

Early Failure Prediction from Near-Anomaly Detection: A Proactive Approach

Léa Billet ✉ 

Schaeffler, Toulouse, France

LAAS-CNRS, University of Toulouse, INSA, France

Louise Travé-Massuyès ✉ 

LAAS-CNRS, University of Toulouse, CNRS, France

Elodie Chanthery ✉ 

LAAS-CNRS, University of Toulouse, INSA, France

Alexandre Gaffet ✉ 

Schaeffler, Toulouse, France

Abstract

Anomaly detection methods often have uncertain behavior with respect to samples near the distribution boundary, limiting their ability to anticipate future anomalies. This work introduces the concept of near-anomalies that, while not yet anomalous, lie close to the boundary and are likely to transition into anomalies in the near future. To address this, we propose an unsupervised method, named Christoffel-based ANomaly Anticipation for eaRly dIScovery (CANARI), which leverages the strong theoretical foundations of the Christoffel function to detect near-anomalies. The method is validated on industrial in-circuit testing data from printed circuit boards, with synthetically generated near-anomaly samples due to the lack of real-world data labeling. Experimental results show that CANARI outperforms the compared baselines that generally use a dual-threshold mechanism (one for anomalies and one for near-anomalies). It therefore provides a proactive solution for anticipating anomalies before they occur, offering a promising approach for resilience, predictive maintenance, and quality control.

2012 ACM Subject Classification Computing methodologies → Anomaly detection

Keywords and phrases Unsupervised anomaly detection, Christoffel function, Near-anomalies

Digital Object Identifier 10.4230/OASICS.DX.2026.4

Funding Our work has benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the France 2030 program under the Grant agreements n°ANR-19-P3IA-0004 and n°ANR-23-IACL-0002.

Acknowledgements The authors would like to thank Christophe Merle, Head of the Manufacturing Intelligent 5 team at Schaeffler and ANITI Industrial Coordinator, for sharing his expertise in robotized production lines.

1 Introduction

Outliers, or anomalies, are typically defined as samples that are rare or out-of-distribution compared to inliers [11]. However, in many real-world applications, potentially problematic samples may not yet qualify as true anomalies at the time of observation. Instead, they may lie near the boundary of the normal distribution, remaining technically within its limits. We refer to such samples as near-anomalies. While not classical outliers, they are of critical interest because their detection enables the early anticipation of future anomalies, thereby enhancing system resilience. In this work, we focus on identifying these near-anomaly samples – those that, though still part of the distribution, are precariously close to its edge. The method is applied to data from industrial manufacturing plant product test stations.



© Léa Billet, Louise Travé-Massuyès, Elodie Chanthery, and Alexandre Gaffet; licensed under Creative Commons License CC-BY 4.0

37th International Conference on Principles of Diagnosis and Resilient Systems (DX 2026).

Editors: Ingo Pill, Marina Zanella, and Gregory Provan; Article No. 4; pp. 4:1–4:18

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Near-anomaly detection has been explored in image anomaly detection, where only a few pixels deviate from the normal distribution [13]. The generation of near-anomaly images has also been proposed for method evaluation [9], notably using diffusion-based models. However, near-anomaly detection for high-dimensional tabular data remains relatively unexplored, despite its strong potential for predicting later problems.

Tabular anomaly detection methods typically score samples within a distribution and apply a threshold to classify them as inliers or outliers [12]. They are usually trained in a semi-supervised manner, using only inlier data. In principle, these methods can be adapted to detect near-anomalies by introducing a second threshold closer to the anomaly boundary. This near-anomaly threshold may be set in a supervised manner, for example, using the dataset’s known contamination rate or defining it as a percentage of the primary anomaly threshold. However, in an unsupervised setting, such a hyperparameter cannot be automatically determined for all datasets.

