THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Machine Learning Analysis of State of Polarization Changes to Detect Optical Fiber Tampering

Leyla Sadighi

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2024

Machine Learning Analysis of State of Polarization Changes to Detect Optical Fiber Tampering

Leyla Sadighi

ISSN 0346-718X
This thesis has been prepared using LaTeX.

Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Reproservice
Gothenburg, Sweden, November 2024

To my family.

Abstract

Fiber optic networks are the backbone of modern communications, supporting a vast range of services, from internet connectivity to critical infrastructure operations, such as defense, healthcare, and finance. Their ability to transmit data at ultra-high rates over long distances with minimal loss makes them the preferred medium for secure and efficient data transmission. However, fiber optic installations face various security and physical damage threats which can compromise the reliability, integrity, and confidentiality of the transmitted data. The threats range from physical damage caused by accidental fiber cuts or mechanical vibrations to sophisticated eavesdropping attacks that exploit the physical properties of optical fibers to gain unauthorized access to the transmitted data. Given the critical role of fiber optic networks in today's interconnected world, ensuring their security, reliability, and resilience is paramount. Effective monitoring is a key aspect of maintaining network security, as it enables the early detection of potential threats and disturbances. Traditional monitoring systems are often limited in scope, costly, and struggle to detect more subtle disturbances like unauthorized tapping or eavesdropping. Recent advances in Machine Learning (ML) offer new avenues for enhancing the detection and diagnostics of anomalies in optical networks.

This thesis investigates the use of the State of Polarization (SOP) of light within optical fibers as a novel technique for monitoring environmental changes and detecting security threats. By employing a range of ML techniques, including supervised, semi-supervised, and unsupervised learning, this research aims at identifying and classifying disturbances that may indicate mechanical damage or security breaches. The work presented in this thesis demonstrates how SOP analysis, enhanced by advanced ML models, can improve the detection capabilities of fiber optic cables as sensing devices, providing a cost-effective and scalable solution for safeguarding data integrity, confidentiality, and network availability. The findings of this research contribute to the development of intelligent and adaptive security systems for fiber optic infrastructure.

Keywords: State of Polarization, Polarization Signature, Machine Learning, Supervised Learning, Semi-supervised Learning, Unsupervised Learning, Anomaly Detection, Mechanical Vibrations, Eavesdropping.

List of Publications

This report is based on the following publications:

[A] **Leyla Sadighi**, Stefan Karlsson, Carlos Natalino, Marija Furdek, "Machine Learning-Based Polarization Signature Analysis for Detection and Categorization of Eavesdropping and Harmful Events". Published in Optical Fiber Communications Conference and Exhibition (OFC), May, 2024.

[B] **Leyla Sadighi**, Stefan Karlsson, Lena Wosinska, Marija Furdek, "Machine Learning Analysis of Polarization Signatures for Distinguishing Harmful from Non-harmful Fiber Events". Published in 24th International Conference on Transparent Optical Networks (ICTON), July, 2024.

[C] **Leyla Sadighi**, Stefan Karlsson, Carlos Natalino, Lena Wosinska, Marco Ruffini, Marija Furdek, "Detection and Classification of Eavesdropping and Mechanical Vibrations in Fiber Optical Networks by Analyzing Polarization Signatures Over a Noisy Environment". Presented in European Conference on Optical Communication (ECOC), September, 2024.

[D] **Leyla Sadighi**, Stefan Karlsson, Carlos Natalino, Marija Furdek, "Anomaly Detection in Optical Fibers: Polarization Signature Analysis with Unsupervised and Semi-supervised Learning". Submitted to Journal of Optical Communications and Networking (JOCN), October, 2024.

Acknowledgments

The journey to this point in my academic career has been shaped and brightened by the contributions of so many people, without whom this work would not have been possible. First and foremost, I wish to thank my supervisor Dr. Marija Furdek Prekratic for her unwavering support, insightful guidance, and tireless encouragement. Your belief in my potential has pushed me to reach new heights, and your patience has been a source of strength throughout this journey. I am sincerely grateful to my co-supervisor, Dr. Carlos Natalino Da Silva, for the many insightful discussions and the guidance provided throughout this journey. Your willingness to be available whenever I needed to discuss ideas or seek advice has been invaluable, and I truly appreciate the time and support you have generously offered.

Special thanks to Prof. Lena Wosinska and Prof. Paolo Monti for all the support they provided. Your encouragement and assistance have been a great source of motivation, and I am deeply appreciative of the impact you have had on both my work and my personal development. To my colleagues and mentors at optical networks and at Chalmers who have challenged my thinking and shared their expertise, I owe a debt of gratitude.

Finally, to my family, whose love and support have been my foundation and my constant motivation. Your unwavering belief in me, through all the highs and lows, has been my greatest source of strength. I am forever grateful for your sacrifices, encouragement, and unconditional love.

This acknowledgment is dedicated to all of you. You have not only contributed to this thesis but to who I am today.

