between the sequence features and the corresponding label. After the training, the LSTM neural network classifier classifies the testing data that are sequence features never used during training. The predicted results are PD/NonPD labels to which the testing data belong. They are predicted according to the mapping relationship established in the training process.

As for the PD class, the ratio of the signals used for training-validation and testing is 7:3. This ratio is determined after considering the efficiency, effect, and confidence level of training, validation, and testing. The training set is expected to be sufficient to avoid under-fitting and improve the quality of the model. At the same time, the testing set is expected to be as large as possible to increase the reliability of the evaluation. The ratio of 7:3 is considered a balanced choice for the two conditions. The oversampling method is used in the training-validation data set to minimize the impact of the class imbalance problem. As for the NonPD class, the data are allocated to match the training-validation data set of the PD class. The remaining data are adopted for testing. In order to improve credibility, the training-validation set and testing set are always divided randomly. Multiple groups of repeated experiments are conducted. All the evaluation parameter values in this chapter are the average of the results of the repeated experiments.

3.2.4 Evaluation Parameters

The quality of noise reduction methods and feature extraction methods is reflected in the final classification performance. Classification performance refers to whether the LSTM neural network classifier with such features as inputs can make correct classification decisions for the testing samples after the training. This thesis intro-

duces multiple evaluation parameters to evaluate the classification performance of the proposed LSTM-based resolution. The terms PD Target, PD Prediction, True PD Prediction, NonPD Target, NonPD Prediction, and True NonPD Prediction are used to explain the meaning of the evaluation parameters. The description for the PD Target, PD Prediction, and True PD Prediction are presented below. The NonPD Target, NonPD Prediction, and True NonPD Prediction can be deduced similarly.

PD Target: The PD Target refers to the voltage signal labelled by experts as PD.

PD Prediction: If a signal is determined by a classifier as PD, this signal is a PD Prediction for this classifier. The classifier may misjudge and predict a NonPD Target as a PD Prediction.

True PD Prediction: If a PD Target is predicted by a classifier as a PD Prediction, this sample is a True PD Prediction for this classifier. In other words, the classifier classifies this sample correctly. The True PD Prediction is the intersection of the PD Target and the PD Prediction.

The following evaluation parameters are adopted to evaluate the classification performance.

Accuracy: In binary classification, the terminology Accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. Accuracy is related to the classification performance of both classes of samples. The higher the Accuracy, the better the overall classification effect. In this thesis, Accuracy $Acc \in [0, 1]$ is related to the classification performance of both PD and NonPD classes. It is calculated as follows:

$$Acc = \frac{TP_{PD} + TP_{NPD}}{T_{PD} + T_{NPD}} \tag{3.14}$$

where TP_{PD} is the number of the True PD Prediction; TP_{NPD} is the number of

the True NonPD Prediction; T_{PD} is the number of the PD Target; and T_{NPD} is the number of the NonPD Target.

Recall: The terminology Recall is the proportion of the retrieved relevant instances in all relevant instances. The higher the Recall, the better the classification effect. This thesis adopts both $R_{PD} \in [0, 1]$, the Recall for the PD class, and $R_{NPD} \in [0, 1]$, the Recall for the NonPD class, to evaluate the classification performance. The calculation formulas are as follows:

$$R_{PD} = \frac{TP_{PD}}{T_{PD}} \tag{3.15}$$

$$R_{NPD} = \frac{TP_{NPD}}{T_{NPD}} \tag{3.16}$$

Precision: The terminology Precision is the fraction of relevant instances among the retrieved instances. The higher the Precision, the better the classification effect. This thesis adopts both $Pr_{PD} \in [0, 1]$, the Precision for the PD class, and $Pr_{NPD} \in [0, 1]$, the Precision for the NonPD class, to evaluate the classification performance. The calculation formulas are as follows:

$$Pr_{PD} = \frac{TP_{PD}}{P_{PD}} \tag{3.17}$$

$$Pr_{NPD} = \frac{TP_{NPD}}{P_{NPD}} \tag{3.18}$$

where P_{PD} is the number of the PD Prediction, and P_{NPD} is the number of the NonPD Prediction.

F1-score: The F1-score is the harmonic mean of the Recall and the Precision. The higher the F1-score, the better the classification effect. Both $F_{PD} \in [0, 1]$, the F1-score for the PD class, and $F_{NPD} \in [0, 1]$, the F1-score for the NonPD class, have been analyzed. The calculation formulas are as follows:

$$F_{PD} = \frac{2R_{PD}Pr_{PD}}{R_{PD} + Pr_{PD}} \tag{3.19}$$

$$F_{NPD} = \frac{2R_{NPD}Pr_{NPD}}{R_{NPD} + Pr_{NPD}} \tag{3.20}$$

Training time: The training time Tr of the LSTM neural network classifier is also adopted as an evaluation parameter for the feature length analysis. Under the same hardware conditions, the less the training time, the faster the computation.

3.3 Result and Analysis

3.3.1 Analysis of Noise Reduction Methods

Different mother wavelets and decomposition levels for the DWT noise reduction are analyzed for the optimization of the noise reduction method. Specifically, three choices of the mother wavelets, the Daubechies 1 (DB1), Daubechies 4 (DB4), and Symlets 8 (Sym8) wavelets, and four choices of decomposition levels for the DWT noise reduction, the levels 2, 5, 8, and 11, are evaluated. The wavelets are randomly selected from the Daubechies series and Symlets series that are recommended to perform noise reduction for voltage signals [78]. The Daubechies wavelets, based on the work of Ingrid Daubechies, are a family of orthogonal wavelets, whereas the Symlets wavelets are a kind of nearly symmetrical wavelets. The Sym8 wavelet has eight vanishing moments; the DB1 wavelet has a maximal one vanishing moment; and the DB4 wavelet has maximal four vanishing moments. The four different decomposition levels are also randomly selected from the natural numbers.

Table 3.1 lists the evaluation parameters of the different noise reduction methods. The second column of Table 3.1 is the mother wavelets used for the DWT noise reduction of the respective noise reduction methods, whereas the third column is the respective decomposition levels. The first method is the control group without noise reduction. The values of the parameters in Table 3.1 are the average of the results of ten groups of repeated experiments. Fig. 3.7 intuitively shows the change of *Acc* caused by different noise reduction methods. The nine numbers of

Table 3.1: Analysis of Different Noise Reduction Methods

	Wavelet	Level	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
1	-	-	0.901	0.870	0.898	0.933	0.904
2	DB4	2	0.933	0.907	0.931	0.959	0.935
3	DB4	8	0.930	0.898	0.928	0.962	0.933
4	DB1	2	0.914	0.882	0.911	0.946	0.916
5	DB1	8	0.917	0.890	0.915	0.944	0.919
6	Sym8	2	0.936	0.906	0.934	0.965	0.937
7	Sym8	5	0.915	0.884	0.912	0.945	0.917
8	Sym8	8	0.939	0.911	0.937	0.967	0.941
9	Sym8	11	0.947	0.923	0.945	0.970	0.948

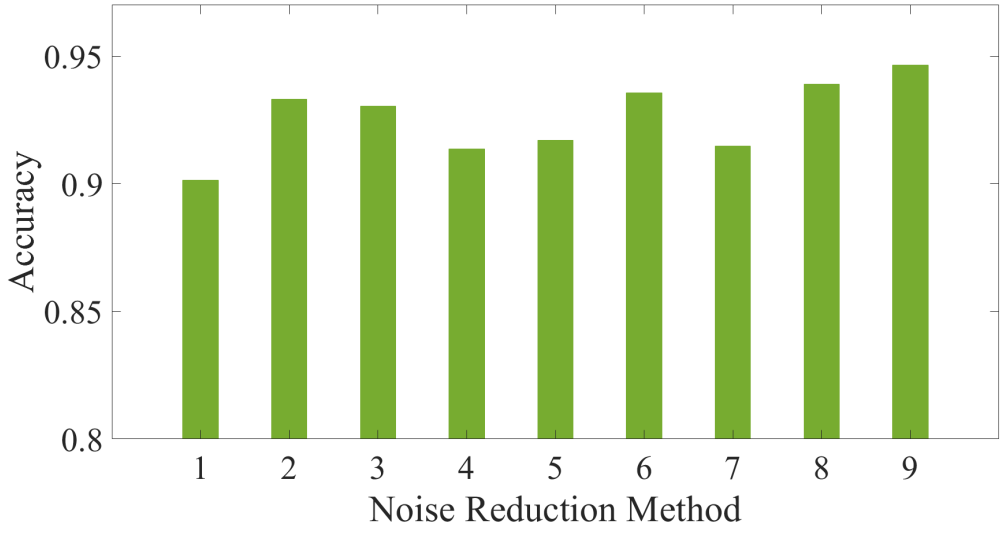


Figure 3.7: Accuracy Acc of different noise reduction methods.

abscissa correspond to the nine different methods described in Table 3.1.

The variation of the values of the evaluation parameters in Table 3.1 and Fig. 3.7 proves the significance and necessity of noise reduction. This preliminary analysis shows that among the DB4, DB1, and Sym8 wavelets, the Sym8 wavelet is more suitable for the PD detection task under study. With the same decomposition level, the values of Acc, R_{NPD}, F_{PD}, and F_{NPD} are higher when the Sym8 wavelet is used. Among the four different decomposition levels of 2, 5, 8, and 11, the values of Acc, R_{PD}, R_{NPD}, F_{PD}, and F_{NPD} are the highest when level 11 is adopted.

3.3.2 Analysis of the Feature Length

The feature length significantly affects the classification performance and the amount of computation. Six different lengths: 10, 25, 50, 100, 200, and 400 are compared. They are randomly selected from the numbers by which the length of the voltage signal is divisible. The features with over 400 elements lead to the requirement of high-performance computing resources. This thesis controls the sequence features within 400 elements to ensure the economic practicability of the proposed resolution.

Table 3.2 lists the evaluation parameters of the six different feature lengths. The first column of Table 3.2 lists the feature lengths. The values of the parameters in Table 3.2 are the average of the results of seven groups of repeated experiments. Fig. 3.8 shows the change of Tr caused by the different feature lengths. Tr increases as the feature length increases. The increase of Tr means the intensity of calculation. Fig. 3.9 shows the change of Acc caused by different feature lengths. On the one hand, if the feature length is too short, the feature will contain insufficient valid information. As the feature length increases from 10 to 100, Acc increases from 0.697 to 0.878. On the other hand, if the feature length is too long, it would add an additional burden to the LSTM neural network classifier. As the feature length increases from 100 to 400, Acc decreases from 0.878 to 0.661. The length of 100 outperforms other lengths in this preliminary analysis. Tr for the 100-element feature is acceptable. And it corresponds to the highest values of Acc, R_{PD}, F_{PD}, and F_{NPD}. It is considered to have greater potential with the optimization of noise reduction and feature extraction methods.

3.3.3 Analysis of 13 Different Features

The proposed LSTM-based resolution adopts the 13 sequence features described in Section 3.2.2 as the inputs to the LSTM neural network classifier. As an example,

Table 3.2: Analysis of Different Feature Lengths

Length	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}	Tr (s)
10	0.697	0.643	0.680	0.750	0.712	308.368
25	0.850	0.836	0.848	0.864	0.852	326.344
50	0.854	0.816	0.848	0.891	0.859	369.475
100	0.878	0.894	0.880	0.862	0.876	511.941
200	0.727	0.678	0.713	0.777	0.740	924.404
400	0.661	0.537	0.613	0.785	0.698	2030.125

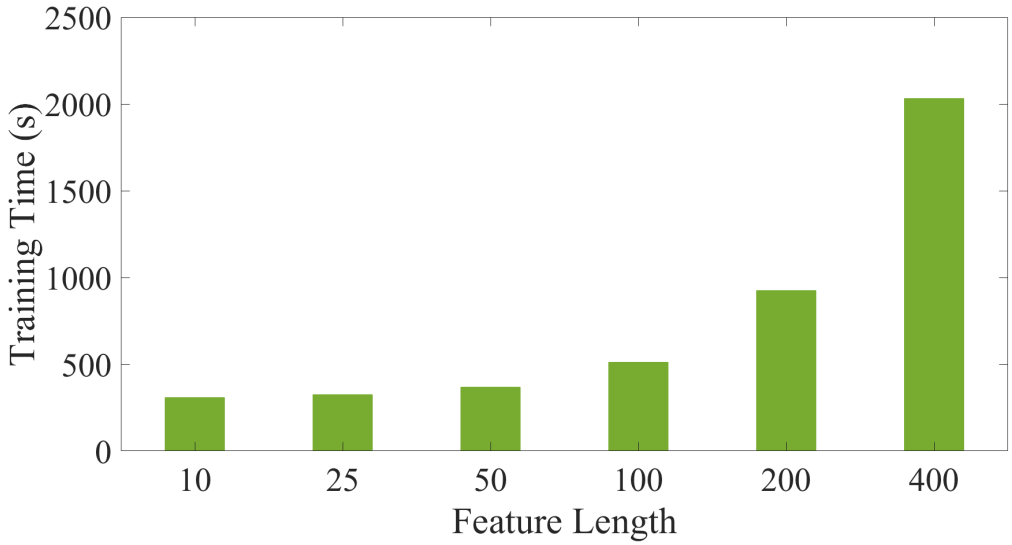


Figure 3.8: Training time Tr of different feature lengths.

