On Explaining and Reasoning about Optical Fiber Link Problems

George Theodorou[1], Sophia Karagiorgou[1], Annamaria Fulignoli[2], and Roberto Magri[2]

[1] UBITECH LTD,
Thessalias 8 & Etolias 10,
Chalandri, Athens, PC 15231, Greece
{gtheodorou, skaragiorgou}@ubitech.eu
[2] Ericsson Telecomunicazioni Spa,
Via Anagnina 203, Roma, PC 00118, Italy
{annamaria.fulignoli,roberto.magri}@ericsson.com

Abstract. Optical fiber links are known for their high bandwidth and reliable data transmission. However, problems may still arise, affecting signal quality and network performance. These problems are usually happening due to external physical extrusion or excessive bending, insufficient transmission power, damaged connectors causing signal loss; or failures of splice tray connector. In response to increasing optical fiber link problems transparency and interpetability, various attempts have been made to bring explainability in Artificial Intelligence (AI) decision-making and reasoning processes. This paper tackles a crucial and timely topic, i.e., understand the various factors contributing to optical link problems by explaining opaque AI models with two goals: (i) either providing instance explanations for a given decision by using a local and model agnostic approach; or (ii) providing global explanations able to describe the overall logic assuming knowledge of the black box model or its internals. The scientific contribution of this paper entails novel explainable AI (XAI) models harvesting data from optical fiber link events to first derive local explanations, and then apply a hierarchical approach to educe global explanations from the local ones. The proposed approach shows that we can efficiently tackle both explanation complexity and fidelity to reason about the causes that have resulted in optical fiber link problems.

Keywords: Explainable AI · SHAP · Reasoning · Optical Fiber Links · Fault Interpretation

1 Introduction

Optical fibers have emerged as the most commonly used transmission medium in telecommunications. Service Providers depend on optical fibers to reliably transport a continuously growing amount of data traffic, thanks to the medium

huge bandwidth and low attenuation. These key attributes make fiber the predominant choice for multiple communication applications ranging from telecommunication backbone infrastructure to metro aggregation segments down to the access network serving fixed users and mobile access infrastructure (e.g. 5G Radio Access Network).

The electronic process miniaturisation paired with the advent of photonic integration has allowed the manufacturing of high-rate optical transceivers (currently up to 800 Gbps) in small pluggable formats (e.g. Small Form-factor Pluggable (SFP) or Quad Small Form-Factor Pluggable (QSFP) [19] standardised by the SFF committee [18]) directly interconnected by optical fibers. Many of these transceivers can be housed in a single telecommunication or data centre equipment allowing scaling of capacity provisioning as requested by the traffic demand trends. The huge capacity transported by these optical links requires great attention to the telecommunication infrastructure monitoring to prevent faults when possible and provide fast and reliable diagnostics when an optical link goes down. Long traffic disruptions and cumbersome troubleshooting and repair procedures would cause high operational costs for a Service Provider. Optical cables damage is one of the major issues: underground and overhead cables are exposed to numerous threats, as they traverse urban environments or travel long distances across remote sites. Construction works is, for example, a relevant cause for fiber cuts. Fiber issues, however, are not the only ones that can impact an optical link connection. For instance, optical link loss can be caused by optical disconnections at the optical distributor frames by operation personnel or by malicious actions in a cabinet, optical transceiver disconnections in a central office, or power failures either in the transceiver itself or in the hosting equipment or on-site. The classification of the root cause for a fiber link fault due to a Loss-of-Signal (LOS) is the first step in the troubleshooting of a network outage. The more rapid is the classification the faster the network operator can fix the issue or put in place further investigation actions. Current level of information to the operator is a binary information, i.e., loss-of-signal reports an alarm on/off. At current state-of-the-art, different tools exist in order to check the root-cause, but they are complex and require dedicated and expensive solutions at the central office, as well as at remote sides, i.e., Remote Fiber Monitoring System (RFTS). However, a data-driven approach can offer simpler solutions providing more precise information for a first estimation of the root cause together with the basic alarm information. To this respect, transceivers support standard Digital Diagnostics Monitoring (DDM) functions [13]. This capability allows monitoring of the transceiver operating parameters in real-time. Parameters include optical output power, optical input power, temperature, laser bias current, and transceiver supply voltage, and many other. This data is sampled by the transceiver internal sensors and made available through standardised digital interfaces to the host equipment and then, to any software application that can exploit them for diagnostic purposes. The refresh rate of this parameters is typically of several milliseconds, allowing for fast variation detections. Magri et al. [23] have shown that the received optical power from a transceiver is a

precious source of data to help the root cause analysis of an optical link fault. In particular, they have shown that different types of fault induce a different transient behavior of the received power from its steady state normal operation value to its loss due to the link disruption. In other words, each root cause leaves a different fingerprint in terms of optical power drop transient than can be captured in the form of a time series when the onset of a disruption is detected. The time series of samples can be fed to a Machine Learning (ML) classifier trained to distinguish between the different root causes. Figure 1 depicts the different optical link fault root causes and the relevant operational actions. The classification performance has been proven to be very satisfactory in distinguishing between four classes of issues (but the concept can be further generalised to other source of disruptions):

1. Fiber stress or cut: in this case the problem is on the deployed fiber and proper personnel must be sent to check the fiber with proper dedicated investigation instrumentation like Optical Time Domain Reflectometer (OTDR);
2. Connector disconnection: this may highlight that some personnel or malicious intruder is operating in part of the installation where distributor frames or patch panels are present;
3. SFP connector: this is the case when someone is disconnecting the fiber directly on the optical transceiver on the host equipment; and
4. Shutdown: in this case the issue is cause by the loss of power supply to the host equipment or within the host equipment causing the transceiver to shutdown. Dedicated personnel should then check directly in the equipment room to verify the power supply.

Summarising, each identified root cause would trigger a different intervention to solve the issue. Leveraging ML simplifies the personnel trouble-shooting task allowing for less skilled personnel to take a proper action much faster with great cost saving for the infrastructure owner. A typical scenario where this ML-driven fault classification is valuable is the radio access network for mobile communication (e.g. 5G). In this scenario, for example troubleshooting may require climbing an antenna tower with very high costs and this should be avoided if the problem is elsewhere. Besides the fault classification itself induced by ML, it is indeed important to add explanatory information to support the decision-making process and select the proper fixing or investigating actions since the classification explanation can complement the root cause feedback with the necessary reliability, trust and transparency required to reduce the risk of misinterpretation. In the following section, we investigate how to add Artificial Intelligence derived explanations (XAI) and feedback to the optical link fault classification problem. The proposed approach shows how XAI can be interpreted to highlight relevant features of the time series or time dependent behavior of the receiver power samples that had contributed to drive an ML model towards the determined class.

The reference dataset has been collected in a dedicated laboratory experimental setup. An optical link based on SFP pluggable transceivers is hosted on a commercial transponder equipment. The link includes fiber patch-cords and

optical connectors to provide means to emulate real optical link faults. Figure 1 illustrates the laboratory experimental setup for the collection of the dataset.