Leyla Sadighi
October 2024

Acronyms

ACC:	Accuracy
AI:	Artificial Intelligence
ARS:	Adjusted Rand Score

DBSCAN:	Density-Based Spatial Clustering of Applications with Noise
FN:	False Negatives
FNR:	False Negative Rate
FP:	False Positives
FPR:	False Positive Rate
FOCS:	Fiber Optical tactical Cable System
ML:	Machine Learning
NMS:	Network Management System
OCSVM:	One-Class Support Vector Machine
SL:	Supervised Learning
SOP:	State of Polarization
SS:	Silhouette Score
SSL:	Semi-supervised Learning
TN:	True Negatives
TNR:	True Negative Rate
TP:	True Positives
TPR:	True Positive Rate
USL:	Unsupervised Learning
XGBoost:	eXtreme Gradient Boosting

Contents

Part I

Overview

CHAPTER 1

Introduction

1.1 Introduction

Fiber optic networks have emerged as the foundational technology for modern telecommunications, transforming the way data is transmitted across vast distances. With their exceptional data transmission speeds and capability to cover long communication ranges without significant signal loss, fiber optic cables play a crucial role in interconnecting regions, cities, and entire nations. As the backbone of global telecommunications infrastructure, these networks are indispensable for facilitating the rapid, large-scale data exchange required by contemporary digital communication. The vast web of fiber optic cables spans continents, ensuring that information flow is fast, reliable, and uninterrupted, effectively forming the core of both the Internet and global telecommunication systems [1]–[3]

Due to their vital role in supporting modern communications, optical network security and integrity is more concerning than ever. The information transmitted through these networks is often highly sensitive, encompassing private communications, financial transactions, and critical government data. Consequently, fiber optic networks are enticing targets for cyber-attacks tar-

geting eavesdropping or service disruption. In eavesdropping, unauthorized light coupling techniques such as evanescent coupling, V-groove cuts, and micro/macro bending [4], [5] can be used, allowing malicious actors to gain access to sensitive data without disrupting the network operation. Such breaches can compromise the confidentiality of the transmitted data, and unauthorized access can also compromise their integrity.

In addition, optical fibers are vulnerable to a variety of disturbances that can compromise network performance, leading to degradation or interruption of overlay services. Physical damage, such as fiber cuts [6], is among the most severe threats, representing, for example, nearly 60 percent of all failures in France Telecom cable infrastructure [7]. These cuts can be caused by deliberate attempts, construction activities, or natural disasters. For instance, mechanical vibrations from heavy machinery, such as excavators operating near fiber optic installations, can put stress on the fibers, causing them to break. A single fiber cut can disrupt connections among thousands of users, resulting in substantial service outages and economic losses. According to the North American Telecommunications Damage Prevention Council report, the cost of repairing fiber optic cables in rural areas closely resembles the installation costs averaging at an impressive $75,000 per mile [8].

Ensuring confidentiality, integrity, and availability of the data transmitted over the optical network is a crucial priority, requiring robust analysis and monitoring techniques. Effective monitoring plays an important role in detecting potential security breaches and maintaining network health by enabling the identification of anomalies or tampering. Unlike traditional metal-based cables that carry electrical signals, optical fibers transmit data using light, making unauthorized access more difficult [9]. However, advanced tapping techniques have been developed that can extract data without causing noticeable signal loss [10], underscoring the need for sophisticated real-time monitoring to protect the transmitted information [2]. Challenges related to optical network security monitoring also include the need for costly specialized devices and their sparse deployment, as well as the requirement of expert knowledge to analyze the subtle changes in the signal parameter values incurred by breaches. The advances in Machine Learning (ML) and Artificial Intelligence (AI) techniques and the recent proliferation of fiber-based sensing offer a powerful tool for detection and identification of physical-layer tampering.

Optical fibers are increasingly being employed as advanced sensing devices capable of monitoring a wide range of environmental changes [11]. This sensing capability is based on the intrinsic properties of light, which is transmitted through the glass core of the fiber. External conditions, such as temperature fluctuations, mechanical vibrations, pressure changes, or radiation, can influence the fiber properties and the behavior of the light propagating along the fiber. By detecting and measuring these alterations, the fiber effectively functions as a precise sensing system. A key advantage of this technology is that every segment of the fiber optic cable can act as a potential sensor node, allowing for continuous, real-time monitoring across vast geographical areas. This eliminates the need for a dense network of individual sensors, making fiber optic sensing a powerful tool for applications that require large-scale and efficient environmental monitoring.

In the context of fiber optic sensing and security, the State of Polarization (SOP) analysis emerges as a powerful technique for detecting and analyzing disturbances within the fiber [12]. SOP refers to the orientation and characteristics of the electric field vector of light as it propagates through the optical fiber. Various environmental factors can alter the SOP [7]. These alterations serve as indicators of external influences on the fiber, providing critical data that can be leveraged for monitoring and security purposes. By continuously monitoring the SOP in real time, it is possible to detect even subtle changes in the fiber's environment, which makes SOP-based analysis particularly effective for both intrusion detection and structural health monitoring [13].

The fiber optic-based security systems are becoming more intelligent and adaptive. Recent advancements in ML/AI allow for the development of systems that not only detect disturbances but also analyze and classify them [14]. Machine learning algorithms can be trained to recognize specific patterns of activity. By analyzing the nature of changes imposed by external events onto the optical signal, these systems can differentiate between harmless disturbances and potential security threats. This ability to classify and respond to different types of disturbances can boost the effectiveness of optic network security systems, contributing to more efficient security policies and resource usage. ML models allow for the detection of potential threats with high accuracy, improving the response times and decision-making capabilities of security teams. Such intelligent systems can be deployed in various settings—from protecting critical infrastructure to monitoring large public events, where the need for

rapid detection and accurate classification of potential threats is crucial.