Fig. 3.10 shows the 13 features extracted from the denoised signal shown in Fig. 3.3b. Table 3.3 lists the evaluation parameters of the different features. The values of the parameters in Table 3.3 are the average of the results of five groups of repeated experiments. It can be seen from Table 3.3 that the features of extremum values (Features 1-5) have nice performance, while the features of the data distribution are less capable of handling the PD class.

3.3.4 Analysis of Loop Optimization

Sections 3.3.1, 3.3.2, and 3.3.3 have discussed the performance of the different noise reduction methods, feature lengths, and feature extraction methods, respectively. The above discussion has preliminarily reflected the general influence of different items on the PD detection performance of the whole proposed LSTM-

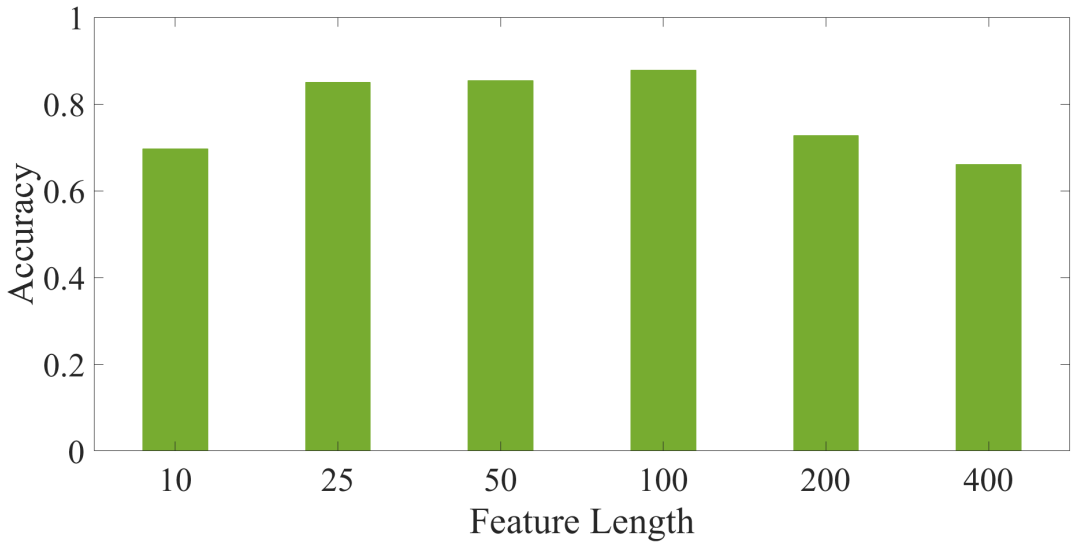


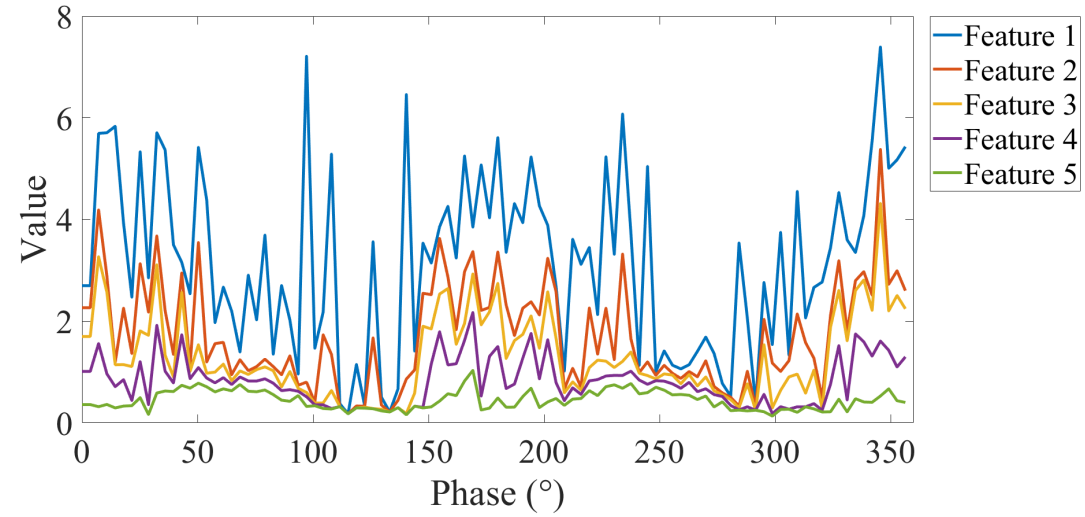
Figure 3.9: Accuracy *Acc* of different feature lengths.

Table 3.3: Analysis of 13 Different Features

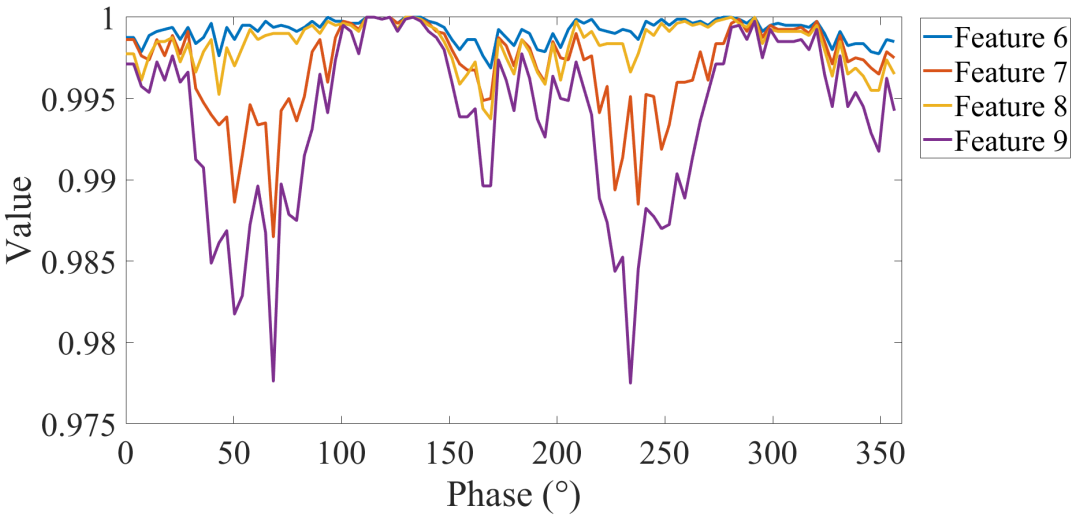
	Acc	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
Feature 1	0.905	0.893	0.904	0.918	0.907
Feature 2	0.871	0.856	0.869	0.885	0.873
Feature 3	0.873	0.862	0.872	0.885	0.875
Feature 4	0.854	0.808	0.847	0.900	0.861
Feature 5	0.779	0.696	0.758	0.864	0.796
Feature 6	0.532	0.103	0.180	0.960	0.672
Feature 7	0.563	0.210	0.325	0.917	0.677
Feature 8	0.646	0.586	0.623	0.706	0.666
Feature 9	0.585	0.293	0.414	0.877	0.679
Feature 10	0.562	0.472	0.519	0.651	0.598
Feature 11	0.601	0.363	0.477	0.839	0.678
Feature 12	0.602	0.368	0.481	0.836	0.678
Feature 13	0.644	0.575	0.618	0.713	0.667

based resolution. However, as the internal comparison of the single items controls other items for the single variable, the impact of other items is deficient in the analysis. In this section, the influence of the different combination schemes for the noise reduction, feature lengths, and different feature extraction methods are analyzed.

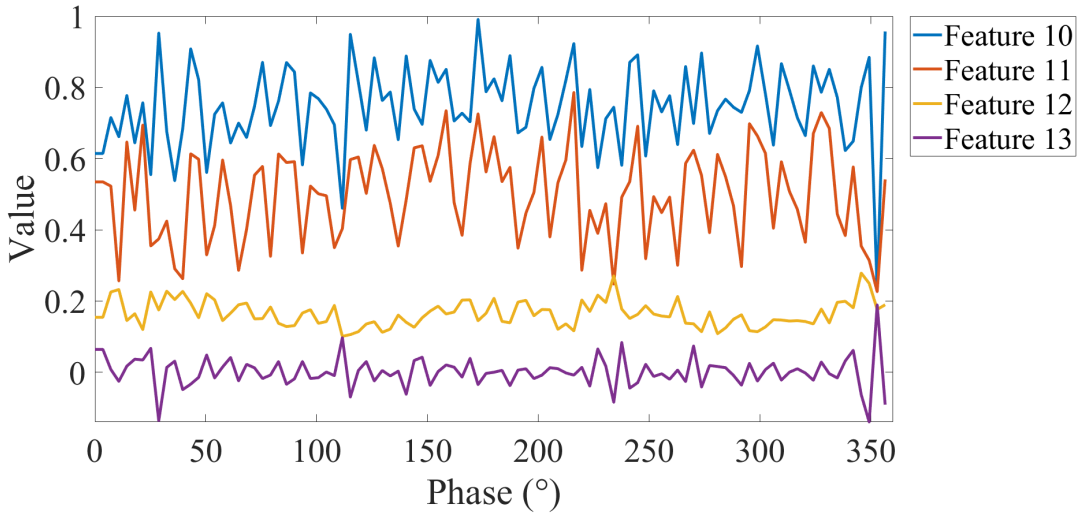
Table 3.4 presents the analyzed seven combination schemes of the noise reduction methods and feature lengths. The results in Section 3.3.1 indicates that the DWT noise reduction with the Sym8 wavelet is more likely to obtain a better



(a) The Features 1-5 of the voltage signal shown in Fig. 3.3b.



(b) The Features 6-9 of the voltage signal shown in Fig. 3.3b.



(c) The Features 10-13 of the voltage signal shown in Fig. 3.3b.

Figure 3.10: The 13 features of the voltage signal shown in Fig. 3.3b.

Table 3.4: Details of 7 Combination Schemes (Set No. 1-7)

Set No.	Wavelet	level	Feature length
1	DB1	8	100
2	DB4	8	100
3	Sym8	5	100
4	Sym8	11	100
5	Sym8	11	50
6	Sym8	11	25
7	Sym8	11	10

performance than the wavelets of the DB series with the same decomposition level. Among all the analyzed noise reduction methods, the 11-level DWT noise reduction with the Sym8 wavelet obtains the best performance in the preliminary analysis. Therefore, in this section, the 11-level DWT noise reduction with the Sym8 wavelet is employed to combine with all the feature lengths. As the results in Section 3.3.2 show that the feature length of 100 outperforms the other lengths in the preliminary analysis, the feature length of 100 is employed in more comparisons of different noise reduction methods.

Table 3.5 presents the analyzed five feature groups (Groups A-E) which combine different feature extraction methods (Features 1-13). The results in Section 3.3.3 indicate that the features of extremum values have nice performance in the preliminary analysis, while the features of the data distribution are less capable of handling the PD class. Groups A, B, and C combine the features of extremum values, data distributions, and statistics, respectively. They are analyzed if the combination of features of the same category outperforms the single feature of that category. Group D selects the features with the best preliminary performance from each category. Group E combines all the 13 features.

Table 3.6 lists the evaluation parameters of the different combination schemes with different feature groups. The values of the parameters in Table 3.6 are the average of the results of five groups of repeated experiments. Fig. 3.11 intuitively shows the change of *Acc* caused by different combinations. It can be seen from Table 3.6

Table 3.5: Details of 5 Different Feature Groups (Groups A-E)

	A	B	C	D	E
1	✓			✓	✓
2	✓			✓	✓
3	✓			✓	✓
4	✓				✓
5	✓				✓
6		✓		✓	✓
7		✓			✓
8		✓			✓
9		✓			✓
10		✓			✓
11		✓			✓
12			✓	✓	✓
13			✓		✓

and Fig. 3.11 that with the same combination scheme, Group A always performs better than Group B and Group C. This is consistent with the performance of the single feature of the three categories. Compared the values of the parameters in Table 3.3 and Table 3.6, Group A outperforms the single Features 1-5. Group B outperforms the single Features 6-11. And Group C outperforms the single Features 12-13.