In this paper, we propose a method, named Christoffel-based ANomaly Anticipation for eaRly dIScovery (CANARI), based on the Christoffel Function (CF), where a second threshold to detect near-anomaly samples is derived by leveraging the theoretical properties of the CF. While we use the CF-based method CLOE (CLOE) [2] for its scalability with respect to the feature space dimension, any other CF-based method, like DyCF [4], could equally apply even if it is more limited with respect to the dimension. CLOE relies on an autoencoder trained with a CF regularized loss to construct a well-structured, low-dimensional representation of the data, which can then be advantageously used for anomaly detection. We extend the CLOE framework to identify samples that, while not yet outliers, lie close enough to the decision boundary to be classified as near-anomalies – potential precursors to future anomalies.

Experiments are conducted on printed circuit boards (PCBs) using in-circuit testing (ICT) tabular data acquired on industrial manufacturing plant product test stations. In manufacturing plants, products undergo extensive testing during production to ensure quality, with results typically recorded as high-dimensional tabular data. Quality engineers define knowledge-based thresholds for each test, and products whose values remain near these thresholds may still pass inspection and proceed through the manufacturing process. Nevertheless, these borderline products can later exhibit issues in subsequent production steps or, more seriously, after delivery to the customer. Because thresholds are defined before all production features and distribution characteristics are fully known, such products are not necessarily labeled as anomalies at the time of testing. A possible approach would be to lower the test station thresholds in order to classify more products as anomalies. However, the appropriate reduction for each threshold is unknown and cannot be reliably determined in advance. Moreover, dependencies may exist between different testing station thresholds, which makes manual adjustment even more difficult. As a result, a more flexible method is needed to identify products that are not yet anomalous but are sufficiently close to abnormal conditions to warrant attention. Detecting them as near-anomalies would therefore provide an opportunity to raise early warnings of potential future quality problems.

The paper is organized as follows: background about the CF is presented in Section 2. Then the method CANARI is described in Section 3 introducing the concept of near-anomaly and its detection. In Section 4, the set of conducted experiments and the associated results are presented with a description of the industrial use case. Finally, the conclusion and future work are given in Section 5.

2 Background - The Christoffel function for anomaly detection

This section defines the CF and establishes the background for our method [6], [7].

2.1 The Christoffel Function

Let μ be a finite Borel measure¹ supported on $\Omega \subset \mathbb{R}^d$, a compact set with a non-empty interior. Let $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ and let $\alpha = (\alpha_i)_{i=1\dots d} \in \mathbb{N}^d$ be the vector of degrees associated to each variable for the monomial $\mathbf{x}^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}$ of total degree $\deg(\alpha) = \sum_{i=1}^d \alpha_i$. Let $\mathbb{N}_n^d = \{\mathbb{N}^d; \deg(\alpha) \leq n\}$ and $\mathbf{v}_n(\mathbf{x}) := (\mathbf{x}^\alpha)_{\alpha \in \mathbb{N}_n^d}$ be the vector of all monomials of degree less or equal to n graded in the lexicographic order². The size of the vector $\mathbf{v}_n(\mathbf{x})$ is equal to $s_d(n) = \binom{d+n}{n}$.

Let $\mathbf{M}_n(\mu)$ be the moment matrix of μ . This matrix encodes information that quantifies the location, the shape, and the scale of Ω . $\mathbf{M}_n(\mu)$ is a real symmetric, positive definite, and non-singular matrix for all n , and is given by

$$\mathbf{M}_n(\mu) = \int_{\mathbb{R}^d} \mathbf{v}_n(\mathbf{x}) \mathbf{v}_n(\mathbf{x})^T d\mu(\mathbf{x}) \quad (1)$$

Let us introduce the Christoffel-Darboux kernel K_n^μ associated with μ . Given any orthonormal basis $(p_i)_{i=1}^{s_d(n)}$ of $\mathbb{R}_n[\mathbf{x}]$, orthonormal with respect to the inner product induced by $\mathbf{M}_n(\mu)$, K_n^μ is defined as:

$$(\mathbf{x}, \mathbf{y}) \mapsto K_n^\mu(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{s_d(n)} p_i(\mathbf{x}) p_i(\mathbf{y}). \quad (2)$$

This kernel can also be expressed in terms of the moment matrix:

$$K_n^\mu(\mathbf{x}, \mathbf{y}) := \mathbf{v}_n(\mathbf{x})^T \mathbf{M}_n(\mu)^{-1} \mathbf{v}_n(\mathbf{y}). \quad (3)$$