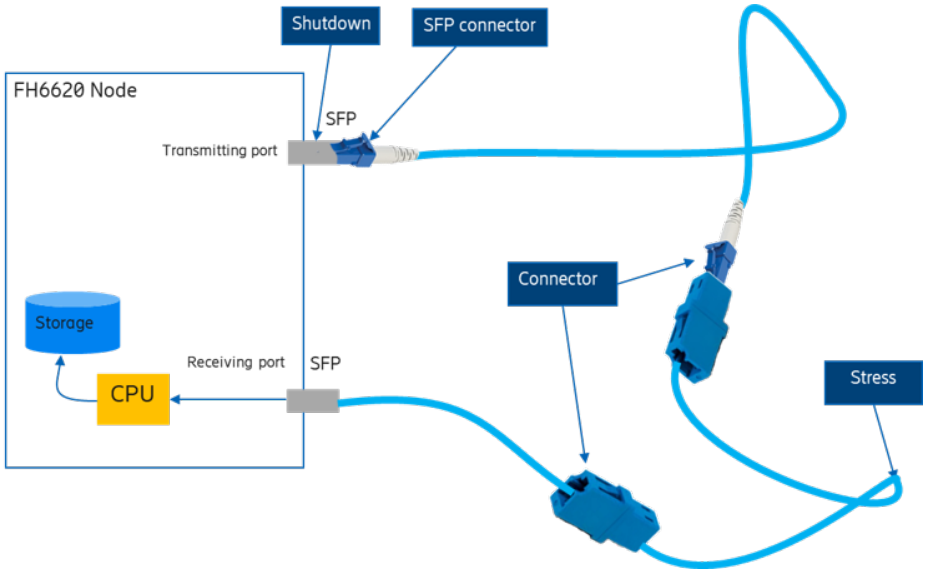

Fig. 1: Laboratory Setup for Dataset Collection.

Four possible faults have been reproduced and repeated hundreds of times in various manners. The received optical power time series have been sampled by the host node micro-controller reading the relevant information from the module memory map, stored and labelled with the corresponding fault class:

- Connector: the loss of signal is caused by the disconnection of one of the connectors between patch-cords;
- SFPConnector: the loss of signal is caused by unplugging the optical connector of the SFP module on the transmitting side;
- Shutdown: the SFP module on the transmitting side has powered off; and
- Stress: the fiber patch-cord manually stressed and broken in random ways by different persons.

This paper investigates how data collected from the laboratory experimental setup can be used to explain and reason about different aspects that cause problems to optical fiber links. The goal of this work is to harvest data from optical fiber link events to first derive local explanations, and then apply a hierarchical approach to derive global explanations from the local ones. The contributions of this paper are, as follows:

- Harvests information from events occurring within fiber optic links. These events indicate Loss-of-Signal (LOS) occurrences caused by disconnections, unplugging, normal shutdowns and stress;

- Analyse the collected data to derive local explanations for each event to identify the root cause of the issue;
- Apply a hierarchical approach towards deriving global explanations after interpreting individual events to identify broader patterns and trends across end-to-end network cases; and
- Tackle both explanation complexity and fidelity via XAI methods within an ML pipeline to gain insights into how the ML models arrive at their conclusion and reason about the causes that have resulted in optical fiber link problems.

The rest of the paper is organised, as follows: Section 2 provides the current literature review around innovative applications of AI for proactive and precise optical link fault detection in optical communications. Section 3 offers a detailed overview of the dataset. Section 4 presents the technical architecture of the explanations pipeline. Section 5 discusses our first experimental results, and Section 6 concludes the paper and provides insights to be pursued in the near future.

2 Literature Review

Numerous studies have delved into innovative applications of AI for fault detection in optical communications. Among the notable contributions, researchers have explored advanced AI techniques such as ML algorithms, Deep Learning (DL) models, and hybrid approaches combining AI with signal processing methods. These efforts have resulted in significant advancements in fault detection accuracy, response time, and robustness. The breadth of research underpins the importance of leveraging AI technologies to mitigate failures or errors in optical communication systems and improve the reliability and resilience of optical networks.

Fan, Z et al. [15] focus on the crucial role of optical networks in modern information society, where approximately 50% of optical service faults originate from errors in the optical transmission system links, monitored through frame overheads. Current maintenance methods often fail to accurately predict these errors, leading to potential disruptions in network services. To tackle this challenge, the study introduces a Gini importance-based feature selection approach and a maximum likelihood-based logistic regression algorithm for predicting performance degradation in optical transmission system links. Moreover, the study proposes a Network AI (NAI) three-level intelligent architecture to facilitate real-world application, culminating in the development of an optical transmission performance degradation prediction system based on this architecture. Experimental results from network experiments validate the effectiveness of the proposed performance degradation model and the NAI-based system.

Li X. et al [16] address the critical need for fiber fault detection in optical communication systems, particularly within established optical access networks facing challenges such as increased insertion loss. Traditional fault-detection methods struggle to capture fault echoes effectively under these circumstances.

To overcome these limitations, the paper introduces a modulation-enhanced external-cavity-resonant-frequency method utilising a laser for fault echo reception, offering superior sensitivity compared to photo-detector-based methods. While previous studies focused on sensitivity and spatial resolution, this research develops a practical model based on Lang–Kobayashi rate equations and validates it through experiments, identifying optimal detection performance associated with specific modulation parameters, including a modulation depth of 0.048, a frequency sweeping range of 0.6 times the laser relaxation oscillation frequency, and a sweeping step of 0.1 times the external cavity resonant frequency. The conclusion highlights the practical analysis of this high-sensitive fiber fault detection method, emphasising the optimal frequency-sweeping approach while considering various practical factors, recommending specific parameter settings crucial for high-sensitivity fiber fault detection and serving as a foundation for practical prototype development.

Liu, P. et al [17] introduce a customized AI module integrated into an Optical Time Domain Reflectometer (OTDR) device, along with an optical power monitoring module, to establish an AI-assisted optical network fault location mechanism tailored for high-density data center interconnections. This mechanism optimally utilizes optical link data by leveraging the AI module to predict potential link failures, subsequently monitored by the optical power module for rapid fault localization and response. Experimental testing demonstrates a significant improvement in link fault detection efficiency, with the AI model enhancing average fault detection efficiency by 98.41%, which is a significant increase in failure detection and localization within data centers.

Karandin, Oleg et al [21] demonstrate the application of XAI in fault localization within optical networks, showcasing how it enhances the comprehensibility of ML-driven decisions. Specifically, it compares the performance and interpretability of ML models trained on Optical Signal-to-Noise Ratio (OSNR) measurements collected from a single receiver monitor versus multiple monitors distributed along the network path. The conclusion drawn from the analysis highlights the stark difference in explainability between the two scenarios. While the model utilizing telemetry from a single monitor successfully identifies faults, its reasoning remains opaque due to the intricate nature of OSNR statistics. Conversely, deploying additional monitors improves both fault-localization accuracy and model explainability, albeit at increased cost. The study suggests that while comprehensive monitoring may be impractical for expansive networks, a judicious deployment of additional monitors enhances both the trustworthiness and utility of ML-based fault management tools.