As for the Group D and Group E that combines the features from different categories, all the seven combination schemes obtain nice results with these two groups of features. Except Set No. 1, all the other six combination schemes obtain better performance when the features of different categories are combined. And Group E always obtains better performance than Group D no matter the change of noise reduction methods and feature length. Comparing Group A, Group D, and Group E, Group E adds some features that are relatively weak in the preliminary analysis. However, the complete combination outperforms the groups that only combine the strong features. The combination of the different features has combined advantages and can overcome each individual's drawbacks.

Set No. 1-4 compare the different noise reduction methods with different combination schemes of the 100-elements features. The results of their Group A, Group

Table 3.6: Analysis of Different Combination Schemes and Feature Groups

Set No.	Feature No.	*Acc*	R_{PD}	F_{PD}	R_{NPD}	F_{NPD}
1	A	0.888	0.877	0.887	0.899	0.889
	B	0.735	0.660	0.714	0.810	0.753
	C	0.558	0.163	0.270	0.954	0.683
	D	0.803	0.767	0.796	0.839	0.810
	E	0.911	0.885	0.908	0.936	0.913
2	A	0.914	0.892	0.912	0.936	0.916
	B	0.773	0.821	0.783	0.725	0.762
	C	0.589	0.786	0.656	0.393	0.489
	D	0.919	0.887	0.917	0.951	0.921
	E	0.942	0.922	0.941	0.961	0.943
3	A	0.913	0.873	0.909	0.952	0.916
	B	0.813	0.674	0.783	0.952	0.836
	C	0.779	0.624	0.739	0.933	0.809
	D	0.917	0.895	0.915	0.938	0.919
	E	0.926	0.886	0.922	0.965	0.928
4	A	0.939	0.921	0.938	0.957	0.940
	B	0.758	0.883	0.785	0.632	0.723
	C	0.660	0.525	0.607	0.796	0.701
	D	0.947	0.926	0.946	0.968	0.948
	E	0.954	0.937	0.953	0.971	0.955
5	A	0.937	0.913	0.936	0.961	0.939
	B	0.781	0.763	0.777	0.798	0.785
	C	0.817	0.797	0.814	0.838	0.821
	D	0.923	0.897	0.921	0.948	0.925
	E	0.935	0.909	0.934	0.962	0.937
6	A	0.912	0.881	0.909	0.944	0.915
	B	0.826	0.844	0.829	0.808	0.823
	C	0.744	0.743	0.743	0.744	0.744
	D	0.899	0.876	0.896	0.922	0.901
	E	0.916	0.887	0.913	0.944	0.918
7	A	0.726	0.606	0.689	0.846	0.755
	B	0.714	0.614	0.683	0.814	0.740
	C	0.669	0.503	0.603	0.835	0.716
	D	0.764	0.733	0.757	0.795	0.771
	E	0.798	0.743	0.786	0.854	0.809

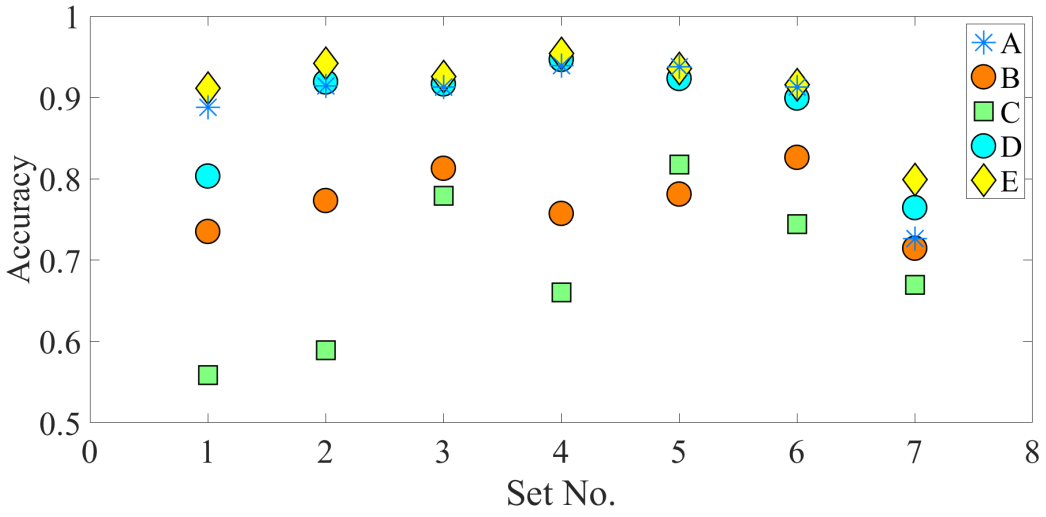


Figure 3.11: Accuracy *Acc* of different combination schemes (Set No. 1-7) with different feature groups (Groups A-E).

D, and Group E are consistent with the results of Table 3.1. Set No. 4-7 compare the different feature lengths with the same noise reduction method and different feature combination schemes. With the combination of different features, the advantages of the advanced feature length are further expanded. With the analysis of Sections 3.3.1, 3.3.2, 3.3.3, and 3.3.4, this chapter determines the comprehensive resolution for the PD detection of MV overhead power lines as the LSTM neural network classifier with the Sym8-11-level DWT noise reduction and 13 combined 100-element features.

The analysis of Table 3.6 and Fig. 3.11 also leads to more common discussions of the optimization of PD detection methods. The single features and noise reduction methods discussed in this chapter are based on the essential characteristics of the PD. They have low hardware implementation and computation requirements. Hence, they have advantages of adapting to wider PD detection tasks. And the improvement of combination and loop optimization analysis of such items are beneficial to practical applications. As combining the noise reduction, feature extraction, and the LSTM neural network classifier as a whole with internal linkages, the comprehensive optimization becomes a problem of permutation and combination. The exhaustive method does not meet the actual economic requirements,

while the one-way selection method lacks the internal linkages. Therefore, a loop optimization mechanism is proposed based on the aforementioned experimental results and analysis.

1. The principle is that it can continuously conduct the cyclic comparison of different items until the overall PD detection performance satisfies the task requirements. Optimization can stop at any step of a loop. The loop optimization has great improvement potential.

2. It is suggested to first determine the feature length preliminarily based on the computation requirements and the detection performance of the single feature. The advanced feature length is likely to have more improvement potential with the further optimization.

3. It is suggested to appropriately increase the number of features.

4. It is suggested to appropriately increase the decomposition level of the noise reduction methods. The noise reduction methods are suggested to be optimized after selecting the strong features in each loop.

3.3.5 Comparison with Other Techniques

The proposed LSTM-based resolution has achieved a comprehensive and complete resolution for PD detection for MV overhead power lines. It is compared with the noise reduction methods, feature extraction methods, and AI classifiers in the literature as shown in Table 3.7 [49, 50].

The "STL" method employs the seasonal and trend decomposition using locally weighted regression (STL). STL transforms the voltage signal into the trend component, seasonal component, and residual component. The peak value, standard deviation, and accumulated peak height of the residual component are used as features. The LSTM neural network is adopted as the classifier.

The "DWT" method also employs the LSTM neural network. It employs the

Table 3.7: Comparison of the Proposed LSTM-Based Resolution and Other Methods in the Literature

Method	*Acc*	R_{PD}	R_{NPD}
STL	0.805	0.838	0.772
DWT	0.871	0.867	0.882
SVM	0.701	0.665	0.737
CNN	0.775	0.742	0.809
Proposed Resolution	0.954	0.937	0.971

DWT noise reduction to the voltage signals. The decomposition components of the DWT are then obtained as features.

The "SVM" method uses six features. The features are the energy of the voltage signal, the energy of the voltage signal after the noise reduction, the energy of the abnormal parts of the signal, the average of the segmental maximum values, the amount of the abnormal measurements, and the standard deviation. The Support Vector Machine (SVM) is adopted for the task of classification.

As for the "CNN" method, the voltage signal is firstly denoised. Then the maximum value is extracted as a feature. The convolutional neural network (CNN) is adopted for the task of classification.

It can be seen from Table 3.7 that R_{PD} and R_{NPD} of the "STL" method differ greatly. The "STL" method has obvious shortcomings for the NonPD class. And the STL processing costs more calculation than other noise reduction and feature extraction methods. As for the "DWT" method, its classification results for the PD and NonPD classes are relatively balanced. With the same LSTM neural network, the proposed resolution greatly improves the PD detection performance with the coordination of the noise reduction, feature combination, and the LSTM based classifier. As for AI techniques, the proposed resolution is compared with the SVM and CNN. Both CNN and LSTM are considered to have the excellent sequence classification ability. However, with different feature extraction methods, the performance of the two neural networks differs a lot for the same PD detection task. This also demonstrates the significance of the coordination of the feature

extraction and the machine learning techniques. The proposed LSTM-based resolution experimentally demonstrates its superior performance.

3.4 Summary

In this chapter, the thesis has presented a LSTM-based classifier with loop optimization schemes for PD detection of MV overhead power lines. The proposed LSTM-based resolution greatly coordinates the multiple processing and interpretation procedures for the large-scale data disturbed with measurement noise. The loop optimization mechanism is established to combine the noise reduction, feature extraction, and the LSTM neural network as a whole with internal linkages. The optimal combination has combined advantages and can overcome each individual's drawbacks. Extensive experiments have been conducted based on the huge real trial data set of large-scale high-frequency voltage signals measured from the MV overhead power lines. The proposed LSTM-based resolution outperforms other techniques in the literature based on the experimental results. It has achieved a comprehensive and complete resolution for automatic PD detection of MV overhead power lines. It is a valuable tool for insulation monitoring and power maintenance. Moreover, this thesis summarizes the common loop optimization rules for PD detection resolutions based on extensive experiments and analysis. The loop optimization rules are competent for the optimization of wider engineering applications of PD detection.

Chapter 4

A Novel Transformer-Based Multilevel Filtering Framework for Partial Discharge Detection

4.1 Introduction

In the previous chapters, the importance and necessity of the research of the intelligent and effective PD detection methods have been introduced. There have been much research applying the machine learning approaches on the PD detection field. The LSTM-based resolution proposed in Chapter 3 has proved its superior performance for the PD detection task of the MV overhead power lines. However, the previous machine learning-based PD detection approaches, including the previous LSTM-based resolution, are either applicable to short measurement signals or rely heavily on feature extraction for long measurement signals. The feature extraction condenses information from millions of measurements with at most a few hundred feature elements. The effectiveness of such information refinement is highly dependent on the feature extraction methods manually decided by the designer. It has low adaptability and robustness to changes in data and measurement

environments. For long-term online monitoring of complex power systems, such PD detection methods cannot effectively utilize the intelligence of the machine learning algorithms. It is necessary to study the novel machine learning-based PD detection methods with high feasibility and adaptability and less manual intervention for large-scale measurement data.

Recently, a machine learning algorithm named transformer has set off a wave of research and applications. The transformer-series algorithms are known for achieving fully parallel computing by solely depending on attention mechanisms while dispensing with recurrence or convolution mechanisms entirely. Previous research works on the transformer have rarely involved power fields such as PD detection. For the PD detection of practical power systems, the existing transformer models still have defects due to the complex background noise interference and the difficulties of acquiring the ground truth.

This thesis dedicates to a novel transformer-based multilevel filtering framework for PD detection of MV overhead power lines. The aim is to effectively and automatically detect PD from large-scale online measurement data while reducing the manual intervention and computation requirements. The innovation of the proposed transformer-based multilevel filtering framework is that it introduces the advanced computing performance and intelligent mechanism of the transformer algorithm into the PD detection field, and establishes the multilevel filtering framework to make the data-driven approach adapt to practical engineering requirements. It can flexibly analyze the PD-related measurements and their phase information without relying on the position relationship of the original measurement sequence. It applies the advanced and intelligent machine learning algorithms on the highly engineered framework. The structural design is easy-adjusting and adapted for various practical applications.