► **Definition 1** (The Christoffel Function). *The Christoffel Function (CF) of degree $n \in \mathbb{N}$ associated with the measure μ , denoted by Λ_n^μ , is given for any $\mathbf{x} \in \mathbb{R}^d$ as:*

$$\Lambda_n^\mu(\mathbf{x}) = K_n^\mu(\mathbf{x}, \mathbf{x})^{-1}, \quad (4)$$

which can be shown to be the solution of this minimization problem:

$$\Lambda_n^\mu(\mathbf{x}) = \min_{P \in \mathbb{R}_n[\mathbf{x}]} \left\{ \int_{\Omega} P(\mathbf{z})^2 d\mu(\mathbf{z}), \quad P(\mathbf{x}) = 1 \right\}, \quad \forall \mathbf{x} \in \mathbb{R}^d. \quad (5)$$

The intuition behind the Christoffel function can be understood from the minimization problem (5). The solution polynomial P of (5) is constrained to satisfy $P(\mathbf{x}) = 1$ at a given point $\mathbf{x}$ while minimizing the L^2 norm $\int_{\Omega} P^2 d\mu$ over the support Ω . If $\mathbf{x} \notin \Omega$ (outside the support), P can be nearly zero on Ω while still satisfying $P(\mathbf{x}) = 1$, resulting in a small minimal value. Conversely, if $\mathbf{x} \in \Omega$ (inside the support), the constraint $P(\mathbf{x}) = 1$ forces P to have significant magnitude on Ω , preventing it from being close to zero and thus yielding a large minimal value.

One of the main and salient features of the CF is its ability to encode the support Ω . The geometric shape of Ω can indeed be captured by a specific level set $\Omega_\gamma := \{\mathbf{x} : \Lambda_n^\mu(\mathbf{x})^{-1} \leq \gamma\}$, defined for some $\gamma \in \mathbb{R}_+$, even for low degrees n [4].

¹ A *Borel measure* is a way to assign a “size” (like length, area, or probability) to subsets of a topological space (here Ω), where the subsets we can measure are all those that can be built from open sets using countable unions and complements. On the real line, the usual *length* is a Borel measure which tells how long any interval is, and by extension, how long more complex sets (like unions of intervals) are.

² Lexicographic order means that monomials are first sorted by degree; then, within each degree, they are ordered lexicographically by the variables, with $X_1 = a$, $X_2 = b$, etc.

2.2 The empirical Christoffel Function

In practical applications, the theoretical measure μ is unknown. Instead, one has access to a cloud of data points $\mathbb{X} = \{\mathbf{x}_i, i = 1, \dots, N\}$ sampled from μ and associated with a discrete measure $\mu_N := \frac{1}{N} \sum_{\mathbf{x} \in \mathbb{X}} \delta_{\mathbf{x}}$ supported on $\mathbb{X}$ (where $\delta_{\mathbf{x}}$ stands for the Dirac measure at $\mathbf{x}$).

The empirical version of the moment matrix writes as

$$\mathbf{M}_n(\mu_N) = \frac{1}{N} \sum_{\mathbf{x} \in \mathbb{X}_e} \mathbf{v}_n(\mathbf{x}) \mathbf{v}_n(\mathbf{x})^T. \quad (6)$$

To guarantee the invertibility of the matrix $\mathbf{M}_n(\mu_N)$, the size of $\mathbb{X}$ must be greater than $s_d(n)$ according to [7] (Corollary 6.3.5). Under the condition $|\mathbb{X}| = N > s_d(n)$, the inverse of the empirical CF (eCF) is defined as:

$$\Lambda_n^{\mu_N}(\mathbf{x})^{-1} := \mathbf{v}_n(\mathbf{x})^T \mathbf{M}_n(\mu_N)^{-1} \mathbf{v}_n(\mathbf{x}), \mathbf{x} \in \mathbb{X}. \quad (7)$$

It has been proven that the eCF converges to the CF as N tends to infinity [7] (Theorem 6.2.3). This convergence ensures that the properties of the CF also hold (under some conditions) for the eCF.