Ayoub, Omran et al [22] focus on the application of XAI, particularly Shapley Additive Explanations (SHAP), in the context of light-path Quality-of-Transmission (QoT) estimation. It provides an overview of XAI and SHAP, discusses their benefits in networking, and surveys existing studies employing XAI in networking tasks. The study formulates the QoT estimation problem as a supervised binary classification task and develops an ML model using eXtreme Gradient Boosting (XGB). Through the application of SHAP, the paper demon-

strates how insights into the model's behavior can be extracted and misclassifications inspected, thereby enhancing transparency and trustworthiness. The conclusion highlights the efficacy of SHAP in providing explanations for model decisions and its utility in feature selection to improve model efficiency. Additionally, it underscores the importance of scrutinizing dataset biases to mitigate inconsistencies in model behavior. Overall, the study underscores the significance of XAI in enhancing the interpretability and trustworthiness of ML models.

However, the practical application of AI on a large scale necessitates the ability to understand the predictions and decisions made by ML or DL models. This requirement underpins the pivotal role of explainable AI (XAI) in fostering trust between AI systems and human actors. By providing transparency and clarity in the decision-making process, XAI bridges the gap between the complex algorithms employed by AI and the understanding of end users. This transparency cultivates trust and confidence to AI systems, thereby facilitating their widespread adoption and integration into various domains with high and safe guarantees. In essence, the ability of AI to offer interpretable explanations for its decisions is paramount for realising its full potential and ensuring harmonious collaboration between humans and machines.

Jacovi et al. [8] address the critical role of trust in human-AI interactions, particularly within the XAI contexts, aiming to provide clarity and structure to the concept of trust. They introduce a model of trust inspired by interpersonal trust but tailored for AI interactions, emphasising the user's vulnerability and the anticipation of AI decisions. The study advocates for explicit contracts between users and AI models to foster warranted trust, highlighting the importance of assessing risk and distinguishing between warranted and unwarranted trust. It asserts the necessity of evaluating whether trust is warranted and identifies explanation as a key factor in fostering intrinsic trust. They conclude by offering guidelines for designing trustworthy AI and emphasise the need for further research in evaluating warranted trust in human-AI interactions.

Adadi, Amina et al. [9] explore the emerging field of XAI in the context of the rapidly evolving landscape of AI applications. Recognising the critical need for transparency and trust in AI systems, the study provides a comprehensive overview of existing research on XAI, addressing fundamental questions such as what, who, when, why, where, and how. Through an interdisciplinary lens, the survey examines various approaches to XAI and identifies key trends and challenges. While acknowledging the significant impact of XAI across diverse domains, the paper also highlights the lack of formalism in problem formulation and the need for more thorough exploration of human involvement in explainability approaches.

Ribeiro et al. [12] introduce the Local Interpretable Model-Agnostic Explanations (LIME), a method for explaining the predictions of ML models in an interpretable and trustworthy manner. By learning an interpretable model locally around each prediction, LIME provides insights into the reasons behind model predictions, crucial for assessing trust in decision-making scenarios. Additionally, the paper proposes SP-LIME, a method for selecting representative

and non-redundant predictions to offer a global view of the model's behavior. Through experiments in text and image classification domains, involving both expert and non-expert users, they demonstrate the utility of explanations in various trust-related tasks, such as choosing between models, assessing trust, and improving untrustworthy models.

Lundberg et al. [10] introduce SHapley Additive exPlanations (SHAP), a unified framework for interpreting predictions, aiming to address the tension between accuracy and interpretability in complex models like ensemble or DL models. SHAP assigns importance values to each feature for a given prediction, offering a novel class of additive feature importance measures with desirable properties. By unifying six existing methods, SHAP provides insights for improved computational performance and better consistency with human intuition. The framework demonstrates a promising step towards reconciling the accuracy-interpretability trade-off in model predictions, offering a foundation for future development of interpretation methods.

Existing works are limited either by only addressing the supervised or unsupervised learning aspects of optical links failure detection, or by not leveraging the strengths of XAI to derive interpretable cause-effect insights. Compared to the above-mentioned approaches, the scientific contribution of this work is the combinatorial approach towards optical links fault detection and XAI-fuelled interpretation of the detected faults to derive more informed decisions. The differentiation of the proposed approach results in faster root cause identification, by understanding the underlying reasons behind the optical link fault detection. It also contributes to reduced downtime and minimises service disruptions, specifically in critical applications. Last, by analysing trends and patterns in explanations across different events, recurring problems within the optical network infrastructure can support proactive maintenance and reasoning about potential causes to be considered before failures occur.

3 Dataset Overview

The dataset utilized for the analysis in this paper stems from a dedicated test setup, owned by Ericsson, designed to investigate LOS events within optical networks. At the receiving side of the setup, the CPU samples the received optical power at a rate of approximately 5 milliseconds. These sampled readings are then stored in a memory buffer, capable of holding up to 1000 samples. As a consequence, each data instance within the dataset constitutes a time series of the received optical power with a maximum length of 1000 timestamps.

The time series stored within the dataset underwent a rigorous filtering process to ensure data quality and relevance for subsequent analyses. Among a larger pool of files, specific selection criteria were applied within the optical domain to identify and retain time series that meet stringent validation standards.

Firstly, the mean time delta, represented by the first column of the file, was scrutinized to ensure adherence to the nominal value of 5 milliseconds, with a tolerance of $\pm 50\%$. This criterion ensures that the temporal resolution of the time

series aligns closely with the intended sampling rate, maintaining consistency and accuracy in capturing temporal dynamics.

Furthermore, additional criteria were applied to the optical power readings within the time series. The minimum value of the time series was required to fall below the low threshold of -30 dBm, indicative of sufficient sensitivity to detect low-power signals. Conversely, the maximum value of the time series was mandated to surpass the high threshold of -17 dBm, ensuring that the dataset encompasses a diverse range of power levels encountered within optical networks.

Moreover, each retained time series was required to exhibit at least one recorded power drop, signifying a transition from a maximum value to a minimum value. This criterion ensures that the dataset captures instances of significant signal attenuation, characteristic of loss of signal events within optical networks.

