Isolation Forest as a Tool for Entangled Photon Detection

Dmytro Babets, Zbigniew Opilski, Volodymyr Hnatushenko, Kashtan Vita, Agnieszka Michalczuk,
Olena Sdvyzhkova, Erwin Maciak and Krzysztof A. Cyran

Abstract — **Entangled photon detection is essential for advancements in quantum communication, cryptography, and fundamental quantum mechanics experiments. This study introduces a novel application of unsupervised machine learning for identifying potential entangled photon events by analyzing voltage signals recorded from Silicon Multiplier Amplified Detectors (SiMPs). By framing photon detection as an anomaly detection problem, we employ the Isolation Forest (iForest) algorithm to isolate rare and distinctive signal patterns within large, noisy datasets without requiring labeled training data. This is the first application of iForest in the context of entangled photon detection. The method enables automated identification of anomalous events exhibiting time correlations across multiple measurement channels, offering a scalable and computationally efficient solution for real-time processing of experimental data in quantum optics.**

I. Introduction

The detection of entangled photon pairs is a critical task in quantum optics, underpinning advancements in quantum communication, cryptography, and foundational tests of quantum mechanics [1]. In our experiments, entangled photon pairs are generated using a periodically poled Potassium Titanyl Phosphate (ppKTP) crystal, which facilitates the type-II spontaneous parametric down-conversion (SPDC) process. This process results in the splitting of a 405 nm photon into two 810 nm photons that are nominally entangled in the polarization domain. Photon signals are captured via Silicon Multiplier Amplified Detectors (SiMPs) connected to oscilloscope channels [2]. These devices record voltage signals produced by photon interactions, enabling a detailed analysis of polarization correlations.

Detecting these photons, however, presents several challenges. The generated entangled photon pairs are inherently rare events, deeply embedded within a noisy signal environment. This noise arises from various sources, including dark current in the SiMP detectors, environmental photons, and spurious avalanche events. Consequently, a reliable method for isolating true photon detection events from noise is essential. The complexity of the task is further compounded by the need to simultaneously analyze signals from two oscilloscope channels corresponding to orthogonal polarizations, which requires precise time and voltage correlations to reliably infer entanglement.

Traditional photon detection methods, such as simple threshold-based approaches, often struggle to distinguish genuine photon events from noise in such challenging environments. This limitation motivates the exploration of advanced techniques, including machine learning algorithms that can adapt to complex signal patterns and improve detection accuracy. Among these techniques, the Isolation Forest (iForest) algorithm, introduced by Liu et al. in 2008 [3], has emerged as a powerful tool for anomaly detection due to its efficiency, scalability, and its ability to handle large datasets with low memory requirements. Unlike conventional approaches that rely on profiling normal data, iForest isolates anomalies through recursive partitioning in an ensemble of binary trees, making it particularly effective for identifying rare and unusual patterns in data [4]. Although our approach is designed to detect entangled photon events, it should be noted that the anomalies identified by iForest may also reflect time-correlated photon detections. Thus, while promising, the results must be interpreted with caution until further analyses can conclusively confirm photon entanglement.

The versatility of the iForest algorithm is well documented across various domains, including cybersecurity, finance, healthcare, and web traffic analysis. In cybersecurity, iForest has proven effective in intrusion detection systems by identifying malicious activities within network traffic. For example, Laskar et al. [5] demonstrated the integration of iForest with K-Means clustering for anomaly detection in industrial big data scenarios, highlighting its capability in monitoring and securing computer networks. In the financial sector, iForest has been employed to detect fraudulent transactions by identifying deviations from typical patterns, making it suitable for real-time fraud detection [6]. In healthcare, iForest has been used to monitor physiological signals, enabling the detection of anomalies that may signify underlying medical conditions [7]. Additionally, in web traffic analysis, iForest has been applied to distinguish anomalous patterns from normal traffic, further emphasizing its versatility and effectiveness in managing complex datasets [8].