In conclusion, the evolving capabilities of fiber optic networks extend well beyond their traditional role for data transmission in telecommunications. Their applications in security are transforming how we monitor and protect physical and digital assets. By interpreting the changes in inherent properties of light within fibers, they provide a unique platform for real-time, continuous monitoring across vast areas. This potential, combined with the advancements in AI and ML, paves the way for the development of highly responsive, intelligent security systems capable of safeguarding optical fiber infrastructures [15]. As reliance on optical networks continues to grow, the development of robust security frameworks will be critical for ensuring the reliability and safety of modern communication and security infrastructures.

1.2 Thesis outline

This thesis is structured as follows: Chapter 2 introduces the foundational concepts of light propagation within optical fibers. This chapter delves into the SOP's theoretical framework, which is crucial for comprehending the sensing approach developed in this work. Chapter 3 discusses the application of various ML algorithms and models designed to process SOP data, improving the detection capabilities of various threats affecting optical network through techniques such as classification and anomaly detection models. Finally, Chapter 4 provides a summary of the appended papers included in this thesis.

CHAPTER 2

State of polarization variations for optical fiber sensing

Fiber optic systems are highly sensitive to external mechanical and environmental factors, which can significantly impact their performance and data transmission capabilities. Various mechanical deformations such as stress, vibration, pressure, or temperature fluctuations, influence the optical fiber by causing disturbances in the propagating light and altering the optical signal parameter values. These disturbances are often very subtle and require precise monitoring and measurement to maintain the performance and integrity of the optical network.

One of the key optical performance indicators affected by mechanical deformations and temperature variations is the state of polarization of light. Manipulations of the fiber induce distinct changes in the SOP, which can be translated into valuable data for sensing the environment and detecting tampering or other disturbances. In this chapter, we explore the concept of light propagation and SOP. We discuss how SOP is influenced by external factors in optical fibers, and how these effects can be utilized in optical network sensing applications.

2.1 The nature of light as an electromagnetic wave

In addition to its behavior as a particle, light also behaves as a wave. Light is a type of electromagnetic radiation, with a portion of its spectrum visible to the human eye. As it travels through vacuum and different materials, light behaves like a wave, exhibiting characteristics such as wavelength, frequency, and amplitude. The nature of a light wave is often characterized by its wavelength (λ) and frequency (ν), which are inherently connected: the wavelength is the spatial period of the wave, or the distance over which the wave pattern repeats itself, while the frequency refers to the number of oscillations that occur per unit time. These two properties are inversely related, with their product equaling the speed of light ($c = \lambda\nu$) [16]. The behavior and properties of light are fundamentally described by Maxwell's equations, which link electric and magnetic fields together in a comprehensive electromagnetic theory [17]. From these equations, the optical wave equation is derived, providing a mathematical description of how the electric and magnetic fields of light evolve across both space and time [18].

2.2 Understanding the state of polarization

In an optical fiber, light is guided as an electromagnetic wave, with its electric field oscillating in a specific direction, referred to as the polarization [19]. The SOP describes the orientation of the electric field vector relative to the direction of light propagation. This orientation can vary over time and space due to external factors such as stress, bending, or temperature changes in the fiber. Maintaining control over the SOP is crucial in many applications, as changes in polarization can impact signal quality and degrade transmission efficiency, but can also be used to detect environmental changes. Understanding and managing SOP is therefore essential for optical communication systems and accurate fiber-based sensing.

Mathematical representation of the SOP

In describing the SOP, one must consider the direction in which the electric field oscillates relative to the direction of wave propagation. For a light wave

traveling along the z-axis, its electric field, $\mathbf{E}(z,t)$, can be decomposed into its orthogonal components in the x and y directions. The electric field is generally expressed as [20]:

$$\mathbf{E}(z,t) = \hat{e}_x E_x(z,t) + \hat{e}_y E_y(z,t),$$

where $\hat{e}_x$ and $\hat{e}_y$ are unit vectors along the x and y axes, respectively, and $E_x(z,t)$ and $E_y(z,t)$ are the electric field components along these directions. These components oscillate over time and space, reflecting how the electric field changes as the light wave travels through the fiber.

Since light is a time-varying wave, these components are typically represented as sinusoidal functions:

$$E_x(z,t) = E_{0x}\cos(\omega t - kz),$$

$$E_y(z,t) = E_{0y}\cos(\omega t - kz + \delta),$$

where E_{0x} and E_{0y} are the amplitudes of the x and y components, ω is the angular frequency of the wave, k is the wave number, and δ is the phase difference between the x and y components. The SOP is thus determined by both the relative amplitudes E_{0x}, E_{0y} and the phase difference δ, which together define the path traced by the tip of the electric field vector over time.

Types of polarization: linear, circular, and elliptical

The nature of the SOP can vary significantly based on the amplitudes of the electric field components and their phase relationship. These variations result in three primary forms of polarization: linear, circular, and elliptical.

