CHALMERS
UNIVERSITY OF TECHNOLOGY

Root Cause Analysis for Autonomous Optical Network Security Management

Citation for the original published paper (version of record):
Natalino Da Silva, C., Schiano, M., Di Giglio, A. et al (2022). Root Cause Analysis for Autonomous Optical Network Security Management. IEEE Transactions on Network and Service Management, 19(3): 2702-2713. http://dx.doi.org/10.1109/TNSM.2022.3198139

Root Cause Analysis for Autonomous Optical Network Security Management

Carlos Natalino, *Member, IEEE,* Marco Schiano, *Senior Member, IEEE,* Andrea Di Giglio, *Member, IEEE,* Marija Furdek, *Senior Member, IEEE, Optica*

***Abstract*—The ongoing evolution of optical networks towards autonomous systems supporting high-performance services beyond 5G requires advanced functionalities for automated security management. To cope with evolving threat landscape, security diagnostic approaches should be able to detect and identify the nature not only of existing attack techniques, but also those hitherto unknown or insufficiently represented. Machine Learning (ML)-based algorithms perform well when identifying known attack types, but cannot guarantee precise identification of unknown attacks. This makes Root Cause Analysis (RCA) crucial for enabling timely attack response when human intervention is unavoidable. We address these challenges by establishing an ML-based framework for security assessment and analyzing RCA alternatives for physical-layer attacks. We first scrutinize different Network Management System (NMS) architectures and the corresponding security assessment capabilities. We then investigate the applicability of supervised and unsupervised learning (SL and UL) approaches for RCA and propose a novel UL-based RCA algorithm called Distance-Based Root Cause Analysis (DB-RCA). The framework's applicability and performance for autonomous optical network security management is validated on an experimental physical-layer security dataset, assessing the benefits and drawbacks of the SL- and UL-based RCA. Besides confirming that SL-based approaches can provide precise RCA output for known attack types upon training, we show that the proposed UL-based RCA approach offers meaningful insight into the anomalies caused by novel attack types, thus supporting the human security officers in advancing the physical-layer security diagnostics.**

***Index Terms*—Optical networks, physical layer security, supervised learning, unsupervised learning, interpretability, explainability.**

I. Introduction

As the key technology underpinning the global communication network infrastructure, secure and reliable operation of optical networks is a necessary enabler for the network evolution beyond 5G. Due to the high data rates carried over the optical fiber, even short disruptions at the physical layer can cause large data losses and affect a multitude of aggregated upper layer services, which makes the optical network an enticing target of man-made attacks. The current trends of optical/wireless integration and extension of the optical domain towards the network edge – in order to reduce latency – further aggravate the damaging potential of physical-layer attacks compared to current architectures with optical-electrical-optical conversions at segment exchange points.

Attacks aimed at physical layer disruption may vary in their levels of disturbance, sophistication, or traceability, among other traits. One of the well-known service degradation techniques reported in the literature is jamming [1], where a harmful signal is inserted in the fiber to either add un-filterable noise to a signal at the same wavelength (in-band jamming), or to deprive the co-propagating useful signals of optical amplifier gain (out-of-band jamming). Another technique is a polarization modulation attack, where the fiber is squeezed at a high frequency, resulting in very fast changes of the polarization of light that cause errors when the coherent receiver's polarization recovery algorithm cannot compensate for the fast fluctuations any longer [2].

Coping with evolving security threats, in the frame of network transformation into automated high-performance systems, requires intelligent approaches for autonomous security management embedded into services' life cycles and incorporated into a cognitive network management system (C-NMS) paradigm [3]. Diagnosing physical-layer security threats is a challenging task due to several factors. Firstly, there are no exact analytical models of physical-layer attack effects to date. Consolidated analytical modeling of physical-layer impairments for, e.g., Quality of Transmission (QoT) estimation or margin reduction purposes, is complex even under normal operating conditions, which spawned an entire field of Machine Learning (ML) applications [4], [5]. Attacks can violate the normal operating conditions, which limits the applicability of known analytical models. Moreover, attacks can cause intricate interplay among the Optical Performance Monitoring (OPM) parameters, and subtle changes in their values may accumulate into service degradation, rendering analytical, e.g., threshold-based approaches ineffective [3]. Finally, the security threat landscape keeps transforming, with new threats targeting the vulnerabilities of evolving systems.

ML has already been demonstrated as a powerful tool to meet the need for intelligent and adaptive security management approaches [2], [6], [7]. Today's commercially available Digital Signal Processing (DSP)-enabled coherent receivers provide the Network Management System (NMS) with a rich OPM dataset that can be processed with ML-based tools and provide insight into the network state without the need for costly, specialized monitoring equipment. In this case, the system observed by means of the OPM data includes the

Carlos Natalino and Marija Furdek are with the Electrical Engineering Department, Chalmers University of Technology, Gothenburg, Sweden. Marco Schiano and Andrea Di Giglio are with Telecom Italia, Turin, Italy.

The results reported in this paper are partially supported by VR (2019-05008), VINNOVA (AI-NET-PROTECT – 2020-03506), and the European Commission 5GPPP TeraFlow (101015857).

We gratefully acknowledge Infinera for providing the Groove G30 transponders used in the experiments.

Manuscript received XXX XX, 2021; revised XXX XX, 2022.

transmission system itself and the measuring system with their own specific physical behavior and processing algorithms. An example of the power of ML techniques is that they are able to perform end-to-end optimization: they extract the information of a feature going beyond the physical meaning of that feature and taking into account the whole measurement system regardless of how specific metrics are obtained.

When applying ML tools to physical-layer security diagnostics, it is not always possible to ensure identification of attacks based on sufficient prior knowledge. In such cases, human intervention is often unavoidable in investigating the root cause of a detected anomaly before triggering the correct countermeasures. The first stage of incorporating ML-supported Root Cause Analysis (RCA) in Security Operations Center (SOC) processes is expected to aid the decision process of a human security operator rather than eliminating the human from the loop [8]. Therefore, it is important to provide insightful information about the detected anomalies. RCA in networks is typically addressed using Supervised Learning (SL) techniques that rely on prior knowledge of anomalies to be detected [9], [10]. However, these approaches may not be applicable to physical-layer security scenarios where a representative labeled dataset covering all attack scenarios may not be available, or new threats may emerge at any time, or models trained for one optical channel may not be applicable to other channels. In this work, we establish an ML-based framework for security assessment. Then, we propose an Unsupervised Learning (UL)-based RCA algorithm named Distance-Based Root Cause Analysis (DB-RCA) that does not require prior knowledge on the anomalies caused by physical-layer attacks. The algorithm, which extends our preliminary study in [11], is validated on an experimental physical-layer security dataset. Our performance assessment includes the RCA outputs of eXtreme Gradient Boosting (XGBoost), a known SL model that enables RCA, in addition to our proposed DB-RCA. By analyzing the output of the SL- and UL-based algorithms we can assert that, although significantly different, the outputs of both approaches can provide meaningful insight when investigating the root cause of an attack.

The contributions of this paper, extensions with respect to [11], and the corresponding sections in this paper, can be summarized as follows:

- We analyze the different NMS architectures and the corresponding ML-based security assessment functionalities (Sec. III).
- We investigate the benefits and drawbacks of the SL- and UL-based RCA, and establish an RCA framework applicable to the security assessment task (Sec. IV).
- We propose a UL-based DB-RCA algorithm that can be used irrespective of the attack detection model, extending the initial study in [11] that was specifically tailored to work with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Sec. V).
- We validate its applicability to autonomous optical network security management using an experimental physical-layer security dataset (Sec. VI).

II. Related Work

The three pillars of network security management, namely risk mitigation, attack detection and recovery, have been addressed in the literature to different extents. Risk mitigation efforts focus on attack-aware network design and service provisioning in an effort to reduce the network exposure to and/or the damaging potential of attacks. Embedding the awareness of physical-layer attacks into connection routing was initially proposed in [12]. Attack-aware routing and wavelength assignment under static and dynamic traffic was performed in [13], [14], respectively. Jamming-aware provisioning of dynamic traffic in Space Division Multiplexing (SDM) networks was studied in [15], [16]. Strategies for cost-efficient placement of dedicated devices that can filter out harmful signals or equalize their excessive power were investigated in [17], [18], respectively. Attack-awareness was also incorporated into advanced networking scenarios, such as connection provisioning in multi-domain elastic optical networks [14], and game theory-assisted control plane design [19].

Detection of faults caused by component failures or attacks is a key aspect of network monitoring. ML has revolutionized the field of optical network monitoring by removing the need for explicit modeling of the physical-layer effects and precise inventory of the optical infrastructure. ML has been shown to successfully tackle challenges related to detection of both failures and attacks. Applications of ML-aided failure management range, e.g., from the early works on integrating SL for failure detection and identification into transport-Software Defined Networking (TSDN) architecture [**?**], to cooperative, self-supervised learning for brokered soft failure detection across multiple domains [20]. For a survey of the various applications of ML for failure management, we refer the reader to [21].