In the context of quantum optics, iForest offers a promising solution for detecting entangled photon pairs due to its ability to identify rare and distinct events without requiring labeled training data – a particularly advantageous feature in experimental settings where labeled data is scarce or difficult to obtain [9]. In this study, we adapt iForest to detect photon events by analyzing voltage signals recorded from the two SiMP-connected oscilloscope channels. The algorithm's ability to rapidly isolate anomalies, coupled with its low computational complexity, makes it an ideal choice for processing the large datasets generated during experiments. By treating photon detection as an anomaly detection problem, we aim to identify time-correlated events

D. Babets, V. Kashtan, V. Hnatushenko, and O. Sdvyzhkova are with the Dnipro University of Technology, Dnipro, Ukraine (e-mail: babets.d.v@nmu.one; ORCID: 0000-0002-5486-9268, 0000-0002-0395-5895, 0000-0003-3140-3788, 0000-0001-6322-7526).

E. Maciak, Z. Opilski, A. Michalczuk, and K. A. Cyran are with the Silesian University of Technology, Gliwice, Poland (ORCID: 0000-0002-0620-4931, 0000-0001-6679-661X, 0000-0002-8963-1030, 0000-0003-1789-4939).

across the channels that may correspond to entangled photon pairs.

The rest of the paper is organized as follows: Section II describes the experimental setup and data acquisition process. Section III details the methodology, including data preprocessing and the application of the Isolation Forest algorithm. Section IV presents and discusses the results. Finally, Section V concludes the paper and outlines future directions.

II. METHODS

A. Isolation Forest

Isolation Forest is a model specifically designed for anomaly detection that isolates observations by constructing random binary trees, known as Isolation Trees (Fig. 1). The core principle of iForest is based on the fact that anomalies are "few and different," making them easier to isolate compared to normal data [3]. Each iTree is built by recursively splitting the data using randomly selected features and thresholds until all instances are isolated. The average path length (i.e., the number of splits required to isolate a data point) serves as the basis for assigning anomaly scores; shorter path lengths correspond to higher anomaly likelihoods.

To detect anomalies effectively, iForest employs an ensemble approach, where multiple iTrees are generated using random subsets of the dataset (Fig. 2). This approach ensures robustness and reduces the impact of random noise, while the algorithm maintains linear computational complexity $O(n\cdot\psi\cdot\log(\psi))$, where n is the dataset size and ψ is the sub-sample size.

B. An Anomaly Score

The anomaly score in the Isolation Forest (iForest) method quantifies the degree of deviation of a data point from the norm. It is derived from the path length $h(x)$, defined as the number of edges traversed from the root node to a terminal node in an isolation tree (iTree). Anomalous points, being sparse and distinct, generally exhibit shorter path lengths compared to normal data points.

For normalization, the average path length for a dataset of n instances, denoted as $c(n)$, is approximated using the harmonic number $H(i)$, defined as:

$$H(i)=ln(i)+\gamma. \tag{1}$$

where γ is the Euler-Mascheroni constant (≈0.577). The average path length for unsuccessful searches in a binary search tree, analogous to termination in iTrees, is given by:

$$c(n) = 2H(n-1) - 2(n-1)/n. \tag{2}$$

Using this, the anomaly score $s(x,n)$ for a data point x is calculated as:

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}} \tag{3}$$

where $E(h(x))$ represents the mean path length over an ensemble of iTrees.

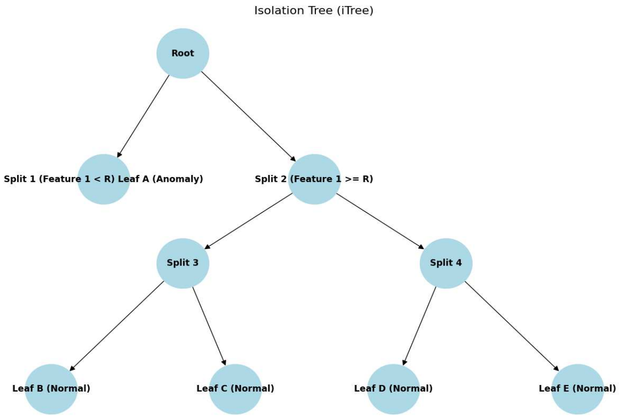

Figure 1. An Isolation Tree (iTree) used in the Isolation Forest algorithm. The tree recursively splits data based on randomly chosen features and thresholds (R). Shorter path lengths (e.g., data point A) correspond to higher anomaly likelihoods, while longer path lengths (e.g., data points B, C, D, E) are indicative of normal instances