Linear polarization Linear polarization occurs when the electric field components are either in phase ($\delta = 0$) or completely out of phase ($\delta = \pi$) [20]. In this case, the electric field vector oscillates along a single line, maintaining a constant orientation relative to the propagation direction. This type of polarization is relatively easy to produce and control, and it is commonly utilized in laser systems and optical communication. The angle of the linear polarization is determined by the ratio of the amplitudes E_{0x} and E_{0y}.

Circular polarization Circular polarization arises when the electric field components have equal amplitudes ($E_{0x} = E_{0y}$), and there is a phase difference

of $\delta = \pm\frac{\pi}{2}$ [20]. This results in the electric field vector rotating in a circular motion in the plane perpendicular to the propagation direction. The rotation can be right-handed (clockwise) or left-handed (counterclockwise) depending on the sign of δ. Circular polarization creates a helical structure of the wave as it travels, and this property is useful in applications requiring rotational symmetry or robustness against scattering, such as in satellite communication.

Elliptical polarization Elliptical polarization is the most general form of polarization and encompasses both linear and circular polarizations as special cases. It occurs when the electric field components have different amplitudes ($E_{0x} \neq E_{0y}$) and the phase difference is neither 0 nor π, nor exactly $\pm\frac{\pi}{2}$ [20]. In this situation, the tip of the electric field vector traces out an ellipse over time, combining aspects of both linear and circular motions. The orientation and shape of the ellipse depend on the relative amplitudes and phase difference of the components. Elliptical polarization is common in natural light sources and can be intentionally produced in fiber optics for advanced signal manipulation.

State of polarization monitoring

The internal structure of an optical fiber is highly sensitive to external factors like bending, acoustic vibrations, and mechanical stress, all of which can induce rapid changes in the SOP. The fluctuations in SOP serve as indicators of external disturbances and are particularly important in the context of fiber security. Shifts in polarization can indicate potential security threats, such as unauthorized access, physical tampering, or attempts to intercept communication. Therefore, continuous monitoring of SOP becomes vital for the early detection of such security breaches, offering a robust mechanism for safeguarding against intrusion, tampering, and other forms of unauthorized interference [21].

2.3 Poincaré sphere mapping

The Poincaré sphere is a powerful geometrical tool used to represent and analyze the SOP of light. It provides a visual and mathematical means to map the orientation and nature of light's polarization onto a three-dimensional

spherical surface. This concept serves as a standard method for describing the behavior of the SOP under various influences within optical fibers.

The Poincaré sphere is a unit sphere where each point on its surface corresponds to a unique SOP. It provides a way to visualize all possible SOPs, including linear, circular, and elliptical. The coordinates on the sphere are determined by the Stokes parameters (S_0, S_1, S_2, S_3), which describe the intensity and polarization characteristics of light. However, the sphere itself is primarily concerned with the normalized Stokes parameters (S_1, S_2, S_3), which define the polarization state independent of intensity, as all SOPs map to a surface of radius 1 [20]. The Poincaré sphere is particularly useful for mapping how the SOP changes as light propagates through an optical fiber under varying conditions. Changes in SOP can be visualized as trajectories or paths on the sphere's surface.

2.4 Optical fiber as a sensing device

The inherent sensitivity of optical fiber to changes in environmental conditions can be exploited for sensing applications. Fiber Optic Sensors (FOS) offer distinct advantages over traditional sensors, making them particularly successful in certain applications, especially where conventional sensors are difficult to deploy, unfeasible, or unable to provide the same level of information. FOS provide a broad range of benefits, including compact size, longer lifetime, immunity to electromagnetic interference (EMI), the capability for multiplexing, and high sensitivity [22]–[24]. These attributes make fiber optic technology the preferred sensing solution in various sectors like healthcare or infrastructure monitoring.

Wide-spread deployment and very broad and diverse geographical coverage of optical networks makes them a valuable source of data used for sensing environmental changes caused by, e.g., earthquakes [25], and security threats [26]. In this work, we consider a system for distinguishing among eavesdropping attempts, harmful and non-harmful vibrations based on the SOP variations, as depicted in Figure 2.1.

The system comprises four main components: a source, a fiber optic transmission line, a receiver with SOP measurement capabilities, and an optical analyzer. The source, usually a laser, generates an optical signal with stable spectral and polarization properties to ensure consistent light transmission.

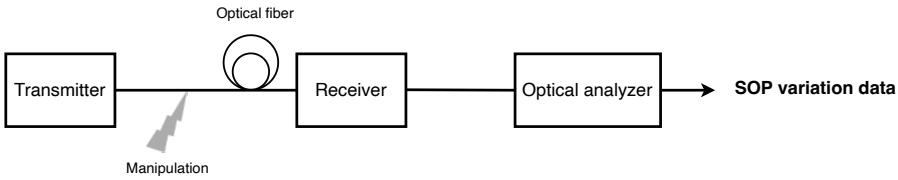

Figure 2.1: A schematic view of a system for anomaly detection based on state of polarisation changes

This light propagates through the fiber optic transmission line, which serves as a transmission medium and a sensor for detecting environmental influences and different kinds of manipulations over fiber optic installations. These external perturbations affect the SOP of the light as it travels, making the transmission line both a carrier and a sensor of SOP changes. These manipulations can be interpreted as unique SOP indicators for each specific action over a fiber installation such as eavesdropping attempts, as well as both harmful and benign mechanical vibrations. The receiver captures the optical signal traveling through the fiber, and this signal is then passed to the optical analyzer to produce meaningful digital SOP data.