The received optical power readings are expressed in decibels relative to one milliwatt (dBm), offering a standardized metric for power levels. To facilitate diverse analyses and interpretations, the dBm readings can be readily converted to milliwatts (mW) utilizing the formula illustrated below. This conversion allows for a seamless integration of the dataset with various analytical methodologies and tools, ensuring its usability across different research contexts.

$$P_{\mathrm{mW}} = 10^{\left(\frac{P_{\mathrm{dBm}}}{10}\right)} \tag{1}$$

During a LOS event within optical networks, the received optical power undergoes a significant drop from its maximum level to a "no-power" state. The precise detection of LOS events is governed by the specific module data-sheet, which outlines parameters such as assert and de-assert thresholds, as well as the relevant hysteresis range. For instance, the Ericsson SFP+ specification provides an example of such thresholds and ranges.

It is important to note that the LOS detection logic embedded within the SFP module may introduce a certain latency with respect to the actual diagnostic reading of received power. Consequently, in this project, a deliberate decision was made to utilize diagnostic values that are not synchronized by the LOS alarm. This approach ensures that the dataset captures the raw received power readings, unaffected by any potential detection delays or logic within the SFP module.

The storage of the received power file is initiated by the sampling software itself, triggered by the sampling interval configured within the experimental setup. This ensures that the dataset accurately reflects the temporal dynamics of received power, enabling detailed analyses of LOS events and their associated characteristics.

For a visual representation of the described process, refer to Figure 2, which provides an illustrative explanation of the sequence of events involved in LOS detection and dataset generation. This figure aids in understanding how the received power readings are captured, stored, and utilized for subsequent analyses within the project framework.

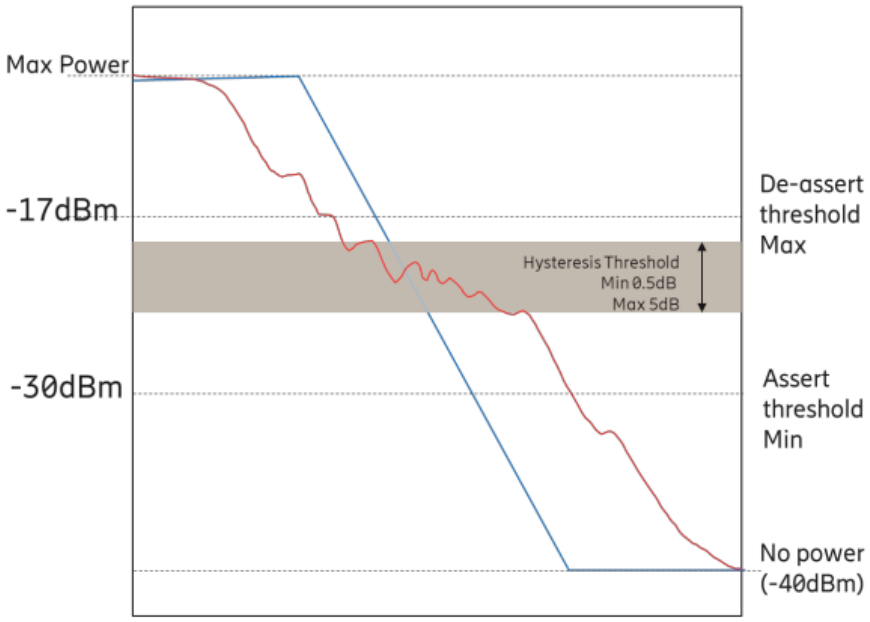

Fig. 2: LOS Event.

Overall, the dataset comprises a total of 3235 individual time series, each documenting the received power readings from an SFP module during a LOS event. Furthermore, each time series is associated with one of four fault categories under examination as shown in Figure 1. For a detailed breakdown of the dataset regarding the specific fault categories, refer to Table 1.

Fault category	Occurrences	% of Occurrences
Connector	906	28%
SFPConnector	833	26%
Shutdown	1190	37%
Stress	306	9%

Table 1: Time Series Breakdown per Fault Category.

This dataset serves as a valuable resource for analyzing LOS events within optical networks, providing insights into the temporal patterns and characteristics associated with different fault categories. Through leveraging this dataset, researchers can gain a deeper understanding of LOS phenomena and develop more effective strategies for network monitoring, fault detection, and mitigation.

4 Explanations Pipeline Architecture

An essential part of learning over the collected data is the Explanations Pipeline Architecture. In this section, we provide the pipeline and the data preparation steps followed before deriving the explanations. The explanations pipeline serves two purposes: (i) harvest the optical power loss time-series data to predict the type of fault that caused it; and (ii) generate explanations that justify the model's decision regarding the reason and type of fault.

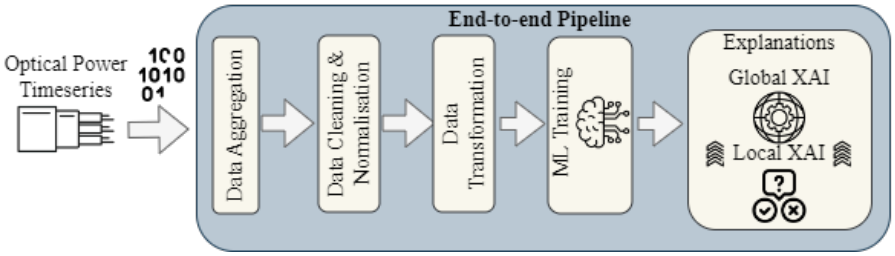

Fig. 3: Explanations Pipeline.

The explanations cultivate trust and transparency to humans and allow for their widespread adoption. As it is depicted in Figure 3, we harvest the optical power loss time-series data by first cleansing, then extracting window-based features in time and finally feeding the data to various ML models to learn patterns about each fault category. Last but not least, SHAP is employed through the Explainer Dashboard [11] to provide meaningful explanations about how models have made their classifications and thus decisions. As shown in the Algorithm 1, the pipeline first performs Data Aggregation (line 3) to create a unified dataset of all optical power time-series. Then, Data Cleansing and Normalization takes place (lines 4-5), followed by Data Transformation (line 6). ML Training is the next thing in the pipeline (line 7). Finally, local and global explanations are derived to explain the model (lines 9-10).

4.1 Data Aggregation

For data aggregation, we consolidate all time-series data pertaining to various fault categories into a unified data structure. This consolidation process enables seamless integration and organisation of diverse fault-related information, ensuring a comprehensive analysis of the system's behaviour. Additionally, we introduce a new column within this unified dataset to denote the specific type of fault associated with each observation. This additional attribute serves as a crucial label for a set of analyses conducted in Section 5, as part of our experimentation. By categorising faults within a structured framework, our pipeline facilitates systematic and targeted investigations into the distinct characteristics

Algorithm 1: Explainable AI Pipeline

Input: Optical power values OP in time-series T
Output: Local and Global Explanations on AI Model about Faults XAI_{F_L} and XAI_{F_G}

```
begin
    foreach (OP ∈ T_i) do
        OP ← Aggregation(OP)
        OP ← Cleansing(OP)
        OP ← Normalization(OP)
        OP ← Transformation(OP)
        OP ← ML_Training(OP)
    end
    XAI_FL ← ExplainLocal(OP)
    XAI_FG ← ExplainGlobal(OP)
end
```

and behaviors of different fault types, ultimately enhancing the interpretability and efficacy of the proposed explanability approach.