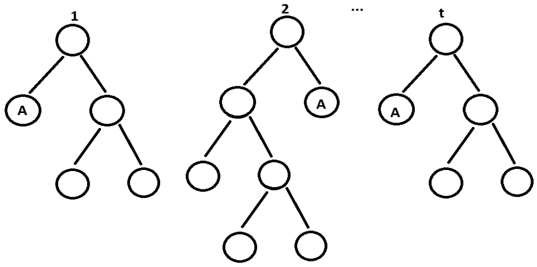

Figure 2. Ensemble of Isolation Trees (iTrees).

This formulation ensures that the anomaly score $s(x,n)$ effectively captures the degree of deviation of a data point x from the general data distribution. Depending on the value of $s(x,n)$, three distinct scenarios can be identified.

- High likelihood of being an anomaly: When the average path length $E(h(x))$ approaches zero, the anomaly score $s(x,n)$ approaches 1. This indicates that the data point x is highly likely to be an anomaly, as it is isolated very early in the construction of the isolation tree.
- Indistinguishability from normal data: When $E(h(x))$ is approximately equal to the average path length $c(n)$, the anomaly score $s(x,n)$ approaches 0.5. This suggests that the data point x cannot be distinguished from the normal data distribution and is not considered anomalous.
- Typicality of the data point: When $E(h(x))$ approaches the maximum possible path length $n-1$, the anomaly score $s(x,n)$ approaches 0. This implies that the data point x is a typical member of the dataset and shares similar characteristics with the majority of the observations.

These three properties enable a precise interpretation of the anomaly score $s(x,n)$, making it a robust metric for identifying data points that deviate from the general distribution.

Therefore, the anomaly score is bounded between 0 and 1, with higher values indicating greater anomaly. By ranking data points based on their anomaly scores, iForest effectively isolates rare, high-value photon detection events amidst noisy observations. This capability is particularly suited to the sparsity and distinctness of entangled photon signals, which are fundamental to quantum communication experiments.

C. *Application to Entangled Photon Detection*

In our experimental setup, entangled photon pairs are generated through a periodically poled Potassium Titanyl Phosphate (ppKTP) crystal using the type-II spontaneous parametric down-conversion (SPDC) process, as shown in Fig. 3. This optical configuration facilitates the generation of orthogonally polarized photon pairs, which are critical for studying entanglement phenomena.

Voltage signals resulting from photon detection are recorded from two oscilloscope channels, each corresponding to one of the two orthogonally polarized photon streams detected by Silicon Multiplier Amplified Detectors (SiMPs) [10]. These signals are inherently noisy, consisting of contributions from environmental photons, dark currents, and spurious avalanche events. To address this, we utilize the Isolation Forest (iForest) algorithm, treating photon detection as an anomaly detection problem to isolate significant photon events.

The experimental setup consists of a pulse generator that provides a reference signal for triggering the laser, while the oscilloscope records three signals: Channel 1 (CH1) captures the output from the first SiMP, Channel 2 (CH2) records the signal from the second SiMP, and Channel 3 (CH3) logs the reference signal from the pulse generator. In the experiments, the time window for detecting photon correlations is determined by the reference signal, which serves as the laser trigger. The laser's rising edge defines the Regions of Interest (ROIs), which encapsulate time intervals where entangled photon events are most likely to occur.

The proposed method adapts the Isolation Forest (iForest) algorithm to detect entangled photons in voltage signals recorded from two oscilloscope channels. Initially, potential anomalies are identified based on their signal characteristics using iForest, trained with an ensemble size of t=100 and a sub-sample size of ψ=256 to balance computational efficiency and detection performance [3]. Following anomaly detection, each identified event is examined to determine whether it falls within a Region of Interest (ROI), defined by rising values in the reference signal (CH3), which indicates laser activation. Events within these ROIs are further analyzed for time correlation across channels, leveraging time-of-flight measurements and signal synchronization to assess the likelihood of entanglement.