The optical analyzer is responsible for deriving the SOP values from the received optical signal and processing these values to quantify the variations in the SOP. Essentially, the analyzer translates the behavior of light into meaningful data, enabling the extraction of information about the sensed parameters.

The SOP variation data can then be subjected to advanced analysis using techniques such as the ML ones described in Chapter 3, for further interpretation and diagnosis of different environmental influences over fiber optical installations. If an anomalous pattern is identified, an alarm is triggered to alert the monitoring system. The detected anomalies are further communicated to the Network Management System (NMS), which oversees network performance and responds to potential threats.

CHAPTER 3

Machine learning for SOP data analysis

SOP data offers detailed signatures of the polarization dynamics corresponding to various disturbances in optical fibers. The frequent analysis of SOP to detect abnormal patterns by human technicians would incur a high labor cost and scalability challenges. The detection of abnormal patterns in SOP data can be significantly enhanced and automated through the application of ML techniques. ML offers a robust framework for analyzing complex datasets, such as those captured through SOP variations in fiber optic-based sensors. By leveraging Supervised Learning (SL), Semi-Supervised Learning (SSL), and Unsupervised Learning (USL) techniques, it is possible to identify (and possibly classify) deviations in polarization patterns that could indicate potentially harmful events or irregularities over fiber optic installations.

The ML techniques capitalize on the distinct characteristics of polarization data, enabling the detection of even subtle changes that might otherwise remain undetected. In the following subsections, we will explore these three ML techniques, detailing how each can be applied to automate the analysis of SOP data. Additionally, we will discuss the various performance assessment metrics that can be used to evaluate the effectiveness of the employed ML algorithms.

3.1 Supervised learning

The use of SL in the context of the detection and identification of SOP patterns can be modeled as a classification problem. SL relies on the availability of labeled data, where the algorithm learns to associate specific SOP data with predefined categories, such as normal operations or particular types of disturbances. The primary benefit of SL is its high accuracy in classification tasks when sufficient labeled data is available. In this context, labeled datasets are used where each data instance is linked to a known event type. The ML model is then trained to recognize and generalize these patterns, enabling it to classify new data samples based on what was learned.

A decision tree is a fundamental ML model used for both classification and regression tasks. It works by splitting the data into subsets based on feature values, forming a tree-like structure where each internal node represents a decision based on a feature, each branch represents an outcome of the decision, and each leaf node represents a final classification or regression result. Decision trees are intuitive and easy to interpret, but they can sometimes be prone to overfitting, especially with complex datasets. Tree boosting [27] builds upon the concept of decision trees and is a highly effective and widely used supervised ML method. By combining multiple decision trees, boosting algorithms iteratively improve their predictions, focusing on reducing the errors of prior models to create a strong, composite model. Descriptions of three notable tree-boosting techniques are provided below.

Gradient boosting

Gradient boosting [27] is a method in ML known as ensemble learning. Ensemble learning combines the outputs of multiple simpler models, called weak learners, to improve overall performance. A weak learner is a model that performs slightly better than random guessing, and while its predictions may not be highly accurate on their own, combining many of them can lead to a strong, accurate model. Unlike traditional decision tree approaches, gradient boosting builds these models in a sequence, where each new tree focuses on correcting the errors made by the previous trees. By optimizing the loss function—a measure of how well the model fits the data—at each stage, the model's accuracy progressively improves. This approach is highly flexible, as gradient boosting can adapt to various loss functions, making it suitable for

classification tasks. In the context of SOP classification, gradient boosting effectively identifies and differentiates between normal and anomalous states by capturing complex patterns within the data, which contributes to more accurate monitoring and detection of disturbances in optical fiber sensing systems.

eXtreme Gradient Boosting (XGBoost)

XGBoost [28] is a scalable and efficient implementation of the gradient boosting framework, recognized for its performance and speed in classification tasks. XGBoost builds decision trees sequentially, with each new tree trained to correct the residual errors made by the ensemble of previously constructed trees. To enhance model performance and avoid overfitting, XGBoost employs regularization techniques and supports parallel and distributed computing. Its robustness in handling missing data and computational efficiency makes it particularly suitable for SOP datasets generated in optical fiber sensing systems.

Histogram Gradient Boosting (HGB)

The HGB [29] is an ensemble learning technique that can assist with problems related to regression and classification. HGB is an advanced variant of the gradient boosting algorithm, designed for efficiency when handling large-scale datasets. HGB discretizes continuous features into histograms, allowing for faster computation and memory optimization without significantly compromising model accuracy. This approach is particularly advantageous in analyzing SOP data, where the dimensionality is high, and the volume is large. By efficiently managing the computational demands, HGB achieves rapid training and prediction, making it suitable for real-time or near-real-time applications in SOP monitoring.

3.2 Semi-supervised learning

SSL is particularly advantageous in situations where labeled data is limited or difficult to obtain. In this approach, the model is trained using a training dataset containing data from normal operating conditions. The model learns the boundaries of these normal operating conditions. During inference, new

samples are analyzed considering the learned boundaries. Samples that fall outside of the learned region are flagged as anomalies. The trade-off, however, lies in the model's potential sensitivity to the quality of the training dataset; if the examples are not representative enough of the overall data distribution, the model's performance may suffer, e.g., by flagging normal samples as anomalies. The following part provides an overview of One-Class Support Vector Machine (OCSVM), a widely utilized semi-supervised algorithm for anomaly detection.