The performance of different SL techniques, including, among others, Artificial Neural Network (ANN) and Support Vector Machine (SVM), for detection and identification of jamming and polarization scrambling attacks was experimentally analyzed in [6]. In [22], SVM was used to analyze the optical spectrum and detect unauthorized transmission. [23] investigated the potential of UL to detect attacks without prior insight or training. Practical implications, advantages and challenges of SL, UL and Semi-Supervised Learning (SSL) for attack detection and/or identification were comparatively investigated in [2]. The role of ML in security management automation was analyzed in [8]. A novel functional block, referred to as the optical security manager, was proposed in [7] to complement the TSDN controller with encompassing optical security functions.

Fast and accurate detection and identification of attacks is crucial for appropriate and effective remediation of security breaches, which should encompass service recovery, threat neutralization and network adaptation. Planning of backup routes to be used for protection in case of jamming was presented in [24], relying on prior knowledge of the attack properties and connection routing. However, analysis of the root cause of a detected anomaly, as a prerequisite for trigger-

ing countermeasures tailored to an evolving threats landscape, has not been investigated in the literature yet.

III. Progress Towards Automated Security Diagnostic Procedures

The introduction of telemetry systems and advanced ML techniques enables NMSs to evolve towards supporting automated security diagnostics [25]. The evolution of the NMS architecture is illustrated in Fig. 1, starting from a legacy scenario characterized by traditional NMS and describing a possible evolution path towards optical networks with embedded telemetry systems and optical security ML tools.

Unauthorized access to the management system of an optical network can cause substantial damage, since the network settings and configurations can be maliciously changed, producing detrimental effects on the services. The protection of management systems, which takes place through increasingly sophisticated authentication systems, including biometric or multi-level, is beyond the scope of this paper and is treated elsewhere [26]. This study aims to evaluate how the data made available by the NMS can feed an ML system to classify the various types of network attacks occurring at physical layer.

A. Network Assurance approach based on traditional NMSs

Today, most optical networks are still operated by traditional NMSs, as illustrated by Fig. 1a. They are characterized by collecting OPM data every 15 minutes with very limited OPM data storage capabilities. The limitations of those technologies force operators to implement elementary reactive strategies based on alarm monitoring and manual intervention as countermeasures to both faults and attacks. In case of an attack, an alarm is typically raised in the NMS and the operator is alerted in case an OPM value exceeds a predetermined threshold. The NMS itself provides to the operator remedy tables that propose plausible causes for each alarm, suggesting also further checks and appropriate remedy actions.

Due to the limited historical records and the requirement of manual checks, security assessment is very difficult and requires highly-trained and experienced operators that understand the specifics of the system. Moreover, the use of predefined thresholds is unscalable and unreliable, as some attack techniques cause only minor variations in OPM values and each optical channel might need a specific fine tuning of its thresholds [2].

B. NMSs with telemetry functions

A first evolutionary step in optical network management comprises the development of telemetry systems embedded in the NMS, that collect large OPM data records every second (or few seconds) and store those records in a database, as illustrated by Fig. 1b. This technology is already proposed by many network system manufacturers for their next generation products. Nevertheless, even with this advanced telemetry system, attack diagnostics is similar to the traditional one, but the checks suggested by remedy tables can be done on OPM historical data series rather then on log files with few tens of

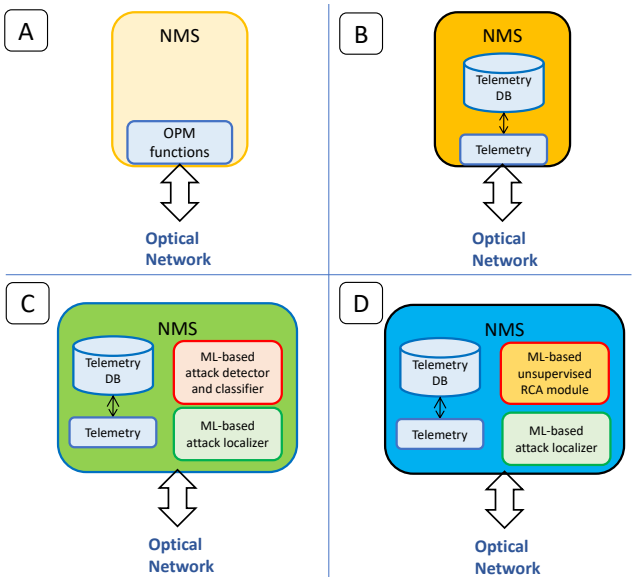


Fig. 1. Network Management System (NMS) architecture evolution with the introduction of telemetry systems and advanced ML techniques for security diagnostics.

entries. In any case, the historical data series analysis, which is performed manually by the operator, is very complex, time consuming, and requires special data processing skills seldom included in the training of Network Operation Center (NOC) operators. Based on their standard fault-oriented mindset, the network assurance operator has no effective tools to analyze what is happening in the network in case of an attack and no criteria for applying specific data analysis tools on the OPM dataset.

C. NMSs with telemetry and attack detection ML functions

A network malfunction is caused by either a failure or a malicious attack. Often, an attack and a failure can have similar *symptoms* but need different *treatments*. For this reason it is very important to have, alongside telemetry systems which continuously monitor the status of the connections and the quality of the optical signals, a system that, based on these data, is able to recognize that the malfunction is caused by a malicious attack and classify such attack based on a previously identified set. Such an NMS architecture, capable of detecting, identifying and locating the source of an attack, is illustrated by Fig. 1c. We have already demonstrated the effectiveness of ML techniques to detect and classify network attacks [2]. When those ML tools are embedded in future generation NMS, they can provide an effective support for the NOC operators: they will gain information on the nature of the network attack directly from the ML system in quasi real time, and they will use their knowledge and experience to apply appropriate countermeasures.

D. NMSs with telemetry, attack detection and Root Cause Analysis ML functions

To support automated security diagnostics, NMS should also have embedded functionalities for tracking newly emerging

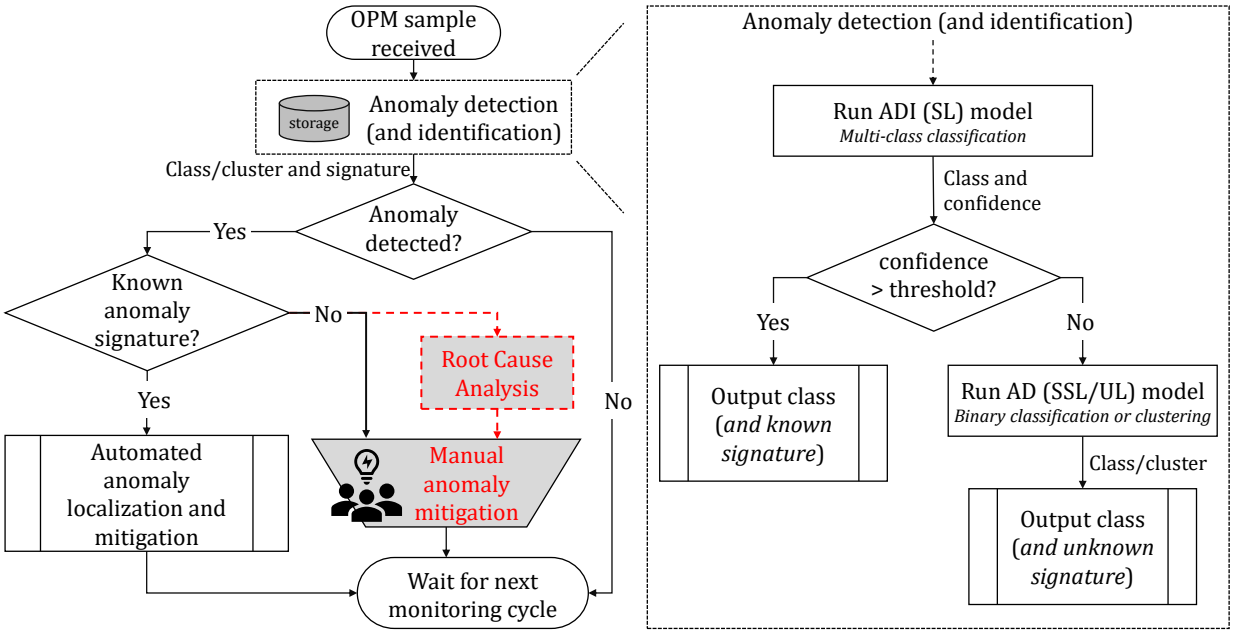


Fig. 2. Workflow of an NMS with ML-based attack detection (and identification) and Root Cause Analysis. The anomaly detection (and identification) (AD/ADI) module outputs the attack class/cluster and known/unknown signature based on the appropriate use of SL and SSL/UL models.

attacks. As an evolution of the previously described scenario, where the NMS has an embedded algorithm capable of discerning between a failure and an *a priori* known attack and, correspondingly, classifying the attack. Leveraging currently available technologies, properly addressing newly emerging attacks unavoidably requires human intervention. However, a step forward is given by an ML-driven RCA functionality due to its ability to provide an initial insight into the effects of a novel threat. NMS architecture with RCA is illustrated in Fig. 1d. This approach does not rely on prior knowledge of attack consequences, but can, nonetheless, provide meaningful insight into their effects on network performance and aid the operators in determining the most effective security countermeasures.