The VSB PD-related data set verifies the practicability and effectiveness of the proposed transformer-based multilevel filtering framework for real-time long-term

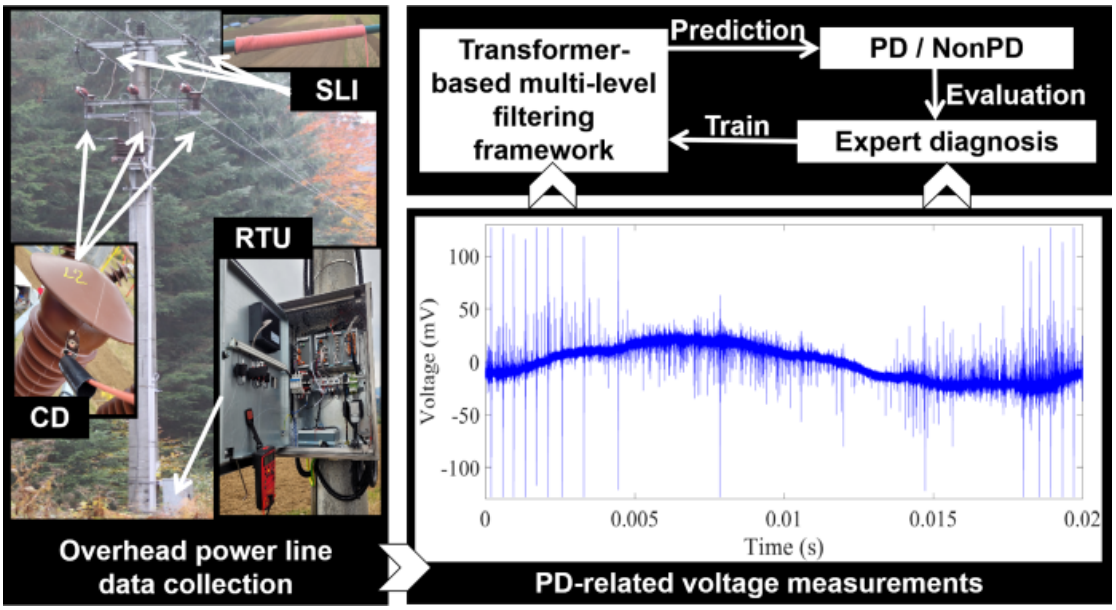


Figure 4.1: The whole process of PD detection for the MV overhead power lines.

online PD detection tasks. The labels marked by experts are used to train the transformer-based classifier and verify the prediction results of the transformer-based multilevel filtering framework. Fig. 4.1 represents the whole process of PD detection in the practical MV overhead power lines using this tool. The detection depends on the large-scale measurement data collected from the cost-effective DAQ platform mainly consisting of the single layer inductor (SLI) sensor, capacitive divider (CD), and remote terminal unit (RTU). The transformer-based multilevel filtering framework can judge whether there are any PD activities occurring in the online measured voltage signals. The experts' judgment is used as evaluation criteria. Extensive experiments are conducted to demonstrate that the proposed transformer-based multilevel filtering framework has great potential to be competent for various practical PD detection tasks, and it can deliver superior performance compared to other PD detection methods based on machine learning.

The remainder of this chapter is organized as follows: Section 4.2 presents the architecture and methodologies of the proposed transformer-based multilevel filtering framework. Section 4.3 presents the experiments and results with analysis and discussions. Section 4.4 summarizes the chapter.

4.2 Architecture and Methodologies

This thesis proposes a novel transformer-based multilevel filtering framework for PD detection for large-scale measurement data disturbed and corrupted significantly by measurement noise. It aims to be an intelligent and effective long-term online monitoring tool for practical electrical systems. Its function is to classify the detected signal into **PD Prediction** and **NonPD Prediction** according to whether the detected signal contains PD activities or not. It is a combination of electrical knowledge and artificial intelligence technologies with a multilevel filtering framework shown in Fig. 4.2. The multilevel filtering mechanism with the transformer-based classifier reduces manual intervention and improves the robustness and intelligence of the framework.

4.2.1 Filtering of Possible PD Measurements

The detected signal is first processed through appropriate noise reduction according to the measurement environment and data characteristics. The concept of **possible PD measurements** is introduced to initially screen the measurements according to the PD characteristics in the denoised signal. The filtering conditions requirements for the possible PD measurement are intentionally simple and loose. The detected signal without enough possible PD measurements under such loose filtering will be directly classified as a NonPD Prediction. The loose filtering of possible PD measurements ensures a very low misjudgment rate for NonPD Prediction since it leaves all the signals containing enough possible PD measurements for the next level of filtering. However, it also provides an economical and fast option so that those detected signals obviously without faults can be directly determined as NonPD Prediction without further analysis. Considering that the occurrence of failure and PD is rather rare in the practical power systems, the filtering of possible PD measurement greatly improves the economic practicability for the long-term online PD monitoring. And it reduces the burden of the

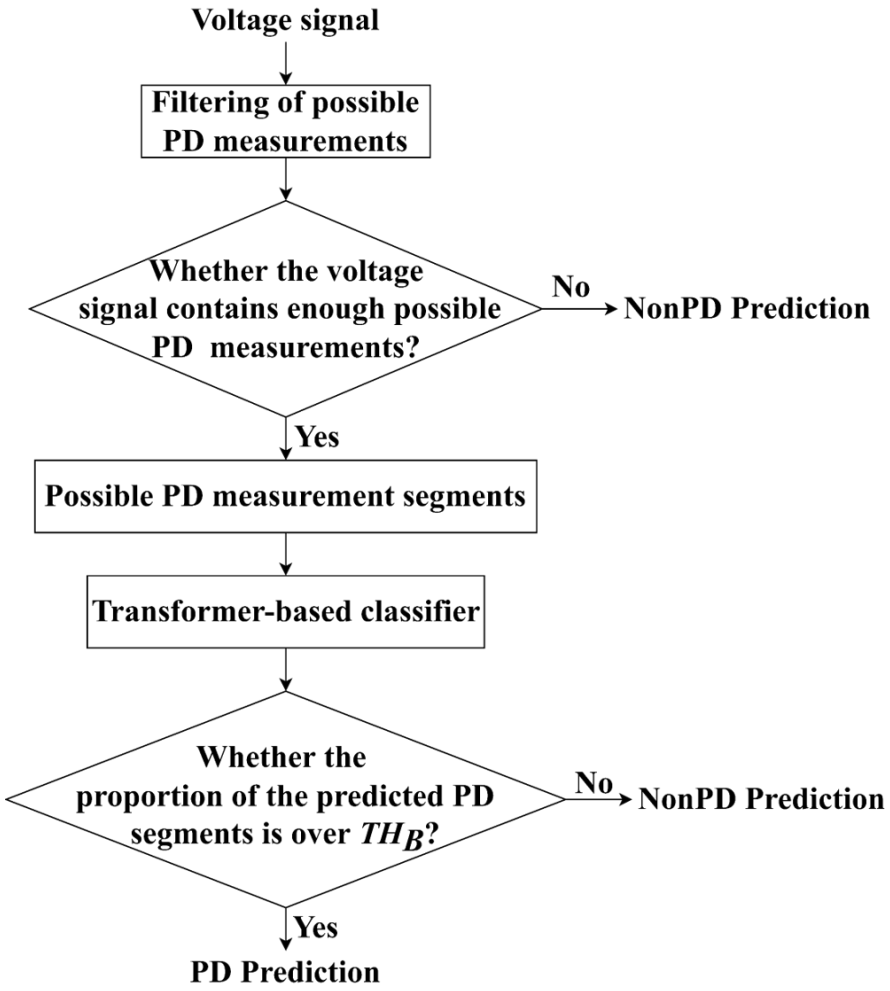


Figure 4.2: The flowchart of the proposed transformer-based multilevel filtering framework.

transformer-based classifier.

4.2.2 Filtering of the Transformer-Based Classifier

If a detected signal has enough possible PD measurements, these possible PD measurements are divided into segments according to the measurement sequence. The possible PD measurement segments are analyzed by the transformer-based classifier. The segmentation step is to regularize the input length of the transformer-based classifier and control the computational cost of the classification and analysis. And it is reasonable to introduce the local analysis of the measurements of phase proximity to the PD detection tasks. Therefore, the transformer-based multilevel filtering framework can process longer signals from a broader scope of

application.

The transformer-based classifier, as shown in Fig. 4.3, has the ability to analyze the possible PD measurements without using the order of inputs to represent their phase relationship. Unlike the traditional RNN which relies on the sequence order of the series input, the transformer-based classifier solely utilizes the attention mechanisms and achieves complete parallel processing. The concatenation input consists of the possible PD measurement segment and the corresponding phase information. The possible PD measurement segments are embedded with their phase information using the positional encoding (PE) technique [51]. The function of PE is as follows:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{4.1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{4.2}$$

where pos is the position; and i is the dimension corresponding to a sinusoid. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. d_{model} is the embedding size, as well as the hidden layer size of the transformer layers. The embedded amplitude and phase information are adopted as the input of the transformer-based classifier. Such input creates a macro perspective similar to the method of manual PD detection. Meanwhile, the amount of data has been greatly reduced.

The input is processed through five transformer layers. Each transformer layer is composed of two sub-layers, as shown in Fig. 4.3. The sub-layers perform either multi-head self-attention mechanism or feed-forward mechanism [51]. They are connected with residual connection and layer normalization.

The attention function is to flexibly map the most relevant parts of the input query and key-value pairs to an output. As for the self-attention mechanism, the query, key, and value are identical vectors of dimension d_v [51]. It analyzes

all the elements of the input vector and calculates their relationship. Different hidden representations capture the global sequence information and highlight the high-relevance elements. In practice, the input vectors are packed together into a matrix V. The matrix of outputs is computed as follows:

$$Attention_{self}(V) = Softmax(\frac{VV^T}{\sqrt{d_v}})V \tag{4.3}$$

where V represents the matrices stacked by multiple input vectors as rows; and d_v is the dimension of the input vectors.

The multi-head self-attention mechanism is represented as (4.4) - (4.5). For a self-attention block with h heads, the inputs are linearly projected h times with different projections respectively. On each head, the self-attention function is then performed on the projected versions of the inputs. The results of all the h heads are concatenated and once again projected, resulting in the final values.

$$MultiHead(V) = Concat(head_1, ..., head_h)W^O \tag{4.4}$$

$$head_i = Attention_{self}(VW_i^V) \tag{4.5}$$

where $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ is the weight matrix of the ith head; and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ is the weight matrix of the final output.

The feed-forward mechanism is performed by two linear transformations with a ReLU activation in between. The equation is shown as follows:

$$FNN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{4.6}$$

where x is the result of the multi-head self-attention sub-layer; W_1 and W_2 are weight matrices; and b_1 and b_2 are the bias terms of the two linear transformations.

A modified classification layer is added after the last transformer layer, which decides whether the possible PD measurement segment is a **predicted PD segment** or **predicted NonPD segment**. A pooling operation is applied to the

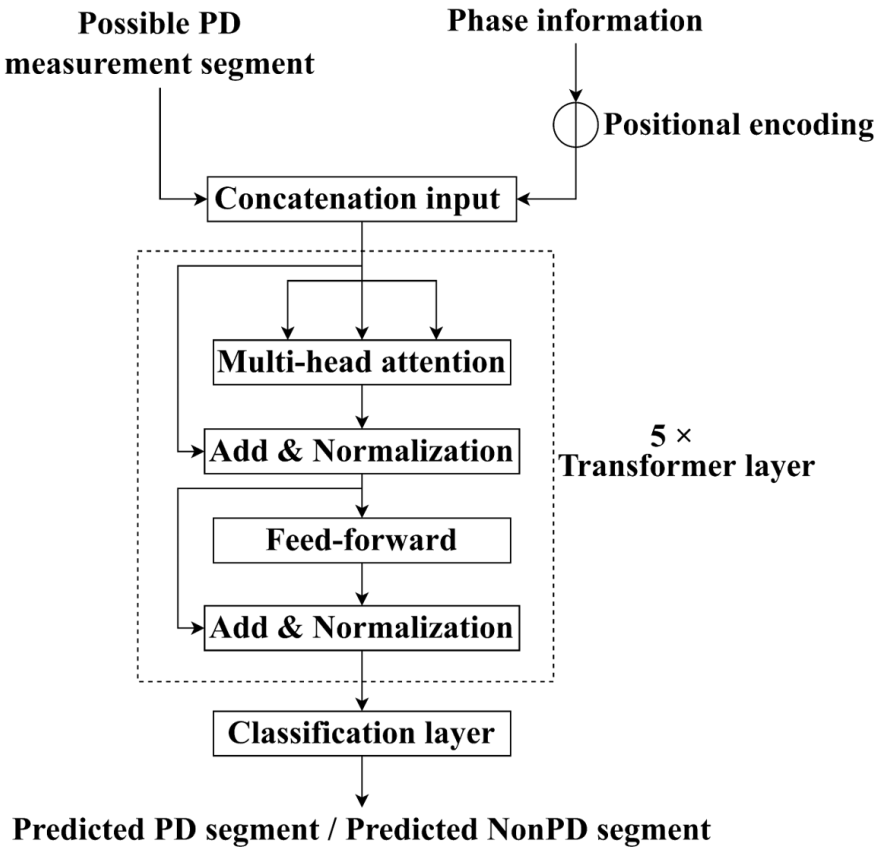


Figure 4.3: The architecture of the transformer-based classifier.

output vectors of the last transformer layer by computing the maximum over-time of the output to aggregate them into a single fixed-size vector. Once the global representation is obtained, the Softmax function is employed to calculate the probability distribution over the two classes. An adjustable threshold TH_A is introduced to improve flexibility in final decision-making. Only if the increase between the probability distributions of predicted PD segment and predicted NonPD segment exceeds TH_A, the possible PD measurement segment will be classified as a predicted PD segment.