D. *Practical Implementation of Isolation Forest*

For the practical implementation of the iForest algorithm, Python was utilized alongside several libraries, including PyOD for anomaly detection, pandas for data manipulation, and matplotlib for visualization. The dataset, stored in a CSV file, consisted of three channels: CH1(V), CH2(V), and CH3(V). The features CH1(V) and CH2(V) were extracted as input variables for the model, while CH3(V) was retained for auxiliary analysis.

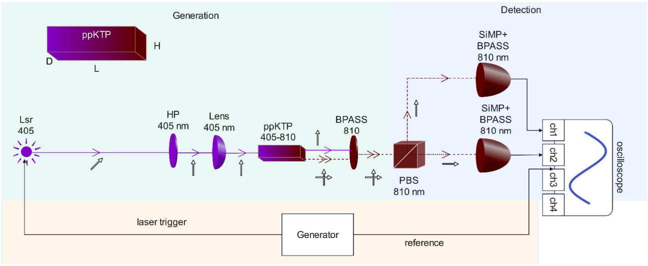

Figure 3. Optical setup for the experiments for photon correlation with ROI determination by laser trigger.

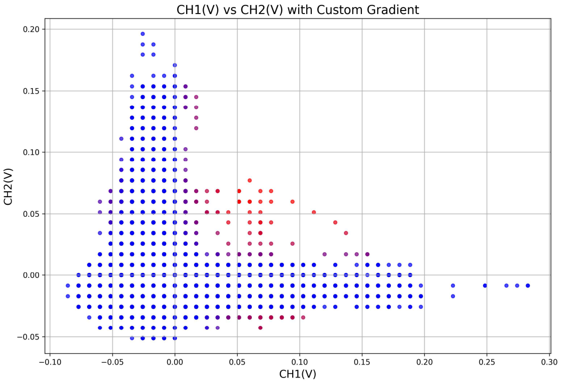

Figure 4. Scatter plot of CH1(V) vs. CH2(V) according to anomaly scores

The iForest model was initialized using the PyOD library's implementation, with a fixed random seed to ensure reproducibility. During training, the model constructed an ensemble of isolation trees, identifying anomalies based on their path lengths within the trees. After training, the model assigned anomaly scores to each data point, ranking observations according to their likelihood of being anomalous.

To visualize the overall distribution of data points and determine an appropriate contamination parameter, which represents the expected proportion of anomalies, we generate a scatter plot where each point corresponds to a measurement in the dataset. The x-axis represents CH1(V), while the y-axis represents CH2(V). The points are color-coded based on their anomaly scores, assigned by the Isolation Forest model (Fig. 4).

Normal points (with anomaly scores below 0.9) are displayed in blue. Potential anomalies are represented using a color gradient from light orange to red, where the intensity of the red hue increases as the anomaly score approaches 1. The most anomalous points, with scores close to 1, are highlighted in red and overlaid on top of normal points to enhance visibility.

This approach enables a smooth transition between normal and anomalous regions, avoiding a strict binary classification. It allows the model to better adapt to the dataset's characteristics and facilitates the selection of an appropriate contamination parameter.

Based on both the scatter plot in Fig. 4 and the proportion of points with anomaly scores exceeding 0.9 (as indicated by gradient coloring), the contamination level was set to 0.000005 (0.0005%) for further analysis.

III. RESULTS AND DISCUSSION

A. Isolation Forest model training and application

The Isolation Forest algorithm was applied to analyze three extensive datasets, each comprising 10 million observations. These datasets contained measurements recorded across three channels, CH1(V), CH2(V), and CH3(V), with values expressed in volts. The analysis aimed to identify anomalous patterns within the data, which may correspond to rare photon detection events.