IV. Root Cause Analysis in Optical Networks

Root Cause Analysis (RCA) is a term used to define methods that aim at identifying the root causes of faults, problems, or anomalies. Application areas include production engineering and management, IT operations, and telecommunications. Depending on the area where RCA is applied, it might include different steps and techniques. With the advent of the use of ML techniques across a wide range of industries, the lack of explainability of many ML models increases the need for RCA specifically tailored for analyzing the output of the models [27]. These ML-based RCA methods analyze the same data that can be input to an ML model, and produce metrics that indicate which aspects or features of the data are more likely to be decisive to the ML output, contributing to the interpretability and explainability of the ML output.

In the context of NMSs, RCA can be used to assist the network management staff with the difficult task of analyzing the OPM parameters and identifying the cause of an anomaly. The difficulty of this task might be exacerbated when the ML algorithms in place raise an alarm without providing further insight such as identifying the problem.

Fig. 2 illustrates the workflow of an NMS and the role of the network management staff in addressing detected anomalies, which can be further enhanced by anomaly identification. OPM samples are received periodically, which triggers the entire process. An ML-based model performs anomaly detection (and possibly identification). The right-hand side of the figure expands the related procedures. The depicted workflow combines SL and SSL/UL models to obtain a reliable anomaly assessment. First, an SL model performs multi-class classification, which results in anomaly detection and identification in addition to a confidence score associated with the output of the model. If the confidence level is above a desired level, the known anomaly signature is reported. On the contrary, if the confidence level is below the desired level, the output of the SL model is not considered reliable enough and, instead, anomaly detection is performed by either SSL or UL. In this case, only binary classification or clustering are possible, and the output specifies whether or not an anomaly is detected. If the SSL/UL model detects an anomaly, it reports the detection flagged as an unknown anomaly signature. Note that the described workflow also functions when no SL model is available to detect and identify an anomaly. In this case, the received OPM sample is used by the SSL/UL model to perform binary classification, always flagging the detected anomaly as unknown. Long-term storage is usually needed in this case to store either a set of the most recent OPM samples, or the ML model itself, or both. The output of this module is evaluated, and if no anomaly is detected, the process idles and waits for the next monitoring cycle. The frequency of the monitoring cycle (e.g., every minute, or every 15 minutes) is decided by the network operator depending on the criticality of the monitored services and the expected reaction time to events.

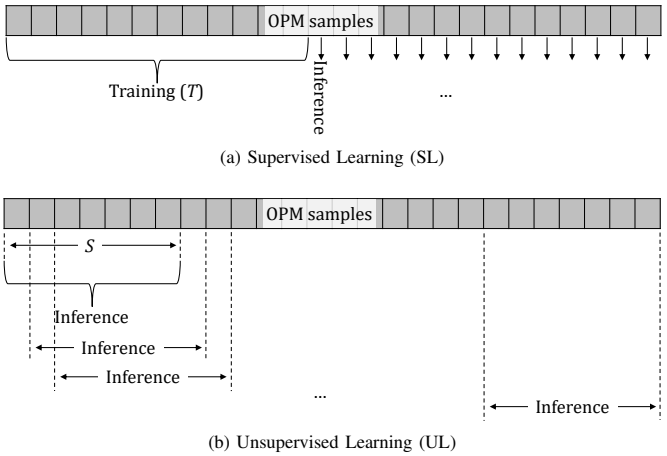


Fig. 3. The operating principle of SL and UL models for training (only in the SL case) and inference. Note that usually $|S| \ll |T|$.

On the other hand, if an anomaly is detected, the process for anomaly identification and mitigation starts. First, if the ML model is capable of anomaly identification, or if a standalone anomaly identification is in place, the NMS checks if the anomaly signature is known. If so, automated anomaly localization and mitigation developed *a priori* are executed.

RCA becomes vital in the case where the anomaly signature is not known. In this case, manual mitigation of the anomaly needs to be tailored by the network management staff. The RCA is helpful at this stage when the network management staff needs to understand what caused the ML model to identify the current network state as anomalous. Without RCA, the network management staff is left with the manual investigation procedures described in Sec. III. RCA provides useful insights, often enriched with graphical representation that can enhance and speed up the analysis phase.

It is important, however, to mention that RCA applied to ML models has some limitations. First and foremost, RCA is based on data, and therefore, can only highlight the possible causes that are represented in the data. This is especially important in optical networks, where minor electrical and temperature variations can affect the correct operation of optical devices [10], [28]–[31]. These variations are usually not represented in the OPM data collected from the coherent transceivers, which usually only contain optical-related parameters (i.e., do not include electrical current, power, or temperature). Therefore, these electrical and/or temperature variations will manifest in the form of OPM parameter changes. These limitations can be addressed by including, integrating and consolidating further monitoring data (e.g., from the electrical power grid and/or cooling system) into the analysis, which is out of the scope of this work. In the next sections, we introduce RCA techniques for SL and UL.

A. Supervised Learning for Root Cause Analysis

Supervised Learning (SL) models comprise two different stages represented in Fig. 3a. First, the model is trained over a training dataset which needs to be labeled, i.e., for each sample, a label specifies what is expected as an output from SL model. During training, the model abstracts the important information present in the dataset that allows it to adjust and store its internal representation to facilitate the association between each input and the label. For instance, traditional Feed-Forward Neural Networks (FFNNs) represent the information in the form of weights and biases. Decision Trees (DTs), on the other hand, store the information in the form of thresholds and conditional statements. Once trained, SL models can perform inference by analyzing only a single sample (i.e., there is no need to analyze previous samples). This means that SL models are usually resource-intensive during training, but lightweight during inference.

One drawback of SL models, in particular in the context of anomaly detection and security assessment, is that their performance expectations can only be drawn over data that is used for training or test purposes. This means that new anomalies or security threats (previously unseen by a model) can potentially remain undetected for long periods of time unless they are mistakenly classified. In the context of optical networks, this issue becomes more critical as each optical channel may need a specific SL model trained just for that channel, given that, to the best of our knowledge, the applicability of a single model addressing multiple optical channels established over a different set of optical devices has not been reported in the literature so far. This is due to the fact that different optical channels use different spectrum and traverse different equipment, which differentiates their behavior enough to render an SL model trained for another optical channel unsuitable. Another potential reason is the fact that different optical channels may be designed for different OPM levels such as Optical Signal-to-Noise Ratio (OSNR), modulation format, etc. During operation, this translates either to a longer deployment time for optical channels (due to the need to collect enough samples for every anomaly/attack case), or running the optical channel with user traffic without proper anomaly/attack detection/classification assessment in place [2].

Another drawback of SL models is their trade-off between accuracy and interpretability. ANNs usually achieve the highest accuracy among the SL models when their hyperparameters are appropriately tuned, but are considered a black-box model. This is due to the fact that their internal representation (in the form of weights and biases) is hard to be interpreted. On the other hand, traditional DTs do not usually reach the same level of accuracy as ANNs, but are easily interpreted and can be implemented with simple conditional statements. This makes DTs ideal for use in the RCA for SL models.

The eXtreme Gradient Boosting (XGBoost) algorithm [32] was proposed to mitigate the usual accuracy limitations of DTs, and reduce the trade-off between accuracy and interpretability. XGBoost implements gradient boosting for DTs, and uses an ensemble of DTs to improve its accuracy. The training procedure uses the gradient descent algorithm, which enables the use of different loss functions depending on the task at hand (e.g., cross-entropy for classification tasks and mean squared error for regression). The gradient descent decides whether to add new nodes/leaves to the current DT structure such that the loss is minimized. XGBoost is in the same category as Random Forests (RFs), i.e., ensemble

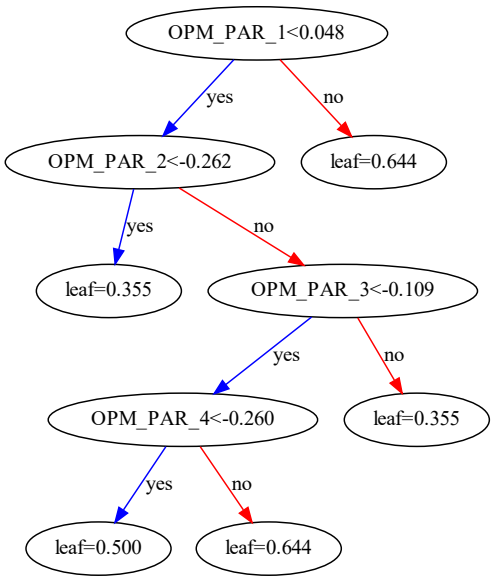


Fig. 4. An illustrative example of a decision tree for binary classification using hypothetical normalized OPM parameters.