Consequently, one can expect the inverse eCF $\Lambda_n^{\mu_N}(\mathbf{x})^{-1}$ to capture the support $\mathbb{X}$ thanks to a specific level set. Defining *inliers* as data points in the support and *outliers* or *anomalies* as data points out of the support, $\Lambda_n^{\mu_N}(\mathbf{x})^{-1}$ hence comes as a good scoring function for outlier detection.

3 Near-anomaly detection

The goal of this work is to automatically detect potentially problematic samples that, while not yet true anomalies, may lie near the boundary of the normal distribution although remaining technically within its limits. To achieve this, we propose to use the dichotomy property of the growth of the CF with the degree n presented in the following section.

3.1 The Christoffel function theoretical growth

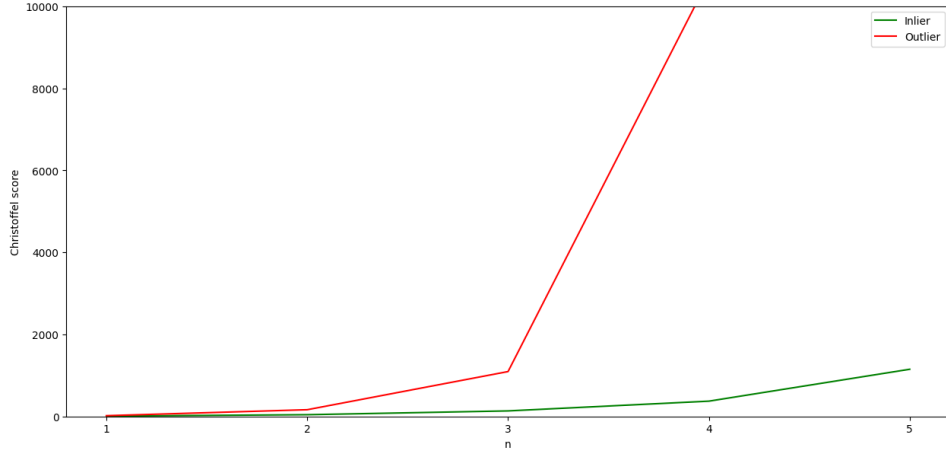
In addition to being able to capture the support of a discrete measure defined by a cloud of data points, the CF has been proven to have drastically different theoretical growths with the degree n for data points in the support of the measure, considered as inliers, and for data points out of the support of the measure, considered as outliers.

Two theorems from [7] characterize the growth of the CF as the degree n increases. The growth with n for data points in the support (inliers) is shown to be polynomial in Theorem 2 whereas it is shown to be exponential for data points out of the support (outliers) in Theorem 3. Both cases are illustrated in Figure 1 (for an outlier sample with the red graph and for an inlier sample with the green graph).

► **Theorem 2** (Lemma 4.3.2, p.51 [7]). *Let μ be a positive Borel measure supported on the compact set Ω , the closure of a bounded domain U with a nice boundary. Assume $\mu|_U$ has a density with respect to the Lebesgue measure λ restricted to U , bounded from below by $c > 0$. Let $\mathbf{x} \in \Omega$ and $\delta > 0$ such that $\text{dist}(\mathbf{x}, \partial U) \geq \delta$. Then, for any $n \in \mathbb{N}, n > 0$, and considering the equality (4), we have:*

$$\Lambda_n^{\mu}(\mathbf{x})^{-1} \leq s_d(n) \frac{2\lambda(\Omega)}{c\delta^d \omega_d} (1+d)^3 \quad (8)$$

where ω_d depends only on d and not on n .



■ **Figure 1** The figure represents the growth of an outlier in the red graph and of an inlier in the green graph as n grows from 1 to 5. The red graph increases much faster than the green graph, showing the dichotomy property stated by Theorems 2 and 3.

In equation 8 of Theorem 2, for fixed d , $s_d(n)$ is a polynomial in n of degree d :

$$s_d(n) = \binom{n+d}{n} = \frac{(n+d)(n+d-1)\cdots(n+1)}{d!} = \frac{n^d}{d!} + (\text{lower-order terms in } n).$$

So, as $n \rightarrow \infty$, it behaves like:

$$\binom{n+d}{n} \sim \frac{n^d}{d!}. \quad (9)$$

Theorem 2 hence shows that $\Lambda_n^\mu(\mathbf{x})^{-1}$ has polynomial growth with n inside the support.