4.2 Data Cleansing and Normalisation

In the data cleansing phase, a crucial step involves pruning each optical power time-series up to the initial point in time when optical power loss is detected. This approach is adopted because only the data preceding the optical power loss event holds relevance in analysing the underlying reasons leading to the outage. By focusing exclusively on the data preceding the optical power loss, the explanations pipeline ensures that the subsequent analysis is centered on identifying the root causes and contributing factors associated with signal power disruptions. This meticulous data pruning strategy not only streamlines the dataset but also facilitates a more targeted and insightful examination of the events leading up to optical power losses, thereby enhancing the overall effectiveness of the proposed architecture.

4.3 Data Transformation

In this step, we adopt a systematic approach where each generated feature encapsulates the average optical received power value within a predefined time window. These time windows are constructed from the precise moment when the optical power loss event is triggered, extending backward over a specified number of time steps. Overall, 15 features are constructed ranging from a window of just the last 5 timestamps to the last 500 timestamps, see the table 5 at the Appendix. By adopting this strategy, our pipeline effectively captures the dynamic evolution of optical power values across short-term, medium-term, and long-term perspectives towards a signal disruption. In essence, our feature extraction process is tailored to encode the nuanced impact of optical power fluctuations leading up to and during the optical power loss event, thereby providing a comprehensive representation of the underlying dynamics within the system. This feature engineering not only enhances the granularity of our analysis but

also empowers our AI models to discern intricate patterns and correlations embedded within the data, facilitating more accurate and insightful predictions.

In addition to facilitating predictive accuracy, our AI pipeline prioritises the interpretability of the generated models. To achieve this, the features representing the average optical power values within specified time windows are transformed into categorical features, each assigned one of six possible values based on their quantile representation. This feature encoding scheme, see Table 2, enhances the interpretability of the features by simplifying their representation while preserving essential information about their distribution. By categorising the features in this manner, our pipeline enhances the transparency of the AI models' decision-making process, enabling end users to better understand the insights provided.

Quantile Value	Encoding
0%-25%	0
25%-50%	1
50%-75%	2
75%-85%	3
85%-95%	4
95%-100%	5

Table 2: Feature Encoding Scheme.

4.4 ML Training

During ML training, we adhere to a standard methodology by partitioning the dataset into distinct training (80%) and test sets (20%) to rigorously assess the performance accuracy of our models. Our objective is to detect the type of optical power fault based on the received power time-series data. To accomplish this task, we deploy several conventional ML models, including Random Forest, AdaBoost, KNN, and Light Gradient Boost. The selection of standard ML models is deliberate, as we prioritise interpretability and explainability, necessitating the use of ML algorithms for which features are directly provided and not inherently constructed as it happens with Deep Learning (DL) algorithms. In that way, clear understanding of how the features impact the model's decisions both at a local and global level is achieved. By employing conventional ML techniques, we ensure that our models remain transparent and interpretable, enabling stakeholders to derive actionable insights from their outputs. Through comprehensive evaluation and analysis, we achieve to develop robust and effective AI solutions capable of accurately detecting the 4 distinct optical fault types while providing meaningful explanations for their predictions.

4.5 AI Explainability

Towards AI explainability, our focus shifts to leveraging the Explainer Dashboard to extract valuable insights and understand the predictions made by these models. Powered by the SHAP TreeExplainer, the Explainer Dashboard facilitates an in-depth analysis of model behavior by offering a comprehensive view of how each feature contributes to the model predictions, enabling stakeholders to interpret model decisions with clarity and confidence both on a per case scenario as well as on a global level. Through this transparent and interpretable approach towards AI explainability, our pipeline ensures that stakeholders can extract actionable insights while fostering trust in the predictive capabilities of our models.

5 Experimental Results

This section presents the experimental results on the ML models performance regarding their precision, recall, f1 score and PR-AUC. We also evaluate the explainability of the ML models using local and then progressing to derive global explanations.

5.1 Model Performance

We conducted experiments with a variety of ML models to compare between models performance and explainability insights. Specifically, we trained four distinct traditional ML models using the time-series dataset. To evaluate the performance of these models, we computed key metrics such as precision, recall, F1-score, and PR-AUC [20]. The results, presented in Table 3 below, provide insights into the effectiveness of each model in capturing the nuances of the data and achieving the desired classification outcomes. Since, we deal with a four-class classification problem, the weighted average has been used to calculate the performance metrics.

Model	Precision	Recall	F1 Score	PRAUC
Random Forest	74%	72%	71%	75%
Ada Boost	49%	53%	49%	45%
LGBM	75%	73%	72%	75%
KNN	74%	72%	71%	71%

Table 3: Performance Metrics of Model Trained with Feature Encoding.

Upon analysing the results presented in the table above, Light Gradient Boosting Machine (LGBM), a powerful and efficient ensemble learning technique, stands out as the top-performing model among the trained models. LGBM leverages gradient boosting to construct decision trees in a sequential manner,

optimising model performance by focusing on misclassified instances. Its ability to handle large datasets efficiently and its high scalability make it a preferred choice for complex classification tasks. Given its exceptional performance in terms of precision, recall, and F1-score, LGBM emerges as a promising candidate for exploring the models' decision-making processes through Explainable AI techniques, providing valuable insights into the underlying mechanisms driving predictive outcomes.

It is worth noting, as mentioned in Section 4 regarding data transformation, that we made a deliberate decision to discretize the average power values into six distinct categories based on quantile values to aid in the interpretation of the results.

In case that we were using the average optical received power without encoding then the corresponding ML models performance would be as shown in Table 4.

Model	Precision	Recall	F1 Score	PRAUC
Random Forest	85%	85%	85%	91%
Ada Boost	54%	33%	23%	55%
LGBM	85%	85%	85%	89%
KNN	83%	83%	83%	83%

Table 4: Performance Metrics of Model Trained without Feature Encoding.

5.2 Model Explainability

We are going to perform the AI Explainability exercise by using the explainer dashboard. Explainerdashboard is a powerful library designed to swiftly construct interactive dashboards tailored for analyzing and clarifying the predictions and operations of machine learning models compatible with scikit-learn, including xgboost, catboost, and lightgbm. Explainerdashboard offers access to a plethora of features, including SHAP values, permutation importances, interaction effects, partial dependence plots, various performance metrics, and the ability to explore individual decision trees within random forests.

In the pursuit of generating comprehensive explanations, the dataset utilized encompasses both accurately classified time series and those that have been misclassified. This inclusive approach ensures a thorough examination of the model's performance across various scenarios, facilitating a deeper understanding of its predictive capabilities and potential areas for improvement.