A threshold TH_B is set up for comprehensively integrating the prediction results of all the possible PD measurement segments in one detected signal. If the proportion of the predicted PD segments in a detected signal is over TH_B, this detected signal will be classified by the transformer-based multilevel filtering framework as a PD Prediction. Otherwise, the detected signal will still be classified as a NonPD

Prediction. For instance, the possible PD measurements in a detected signal are divided into M possible PD measurement segments. Among them, m segments are predicted as the predicted PD segment. The proportion of predicted PD segments in this detected signal is m/M. Only if m/M is over TH_B, this detected signal would be classified by the transformer-based multilevel filtering framework as a PD Prediction.

4.3 Experiments and Results

4.3.1 Experimental Setup

The proposed transformer-based multilevel filtering framework is verified by the VSB PD-related data set of MV overhead power lines. Chapter 3 has compared and analyzed different noise reduction methods for the voltage measurements from the overhead power lines. The proposed transformer-based multilevel filtering framework adopts its optimal noise reduction scheme: Firstly, the digital trap filter is used on the voltage signal to reduce the power line interference. Secondly, the RPI noise pulses are removed from the voltage signal. Thirdly, the DWT noise reduction with Sym8 wavelet is performed on the voltage signal.

Fig. 4.4 shows an example of a voltage signal before and after the above noise reduction processing. As an example, the level 11 DWT noise reduction is conducted in Fig. 4.4. The noise reduction filters the power line interference and irrelevant high-amplitude noise pulses. This results in the asymmetric amplitude reduction. The remaining high amplitude impulse components in the denoised signal have a greater possibility to correlate to PD activities.

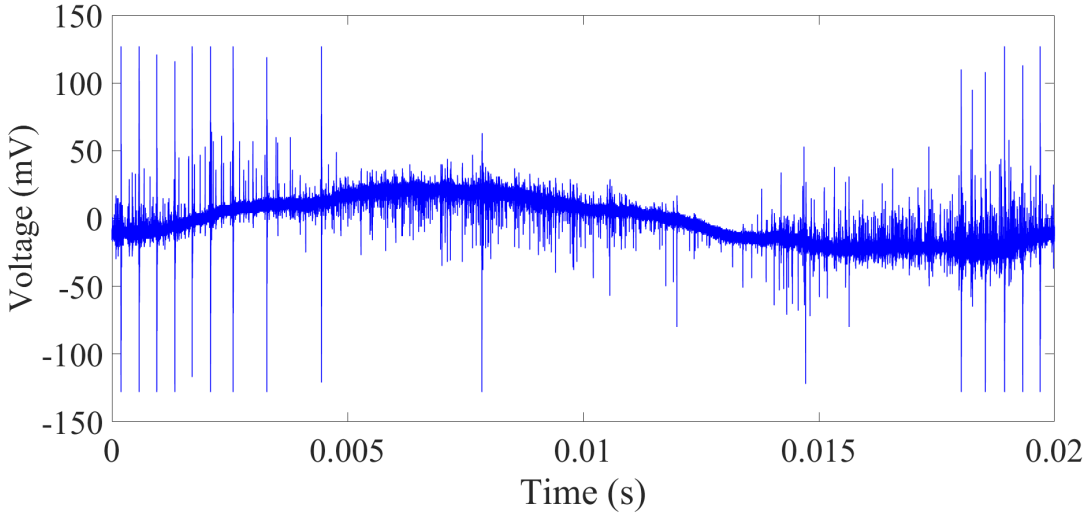
The proposed transformer-based multilevel filtering framework considers the measurements over 1 mV after the noise reduction as the possible PD measurements of the voltage signals in this PD-related data set. The measurements belonging to the interval of [-1 mV, 1 mV] are considered to contain few PD characteristics

with such low amplitude and small energy, while the measurements with negative values are more disturbed by the background noises. The possible PD measurements of the voltage signal in Fig. 4.4a are highlighted as red dots superimposed on the blue line in Fig. 4.4b.

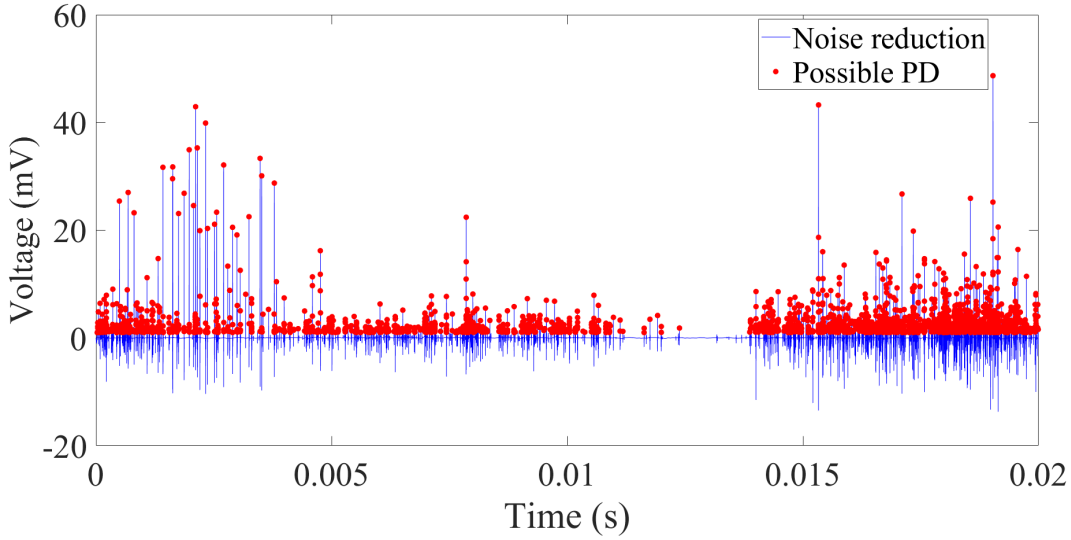
If a denoised voltage signal contains less than 100 possible PD measurements, the voltage signal is almost impossible to contain PD activities considering the huge amount of data in a voltage signal. Such voltage signals will be directly classified as the NonPD Predictions without the analysis of the transformer-based classifier. The possible PD measurements of the remaining voltage signals are divided into segments of less than 2,000 measurements. The maximum length of the segments is expected to be as long as computation conditions permit so that the classifier can comprehensively analyze the relationship between the possible PD measurements and their phases in the segment. The length of 2,000 is a balanced choice comprehensively considering the computation cost and analysis effectiveness of the transformer-based classifier.

Fig. 4.5 uses different colors to represent the possible PD measurement segments of the denoised voltage signal shown in Fig. 4.4. The abscissa of Fig. 4.5 shows the phase information of the possible PD measurements, whereas the ordinate shows their amplitude. The voltage signal contains 5,351 possible PD measurements that are divided into 3 segments: The first segment of 2,000 possible PD measurements highlighted in blue is in the interval of [-293.004°, -292.097°]. The second segment of 2,000 possible PD measurements highlighted in red is in the interval of [-292.096°, 10.199°]. The third segment of 1,351 possible PD measurements highlighted in yellow is in the interval of [10.297°, 66.348°].

It can be seen from Fig. 4.5 that the phase distribution of the possible PD measurements is very uneven. The first segment concentrates on the small interval of [-293.004°, -292.097°]. The denoised signal of this interval is almost always above 1 mV. The proportion of the possible PD measurement is close to 100%. This



(a) An example of a voltage signal labelled by experts as PD.



(b) The signal shown in Fig. 4.4a after the noise reduction.

Figure 4.4: The example of a voltage signal (PD) before and after the noise reduction.

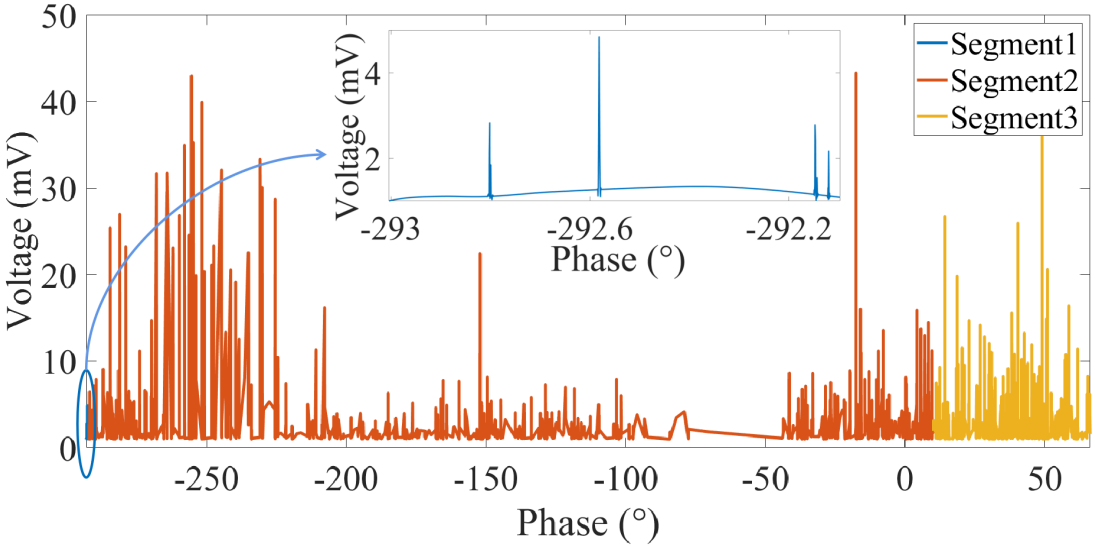


Figure 4.5: The three possible PD measurement segments of the voltage signal shown in Fig. 4.4.

makes the first possible PD measurement segment presented as the blue line very smooth with little interruption. As for the second and third possible PD measurement segments shown in Fig. 4.5, since the denoised signal has fewer possible PD measurements in this interval, the elements of the segments are sampled with a larger phase interval. Although the number of elements in each segment is similar, the second and third segments are more irregular when they are presented in the same figure.

To train the transformer-based classifier, 30% of both the PD Targets and NonPD Targets in the VSB PD-related data set are randomly selected as the testing data set. The possible PD measurements of the other PD Targets and NonPD Targets are divided into segments as the training-validation set of the transformer-based classifier. This arrangement is determined after considering the efficiency, effect, and confidence level of training, validation, and testing. The training-validation set is expected to be sufficient to avoid under-fitting and improve the quality of the transformer-based classifier. At the same time, the testing set is expected to be as large as possible to increase the reliability of the evaluation. This arrangement is considered a balanced choice for the two conditions.

During the training stage, the transformer-based classifier learns that the possible PD measurement segments belonging to PD Targets are predicted PD segments, while the possible PD measurement segments belonging to NonPD Targets are predicted NonPD segments. It should be noted that the number of positive and negative input samples during each training should be balanced. Multiple groups of repeated experiments are conducted to improve the experimental credibility. The oversampling method is used in the training data set to minimize the impact of the class imbalance problem. The cross-entropy loss is used for optimization.

The PD detection performance of the proposed transformer-based multilevel filtering framework is reflected in the correct classification of PD Target and NonPD Target to PD Prediction and NonPD Prediction respectively. We use the evaluation parameters Acc, R_{PD}, and R_{NPD} obtained by the following formulas to quantify the performance:

$$Acc = \frac{TP_{PD} + TP_{NPD}}{T_{PD} + T_{NPD}} \tag{4.7}$$

$$R_{PD} = \frac{TP_{PD}}{T_{PD}} \tag{4.8}$$

$$R_{NPD} = \frac{TP_{NPD}}{T_{NPD}} \tag{4.9}$$

where TP_{PD} is the number of the PD Targets that are correctly classified as the PD Prediction; TP_{NPD} is the number of the NonPD Targets that are correctly classified as the NonPD Prediction; T_{PD} is the number of the PD Targets; and T_{NPD} is the number of the NonPD Targets. All the evaluation parameter values in this chapter are the average of the results of the repeated experiments.