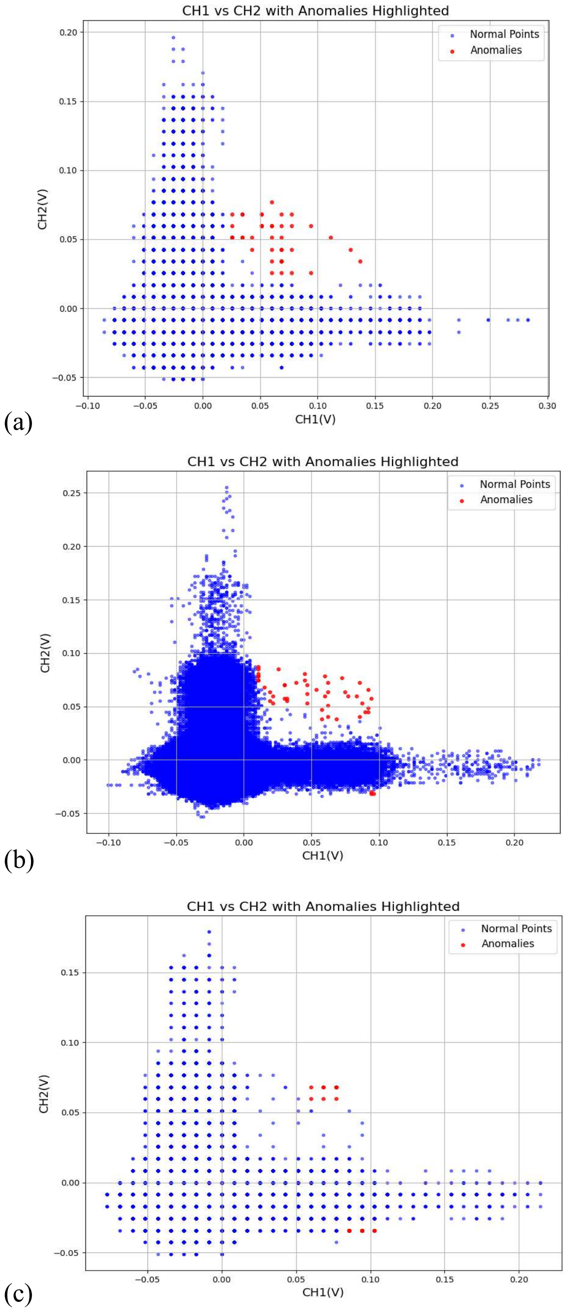

Figure 5. Scatter plots of CH1(V) vs. CH2(V) for contamination level 0.000005. (a): Dataset 1. Anomalies are highlighted in red, while normal data points are shown in blue. (b): Dataset 2. (c): Dataset 3

The algorithm's capacity to isolate anomalies was evaluated under varying contamination thresholds, enabling a detailed examination of its sensitivity and effectiveness in detecting deviations from the normal data distribution.

To refine the anomaly detection process, models were trained on three separate datasets at a contamination level of 0.000005 (0.0005%). These models were subsequently utilized to detect anomalies within their respective datasets, with results visualized through two-dimensional scatter plots of CH1(V) versus CH2(V).

The analysis revealed that the model trained on Dataset 1 (Fig. 5 a) produced satisfactory results, identifying anomalies predominantly in the central region of the scatter plot without any false positives. This outcome demonstrates the model’s capability to isolate statistically significant deviations effectively and with high precision. Conversely, the models trained on Dataset 2 and Dataset 3 exhibited limitations (Fig. 5 b,c). While some anomalies in the central region were correctly identified, the detection process failed to capture all critical deviations. Furthermore, both models flagged a notable number of false positives, indicating reduced specificity.