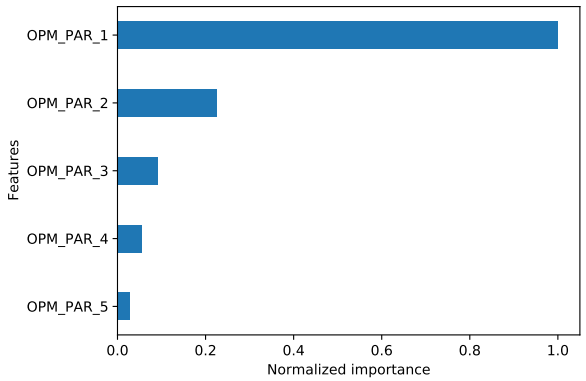


Fig. 5. An illustrative example of feature importance obtained from the XGBoost algorithm using hypothetical OPM parameters.

learning algorithms based on decision trees. The main difference between the two is that in RFs each decision tree is independent from another, while in XGBoost the decision tree $n+1$ focuses on improving the accuracy obtained by decision tree n.

Fig. 4 illustrates an example decision tree for a binary classification task in an optical network. Each node in the tree represents a conditional statement, similar to traditional DTs. However, leaves represent the probabilities (e.g., *logit*, used in the example and throughout the paper) of a sample belonging to a given class, as opposed to traditional DTs where leaves represent decisions. Note that the example shows a tree for binary classification, so a *logit* ≤ 0 (i.e., probability ≤ 0.5) means that the sample is likely of the first class, while a *logit* > 0 (i.e., probability > 0.5) means the sample is likely of the second class. By using an ensemble of DTs, XGBoost collects these probabilities across the different DTs to define the output.

XGBoost is particularly useful for RCA due to its importance score that results from the training process. The importance score is attributed to each feature (or hypothetical OPM parameter in our context) and represents the number of times that the feature is used to add a new split/node. The intuition is that a tree with an additional split/node is more accurate than one without the split. Therefore, the more times a feature is used in splits, the more important it is for the particular effect under analysis. A shortcoming of this approach is that it does not account for the impact of a split to the final output (e.g., a split with both leaves leading to the same output), but alternative approaches such as the Shapley values can be used to mitigate this issue [33]. Fig. 5 illustrates the normalized feature importance resulting from a XGBoost training over a hypothetical dataset collected from optical transceivers for classification of anomalies. We can see that OPM_PAR_1 is the most important feature, with OPM_PAR_2 accounting for approximately 30% of the importance of OPM_PAR_1. OPM_PAR_3, OPM_PAR_4 and OPM_PAR_5 have a minor importance. This importance score can be directly linked to the particular effect under analysis and used to tailor mitigation strategies.

B. Unsupervised Learning and Root Cause Analysis

Unsupervised Learning (UL) models are tailored to identify anomalous samples by considering an observation window containing a number of samples. The observation window usually has a few hundreds of samples. The UL models assume that at the beginning of observation, only a few samples in the observation window will contain anomalous behavior. By considering the distance between samples to group them, it is expected that anomalous samples will be far from the samples that characterize normal behavior, and will, therefore, be flagged as anomalies.

Fig. 3b illustrates the operating principle of UL models. As opposed to SL, UL does not require training, which translates into neither using a labeled (training) dataset nor storing of any internal states. This means that it needs to receive an entire observation window every time an inference is needed. Therefore, inference is costlier in UL than in SL models in terms of complexity and runtime [7].

Compared to SL (illustrated in Fig. 3a), UL models have several benefits important for their use in the autonomous operation of optical networks. The absence of training brings two important advantages: *(i)* these models do not need to be tailored for each optical channel, and a single implementation can be used for all optical channels in a network; and *(ii)* the models can start their operation shortly after the optical channel is set up, i.e., there is no need to collect a training dataset. Finally, given the general ability to detect anomalies, these models can be potentially applied to any sort of anomaly (e.g., attack, malfunction), unlike SL models that need labeled samples and training for each and every anomalous condition.

On the other hand, there are a few drawbacks of UL models when compared to SL. First, UL models can only detect

the anomaly, without any further insight. On the contrary, SL models are able to identify the anomaly as long as they have been trained for it. Moreover, SL models are typically exposed to a larger number of samples during training, which allows them to identify subtle but consistent properties of the system behavior. By analyzing a relatively small number of samples compared to a training dataset for SL models, UL models rely on observing the unstable behavior of the system upon an anomaly occurrence to detect significant deviations from normal operation. As previously mentioned, UL models usually have more complex inference than SL models, although some efforts have been made to combine SL and UL to provide lightweight inference [34]. Finally, UL models usually have a few hyperparameters that define how strictly the relationships among samples will be evaluated, and defining these parameters without knowledge of the anomalies is challenging. One possible way to tune hyperparameters using only normal samples is to start from a more relaxed set of parameters and progressively make them more strict until the algorithm starts presenting false positives. At this point, one can decide on an initially acceptable level of false positives. During network operation, as anomalies are detected and validated through, e.g., an RCA framework, the hyperparameters can be further optimized.

These differences between SL and UL models significantly change the way that RCA is performed in each scenario. With SL models, RCA aims at finding the ground truth importance of features based on extended (and usually far) past behavior (represented by the training dataset). On the other hand, RCA based on UL models aims at finding the importance of features based on a limited number of (close to) real-time observations. In the following section, we describe the algorithm proposed in this paper that aims at providing a viable RCA approach for UL models.

V. Distance-based Root Cause Analysis (DB-RCA)

This section describes the main contribution of this paper, i.e., the algorithm for UL-based RCA called DB-RCA. The main output of the algorithm is the Anomaly Vector (AV). The AV enriches the anomaly detection algorithm by representing the average distance between anomalous samples and their closest cluster of normal samples. The results of the AV can be graphically shown to the network management staff together with other data relevant for anomaly mitigation. Furthermore, the AV can be integrated into automated anomaly localization and mitigation procedures.

Alg. 1 presents the pseudocode of the proposed algorithm. The algorithm takes as input the set F that describes which features are present in the dataset, and the (pre-processed) dataset X containing values of the features for every sample. Additionally, the algorithm also receives the output from the anomaly detection algorithm Y, where the number of elements in Y is equal to the number of elements in X. DBSCAN [35] is one of the UL models that can be used for anomaly detection, i.e., to produce Y, which is also the approach taken in this work. In any case, the UL algorithm used to detect an anomaly will provide at least two clusters for RCA. For a

Algorithm 1 Distance-based RCA

Data: Set of features F, (pre-processed) dataset $X \in \mathbb{R}^{|F|\times|X|}$, anomaly detection output $Y \in \mathbb{Z}^{|X|}$
Result: Anomaly vector AV
1 $N \leftarrow \bigcup_{i=1}^{|X|}\{X_i\} : Y_i \geq 0$
2 $A \leftarrow \bigcup_{i=1}^{|X|}\{X_i\} : Y_i = -1$
3 $P \leftarrow$ cluster in N closest to A
4 $AV_i \leftarrow \frac{\sum_{j=0}^{|A|} A_{i,j}}{|A|} - \frac{\sum_{k=0}^{|P|} P_{i,k}}{|P|}, i{=}0..F$
5 **return** AV

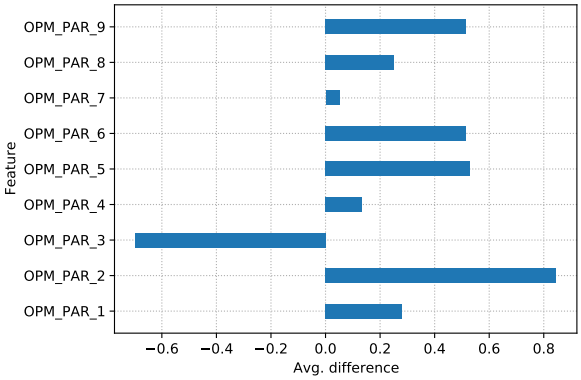


Fig. 6. An illustrative example of the Anomaly Vector (AV) obtained from the proposed RCA framework using hypothetical OPM parameters.

sample with index i, we assume that $Y_i = -1$ when sample X_i is flagged as anomaly, and $Y_i \geq 0$ when the sample is considered normal, in which case the value of Y_i acts also as the cluster index for the sample indexed by i.