► **Theorem 3** (Lemma 4.3.1, p.50 [7]). *Let μ be a positive Borel measure supported on the compact set $\Omega \subset \mathbb{R}^d$. Let $\mathbf{x} \notin \Omega$ and $\delta > 0$ such that $\text{dist}(\mathbf{x}, \Omega) \geq \delta$. Then, for any $n \in \mathbb{N}, n > 0$, and considering the equality (4), we have:*

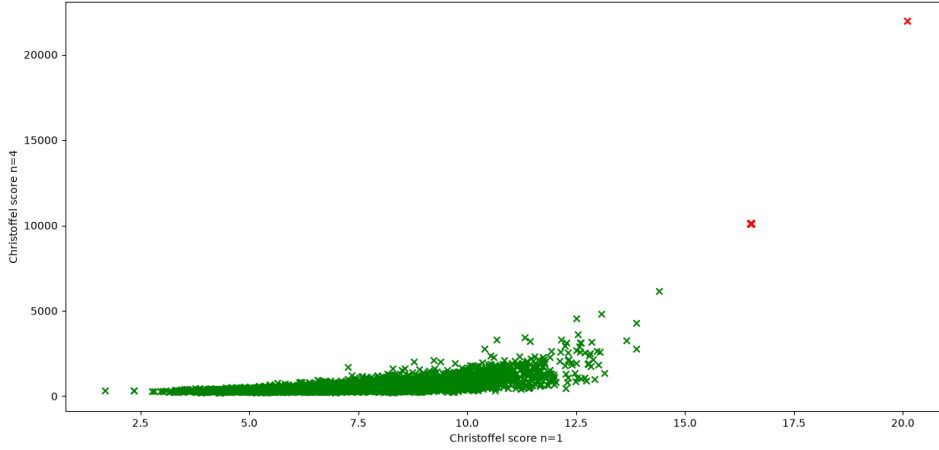
$$\frac{\Lambda_n^\mu(\mathbf{x})^{-1}}{s_d(n)} \geq 2^{\frac{\delta n}{\delta + \text{diam}(\Omega)}} - 3n^{-d} \left(\frac{d}{e}\right)^d \exp\left(-\frac{d^2}{n}\right) \quad (10)$$

Theorem 3 clearly shows that $\Lambda_n^\mu(\mathbf{x})^{-1}$ has exponential growth with n outside the support.

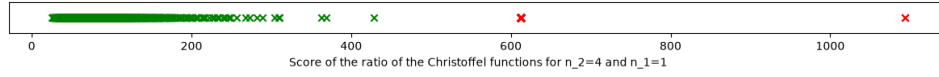
This theoretical growth with respect to the parameter n has already been leveraged by DyCG [4], which proposes a parameter-free outlier detection approach for data streams. DyCG evaluates the difference between two eCFs computed for distinct values of the parameter n , normalized by the difference between the corresponding values of n . For a fixed $x \in \mathbb{X}$, this score computes the slope of the function depending on n .

3.2 The Christoffel function growth ratio

To leverage the dichotomy property stated by Theorems 2 and 3, and given that $\Lambda_n^\mu(\mathbf{x})^{-1}$ is strictly positive, we propose to define the following *CF growth ratio*, based on two different values of n , namely n_1 and n_2 , $n_1 < n_2$ to compare eCF scores.



■ **Figure 2** The figure illustrates the eCF obtained for $n_1 = 1$ alongside that obtained for $n_2 = 4$. Inliers are in green and outliers are in red. Outlier and inlier scores are clearly distinguishable.



■ **Figure 3** The figure presents the values of the ratio between the two eCFs computed for $n_1 = 1$ and $n_2 = 4$. Inliers are represented in green, while the outliers are represented in red. A clear separation is observed, outliers are associated with much higher scores than inliers.