4.3.2 Analysis of the Noise Reduction Decomposition Level

The decomposition level of the DWT significantly affects the noise reduction intensity that closely correlates to the possible PD measurements. This thesis compares the decomposition levels 1-15 to analyze the influence of the decomposition level

on the noise reduction scheme and the construction of the possible PD measurements. We adopt $PPS_{PD} \in [0,1]$, the proportion of the PD Targets that contains more than 100 possible PD measurements, $PPS_{NPD} \in [0,1]$, the proportion of the NonPD Targets that contains more than 100 possible PD measurements, $NPM_{PD} \in [0,800,000]$, the average number of the possible PD measurements contained in one PD Target, and $NPM_{NPD} \in [0,800,000]$, the average number of the possible PD measurements contained in one NonPD Target, to evaluate the influence of the different decomposition levels. The calculation formulas are as follows:

$$PPS_{PD} = \frac{PS_{PD}}{T_{PD}} \tag{4.10}$$

$$PPS_{NPD} = \frac{PS_{NPD}}{T_{NPD}} \tag{4.11}$$

$$NPM_{PD} = \frac{\sum_{n=1}^{T_{PD}} PM_{PD}(n)}{T_{PD}} \tag{4.12}$$

$$NPM_{NPD} = \frac{\sum_{n=1}^{T_{NPD}} PM_{NPD}(n)}{T_{NPD}} \tag{4.13}$$

where PS_{PD} is the number of the PD Targets that contain more than 100 possible PD measurements; PS_{NPD} is the number of the NonPD Targets that contain more than 100 possible PD measurements; T_{PD} is the number of the PD Target; T_{NPD} is the number of the NonPD Target; $PM_{PD}(n)$ is the number of the possible PD measurements contained in the n^{th} PD Target; and $PM_{NPD}(n)$ is the number of the possible PD measurements contained in the n^{th} NonPD Target.

Table 4.1, Fig. 4.6, and Fig. 4.7 show PPS_{PD}, PPS_{NPD}, NPM_{PD}, and NPM_{NPD} of different decomposition levels. For visual clarity, Fig. 4.7 presents the ratio of NPM_{PD}, NPM_{NPD} and 800,000, the total number of the measurements in one detected signal. Table 4.1 and Fig. 4.7 choose different forms of NPM_{PD} and NPM_{NPD} because for a scatter plot, it is more intuitive to present the average proportion of the possible PD measurements in one signal rather than directly

presenting the average number of the possible PD measurements. However, as for Table 4.1, when the decomposition level rises to a certain level, the average proportion of the possible PD measurements in one signal is quite small. The limited number of decimal places cannot simply reflect the continuous decline of the possible PD measurements at this time. For Table 4.1, it is more intuitive and accurate to directly use the average number of the possible PD measurements contained in one signal to represent this continuous change.

The evaluation parameter values are the calculation results of the whole VSB PD-related data set including 2,104 PD Targets and 26,945 NonPD Targets. The first column of Table 4.1 is the list of decomposition level of the performed DWT noise reduction. It can be seen from Table 4.1, Fig. 4.6, and Fig. 4.7 that if the decomposition level is too low, too much irrelevant noise will remain. When the decomposition level is less than 6, NPM_{NPD}, over 60,000, is greater than NPM_{PD}. This will lead to plenty of interference and unnecessary calculations. If the decomposition level is too high, the effective information will be filtered, which will result in the reduction of detection accuracy. Taking the above two factors into consideration, level 11 is employed on the transformer-based multilevel filtering framework in the experiments: NPM_{NPD} is 4058, whereas NPM_{PD} is 7063. PPS_{NPD} is 0.939. Among the 26,945 NonPD Targets of the PD-related data set, 1,643 NonPD Targets can be classified correctly as the NonPD Prediction without further processing.

4.3.3 Analysis of TH_A

The default threshold TH_A is 0, which means the class with a higher probability will be the final output. 400 different values from -2 to 2 with an interval of 0.01 are compared to analyze the influence of the different TH_A on the transformer-based classifier. We adopt $RM_{PD} \in [0, 1]$, the proportion of the correctly classified predicted PD segments, $RM_{NPD} \in [0, 1]$, the proportion of the correctly classified

Table 4.1: Analysis of Different Decomposition Levels

Level	PPS_{PD}	PPS_{NPD}	NPM_{PD}	NPM_{NPD}
1	1	1	93948	152516
2	1	1	71546	137637
3	1	1	50788	133061
4	0.999	1	44272	123424
5	0.997	1	33247	64621
6	0.986	0.993	26752	20535
7	0.982	0.991	22458	15378
8	0.980	0.953	10476	5260
9	0.978	0.947	10065	5106
10	0.979	0.949	8078	4277
11	0.975	0.939	7063	4058
12	0.972	0.922	6942	3972
13	0.971	0.914	6873	3860
14	0.971	0.910	6829	3813
15	0.971	0.908	6704	3751

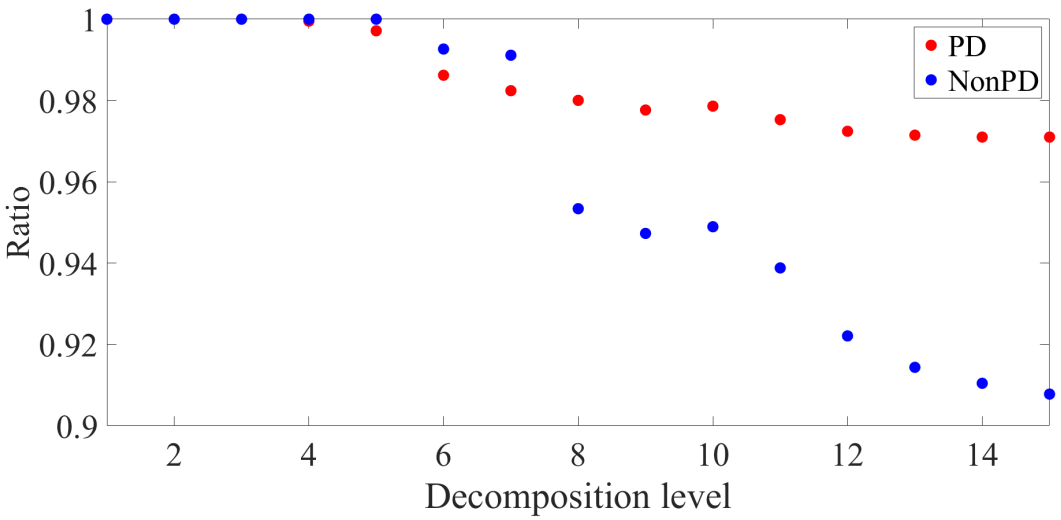


Figure 4.6: PPS_{PD} in red dots and PPS_{NPD} in blue dots of different decomposition levels.

predicted NonPD segments, and their average $Ave_{RM} \in [0, 1]$ to evaluate the performance of the transformer-based classifier as TH_A changes. The calculation formulas are as follows:

$$RM_{PD} = \frac{TPM_{PD}}{PM_{PD}} \tag{4.14}$$

$$RM_{NPD} = \frac{TPM_{NPD}}{PM_{NPD}} \tag{4.15}$$

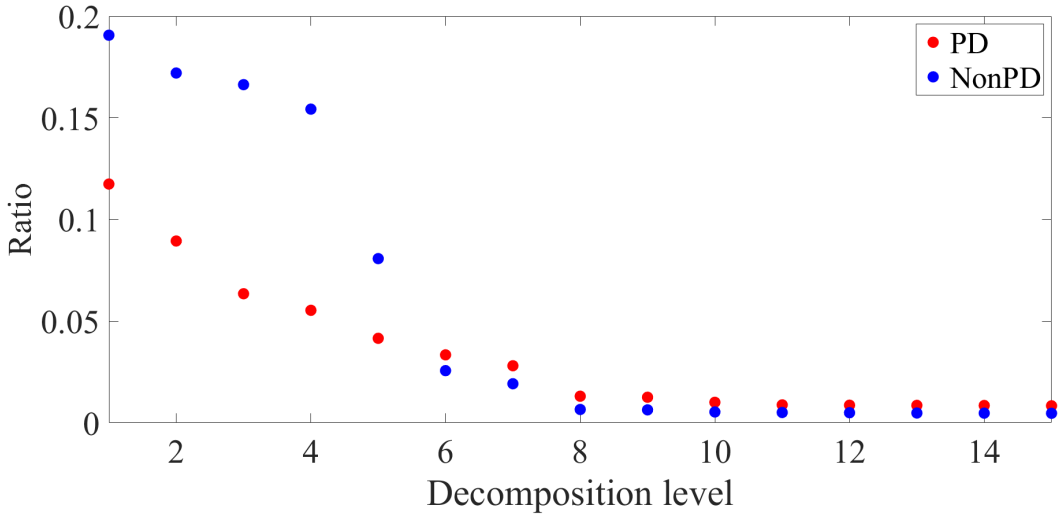


Figure 4.7: The ratio of NPM_{PD} and the total number of the measurements in one detected signal in red dots and the ratio of NPM_{NPD} and the total number of the measurements in one detected signal in blue dots of different decomposition levels.

$$Ave_{RM} = \frac{RM_{PD} + RM_{NPD}}{2} \tag{4.16}$$

where PM_{PD} is the number of the possible PD measurement segments belonging to PD Targets; TPM_{PD} is the number of the possible PD measurement segments correctly classified by the transformer-based classifier as predicted PD segments; PM_{NPD} is the number of the possible PD measurement segments belonging to NonPD Targets; and TPM_{NPD} is the number of the possible PD measurement segments correctly classified by the transformer-based classifier as predicted NonPD segments.

Fig. 4.8 presents RM_{PD}, RM_{NPD}, and Ave_{RM} as TH_A changes. Table 4.2 selects the results of 20 typical TH_A for display. The values of the parameters in Fig. 4.8 and Table 4.2 are the average of the results of six groups of repeated experiments. It can be seen from Fig. 4.8 and Table 4.2 that when TH_A is negative, the possible PD measurement segments are more likely to be classified as predicted PD segments, while positive TH_A leads to the opposite effect. When TH_A is -2, RM_{NPD} is only 0.849, whereas RM_{PD} is 0.994. The high RM_{PD} is at the cost of

the plenty of misjudgments of the NonPD samples. As TH_A approaches zero, the classification rules become less extreme. Ave_{RM} obtains the highest result when TH_A is -0.07. With such TH_A, RM_{PD} still maintains a relatively high value, whereas RM_{NPD} is much higher. As TH_A continues to increase, RM_{PD} starts to decline rapidly, whereas RM_{NPD} does not have much room for growth. Therefore, the transformer-based multilevel filtering framework selects -0.07 as TH_A for the PD-related data set of MV overhead power lines.

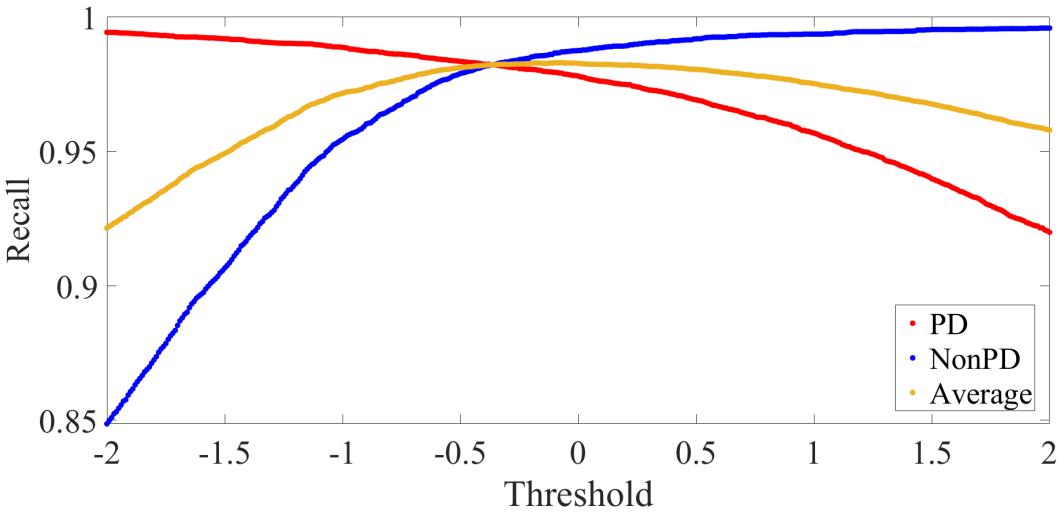


Figure 4.8: RM_{PD} in red dots, RM_{NPD} in blue dots, and Ave_{RM} in yellow dots of different TH_A.

4.3.4 Analysis of TH_B

It is reasonable that after the very loose filtering of possible PD measurements, not all the possible PD measurements are actually related to PD activities. For the PD Targets containing enough possible PD measurements that result in multiple possible PD measurement segments for the transformer-based classifier, it is not necessary that all these segments are classified as predicted PD segments. A more intelligent transformer-based classifier is expected to distinguish those possible PD measurement segments without PD activities from PD Targets as the same category with the predicted NonPD segments. We discuss the influence of 100 different TH_B from 0.01 to 1 with an interval of 0.01 on the performance of the

Table 4.2: Analysis of TH_A

TH_A	RM_{PD}	RM_{NPD}	Ave_{RM}
-1.87	0.994	0.864	0.929
-1.67	0.993	0.889	0.941
-1.47	0.992	0.910	0.951
-1.27	0.990	0.931	0.961
-1.07	0.989	0.950	0.970
-0.87	0.987	0.961	0.974
-0.67	0.985	0.972	0.979
-0.47	0.983	0.980	0.982
-0.27	0.981	0.984	0.982
-0.07	0.979	0.987	0.983
0.13	0.976	0.989	0.982
0.33	0.973	0.991	0.982
0.53	0.967	0.992	0.980
0.73	0.964	0.993	0.978
0.93	0.959	0.994	0.976
1.13	0.952	0.994	0.973
1.33	0.946	0.995	0.970
1.53	0.939	0.995	0.967
1.73	0.931	0.996	0.963
1.93	0.923	0.996	0.959

transformer-based multilevel filtering framework. *Acc*, R_{PD}, and R_{NPD}, as shown in Fig. 4.9, are adopted as evaluation parameters. Table 4.3 selects the results of 10 typical TH_B for display. The values of the parameters in Fig. 4.9 and Table 4.3 are the average of the results of six groups of repeated experiments.