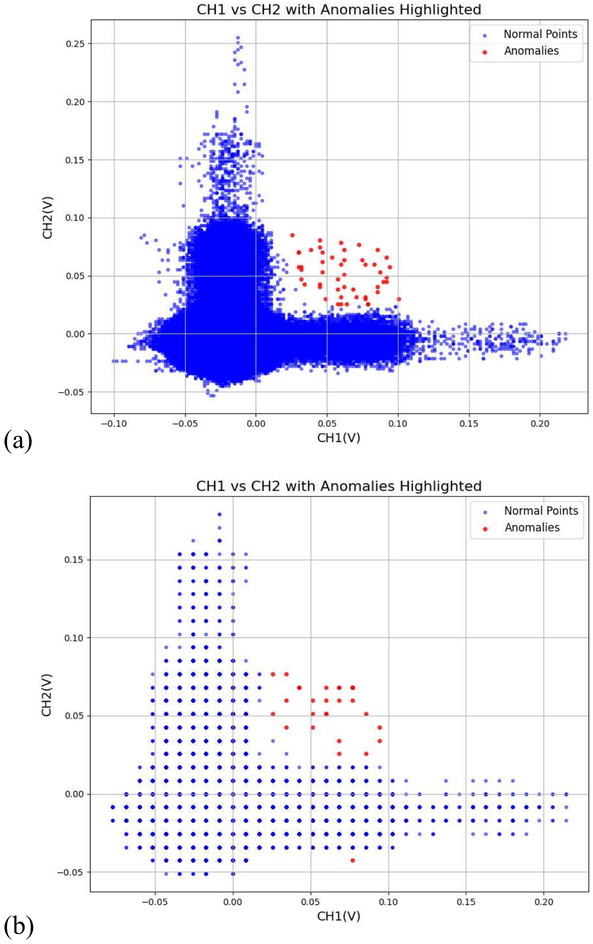

Figure 6. Scatter plots of CH1(V) vs. CH2(V) for contamination level 0.000005. (a): Trained on Dataset 1 applied to Dataset 2. (b): Trained on Dataset 1 applied to Dataset 3

Based on these observations, the decision was made to apply the model trained on Dataset 1 to the anomaly detection tasks for Dataset 2 and Dataset 3. The results, as illustrated in subsequent scatter plots, showed a significant improvement in performance (Fig. 6). Anomalies in the central region of the diagrams were accurately identified, with a marked reduction in false positives, thereby confirming the robustness and generalizability of the model trained on Dataset 1.

B. Visualization of Anomaly Detection Results

For a detailed analysis of potential events of entangled photon generation, we extract and visualize localized segments of the time series. The segmentation is based on detected anomaly clusters, where anomalies occurring within 10 consecutive indices are considered a single event. Each extracted segment consists of 100 consecutive data points, centered around the most significant anomaly within the cluster.

The visualization presents CH1(V) (blue) and CH2(V) (green) as continuous lines, adjusted so that their minimum value within the window is zero, enhancing readability by normalizing signal variations. Anomalous points in CH1(V) and CH2(V) are overlaid in orange and purple, respectively, to highlight deviations from normal behavior. A vertical dashed red line marks the index of the central anomaly within the segment, providing a clear reference for precise localization. In addition to CH1(V) and CH2(V), we include CH3(V), which serves as an external indicator relevant to the experimental conditions. Due to the significantly different magnitude of CH3(V) compared to CH1(V) and CH2(V), a scaling transformation is applied:

The minimum CH3(V) value within the selected segment is subtracted to shift the curve above zero:

$$CH3_{shift}=CH3\text{-}min(CH3_{win}). \tag{4}$$

The resulting values are scaled proportionally to the amplitude range of CH1(V) and CH2(V) to ensure appropriate visualization on the same plot:

$$CH3_{scaled} = CH3_{shift} \cdot \frac{max(X_{win}) - min(X_{win})}{max(CH3_{shift})} \tag{5}$$

The actual CH3(V) values remain interpretable via a secondary y-axis (right-hand side), where the original scale is preserved.

The CH3 signal plays a critical role in identifying potential events of entangled photon generation. This experiment utilizes a signal from a wave generator to define the Regions of Interest (ROIs) in which the correlation can appear. Specifically, the rising slope of this signal triggers the laser, marking the potential time frame for photon pair generation. Consequently, for an entangled photon generation event to be considered valid, the CH3 signal must exhibit an increasing trend within the analyzed window.

C. Detected Potential Events of Entangled Photon Generation

Two potential photon generation events were identified across the analyzed datasets. These events are characterized by peaks in both CH1 and CH2 signals, accompanied by a simultaneous increase in CH3 values within the corresponding time windows, indicating laser activation.