Model Global Explainability sudo service ssh restart In Figure 4, the SHAP global features importance are depicted, revealing crucial insights into the factors influencing optical power loss prediction. Notably, the analysis highlights that events occurring within the last 5 time-steps hold the highest significance,

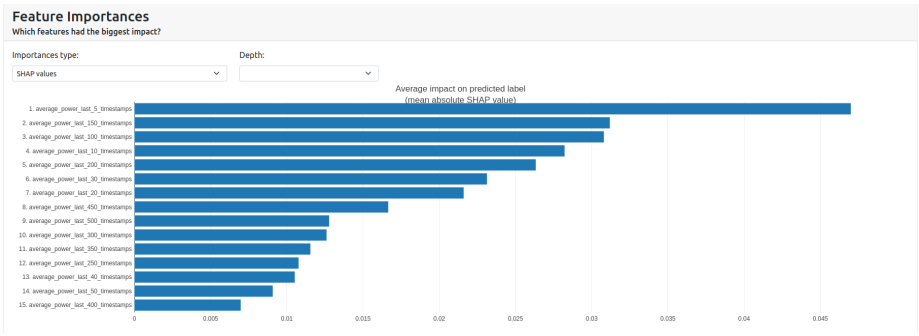

Fig. 4: SHAP Global Features Importance.

underscoring the substantial impact of short-term power drops on the model's decision-making process. Additionally, the second and third most important features correspond to events occurring approximately 150 and 100 time-steps before the optical power loss, indicating that both short-term and longer-term events play pivotal roles as predictors of future optical power loss occurrences. This finding underscores the importance of considering both short-term fluctuations and longer-term trends in understanding and predicting power system dynamics effectively. Furthermore, our analysis delved into the SHAP global feature importance categorised by fault type within our dataset.

For the Connector fault category, see Figure 5, caused by the disconnection of one of the connectors between patch-cords, the most critical feature is events within the last 5 timestamps. Hence, this finding suggests that these faults manifest abruptly and are less influenced by longer-term patterns.

Conversely, the SFP-Connector fault category, see Figure 6, caused by the direct fiber disconnection on the optical transceiver on the host equipment, exhibits a balanced dependence on both short-term (5 timestamps) and longer-term (100 timestamps before the power-loss event) features, indicating a more nuanced relationship between fault occurrence and temporal context.

In contrast, the Stress fault category, see Figure 7, caused by problems on the deployed fiber, showcases a unique pattern where a mixture of short-term and longer-term features contributes to its detection. Notably, while the SFP-Connector category prioritises short-term events within the last 5 timestamps, features spanning the last 20 to 30 timestamps are more informative for detecting fiber stress related issues.

Last but not least, the Shutdown fault category, see Figure 8, caused by the loss of power supply, the most predictive features are primarily concentrated between 5 and 30 timestamps before the power-loss event, highlighting the significance of short-term events in this context.

Our findings suggest that certain fault types exhibit a higher predictability when considering events immediately preceding the optical power loss event, emphasising the significance of short-term contextual factors. Conversely, other

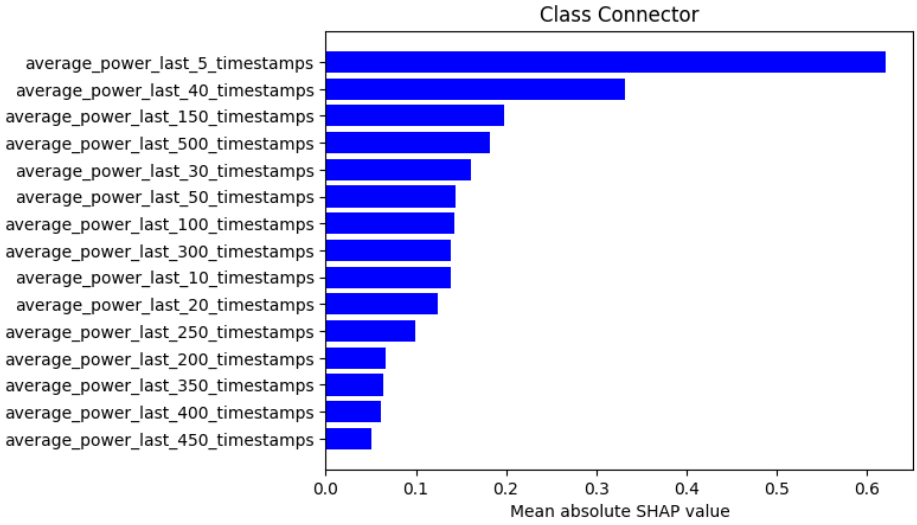

Fig. 5: SHAP Global Features Importance for the Connector Fault Type.

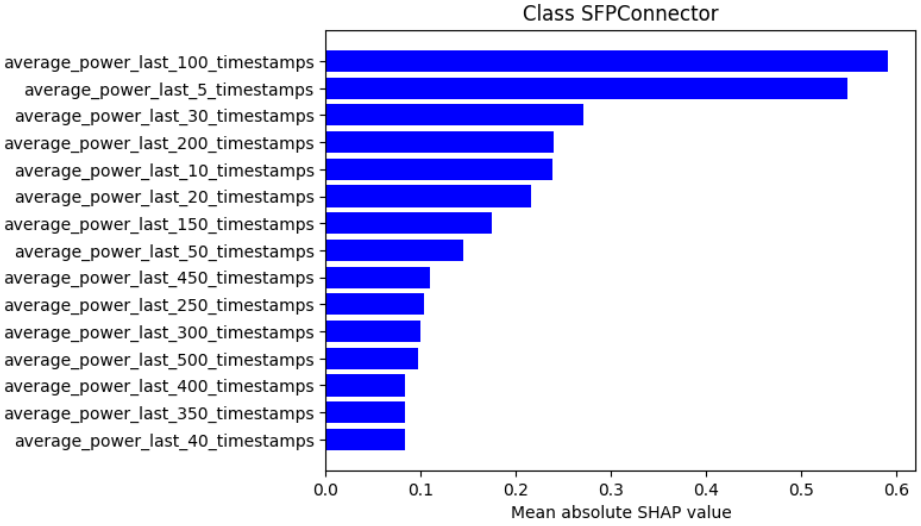

Fig. 6: SHAP Global Features Importance for the SFP-Connector Fault Type.