The algorithm begins by obtaining a cluster that contains normal samples (line 1) and anomalous samples (line 2) from X, denoted by N and A, respectively. Then, the algorithm selects the cluster closest to the anomalous samples (line 3) and computes the anomaly vector as the difference between the average values of the features in the two clusters (line 4). By considering only the closest cluster, the algorithm is more likely to observe smaller variations in the features that led the samples to be positioned outside of the cluster area. Moreover, this limits the number of samples whose distance needs to be computed for each feature. The AV is returned (line 5). Note that our proposed algorithm has a small impact on the overall system runtime, since it is only executed when an anomaly is detected. Note also that the algorithm is extensible enough to be included in pipelines using other models. Finally, the algorithm uses the same data used for DBSCAN (or another UL algorithm), in addition to its outputs, without keeping any state between inferences. This means that the algorithm does not change the fundamental properties of DBSCAN and therefore can also be considered a UL approach. The worst-case complexity of the algorithm can be derived as $O(|X|^2 \times |F| + |X| + |X|)$ since we need to calculate the distance between every pair of samples ($|X|^2 \times |F|$) and traverse the samples twice ($|X|$).

Fig. 6 shows an illustrative example of how the AV would be presented to the network management staff for a hypo-

TABLE I
SUMMARY OF CONSIDERED ATTACK SCENARIOS [2].

Attack scenario		Jamming signal power	Jamming signal frequency	Fiber squeezer driver amplitude
Out of band jamming	Light	P_0+3 dB	195.1 THz	-
	Strong	P_0+8.7 dB	195.1 THz	-
In band jamming	Light	P_0-10 dB	f_0	-
	Strong	P_0-7 dB	f_0	-
Polarization modulation	Light	-	-	0.3 V
	Strong	-	-	1.6 V

P_0 and f_0 denote the power level and frequency of the optical channel under test.

thetical detected attack. Note that the AV enables a network security specialist to identify the level of deviation from the normal for the considered hypothetical OPM parameters (e.g., OPM_PAR_2 and OPM_PAR_3 deviate the most, while OPM_PAR_7 deviates the least in this hypothetical case). It also indicates the increasing or decreasing trends in the parameter variations, represented by positive or negative AV values, respectively.

VI. PERFORMANCE ASSESSMENT

This section presents a detailed analysis of the performance of the SL and UL learning models presented in this work, focusing on their accuracy and RCA properties. The accuracy performance is assessed in terms of the *f1 score*, which balances the importance of false positives and false negatives into a single metric equal to 1 when the model is perfectly accurate in detecting attacks, and smaller otherwise. Then, we analyze the importance metrics output of XGBoost and the proposed UL-based RCA algorithm. Finally, we discuss the characteristics and properties of the evaluated approaches.

A. Use Case

We validate the proposed algorithm on a physical layer security use case where anomalies are characterized by physical layer attacks launched over optical channels. OPM samples are collected from an experimental optical network testbed with coherent transceivers, Reconfigurable Optical Add-Drop Multiplexers (ROADMs) and Erbium-Doped Fiber Amplifiers (EDFAs). A detailed description of the experiments and testbed can be found in our previous work [2]. The monitored channels are two optical 200 Gbit/s polarization multiplexed 16 quadrature amplitude modulation (16QAM) signals at 195.2 and 195.3 THz. Then, three attack strategies are launched in the network, namely In-Band (IB) and Out-of-Band (OOB) jamming, and Polarization Modulation (PM). For each attack strategy, a light and a strong attack condition is used, resulting in 7 different attack scenarios: Light In-Band jamming (INBLGT), Strong In-Band jamming (INBSTR), Light Out-of-Band jamming (OOBLGT), Strong Out-of-Band jamming (OOBSTR), Light Polarization Modulation (POLLGT), and Strong Polarization Modulation (POLSTR).

The characteristics of the considered attack scenarios are summarized in Table I. In IB jamming, a signal at the same frequency as the channel under test is inserted in the network, adding unfilterable noise. This type of attack requires precision in terms of frequency, but can be realized with relatively low power. We set the jamming signal power to 10 and 7 dB below the power of the channel under test to obtain INBLGT and INBSTR intensities, respectively. For OOB jamming, the inserted signal has higher power, but is substantially separated from the spectrum used by the channels under test. We set the jamming signal power to 3 and 8.7 dB above the power of the channel under test to obtain OOBLGT and OOBSTR intensities, respectively. In the polarization modulation attack, the fiber is squeezed with a piezoelectric squeezer at a resonant frequency of 136 kHz. The two different intensities, POLLGT and POLSTR are obtained by using two different values of the sinewave signal driving the squeezer, i.e., 0.4 and 1.6 V peak-to-peak, respectively.

In each minute, and for each attack condition and optical channel, a script collects the following OPM parameters over the course of 24 hours: pre-FEC Bit Error Rate (BER-FEC), post-FEC Bit Error Rate (BER-PF), Loss of Signal (LOS), Optical Power Received (OPR), Chromatic Dispersion (CD), Differential Group Delay (DGD), OSNR, Polarization Dependent Loss (PDL) and Q-factor. For some of the OPM parameters, the minimum and maximum values obtained within the monitoring window are also reported. The resulting dataset is the largest optical security-related dataset reported in the literature, composed of 1440 samples for each scenario, with each sample containing 31 features. The collected dataset is pre-processed by removing samples with missing features and by applying z-score standardization. Finally, for the analysis reported in this section, only the nominal values are considered (we do not consider minimum and maximum values reported by the transceivers).

B. RCA Using XGBoost

For the results presented in this section, we used the Python open-source implementation of the XGBoost model. We configured the model for binary classification (which is appropriate for the attack detection task) and to use a single decision tree. The dataset was split into 50% for training and 50% for testing purposes. For the binary classification, the XGBoost attack detector achieved a f1 score of 0.995, which represents a good performance, with near zero false positives and negatives.

Figures 7 and 8 show the resulting decision tree for classifying IB jamming and PM strong attacks, respectively. Note that the numerical values for the OPM parameters are standardized, and that each leaf represents the probability of a sample being an attack. We can see that for INBSTR attacks (Fig. 7), only BER-FEC and OPR are sufficient to detect an attack. On the other hand, for POLSTR attacks (Fig. 8), BER-PF and OSNR are needed in addition to OPR. These results make it evident that each attack scenario is best represented by a particular set of OPM parameters. More importantly, it shows that collecting a set of OPM parameters as complete as possible is paramount for enabling a reliable and future-proof security assessment. In other words, certain parameters might not be important for the currently known attacks, but might become important to diagnose new attacks in the future.

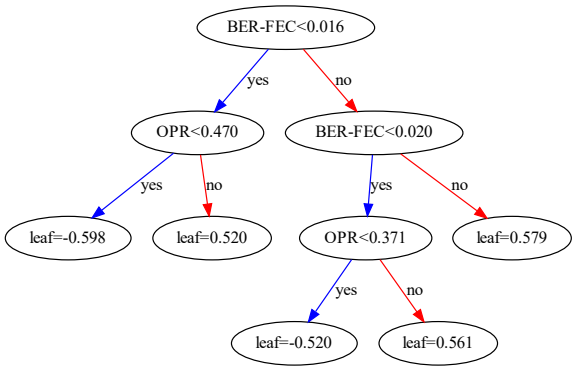


Fig. 7. Resulting tree for identifying strong in-band jamming attacks.

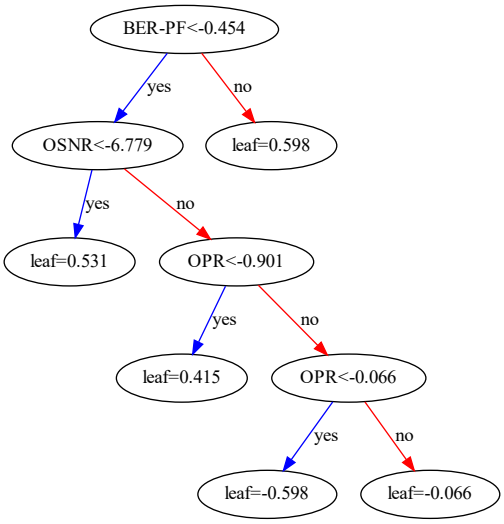


Fig. 8. Resulting tree for identifying strong polarization attacks.

Next, we move our attention to the feature importance scores obtained by XGBoost (in our case, features represent the OPM parameters collected from the transceivers). Fig. 9 shows the feature importance for the models trained over all the attack scenarios (Fig. 9a) and over each individual scenario. When detecting all the attack scenarios (Fig. 9a), OPR and BER-FEC are the two most important features. However, PDL and CD, usually disregarded for most of the analysis in the literature, retain a significant importance for the overall attack detection. When examining an OPM parameter and its impact on ML algorithms, we must consider how this parameter is measured by the coherent transponder optical interface. Recall that the OPM data represents the combination of the transmission and the measurement system. Considering, for instance, the OSNR, the transponder manufacturer does not provide details on how it is measured, but, based on the literature (see, e.g., [36]), we can argue that OSNR is derived from the Error Vector Magnitude (EVM). If we consider the measurement technique, it is not a surprise that OSNR is not an important feature for some kind of attacks (while in general it is). This simply means that the whole OSNR measurement system: EVM estimation from a given number of received symbols, sampling system, OSNR derivation (and perhaps many more intermediate measuring functions) produce a result that is not strongly influenced by some kinds of attacks.