The anomaly group detected around index 9,324,414 in Dataset 1 exhibits peaks in both CH1 and CH2 signals (Fig. 7). The CH1 (blue) peak lags behind the CH2 (green) peak by 2 samples, corresponding to a time delay of 1 ns (each sample representing 0.5 ns). This temporal shift suggests a potential spatial separation, indicating a time correlation between the photons. The algorithm identified peak amplitudes of approximately 0.14 V, while the background noise level remains around 0.06 V.

Additionally, within this time window, the CH3 signal shows a rising trend, increasing from 3.0 V to 3.3 V. This confirms that the laser was active during the observed anomaly, supporting the hypothesis that the event may be related to photon pair generation.

However, the detected peaks are not sharply defined, making visual confirmation challenging. While the presence of time correlation and rising CH3 values suggests a plausible entanglement event, further validation is required.

A second anomaly group was detected around index 7,123,247 in Dataset 2. Similar to the first event, both CH1 and CH2 signals exhibit peaks, with a measured time delay of 3 samples (1.5 ns). The CH3 signal also exhibits a rising trend, increasing from 2.8 V to 3.15 V within the same time window (Fig. 8).

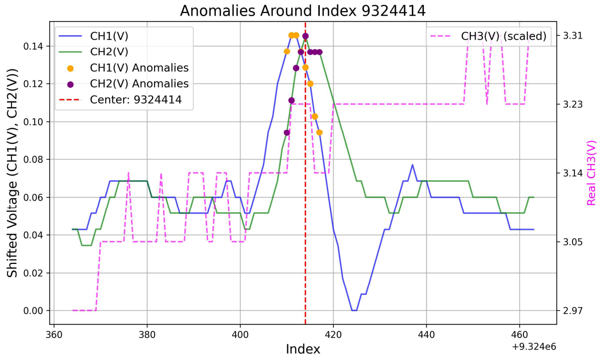

Figure 7. Anomaly group detected around index 9,324,414 in Dataset 1.

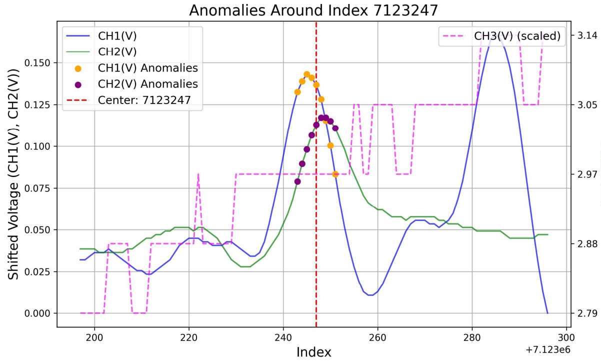

Figure 8. Anomaly group detected around index 7,123,247 in Dataset 2

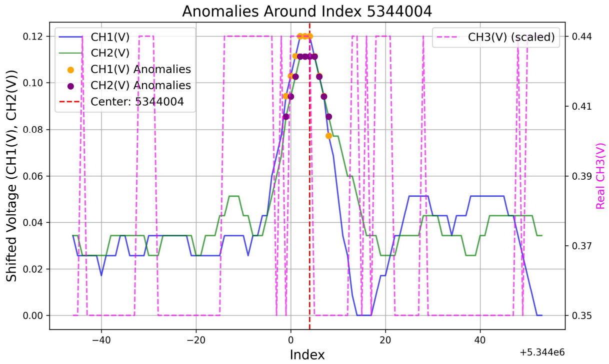

Figure 9. Anomaly group detected around index 5,344,004 in Dataset 3.

The observed 1.5 ns delay between CH1 and CH2 peaks suggests a consistent time correlation, potentially indicative of photon entanglement. However, as with the first event, the peak shapes are not sharply defined, requiring additional investigation to confirm their origin and significance.

Figure 9 illustrates an event from Dataset 3 that initially appeared to be a strong candidate for photon entanglement due to the precise time correlation between CH1 and CH2 signals. The peaks in both channels occur simultaneously, with no measurable delay. Additionally, both maxima exhibit flattened tops, suggesting that the algorithm identified peak positions at the midpoint of these regions. The rising slopes of CH1 and CH2 begin at the same sample and align perfectly, a characteristic behavior expected in single-photon avalanche diode (SiPM) operation, where a photon triggers the avalanche.