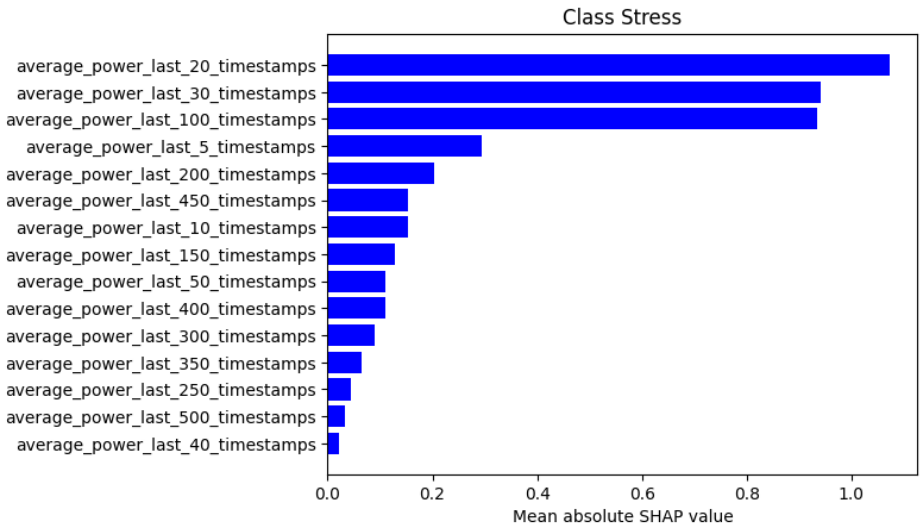

Fig. 7: SHAP-driven Global Features Importance for the Stress Fault Type.

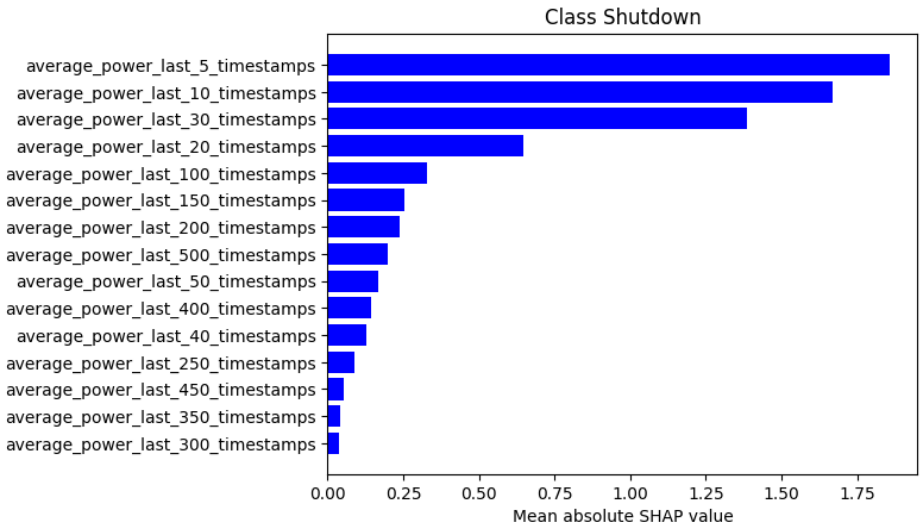

Fig. 8: SHAP-driven Global Features Importance for the Shutdown Fault Type.

fault categories demonstrate a stronger correlation with longer-term optical power level information, highlighting the importance of incorporating longer-term data to capture underlying trends and patterns. By understanding these distinctions in global feature importance across fault categories, we can enhance the accuracy and reliability of our fault prediction models tailored to specific scenarios.

Model Local Explainability Furthermore, in addition to analysing the SHAP global features importance, an equally crucial aspect of explainability in AI is justifying individual predictions, often referred to as local feature importance. The explainer dashboard offers a valuable tool to inspect the most influential features for individual predictions.

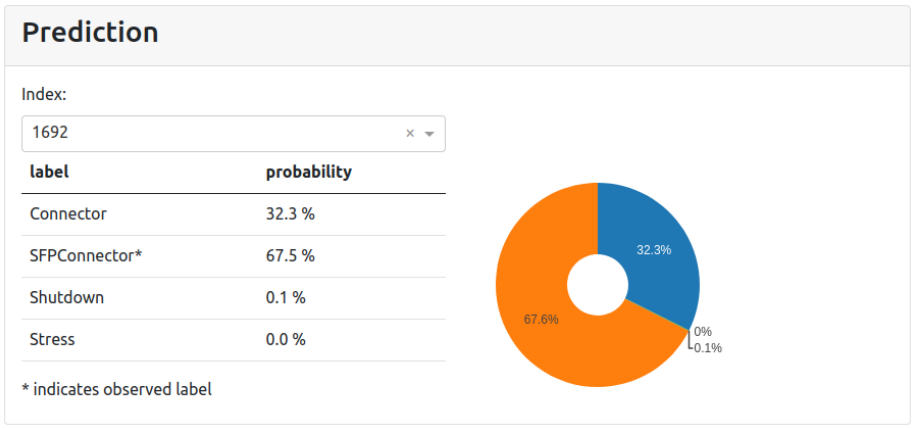

Fig. 9: Model's Individual Prediction.

For instance, in the Figure 9 the model confidently predicts the class as SFPConnector with 67.5% probability. Upon examination Figure 10, we find that this figure depicts the model's features contributions, showing green the positive contributors and red the negative ones. The yellow bar on the left shows the averaged out probability out of all classes in the dataset and the blue bar on the right is the final contribution of all the features to the model's decision. However, even more interestingly, the most significant contributing features to the specific prediction are the events occurring within the last 30, 20 and 10 time-steps. This level of granularity in understanding the model's decision-making process enhances transparency and trust in AI systems, enabling stakeholders to gain deeper insights into the rationale behind specific predictions and improve the model interpretability.

The utilization of Explainable AI (XAI) techniques, such as the Explainer Dashboard and SHAP (SHapley Additive exPlanations), holds immense poten-

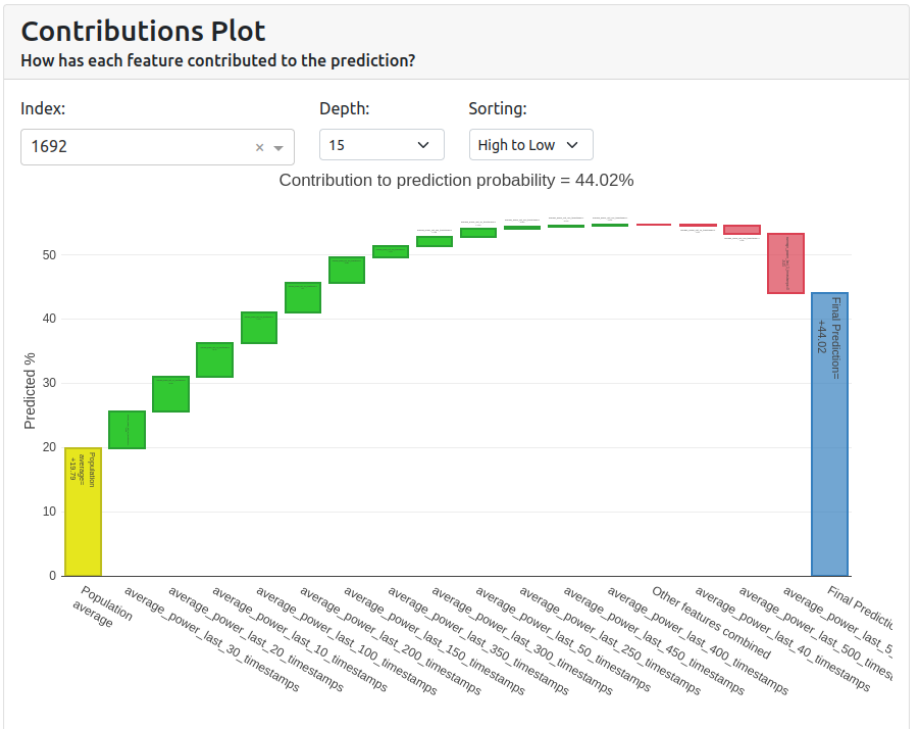

Fig. 10: Features Contributing to the Prediction.

tial for enhancing the practical application of artificial intelligence in optical networks. By providing transparent insights into the predictive behavior of machine learning models, stakeholders within the optical networking domain can make informed decisions with greater confidence and precision. The ability to interpret model decisions at both a global and local level empowers network operators and engineers to understand the underlying factors driving predictions, thereby facilitating proactive maintenance, optimization, and troubleshooting of optical network infrastructure.