It can be seen from Fig. 4.9 and Table 4.3 that as TH_B increases, the conditions under which the detected signal is judged as PD Prediction are becoming increasingly harsh, while the conditions for NonPD Prediction are gradually loose. When TH_B is over 0.51, R_{PD} sharply decreases. This also results in the sharp drop of *Acc*. On the contrary, the rise of R_{NPD} is not obvious. Considering the emphasis on PD detection, it is suggested to reduce TH_B appropriately so as to make the proposed transformer-based multilevel filtering framework conducive to the accurate detection of PD activities. TH_B is easily adjusted in the practical application according to the specific detection environments and tasks. An appropriate reduction of TH_B helps to reduce the missed detection of PD activities,

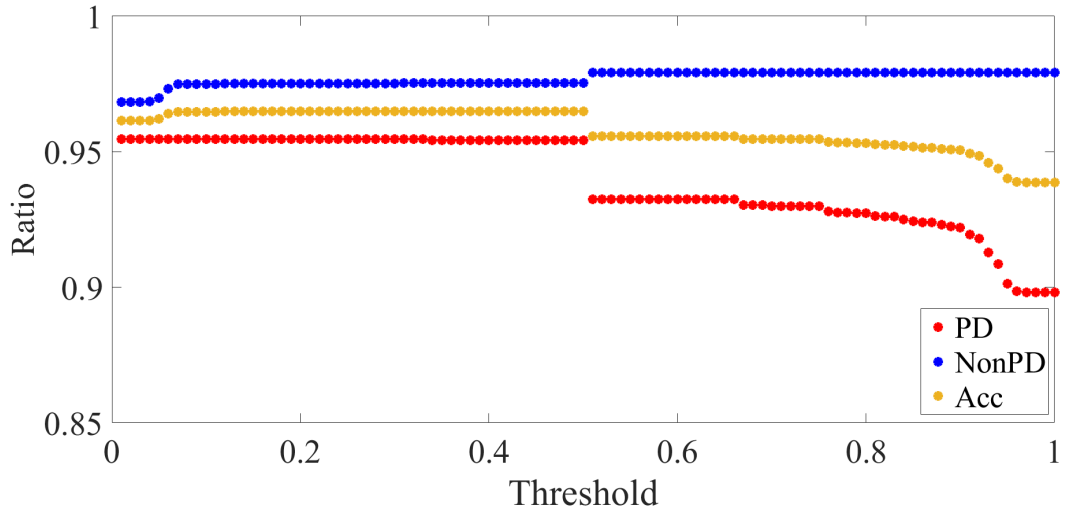


Figure 4.9: R_{PD} in red dots, R_{NPD} in blue dots, and Acc in yellow dots of different TH_B.

Table 4.3: Analysis of TH_B

TH_B	R_{PD}	R_{NPD}	Acc
0.01	0.955	0.968	0.961
0.11	0.955	0.975	0.965
0.21	0.955	0.975	0.965
0.31	0.955	0.975	0.965
0.41	0.954	0.975	0.965
0.51	0.932	0.979	0.956
0.61	0.932	0.979	0.956
0.71	0.930	0.979	0.955
0.81	0.926	0.979	0.953
0.91	0.919	0.979	0.949

while an appropriate increase of TH_B helps to reduce the misjudgment of normal working signals. In this thesis, we select 0.31 as TH_B for the PD-related data set of MV overhead power lines. TH_B of 0.31 leads to the highest Acc and R_{NPD}, while R_{PD} has not decreased sharply.

4.3.5 Comparison with Other Techniques

The experimental results of the transformer-based multilevel filtering framework are compared to other PD detection methods. The comparison of Acc, R_{PD}, and R_{NPD} are listed in Table 4.4. "SVM" and "CNN" are the same methods

under comparison described in Section 3.3.5. The "LSTM" method presents the results from the LSTM-based resolution proposed in Chapter 3. All the four PD detection resolutions under comparison adopt the same noise reduction scheme described in Section 4.3.1. The "SVM", "CNN", and "LSTM" methods employ different feature extraction methods after the noise reduction, while the proposed transformer-based multilevel filtering framework integrates the detection results according to the analysis of all the possible PD measurements. All the four PD detection resolutions are verified by the whole VSB PD-related data set with repeated experiments to obtain the evaluation parameter values listed in Table 4.4.

It can be seen that for the compared machine learning methods, the feature extraction methods significantly affect the function display of the artificial intelligence models. Both CNN and LSTM are considered to have excellent sequence classification ability. However, with different feature extraction methods, the performance of the two neural networks differs a lot for the same PD detection task. Such differences reveal the lack of adaptability of such detection methods to the dynamic PD monitoring environment and measurement data. Compared with the other PD detection methods, the transformer-based multilevel filtering framework obtains better detection performance with less manual intervention. The previous LSTM-based resolution depends on 13 carefully selected features, while the transformer-based multilevel filtering framework proposed in this chapter is more intelligent and it obtains better results in all the three evaluation parameters. It has higher adaptability and robustness for wider applications for PD detection of real-world power systems.

Table 4.4: Comparison of the Transformer-Based Multilevel Filtering Framework and Other Methods

Method	R_{PD}	R_{NPD}	Acc
SVM	0.665	0.737	0.701
CNN	0.742	0.809	0.775
LSTM	0.937	0.971	0.954
Proposed Resolution	0.955	0.975	0.965

4.4 Summary

This thesis has proposed a novel transformer-based multilevel filtering framework for PD detection. It combines the advanced and intelligent computation ability of the transformer algorithm and the multilevel filtering mechanism that makes the whole framework more applicable and robust to complicated and dynamic power systems and various practical engineering applications. The transformer-based classifier exploits the advantages of the data-driven approach with a self-attention mechanism to accurately interpret the PD-related data with less manual intervention. And it is famous for its excellent ability of transfer learning, which means the well-trained classifier for one specific PD detection task can also be a valuable experience for the other detection tasks. The proposed transformer-based multilevel filtering framework is verified by the VSB PD-related data set. Extensive experiments that demonstrate the superior performance of the proposed transformer-based multilevel filtering framework for the large-scale online measurement data of complex real environments have been conducted. It provides an economic and optimal resolution for PD monitoring of MV overhead power lines. The easily adjustable parameters make it more suitable for the promotion of real engineering applications. It provides a valuable tool for insulation inspection which helps to maintain the economical and safe operation of power systems and ensures the quality continuous supply of electrical energy.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

This thesis has dedicated to the novel and effective resolutions for PD detection of MV overhead power lines. The data-driven machine learning-based approaches are employed for the interpretation of the large-scale online PD measurement data. The research is on the basis of the large PD-related data set of MV overhead power lines shared by VSB.

An LSTM-based classifier with an optimal signal processing scheme is firstly proposed as the PD detection resolution. The proposed LSTM-based resolution greatly coordinates the multiple processing and interpretation procedures for the large-scale data disturbed with measurement noises. The loop optimization mechanism is established to combine the noise reduction, feature extraction, and the LSTM neural network as a whole with internal linkages. The optimal combination has combined advantages and can overcome each individual's drawbacks. The proposed method outperforms other techniques in the literature based on the experimental results. Its practicality and effectiveness are verified by the VSB PD-

related data set. Moreover, the common loop optimization rules for PD detection solutions have been summarized based on extensive experiments and analysis. The loop optimization rules are competent for the optimization of wider engineering applications of PD detection.

A transformer-based multilevel filtering framework is then proposed to reduce the manual intervention of the complicated signal processing scheme in the previous LSTM-based PD detection resolution. It combines the advanced and intelligent computation ability of the transformer algorithm and the multilevel filtering mechanism that makes the whole framework more applicable and robust to complicated and dynamic power systems and various practical engineering applications. The proposed transformer-based multilevel filtering framework is also verified by the VSB PD-related data set. Extensive experiments that demonstrate the superior performance of the proposed transformer-based multilevel filtering framework for the large-scale online measurement data of complex real environments have been conducted. It has outperformed the other techniques in the literature based on the experimental results. And it has also made improvements from the previous LSTM-based PD detection resolution. It provides an economic and optimal resolution for PD monitoring of MV overhead power lines. The easily adjustable parameters make it more suitable for the promotion of real engineering applications. It provides a valuable tool for insulation inspection which helps to maintain the economical and safe operation of power systems and ensures the quality and continuous supply of electrical energy.

5.2 Recommendations for Further Research

Intelligent PD detection resolutions have wide application prospects in various fields. Further research for intelligent PD detection resolutions is planned from the following aspects:

1. The quality of the ground truth significantly affects the performance of the supervised learning methods. To develop the supervised learning applications for PD monitoring, the biggest challenge is to balance the complex real-world measurement data and the ground truth for the machine learning models. More research works will be conducted with more data from real-world measurements, experiments, and simulations. The next objective is to not only detect the existence of the PD activities but also recognize the location, type, and severeness level of the PD activities.

2. The development of the self-supervised learning and unsupervised learning methods introduces new resolutions for the fault diagnosis. Next, we will study the unsupervised learning methods combined with model-based PD detection methods or data-driven PD detection methods. The goal is to reduce the limitation imposed by the limited labelled data.

3. The model-based PD detection methods depend on the specific electrical models, while the data-driven PD detection methods depend on the specific PD measurements. With the development of the machine learning technologies such as transfer learning and meta-learning, it is a meaningful research objective to propose adaptive and intelligent PD detection resolutions that are robust for the different monitoring environments and dynamic electrical systems.

List of Publications

The Author's contribution in this M.Eng thesis are summarized in the following publications:

Journal Publications:

1. **Ning Xu**, Hoay Beng Gooi, Lipo Wang, Yuanjin Zheng, Wensong Wang, Jiawei Yang, "Noise-Reduced LSTM Neural Network Based on Signal Processing and Analysis for PD Detection," *IEEE Industry Applications Society* (*Under second-round review*).

Conference Publications:

1. **Ning Xu**, Hoay Beng Gooi, Lipo Wang, Yuanjin Zheng, Jiawei Yang, "Partial Discharge Detection Based on Long Short-Term Memory Neural Network Classifier with Efficient Feature Extraction Methods," *The 2021 IEEE 12th Energy Conversion Congress and Exposition – Asia* (*ECCE-ASIA*), 24-27 May 2021, Singapore.

Bibliography

[1] Hossein Dadashi Ilkhechi and Mohammad Hamed Samimi. Applications of the acoustic method in partial discharge measurement: A review. *IEEE Transactions on Dielectrics and Electrical Insulation*, 28(1):42–51, Feb. 2021.

[2] Shibo Lu, Hua Chai, Animesh Sahoo, and B. T. 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*, 27(6):1861–1888, Dec. 2020.

[3] Michal Krátký, Stanislav Mišák, Petr Gajdoš, Petr Lukáš, Radim Bača, and Peter Chovanec. A novel method for detection of covered conductor faults in medium voltage overhead line systems. *IEEE Transactions on Industrial Electronics*, 65(1):543–552, Jan. 2018.

[4] Yasutomo Kakimoto, Hina Yoshikawa, Tatsuki Jogo, Toshiyuki Wakisaka, Masahiro Kozako, Masayuki Hikita, Hidefumi Sato, Masahiro Soeda, and Hideki Tagashira. Application of novel online partial discharge monitoring system using power line communication to noise and pd source discrimination. In *2020 8th International Conference on Condition Monitoring and Diagnosis (CMD)*, pages 130–132, Oct. 2020.

[5] IEC Standard et al. High-voltage test techniques: partial discharge measurements. *IEC-60270*, pages 13–31, Dec. 2000.

[6] David A Nattrass. Partial discharge. xvii. the early history of partial discharge research. *IEEE Electrical Insulation Magazine*, 9(4):27–31, Jul. 1993.

[7] Dieter Kind and Dieter Konig. Ac breakdown of epoxy resins by partial discharges in voids. *IEEE Transactions on Electrical Insulation*, EI-3(2):40–46, May 1968.