However, further analysis of the CH3 signal reveals that its values remain stable within the range of 0.35–0.44 V, without the characteristic increase associated with laser activation. Since entangled photon generation in this setup is only possible during laser pulses, the absence of a rising CH3 signal strongly suggests that this event is not related to spontaneous parametric down-conversion (SPDC) but is instead a result of dark current or other background noise in the detection system. While this observation exhibits excellent peak alignment, the lack of supporting laser activity indicates that it does not correspond to a genuine entangled photon event.

IV. Conclusion

This study presented a novel application of the Isolation Forest (iForest) algorithm for detecting potential entangled photon events based on voltage signals acquired from Silicon Multiplier Amplified Detectors (SiMPs). By conceptualizing photon detection as an anomaly detection problem, we demonstrated that iForest effectively isolates rare signal patterns within large and noisy datasets – without the need for labeled data. This capability is particularly valuable for identifying time-correlated events across polarization channels, which are indicative of photon entanglement.

This work represents the first known application of iForest to the domain of entangled photon detection and provides compelling evidence of its potential as a scalable and computationally efficient tool for quantum optics. The findings open new directions for real-time, data-driven processing in quantum experiments. Future research will aim to enhance temporal correlation analysis, integrate statistical confidence measures, and explore the fusion of iForest with deep learning models. Expanding this approach to broader quantum systems may further accelerate progress in quantum communication and computation.

References

[1] H. Defienne, M. Reichert, and J. Fleischer, "Adaptive quantum optics with spatially entangled photon pairs," *Phys. Rev. Lett.*, vol. 121, no. 23, p. 233601, 2018. doi: 10.1103/PhysRevLett.121.233601

[2] K. Wereszczyński, A. Michalczuk, M. Paszkuta, and J. Gumiela, "High-precision voltage measurement for optical quantum computation," *Energies*, vol. 15, no. 12, p. 4205, 2022. doi: 10.3390/en15124205

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422.

[4] W. S. Al Farizi, I. Hidayah, and M. N. Rizal, "Isolation forest based anomaly detection: A systematic literature review," in *Proc. 8th Int. Conf. Inf. Technol., Comput. Electr. Eng. (ICITACEE)*, Semarang, Indonesia, 2021, pp. 118–122. doi: 10.1109/ICITACEE53184.2021.9617498

[5] T. R. Laskar *et al.*, "Extending isolation forest for anomaly detection in big data via K-Means," *arXiv preprint*, arXiv:2104.13190, 2021. doi: 10.48550/arxiv.2104.13190

[6] H. Abbassi, S. Mendili, and Y. Gahi, "Digital banking fortification: A real-time isolation forest architecture for detecting online transaction fraud," *Eng. Res. Express*, vol. 6, 2024. doi: 10.1088/2631-8695/ad4958

[7] H. Bansal, B. Chinagundi, P. Rana, and N. Kumar, "Time series generative adversarial network for muscle force prognostication using statistical outlier detection," *Expert Syst.*, 2024. doi: 10.1111/exsy.13653

[8] J. Zhang, K. Jones, T. Song, H. Kang, and D. Brown, "Comparing unsupervised learning approaches to detect network intrusion using NetFlow data," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Charlottesville, VA, USA, 2017, pp. 122–127. doi: 10.1109/SIEDS.2017.7937701

[9] V. Hnatushenko and V. Zhernovyi, "Complex approach of high-resolution multispectral data engineering for deep neural network processing," in *Lect. Notes Comput. Intell. Decis. Making*, Cham: Springer, 2019, pp. 659–672. doi: 10.1007/978-3-030-26474-1_46

[10] F. Corsi *et al.*, "Electrical characterization of silicon photo-multiplier detectors for optimal front-end design," in *IEEE Nucl. Sci. Symp. Conf. Rec.*, 2006, vol. 2, pp. 1276–1280. doi: 10.1109/NSSMIC.2006.356076