Furthermore, the adoption of XAI methodologies can significantly improve the reliability and efficiency of fault management systems in optical networks. By leveraging advanced visualization techniques and interpretable feature analysis, network operators can identify potential fault patterns, anomalies, and degradation trends with greater accuracy and timeliness. This proactive approach enables preemptive measures to be taken to mitigate network disruptions, optimize resource allocation, and enhance overall network resilience. Ultimately, the integration of XAI into optical network management processes can lead to a more robust and adaptive infrastructure, capable of meeting the evolving demands of modern telecommunications while minimizing downtime and maximizing operational efficiency.

6 Conclusion

This paper presents an innovative approach to leveraging state-of-the-art explainable AI techniques for understanding the predictions generated by an ML model designed to detect faults in optical fiber links. We described a comprehensive comparison of four different traditional ML models and found that Light Gradient Boosting Machine (LGBM) outperforms the others. Furthermore, we explored the application of explainable AI in enhancing the model's interpretability, revealing that each type of fault is characterised by a distinct set of features that significantly contribute to explaining the occurrence of the fault. This research sheds light on the importance of transparency and interpretability of AI models deployed in critical infrastructures monitoring and fault detection systems related to optical fiber links.

In the future, we plan to extend the proposed Explanations Pipeline Architecture towards simulating how XAI outputs would change under different input conditions, aiding in better understanding the reasoning of ML models and deriving counterfactual explanations.

References

1. Omitaomu, Olufemi A., and Haoran Niu. "Artificial intelligence techniques in smart grid: A survey." Smart Cities 4.2 (2021): 548-568.
2. Chen, Kunjin, et al. "Fault location in power distribution systems via deep graph convolutional networks." IEEE Journal on Selected Areas in Communications 38.1 (2019): 119-131.

3. Sirojan, Tharmakulasingam, et al. "Sustainable deep learning at grid edge for real-time high impedance fault detection." IEEE Transactions on Sustainable Computing 7.2 (2018): 346-357.
4. Zhang, Senlin, et al. "Data-based line trip fault prediction in power systems using LSTM networks and SVM." Ieee Access 6 (2017): 7675-7686.
5. Wang, Yixing, et al. "Stacked sparse autoencoder with PCA and SVM for data-based line trip fault diagnosis in power systems." Neural computing and applications 31 (2019): 6719-6731.
6. Shafiullah, Md, and Mohammad A. Abido. "S-transform based FFNN approach for distribution grids fault detection and classification." IEEE Access 6 (2018): 8080-8088.
7. Jayamaha, D. K. J. S., N. W. A. Lidula, and Athula D. Rajapakse. "Wavelet-multi resolution analysis based ANN architecture for fault detection and localization in DC microgrids." IEEE Access 7 (2019): 145371-145384.
8. Jacovi, Alon, et al. "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI." Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021.
9. Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.
10. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
11. Explainer Dashboard, https://doi.org/10.5281/zenodo.7633294
12. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
13. SFF-8472: Specification for Management Interface for SFP+, https://members.snia.org/document/dl/25916
14. Cai, Ganhong, et al. "Research on Fault Location and Detection Technology of Optical Network Based on Long Short-Term Memory Neural Network." 2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Vol. 11. IEEE, 2023.
15. Fan, ZhiQiang, et al. "Machine Learning Based Optical Transmission System Link Performance Degradation Prediction and Application." 2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS). IEEE, 2023.
16. Li, Xiuzhu, et al. "Parameter Optimization for Modulation-Enhanced External Cavity Resonant Frequency in Fiber Fault Detection." Photonics. Vol. 10. No. 7. MDPI, 2023.
17. Liu, Pengcheng, et al. "AI-Assisted Failure Location Platform for Optical Network." International Journal of Optics 2023 (2023).
18. SNIA: SFF Specifications, https://www.snia.org/technology-communities/sff/specifications
19. NSys: Mikrotik SFP/QSFP, https://nsys.gr/product-category/mikrotik/sfp-qsfp/
20. Powers, David MW. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061 (2020).
21. Karandin, Oleg, et al. "If not here, there. Explaining machine learning models for fault localization in optical networks." 2022 International Conference on Optical Network Design and Modeling (ONDM). IEEE, 2022.
22. Ayoub, Omran, et al. "Towards explainable artificial intelligence in optical networks: the use case of lightpath QoT estimation." Journal of Optical Communications and Networking 15.1 (2022): A26-A38.

23. Magri, A. and Debenedetti, S. and Morchio, M. and Orsi, P. "Fault Classificaiton Patent," WO2021032292A1, US Patent US11901938B2, February 2021.

Acknowledgment

The work of the authors has been supported by the TALON project funded by the European Union's Horizon Europe Research and Innovation program under the grant agreement No. 101070181.

A Appendix A

Feature	Description
average_power_last_5_timestamps	Last 5 timestamps prior to the LOS event
average_power_last_10_timestamps	Last 10 timestamps prior to the LOS event
average_power_last_20_timestamps	Last 20 timestamps prior to the LOS event
average_power_last_30_timestamps	Last 30 timestamps prior to the LOS event
average_power_last_40_timestamps	Last 40 timestamps prior to the LOS event
average_power_last_50_timestamps	Last 50 timestamps prior to the LOS event
average_power_last_100_timestamps	Last 100 timestamps prior to the LOS event
average_power_last_150_timestamps	Last 150 timestamps prior to the LOS event
average_power_last_200_timestamps	Last 200 timestamps prior to the LOS event
average_power_last_250_timestamps	Last 250 timestamps prior to the LOS event
average_power_last_300_timestamps	Last 300 timestamps prior to the LOS event
average_power_last_350_timestamps	Last 350 timestamps prior to the LOS event
average_power_last_400_timestamps	Last 400 timestamps prior to the LOS event
average_power_last_450_timestamps	Last 450 timestamps prior to the LOS event
average_power_last_500_timestamps	Last 500 timestamps prior to the LOS event

Table 5: List of features used to represent the average optical received power over the different time windows