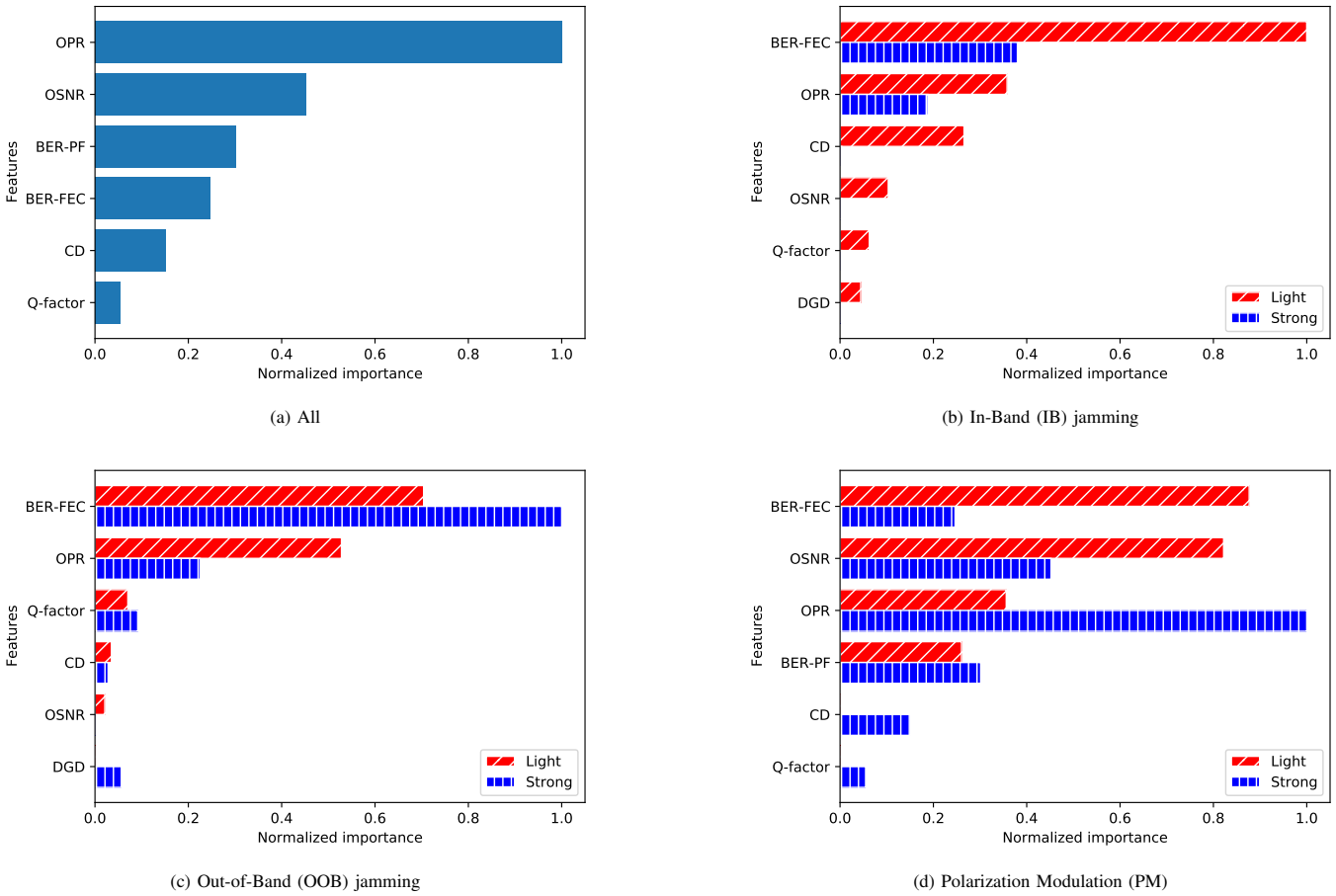
For IB and OOB jamming (Figs. 9b and 9c), the two most important OPM parameters are BER-FEC and OPR. However, starting from the third place, things are different. While Q-factor is the third most important OPM parameter for both intensities of OOB jamming, it is only the fifth for IB. For IB, CD is the third most important OPM parameter. Interestingly, for the IB, the importance for the strong attack is concentrated over BER-FEC and OPR, while for the light attack, the importance is spread over 6 different OPM parameters. Moreover, looking at the physical nature of the attack, CD and DGD features should not be representative of the OOB jamming attack where just OSNR, Q-factor and BER-PF are expected to vary, but still have significant importance for the attack detection.

Fig. 9d shows that for PM, the order of importance changes from the light to the strong intensities. BER-FEC and OPR swap places in terms of importance. BER-FEC is the most important for the light attack while OPR is ranked third. For the strong attack, OPR comes first and BER-FEC in third. Finally, BER-PF, which is not ranked for the IB and OOB jamming attacks, is important for PM, and replaces DGD that was important for the two jamming attacks.

C. DB-RCA using DBSCAN

For the results presented in this section, we assume the use of DBSCAN for the anomaly detection task. DBSCAN [35] is an anomaly detection algorithm that clusters samples based on their pair-wise distances. The algorithm has two parameters: ϵ defines the radius around each sample within which other samples are considered *neighbors* (usually considering Euclidean distance); and M defines the minimum number of neighbors that a sample needs to have in order to be considered a normal sample. Any sample whose number of neighbors is less than M is considered an anomaly. We used the Scikit-Learn Python implementation of the DBSCAN algorithm. We fine tuned the model to obtain the best f1 score according to the dataset and the standardization procedures applied by testing the following sets of parameter values. We performed a hyperparameter search over $\epsilon = \{0.1, 0.5, 1, 1.5, 2, 3, 4, 5, 10\}$ and $M = \{3, 5, 8, 10, 12, 15, 20\}$. The best accuracy was obtained with M=15 and ϵ=1.0, which results in a f1 score equal to 0.8, with 0.073 false positive and 0.251 false negative rates. In a real deployment, strategies such as Window-based Attack Detection (WAD) [2] can be used to mitigate the false positive and false negative rates of DBSCAN, increasing f1 score to levels close to those achieved by XGBoost, e.g., 0.995.

Our algorithm was executed using a 10:1.5 proportion between normal (no attack) and anomalous (attack) samples. The experiments reported in Fig. 10 are obtained by running



(a) All

(b) In-Band (IB) jamming

(c) Out-of-Band (OOB) jamming

(d) Polarization Modulation (PM)

Fig. 9. Feature importance for the decision trees trained to identify all attacks, and each one of them individually.

the algorithm for each attack scenario, averaged over 50 runs. Since in this work the focus is on obtaining a representative AV, DBSCAN was executed with a relatively large number of samples, i.e., 200 normal (no attack) samples and 30 anomalous (attack) samples, which represents a reasonable trade-off between accuracy and runtime.

Fig. 10 shows the AVs obtained for the different attack scenarios. Note that light and strong attack conditions are shown in the same plot due to space reasons, but in a real-world scenario, a plot similar to the one seen in Fig. 6 is expected to be shown. Out of the three attack strategies, IB jamming (Fig. 10a) is the one that incurs the highest relative change on the OPM parameters. However, PM incurs the changes over the highest number of OPM parameters.

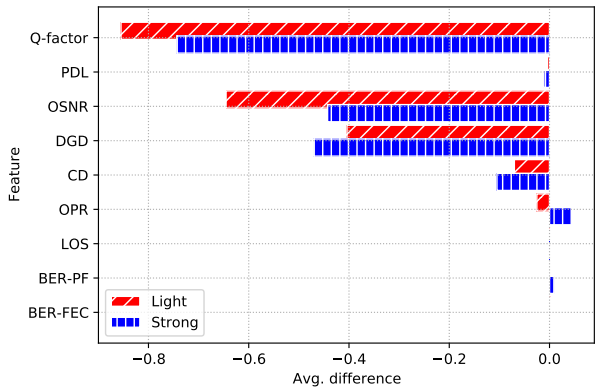
The AVs for IB and OOB jamming (Figs. 10a and 10b), show interesting differences between the two attacks. While for IB the CD shows no variation, OOB incurs a significant change in its value. OPR has a slight variation, while DGD, OSNR and Q-factor are the most affected OPM parameters. BER-FEC, BER-PF, LOS and PDL do not show significant fluctuations for IB and OOB jamming attacks. The PM attack shows a very different AV. It affects BER-PF in addition to other parameters, and the OSNR variations are both in the positive and negative directions, indicating that there might be a fluctuation of this OPM parameter throughout an attack.