One-Class Support Vector Machine

OCSVM [30] is a powerful ML algorithm widely used for anomaly detection, particularly in scenarios where the data (predominantly) represents normal behavior, with the goal of identifying outliers or anomalies that deviate from this norm. OCSVM operates in a semi-supervised manner by learning the boundary that encapsulates the majority of the data points, which are assumed to represent normal working conditions. The algorithm maps the input data into a high-dimensional feature space using a kernel function and then constructs a decision boundary that maximizes the separation between the origin and the data points in this feature space. Data points that fall outside this boundary are detected as anomalies. The choice of kernel function and hyperparameters, such as the kernel coefficient gamma (γ) and the regularization parameter nu (ν), are critical in determining the model's sensitivity to outliers and its overall performance. During the execution of the model, it detects new data points as either normal (if they fall within the boundary) or abnormal (if they fall outside the boundary).

3.3 Unsupervised learning

USL is particularly valuable in scenarios where no labeled data is available, enabling the model to identify patterns or clusters within the data. This approach is ideal for uncovering unknown or unexpected anomalies in the optical network without the need for prior labeling. USL works by analyzing a sequence of data samples, and assessing whether or not there are samples that differ from the majority of other samples based on the density of samples in a certain region. The samples that diverge are then flagged as anomalies. USL is crucial for detecting novel and unforeseen anomalies in SOP data that have

not been previously encountered or labeled. This capability is particularly important for maintaining the security and reliability of optical networks, where new types of disturbances or attacks could emerge unannounced.

The following part provides an overview of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) as a powerful USL technique.

Density-Based Spatial Clustering of Applications with Noise

DBSCAN is an unsupervised ML algorithm particularly effective for clustering data points in high-dimensional spaces. Unlike OCSVM, which requires a training phase to learn the boundary of normal behavior, DBSCAN does not involve any training process. Instead, it analyzes a sequence of data points and identifies clusters based on their density, making it robust in discovering clusters of arbitrary shape and isolating outliers. The DBSCAN algorithm operates by defining two key parameters: the neighborhood radius (ϵ) and the minimum number of points required to form a dense region ($MinPts$). A sample is considered a core point if it has at least $MinPts$ neighbors within the radius ϵ. Clusters are formed by core points that are within ϵ of each other. Points that do not belong to any cluster and have fewer than $MinPts$ neighbors are classified as noise or outliers. Since DBSCAN does not rely on a training process, it can be directly applied to a sequence of consecutive data points, making it particularly flexible and adaptable to various conditions. The number of consecutive samples included in the system depends on the stability of the system under analysis. Highly stable systems can be analyzed by including a few tens of samples. More dynamic systems, such as optical networks which usually have tens of co-propagating channels, may require up to several hundreds of samples to achieve good performance.

The choice of ϵ and $MinPts$ significantly impacts the performance of DBSCAN. A small ϵ may result in many data points being classified as noise, while a large ϵ can lead to formation of fewer clusters, and omission of detecting potential anomalies. Similarly, $MinPts$ determines the minimum number of points to form a cluster, influencing how DBSCAN differentiates between clusters and noise. By tuning these parameters, DBSCAN can be applied to detect anomalies in SOP data, where abnormal patterns manifest as sparse or isolated points in the feature space.

3.4 Assessment metrics in machine learning

In machine learning, especially in anomaly detection and classification problems, several performance metrics are used to assess the effectiveness of models, providing a comprehensive understanding of their predictive capabilities. At the heart of this evaluation is the *confusion matrix*, which offers a detailed breakdown of a model's predictions. It records four key outcomes: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These metrics give a holistic view of how well a model distinguishes between normal and anomalous instances.

A TP is when the model correctly identifies an anomalous instance as an anomaly, whereas an FP occurs when the model incorrectly labels a normal instance as an anomaly. A TN is when the model accurately identifies a normal instance as normal, and an FN arises when the model fails to detect an anomaly, labeling an anomalous instance as normal. Each of these elements is critical for understanding the model's behavior, especially when it comes to balancing the detection of true anomalies while minimizing false alarms.

Key metrics derived from the confusion matrix include:

- **True Positive Rate (TPR)**, also known as *Recall* or *Sensitivity*, measures the proportion of actual anomalies that the model correctly identifies. It is a critical metric for assessing how well the model minimizes missed detections:
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.1}$$
- **False Positive Rate (FPR)** evaluates the proportion of normal instances that are incorrectly classified as anomalies, providing insight into the model's trade-off between detecting true anomalies and avoiding false alarms:
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{3.2}$$
- **Accuracy (ACC)** represents the overall proportion of correct predictions made by the model, considering both normal and anomalous instances:
$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3.3}$$
- **Precision** is particularly informative in anomaly detection as it indi-

cates the proportion of true anomalies among all instances predicted as anomalies. A high precision score means fewer false positives, which is critical for reducing the cost of misclassifying normal data as anomalies:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.4}$$

- The **F1-Score** serves as a harmonic mean of *Precision* and *Recall*, offering a balanced metric that considers both false positives and false negatives. It is particularly useful when a single performance metric is needed to compare models, especially in scenarios where there is a trade-off between precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

In clustering-based anomaly detection algorithms, additional metrics such as the *Silhouette Score (SS)* and the *Adjusted Rand Score (ARS)* are utilized to evaluate clustering performance. The SS measures how well an object fits within its own cluster compared to other clusters, giving insight into both cluster separation and cohesiveness. A higher silhouette score indicates well-defined, distinct clusters, which is crucial for effective anomaly detection. The ARS evaluates the similarity between the clustering results and the ground truth labels, adjusting for chance, and provides a nuanced evaluation by considering both false positives and false negatives.