[8] Xiaohua Zhang, Bo Pang, Yaxin Liu, Shaoyu Liu, Peng Xu, Yan Li, Yifan Liu, Leijie Qi, and Qing Xie. Review on detection and analysis of partial discharge along power cables. *Energies*, 14(22):7692, Nov. 2021.

[9] Mehrdad Majidi, Mohammed Sami Fadali, Mehdi Etezadi-Amoli, and Mohammad Oskuoee. Partial discharge pattern recognition via sparse representation and ann. *IEEE Transactions on Dielectrics and Electrical Insulation*, 22(2):1061–1070, Apr. 2015.

[10] A. Cavallini, G.C. Montanari, and F. Ciani. Analysis of partial discharge phenomena in paper-oil insulation systems as a basis for risk assessment evaluation. In *IEEE International Conference on Dielectric Liquids, 2005. ICDL 2005.*, pages 241–244, Jun. 2005.

[11] P.H.F. Morshuis. Degradation of solid dielectrics due to internal partial discharge: some thoughts on progress made and where to go now. *IEEE Transactions on Dielectrics and Electrical Insulation*, 12(5):905–913, Oct. 2005.

[12] Qasim Khan, Shady S. Refaat, Haitham Abu-Rub, and Hamid A. Toliyat. Partial discharge detection and diagnosis in gas insulated switchgear: State of the art. *IEEE Electrical Insulation Magazine*, 35(4):16–33, Jul. 2019.

[13] Viesturs Zimackis and Sandra Vitolina. Simulation of direct lightning strike in medium voltage covered conductor overhead line with arc protection device. In *2017 IEEE 58th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*, pages 1–4, Oct. 2017.

[14] Junhua Luo, Jikang Shi, and Jian Yuan. Study on surface discharge of composite dielectric in xlpe power cable joints. In *Proceedings: Electrical Insulation Conference and Electrical Manufacturing and Coil Winding Conference (Cat. No.01CH37264)*, pages 341–343, Oct. 2001.

[15] S. Misák, J. Fulnecek, T. Vantuch, T. Buriánek, and T. Jezowicz. A complex classification approach of partial discharges from covered conductors in real environment. *IEEE Transactions on Dielectrics and Electrical Insulation*, 24(2):1097–1104, Apr. 2017.

[16] R. Ghosh, B. Chatterjee, and S. Dalai. A method for the localization of partial discharge sources using partial discharge pulse information from acoustic emissions. *IEEE Transactions on Dielectrics and Electrical Insulation*, 24(1):237–245, Feb. 2017.

[17] E. Lemke. A critical review of partial-discharge models. *IEEE Electrical Insulation Magazine*, 28(6):11–16, Nov. 2012.

[18] A Pedersen, GC Crichton, and IW McAllister. Partial discharge detection: theoretical and practical aspects. *IEE Proceedings-Science, Measurement and Technology*, 142(1):29–36, Jan. 1995.

[19] Shi-You Wu and Shu-Sheng Zheng. Detection of partial discharge in gis and transformer under impulse voltage by fluorescent optical fiber sensor. *IEEE Sensors Journal*, 21(9):10675–10684, May 2021.

[20] Amanda Binotto, Bruno Albuquerque de Castro, Jorge Alfredo Ardila-Rey, and André Luiz Andreoli. Partial discharge detection of transformer bushing based on acoustic emission and current analysis. In *2021 IEEE XXVIII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pages 1–4, Aug. 2021.

[21] Sorokhaibam Nilakanta Meitei, Kunal Borah, and Saibal Chatterjee. Partial discharge detection in an oil-filled power transformer using fiber bragg grating

sensors: A review. *IEEE Sensors Journal*, 21(9):10304–10316, May 2021.

[22] Sajjad Sharifinia, Mehdi Allahbakhshi, Teymoor Ghanbari, Asghar Akbari, and Hassan Reza Mirzaei. A new application of rogowski coil sensor for partial discharge localization in power transformers. *IEEE Sensors Journal*, 21(9):10743–10751, May 2021.

[23] HF Ye, Y Qian, YD Liu, RL Chen, GH Sheng, and XC Jiang. Partial discharge detection technology based on the fibre bragg grating. *Australian Journal of Electrical and Electronics Engineering*, 11(2):217–225, Nov. 2014.

[24] Qing Xie, Xiong Liu, Junhan Tao, Tong Li, Shuyi Cheng, and Fangcheng Lu. Experimental verification of the sparse design of a square partial discharge acoustic emission array sensor. *Measurement Science and Technology*, 26(4):045101, Feb. 2015.

[25] Kun Tang, Jiaqi Tao, Yuan Yan, Junbai Chen, and Xianyu Yue. Realization of rogowski coil for partial discharge measurement. In *2021 6th International Conference on Power and Renewable Energy (ICPRE)*, pages 46–50, Sep. 2021.

[26] Stanislav Mišák, Štefan Hamacek, Petr Bilík, Marek Hofinek, and Petr Petvaldský. Problems associated with covered conductor fault detection. In *11th International Conference on Electrical Power Quality and Utilisation*, pages 1–5, Oct. 2011.

[27] S. Mišák and V. Pokorný. Testing of a covered conductor's fault detectors. *IEEE Transactions on Power Delivery*, 30(3):1096–1103, Jun. 2015.

[28] David Gonzalez-Jimenez, Jon Del-Olmo, Javier Poza, Fernando Garramiola, and Patxi Madina. Data-driven fault diagnosis for electric drives: A review. *Sensors*, 21(12):4024, Jun. 2021.

[29] Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta,

Ashish Khanna, K. Shankar, and Gia Nhu Nguyen. An effective training scheme for deep neural network in edge computing enabled internet of medical things (iomt) systems. *IEEE Access*, 8:107112–107123, Jun. 2020.

[30] Benjamin Maschler and Michael Weyrich. Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning. *IEEE Industrial Electronics Magazine*, 15(2):65–75, Jun. 2021.

[31] Haresh Kumar, Muhammad Shafiq, Ghulam Amjad Hussain, and Kimmo Kauhaniemi. Comparison of machine learning algorithms for classification of partial discharge signals in medium voltage components. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pages 1–6, Oct. 2021.

[32] Ebrahim Balouji, Thomas Hammarström, and Tomas McKelvey. Partial discharge classification in power electronics applications using machine learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, Nov. 2019.

[33] Wong Jee Keen Raymond, Chong Wan Xin, Lai Weng Kin, and Hazlee Azil Illias. Noise invariant partial discharge classification based on convolutional neural network. *Measurement*, 177:109220, Jun. 2021.

[34] Hazlee Illias, Gamil Altamimi, Norrima Mokhtar, and Hamzah Arof. Classification of multiple partial discharge sources in dielectric insulation material using cepstrum analysis–artificial neural network. *IEEJ Transactions on Electrical and Electronic Engineering*, 12(3):357–364, Dec. 2017.

[35] Jinliang Yin, Xuesong Zhou, Youjie Ma, Yanjuan Wu, and Xiaoning Xu. Power transformer fault diagnosis based on multi-class multi-kernel learning relevance vector machine. In *2015 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 217–221, Aug. 2015.

[36] T. Hucker and H.-G. Krantz. Requirements of automated pd diagnosis systems for fault identification in noisy conditions. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2(4):544–556, Aug. 1995.

[37] A. Contin, A. Cavallini, G.C. Montanari, G. Pasini, and F. Puletti. Digital detection and fuzzy classification of partial discharge signals. *IEEE Transactions on Dielectrics and Electrical Insulation*, 9(3):335–348, Jun. 2002.

[38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[39] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, Jul. 2019.

[40] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955, Dec. 2020.

[41] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, Aug. 2015.

[42] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, Aug. 2017.

[43] Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2):68–75, Feb. 2017.

[44] Ayaz Hussain Bukhari, Muhammad Asif Zahoor Raja, Muhammad Sulaiman, Saeed Islam, Muhammad Shoaib, and Poom Kumam. Fractional neuro-sequential arfima-lstm for financial market forecasting. *IEEE Access*, 8:71326–71338, Apr. 2020.

[45] Musaed Alhussein, Khursheed Aurangzeb, and Syed Irtaza Haider. Hybrid cnn-lstm model for short-term individual household load forecasting. *IEEE Access*, 8:180544–180557, Oct. 2020.

[46] Tae-Young Kim and Sung-Bae Cho. Predicting residential energy consumption using cnn-lstm neural networks. *Energy*, 182:72–81, Sep. 2019.

[47] Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons & Fractals*, 140:110212, Nov. 2020.

[48] Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, 135:109864, Jun. 2020.

[49] Na Qu, Zhongzhi Li, Jiankai Zuo, and Jiatong Chen. Fault detection on insulated overhead conductors based on dwt-lstm and partial discharge. *IEEE Access*, 8:87060–87070, May 2020.

[50] Ming Dong and Jessie Sun. Partial discharge detection on aerial covered conductors using time-series decomposition and long short-term memory network. *Electric Power Systems Research*, 184:106318, Jul. 2020.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Dec. 2017.

[52] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, Feb. 2021.

[53] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

arXiv preprint arXiv:1810.04805, May 2018.

[54] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, Mar. 2020.

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, Feb. 2019.

[56] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, Jul. 2019.

[57] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, Oct. 2020.

[58] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, Nov. 2020.

[59] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, Jun. 2020.

[60] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829, Feb. 2021.

[61] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, Oct. 2021.

[62] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*, Aug. 2021.

[63] Zhengtao Li, Guokun Chen, and Tianxu Zhang. A cnn-transformer hybrid approach for crop classification using multitemporal multisensor images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:847–858, Feb. 2020.

[64] Zhou Lei, Congcong Zhou, Shengbo Chen, Yiyong Huang, and Xianrui Liu. A sparse transformer-based approach for image captioning. *IEEE Access*, 8:213437–213446, Sep. 2020.

[65] Fuping Zhang, Pengcheng Zhao, and Jianming Wei. Channel transformer network. *IEEE Access*, 8:220762–220778, Dec. 2020.

[66] Alvaro Arcos-Garcia, Mario Soilan, Juan A Alvarez-Garcia, and Belén Riveiro. Exploiting synergies of mobile mapping sensors and deep learning for traffic sign recognition systems. *Expert Systems with Applications*, 89:286–295, Dec. 2017.

[67] Ji He, Lina Zhao, Hongwei Yang, Mengmeng Zhang, and Wei Li. Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):165–178, Sep. 2019.

[68] Yuan Yuan and Lei Lin. Self-supervised pretraining of transformers for satellite image time series classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:474–487, Nov. 2021.

[69] Xiu Zhou, Xutao Wu, Pei Ding, Xiuguang Li, Ninghui He, Guozhi Zhang, and Xiaoxing Zhang. Research on transformer partial discharge uhf pattern recognition based on cnn-lstm. *Energies*, 13(1):61, Dec. 2019.

[70] Hui Song, Jiejie Dai, Gehao Sheng, and Xiuchen Jiang. Gis partial discharge pattern recognition via deep convolutional neural network under complex data source. *IEEE Transactions on Dielectrics and Electrical Insulation*, 25(2):678–685, Apr. 2018.

[71] Angran Gao, Yongli Zhu, Weihao Cai, and Yi Zhang. Pattern recognition of partial discharge based on vmd-cwd spectrum and optimized cnn with cross-layer feature fusion. *IEEE Access*, 8:151296–151306, Aug. 2020.

[72] Feng-Chang Gu. Identification of partial discharge defects in gas-insulated switchgears by using a deep learning method. *IEEE Access*, 8:163894–163902, Aug. 2020.

[73] Manel Rhif, Ali Ben Abbes, Imed Riadh Farah, Beatriz Martínez, and Yanfang Sang. Wavelet transform application for/in non-stationary time-series analysis: a review. *Applied Sciences*, 9(7):1345, Mar. 2019.

[74] Gheyath Othman and Diyar Qader Zeebaree. The applications of discrete wavelet transform in image processing: A review. *Journal of Soft Computing and Data Mining*, 1(2):31–43, Dec. 2020.

[75] Dengsheng Zhang. Wavelet transform. In *Fundamentals of Image Data Mining*, pages 35–44. Springer, May 2019.

[76] Renxiang Chen, Xin Huang, Lixia Yang, Xiangyang Xu, Xia Zhang, and Yong Zhang. Intelligent fault diagnosis method of planetary gearboxes based on convolution neural network and discrete wavelet transform. *Computers in industry*, 106:48–59, Apr. 2019.

[77] Peng Song, Yumei Tan, Xiaojun Geng, and Taifei Zhao. Noise reduction on

received signals in wireless ultraviolet communications using wavelet transform. *IEEE Access*, 8:131626–131635, Jul. 2020.

[78] Wai Keng Ngui, M Salman Leong, Lim Meng Hee, and Ahmed M Abdelrhman. Wavelet analysis: mother wavelet selection methods. In *Applied mechanics and materials*, volume 393, pages 953–958. Trans Tech Publ, Sep. 2013.