D. Discussion

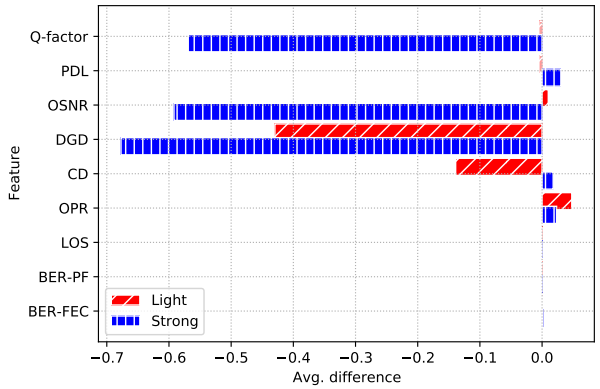
This work analyzes two fundamentally different ways to implement RCA while using ML models. One is the XGBoost, which is a SL algorithm that, based on the training dataset, builds decision trees capable of performing regression or classification. The other is DBSCAN, which is an UL algorithm that has no training and is executed over a set of samples every time an anomaly detection needs to be performed.

XGBoost is designed to capture the behavior of a phenomena (optical physical layer attacks in our case) by looking at a (historical) dataset that has enough samples to properly represent how the different features (or OPM parameters in our case) are affected by the phenomena. In our case, we use 50% of our dataset to train XGBoost, which represents 12 hours of OPM monitoring for each attack scenario. However, using current ML models, this training needs to be performed for each optical channel, and having these long monitoring windows prior to the optical channel use (i.e., prior to start sending user traffic) are impractical in real world deployments.

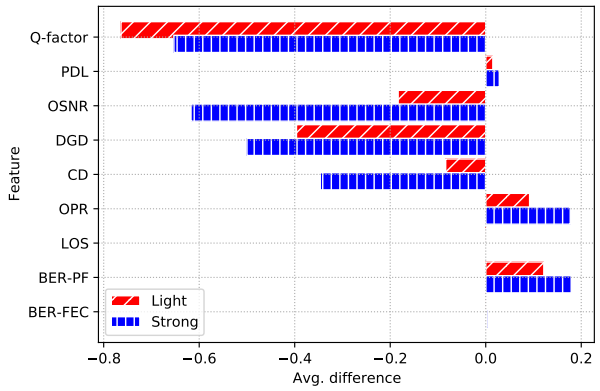
DBSCAN, on the other hand, is designed to identify changes on the feature (or OPM parameters in our case) trends as they happen. This means that it does not need any prior training, and can be used shortly after the optical channel establishment. In our case, we used 200 samples, which represents 3 hours and 20 minutes considering a 1-minute monitoring window.



(a) In-Band (IB) jamming



(b) Out-of-Band (OOB) jamming



(c) Polarization Modulation (PM)

Fig. 10. Anomaly Vector (AV) for each attack scenario.

However, this window can be shortened by increasing the monitoring frequency at the beginning of the optical channel operation, or by using a lower number of samples as input to DBSCAN at the expense of potentially lower accuracy. Such lower accuracy can be mitigated by, e.g., leveraging WAD [2].

From their very different approaches to training and inference, it can be expected for the RCA results from XGBoost and DBSCAN to diverge to some extent. Fig. 9 shows that one of the most important features found by XGBoost is BER-FEC, which presents negligible variations in the AVs for all attack scenarios in Fig. 10. On the contrary, OPM parameters such as DGD show a significant variation in the AVs (Fig. 10), but do not figure among the top five most important features in Fig. 9. These divergences are explained by the different time scales that the algorithms have access to. For instance, the variations caused to BER-FEC by attacks can be very small, but consistent. By analyzing a large number of samples, XGBoost can consistently and confidently identify these small variations that are present in the attack samples and leverage this to build the decision trees. Conversely, OPM parameters such as DGD can have inconsistent patterns during an attack, making it harder to consistently and confidently use it to determine whether there is an attack or not. In this case, using BER-FEC to identify a small but consistent difference is more effective than using DGD. This effect can also be observed with respect to OPR, which shows a small variation in the AVs (Fig. 10a), but figures among the top two most important OPM parameters for the IB jamming (Fig. 9b). However, few OPM parameters are identified as valuable by both algorithms. For instance, BER-PF shows variations in the AV only for the PM attack in Fig. 10. Likewise, Fig. 9 shows BER-PF among the important features only for the PM attack.

This demonstrates that the DB-RCA algorithm proposed in this work is an effective way to illustrate the important features that were decisive for the samples to be considered anomalies. Moreover, in some cases, it shows similar results as other algorithms that require much more data to be trained. The AV visualization equips the operators with deeper insight into the anomaly structure and eases the physical interpretation of the anomaly thus complementing the ML-assisted RCA tool. AV visualization goes beyond the typical historical data plot that is provided by NMSs today, where simple time series of OPM parameters are presented to the operators. This is especially significant when a new, previously undetected anomaly is analyzed.

VII. Conclusions

This work investigated ML-based techniques for RCA as an important enabler of autonomous optical network security management. The considered RCA framework may rely on either existing SL-based approaches, such as XGBoost, or on the newly proposed UL-based DB-RCA algorithm, or both. The proposed DB-RCA algorithm also enables a useful graphical representation of a detected anomaly to the network operator in the form of an anomaly vector. An in-depth analysis reveals the performance of the two fundamentally different approaches and uncovers their advantages and drawbacks when applied to the physical-layer security diagnostics. Although differences are observed, we can see that both methods can show significant insights into the properties of the experienced attack. For future work, we plan to investigate the practical implementation aspects of the frameworks in real-world NMSs and evaluate the applicability of the approaches to other anomaly detection use cases. For instance, the proposed framework and algorithm has potential for application to fault management, where interpretability is key to define mitigation actions.

REFERENCES

[1] N. Skorin-Kapov, M. Furdek, S. Zsigmond, and L. Wosinska, "Physical-layer security in evolving optical networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 110–117, Aug 2016, DOI: 10.1109/MCOM.2016.7537185.

[2] M. Furdek, C. Natalino, F. Lipp, D. Hock, A. D. Giglio, and M. Schiano, "Machine learning for optical network security monitoring: A practical perspective," *J. Lightwave Techn.*, vol. 38, no. 11, pp. 2860–2871, 2020, DOI: 10.1109/JLT.2020.2987032.

[3] D. Rafique *et al.*, "Cognitive assurance architecture for optical network fault management," *J. Lightwave Technol.*, vol. 36, no. 7, pp. 1443–1450, Apr 2018, DOI: 10.1109/JLT.2017.2781540.

[4] R. M. Morais and J. Pedro, "Machine learning models for estimating quality of transmission in dwdm networks," *Journal of Optical Communications and Networking*, vol. 10, no. 10, pp. D84–D99, 2018, DOI: 10.1364/JOCN.10.000D84.

[5] E. Seve, J. Pesic, C. Delezoide, S. Bigo, and Y. Pointurier, "Learning process for reducing uncertainties on network parameters and design margins," *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A298–A306, 2018, DOI: 10.1364/JOCN.10.00A298.

[6] C. Natalino, M. Schiano, A. Di Giglio, L. Wosinska, and M. Furdek, "Experimental study of machine-learning-based detection and identification of physical-layer attacks in optical networks," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4173–4182, 2019, DOI: 10.1109/JLT.2019.2923558.

[7] M. Furdek, C. Natalino, A. Di Giglio, and M. Schiano, "Optical network security management: requirements, architecture, and efficient machine learning models for detection of evolving threats [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 13, no. 2, pp. A144–A155, 2021, DOI: 10.1364/JOCN.402884.

[8] C. Natalino, A. Di Giglio, M. Schiano, and M. Furdek, "Autonomous security management in optical networks," in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2021, p. Tu1I.1.

[9] J. M. N. Gonzalez, J. A. Jimenez, J. C. D. Lopez, and H. A. Parada G, "Root cause analysis of network failures using machine learning and summarization techniques," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 126–131, 2017, DOI: 10.1109/MCOM.2017.1700066.

[10] C. Zhang, D. Wang, C. Song, L. Wang, J. Song, L. Guan, and M. Zhang, "Interpretable learning algorithm based on xgboost for fault prediction in optical network," in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, p. Th1F.3.

[11] C. Natalino, A. Di Giglio, M. Schiano, and M. Furdek, "Root cause analysis for autonomous optical networks: A physical layer security use case," in *European Conference on Optical Communications (ECOC)*, 2020, pp. We2K–1.

[12] N. Skorin-Kapov, J. Chen, and L. Wosinska, "A new approach to optical networks security: Attack-aware routing and wavelength assignment," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 750–760, 2010, DOI: 10.1109/TNET.2009.2031555.

[13] M. Furdek, N. Skorin-Kapov, and M. Grbac, "Attack-aware wavelength assignment for localization of in-band crosstalk attack propagation," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 2, no. 11, pp. 1000–1009, 2010, DOI: 10.1364/JOCN.2.001000.