CHAPTER 4

Summary of included papers

4.1 Paper A

Leyla Sadighi, Stefan Karlsson, Carlos Natalino, Marija Furdek
Machine Learning-Based Polarization Signature Analysis for Detection and Categorization of Eavesdropping and Harmful Events
Published in Optical Fiber Communications Conference and Exhibition (OFC), 24-28 March, 2024, pp. 1-3, San Diego, CA, USA.
 ISBN:979-8-3503-7758-3.

Paper A presents a Machine Learning-based approach to enhance security in optical fiber networks by detecting and categorizing eavesdropping and potentially harmful and non-harmful vibrations events over fiber optic installations. It utilizes the state of polarization variations to identify unique signatures caused by different physical manipulations of the fiber optic transmission line. The methodology includes data collection from 13 experimental scenarios from three different optical cables, including military Fiber Optical tactical Cable Systems (FOCS), indoor cables, and bare single-mode G.675 bend-insensitive fiber. We conducted experiments over a number of ML algo-

rithms to select the most appropriate classifier for this 13-class classification problem. The XGBoost classifier achieves the best performance with a 92.3% accuracy in distinguishing between normal operations and potentially harmful activities. This approach automates threat detection, providing a scalable and effective solution for securing optical networks.

4.2 Paper B

Leyla Sadighi, Stefan Karlsson, Lena Wosinska, Marija Furdek
Machine Learning Analysis of Polarization Signatures for Distinguishing Harmful from Non-harmful Fiber Events
Published in 24th International Conference on Transparent Optical Networks (ICTON), 14-18 July, 2024, pp. 1-5, Bari, Italy.
 DOI:10.1109/ICTON62926.2024.10648140.

This paper introduced a method for detecting and classifying harmful and non-harmful events in optical fiber networks by leveraging machine learning to analyze changes in the state of polarization. We collected state of polarization signatures by manipulating indoor cables to mimic real-world attack scenarios. Five scenarios were examined, including normal conditions, non-harmful vibrations, eavesdropping attempts, potentially harmful vibrations, and dual-frequency vibrations (both harmful and non-harmful). By generating unique polarization signatures for each event type, we employed various machine learning classifiers to differentiate these scenarios, with the Histogram Gradient Boosting classifier achieving a high accuracy of 97.94%. This approach significantly improves the identification of physical layer anomalies in optical networks, particularly harmful events such as mechanical vibrations caused by heavy machinery activities.

4.3 Paper C

Leyla Sadighi, Stefan Karlsson, Carlos Natalino, Lena Wosinska, Marco Ruffini, Marija Furdek
Detection and Classification of Eavesdropping and Mechanical Vibrations in Fiber Optical Networks by Analyzing Polarization Signatures Over a Noisy Environment

Presented in European Conference on Optical Communication, (ECOC), 23-26 September, 2024, Frankfurt, Germany
.

In this paper, we study how polarization signatures can be recorded and classified, originating from an installed transmission line in a real-life network OpenIreland, operated by Trinity College and located under the street in Dublin, Ireland. Real-world data from two separate installations in Dublin with link lengths of 0.15 km and 10.5 km are used to evaluate the method. Our ML analysis uses data from seven real-life network signatures to differentiate between polarization patterns obtained during normal operation and those suggesting malicious vibrations and eavesdropping. We evaluate several ML algorithms to determine a suitable classifier for our seven-class classification problem. The Histogram Gradient Boosting classifier outperforms other models in the real-world dataset, achieving an accuracy of 86.5% and an F1-score of 0.866.

4.4 Paper D

Leyla Sadighi, Stefan Karlsson, Carlos Natalino, Marija Furdek
Anomaly Detection in Optical Fibers: Polarization Signature Analysis with Unsupervised and Semi-supervised Learning
Submitted in Journal of Optical Communications and Networking (JOCN), October, 2024.

In this paper, we employ OCSVM as an SSL technique and DBSCAN as an USL method to detect a wide range of anomalies in polarization signatures, including eavesdropping, harmful and non-harmful vibrations, and overlapping events. This study analyzes 13 polarization signatures across three cable types (bare fiber, FOCS, and indoor cables), simulating various normal and abnormal scenarios. The results demonstrate that OCSVM achieves high F1-scores and is effective in detecting both normal and abnormal events, particularly in complex, overlapping scenarios. DBSCAN, though less accurate, shows potential for scenarios lacking labeled data. This work underscores the potential of SSL and USL techniques to provide scalable and cost-effective anomaly detection in optical networks, enhancing security by identifying both subtle and significant disturbances and reducing the need for manual monitoring

References

[1] M. Azadeh and M. Azadeh, "Fiber optic communications: A review," *Fiber Optics Engineering*, pp. 1–27, 2009.