[14] J. Zhu, B. Zhao, W. Lu, and Z. Zhu, "Attack-aware service provisioning to enhance physical-layer security in multi-domain EONs," *Journal of Lightwave Technology*, vol. 34, no. 11, pp. 2645–2655, 2016, DOI: 10.1109/JLT.2016.2541779.

[15] J. Zhu and Z. Zhu, "Physical-layer security in MCF-based SDM-EONs: Would crosstalk-aware service provisioning be good enough?" *J. Lightwave Techn.*, vol. 35, no. 22, pp. 4826–4837, 2017, DOI: 10.1109/JLT.2017.2757956.

[16] G. Savva, K. Manousakis, J. Rak, I. Tomkos, and G. Ellinas, "High-power jamming attack mitigation techniques in spectrally-spatially flexible optical networks," *IEEE Access*, pp. 1–1, 2021, DOI: 10.1109/ACCESS.2021.3058259.

[17] K. Manousakis and G. Ellinas, "Crosstalk-aware routing spectrum assignment and wss placement in flexible grid optical networks," *Journal of Lightwave Technology*, vol. 35, no. 9, pp. 1477–1489, 2017, DOI: 10.1109/JLT.2017.2681943.

[18] N. Skorin-Kapov, A. Jirattigalachote, and L. Wosinska, "An integer linear programming formulation for power equalization placement to limit jamming attack propagation in transparent optical networks," *Security and Communication Networks*, vol. 7, no. 12, pp. 2463 – 2468, 11 2014, DOI: 10.1002/sec.958.

[19] P. Lu, Z. Yan, L. Yi, J. Zhu, and Z. Zhu, "Resilient control plane design for inter-datacenter cloud network with various attacks," in *2021 7th International Conference on Computer and Communications (ICCC)*, 2021, pp. 1463–1468, DOI: 10.1109/ICCC54389.2021.9674273.

[20] X. Chen, C.-Y. Liu, R. Proietti, J. Yin, Z. Li, and S. J. B. Yoo, "On cooperative fault management in multi-domain optical networks using hybrid learning," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 4, pp. 1–9, 2022, DOI: 10.1109/JSTQE.2022.3151878.

[21] F. Musumeci, C. Rottondi, G. Corani, S. Shahkarami, F. Cugini, and M. Tornatore, "A tutorial on machine learning for failure management in optical networks," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4125–4139, 2019, DOI: 10.1109/JLT.2019.2922586.

[22] Y. Li, N. Hua, Y. Yu, Q. Luo, and X. Zheng, "Light source and trail recognition via optical spectrum feature analysis for optical network security," *IEEE Communications Letters*, vol. 22, no. 5, pp. 982–985, 2018, DOI: 10.1109/LCOMM.2018.2801869.

[23] M. Furdek, C. Natalino, M. Schiano, and A. D. Giglio, "Experiment-based detection of service disruption attacks in optical networks using data analytics and unsupervised learning," in *Metro and Data Center Optical Networks and Short-Reach Links II*. SPIE, 2019, pp. 73 – 82, DOI: 10.1117/12.2509613.

[24] M. Furdek, N. Skorin-Kapov, and L. Wosinska, "Attack-aware dedicated path protection in optical networks," *Journal of Lightwave Technology*, vol. 34, no. 4, pp. 1050–1061, 2016, DOI: 10.1109/JLT.2015.2509161.

[25] R. Casellas, R. Martínez, R. Vilalta, and R. Muñoz, "Control, management, and orchestration of optical networks: Evolution, trends, and challenges," *Journal of Lightwave Technology*, vol. 36, no. 7, pp. 1390–1402, 2018, DOI: 10.1109/JLT.2018.2793464.

[26] A. Sadasivarao, S. Bardhan, S. Syed, B. Lu, and L. Paraschis, "Optonomic: Architecture for secure autonomic optical transport networks," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, pp. 321–328.

[27] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021, DOI: 10.3390/e23010018.

[28] Z. Wang, M. Zhang, D. Wang, C. Song, M. Liu, J. Li, L. Lou, and Z. Liu, "Failure prediction using machine learning and time series in optical network," *Opt. Express*, vol. 25, no. 16, pp. 18 553–18 565, Aug 2017, DOI: 10.1364/OE.25.018553.

[29] D. Rafique, T. Szyrkowiec, A. Autenrieth, and J.-P. Elbers, "Analytics-driven fault discovery and diagnosis for cognitive root cause analysis," in *Optical Fiber Communication Conference*, 2018, p. W4F.6, DOI: 10.1364/OFC.2018.W4F.6.

[30] D. Das, M. F. Imteyaz, J. Bapat, and D. Das, "A non-intrusive failure prediction mechanism for deployed optical networks," in *International Conference on COMmunication Systems NETworkS (COMSNETS)*, 2021, pp. 24–28, DOI: 10.1109/COMSNETS51098.2021.9352868.

[31] C. Natalino, A. Udalcovs, L. Wosinska, O. Ozolins, and M. Furdek, "Spectrum anomaly detection for optical network monitoring using deep unsupervised learning," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1583–1586, 2021, DOI: 10.1109/LCOMM.2021.3055064.

[32] 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*, New York, NY, USA, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.

[33] S. M. Lundberg, G. G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable ai for trees: From local explanations to global understanding," *CoRR*, 2019, arXiv abs/1905.04610.

[34] X. Chen, B. Li, R. Proietti, Z. Zhu, and S. J. B. Yoo, "Self-taught anomaly detection with hybrid unsupervised/supervised machine learning in optical networks," *J. Lightwave Techn.*, vol. 37, no. 7, pp. 1742–1749, April 2019, DOI: 10.1109/JLT.2019.2902487.

[35] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, Jul. 2017, DOI: 10.1145/3068335.

[36] R. Schmogrow, B. Nebendahl, M. Winter, A. Josten, D. Hillerkuss, S. Koenig, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, W. Freude, and J. Leuthold, "Error vector magnitude as a performance measure for advanced modulation formats," *IEEE Photonics Technology Letters*, vol. 24, no. 1, pp. 61–63, 2012, DOI: 10.1109/LPT.2011.2172405.

Carlos Natalino (M'16) received the M.Sc. and Ph.D. degrees in electrical engineering from the Federal University of Pará, Brazil, in 2011 and 2016, respectively. From June 2016 to March 2019, he was a Postdoctoral Fellow with KTH Royal Institute of Technology, Stockholm, Sweden, where he was previously a visiting researcher from 2013 to 2014. He is currently a Postdoctoral Researcher with Chalmers University of Technology, Gothenburg, Sweden. He has authored or coauthored more than 50 papers published in international conferences and journals. His research focuses on the application of machine learning techniques for the design and operation of optical networks. He has served as a TPC member of several international conferences and workshops.

Marco Schiano is in the Transport Innovation group of Telecom Italia where he works on the progress of backbone and metro networks. He has been the Coordinator of the IST-NOBEL project funded by the European Commission. His current interests include transparent optical networks, ultra-high capacity transmission systems, and network dimensioning and optimization. He is co-author of more than forty publications and four patents.

Andrea Di Giglio received a Dr. Ing. degree in Electronic Engineering from the University of Pisa, Italy, and the Engineering License degree from Scuola S. Anna. He joined Telecom Italia Lab (formerly CSELT), that is the Telecom Italia Group's Company for study, research, experimentation and qualification in the field of Telecommunications and Information Technology. The fields of his research are addressed toward Internet Security, Storage Area Networks and Optical Networks Architecture and 5G. He was involved in EURESCOM and European Projects on IP over optical networks; he was involved in the architectural Work Package of the IST Project Nobel (Phase 1 and 2) "Next generation Optical networks for Broadband European Leadership" as Work Package leader and ICT STRONGEST and H2020 ICT-19 5G-SOLUTIONS as project coordinator. Dr. Di Giglio is the author of dozens of publications, including books, conference papers and workshops.

Marija Furdek (M'09-SM'17) obtained her Docent degree in optical networking from KTH Royal Institute of Technology in Stockholm, Sweden, and her PhD and Dipl. Ing. degrees in electrical engineering from the University of Zagreb, Croatia, in 2012 and 2008, respectively. She has been an Assistant Professor at Chalmers University of Technology in Gothenburg, Sweden, since 2019. From 2013 to 2019 she was a Postdoc and then Senior Researcher at KTH. She was a visiting researcher at Telecom Italia, Italy, Massachusetts Institute of Technology, USA, and Auckland University of Technology, New Zealand. Her research interests encompass optical network design and automation, with a focus on resiliency and physical-layer security.

Dr. Furdek is the PI of the project Safeguarding optical communication networks from cyber-security attacks, funded by the Swedish Research Council. She has co-authored more than 100 scientific publications in international journals and conferences, 5 of which received best paper awards. She is an Associate Editor of IEEE/Optica Journal of Optical Communications and Networking and Photonic Network Communications, and has been a Guest Editor of IEEE/Optica Journal of Lightwave Technology and IEEE Journal of Selected Topics in Quantum Electronics. She is a Senior Member of IEEE and Optica.