[2] M. P. Fok, Z. Wang, Y. Deng, and P. R. Prucnal, "Optical layer security in fiber-optic networks," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 725–736, 2011.

[3] N. Skorin-Kapov, M. Furdek, S. Zsigmond, and L. Wosinska, "Physical-layer security in evolving optical networks," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 110–117, 2016.

[4] A. Harris and P. Castle, "Bend loss measurements on high numerical aperture single-mode fibers as a function of wavelength and bend radius," *Journal of Lightwave Technology*, vol. 4, no. 1, pp. 34–40, 1986.

[5] M. Zafar Iqbal, H. Fathallah, and N. Belhadj, "Optical fiber tapping: Methods and precautions," in *8th International Conference on High-capacity Optical Networks and Emerging Technologies*, 2011, pp. 164–168.

[6] M. Hoffman, *Cable cuts*, `http://all.net/CID/Attack/papers/CableCuts.html`.

[7] J. Pesic, E. Le Rouzic, N. Brochier, and L. Dupont, "Proactive restoration of optical links based on the classification of events," in *15th International Conference on Optical Network Design and Modeling-ONDM 2011*, IEEE, 2011, pp. 1–6.

[8] Urbint, *Telecom fiber cuts: Causes, consequences, and prevention*, `https://www.urbint.com/blog/telecom-fiber-cuts-consequences`.

[9] FiberMall, *What is fiber optics? definition, advantages and uses*, `https://www.fibermall.com/blog/what-is-fiber-optics.htm?srsltid=AfmBOopuXRfSdMLHPAOQUj7WhVmT_m8ep7KrVcsqmH4eK51NN8EsEs7r`, 2024.

[10] S. Karlsson, R. Lin, L. Wosinska, and P. Monti, "Eavesdropping g. 652 vs. g. 657 fibres: A performance comparison," in *2022 International Conference on Optical Network Design and Modeling (ONDM)*, IEEE, 2022, pp. 1–3.

[11] N. Sabri, S. A. Aljunid, M. S. Salim, R. B. Ahmad, and R. Kamaruddin, "Toward optical sensors: Review and applications," *Journal of Physics: Conference Series*, vol. 423, no. 1, p. 012 064, Apr. 2013.

[12] N. Brochier, "Field measurement of polarization fluctuation dynamics and related impact for 40gbit/s submarine systems," in *Proc. SubOptic 2010 Conference, Yokohama, Japan, May*, 2010.

[13] F. Boitier, V. Lemaire, J. Pesic, *et al.*, "Proactive fiber damage detection in real-time coherent receiver," in *2017 European Conference on Optical Communication (ECOC)*, IEEE, 2017, pp. 1–3.

[14] S. Ray, "A quick review of machine learning algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 35–39.

[15] D. F. Kandamali, X. Cao, M. Tian, Z. Jin, H. Dong, and K. Yu, "Machine learning methods for identification and classification ofevents in φ-otdr systems: A review," *Appl. Opt.*, vol. 61, no. 11, pp. 2975–2997, Apr. 2022.

[16] T. Blaney, C. Bradley, G. Edwards, *et al.*, "Measurement of the speed of light," *Nature*, vol. 251, no. 5470, pp. 46–46, 1974.

[17] G. P. Agrawal, *Fiber-optic communication systems*. John Wiley & Sons, 2012.

[18] G. R. Fowles, *Introduction to modern optics*. Courier Corporation, 1989.

[19] I. Kaminow, "Polarization in optical fibers," *IEEE Journal of Quantum Electronics*, vol. 17, no. 1, pp. 15–22, 1981.

[20] E. Collett, *Polarized light in fiber optics*. SPIE Press, 2003.

[21] A. Tomasov, P. Dejdar, P. Munster, T. Horvath, P. Barcik, and F. Da Ros, "Enhancing fiber security using a simple state of polarization analyzer and machine learning," *Optics Laser Technology*, vol. 167, p. 109 668, 2023, ISSN: 0030-3992.

[22] N. Sabri, S. Aljunid, M. Salim, and S. Fouad, "Fiber optic sensors: Short review and applications," *Recent trends in physics of material science and technology*, pp. 299–311, 2015.

[23] A. K. Ghatak and K. Thyagarajan, *An introduction to fiber optics*. Cambridge university press, 1998.

[24] J. Castrellon-Uribe, *Optical fiber sensors: an overview*. IntechOpen, 2012.

[25] F. Usmani, H. Awad, E. Virgillito, *et al.*, "Earthquake early warning through terrestrial optical networks: A bi-gru attention model approach on sop data," in *Optical Fiber Communication Conference (OFC) 2024*, Optica Publishing Group, 2024, Tu3J.2.

[26] C. Natalino, M. Schiano, A. D. Giglio, and M. Furdek, "Root cause analysis for autonomous optical network security management," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 2702–2713, 2022.

[27] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[28] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[29] K. A, R. Khilar, A. U, S. R, S. G, and A. R, "Breast cancer prediction using histogram gradient boosting classifier," in *2023 3rd International Conference on Advancement in Electronics Communication Engineering (AECE)*, 2023, pp. 161–165.

[30] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of machine Learning research*, vol. 2, no. Dec, pp. 139–154, 2001.