► **Definition 4** (CF growth ratio). *Given $\mathbf{x} \in \mathbb{X}$, the CF growth ratio is defined for $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$ as follows:*

$$\frac{\Lambda_{n_2}^\mu(\mathbf{x})^{-1}}{\Lambda_{n_1}^\mu(\mathbf{x})^{-1}} \quad (11)$$

A ratio was preferred to a normalized difference, such as the one used in DyCF, for two main reasons:

1. The ratio eliminates the effect of the absolute scale of the scores and avoids normalization, enabling a direct comparison of the degree's impact on the metric without bias related to the magnitude of the raw values,
2. The CF growth ratio comes with two interesting asymptotic properties.

The relative magnitude of the eCFs obtained for two different values of n can be captured through their ratio. Figure 2 illustrates the eCF score for $n_1 = 1$ along the horizontal axis and the eCF score for $n_2 = 4$ along the vertical axis. This figure highlights high values for outliers, denoted by red crosses, while inliers, denoted by green crosses, have lower values. The growth ratio characterizes the relative change of the CF as n increases, whereas DyCG [4] characterizes the absolute change. Figure 3 displays the ratio between two eCFs for $n_1 = 1$ and $n_2 = 2$ for inliers, in green, and outliers, in red. Consistently to the dichotomy property and Figure 2, outlier samples have the highest score.

► **Corollary 5.** *Given $\mathbf{x} \in \mathbb{X}$, and $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, the CF growth ratio of $\mathbf{x}$ behaves like:*

$$\frac{\Lambda_{n_2}^\mu(\mathbf{x})^{-1}}{\Lambda_{n_1}^\mu(\mathbf{x})^{-1}} \sim \left(\frac{n_2}{n_1} \right)^p, \quad (12)$$

where $p \in \mathbb{N}^*$ is a constant.

Proof. Given $\mathbf{x} \in \mathbb{X}$ and $n \in \mathbb{N}^*$, from Theorem 2, there exist $C_d \in \mathbb{R}^{*+}$ and $p \in \mathbb{N}^*$ such that, $\Lambda_n^\mu(\mathbf{x})^{-1}$ behaves like:

$$\Lambda_n^\mu(\mathbf{x})^{-1} \sim C_d n^p \quad (13)$$

Taking $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, we have $\Lambda_{n_1}^\mu(\mathbf{x})^{-1} \sim C_d(n_1)^p$ and $\Lambda_{n_2}^\mu(\mathbf{x})^{-1} \sim C_d(n_2)^p$, which proves Corollary 5. $\blacktriangleleft$

► **Remark 6.** Given $\mathbf{x} \in \mathbb{X}$, and $n_1, n_2 \in \mathbb{N}^*$, $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$, the CF growth ratio of $\mathbf{x}$ behaves asymptotically like:

$$\frac{\Lambda_{n_2}^\mu(\mathbf{x})^{-1}}{\Lambda_{n_1}^\mu(\mathbf{x})^{-1}} \sim \left(\frac{n_2}{n_1}\right)^d, \quad (14)$$

► **Corollary 7.** Given $\mathbf{x} \notin \mathbb{X}$, and $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, the CF growth ratio of $\mathbf{x}$ behaves like:

$$\frac{\Lambda_{n_2}^\mu(\mathbf{x})^{-1}}{\Lambda_{n_1}^\mu(\mathbf{x})^{-1}} \sim e^{(n_2-n_1)b}, \quad (15)$$

where $b \in \mathbb{R}^{*+}$ is a constant.

Proof. Given $\mathbf{x} \notin \mathbb{X}$ and $n \in \mathbb{N}^*$, from Theorem 3, there exist $a, b \in \mathbb{R}^{*+}$ such that $\Lambda_n^\mu(\mathbf{x})^{-1}$ behaves like:

$$\Lambda_n^\mu(\mathbf{x})^{-1} \sim ae^{nb} \quad (16)$$

Taking $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, we have $\Lambda_{n_1}^\mu(\mathbf{x})^{-1} \sim ae^{n_1b}$ and $\Lambda_{n_2}^\mu(\mathbf{x})^{-1} \sim ae^{n_2b}$, which proves Corollary 7. $\blacktriangleleft$

3.3 Near-anomaly definition

Given that the eCF inherits the properties of the CF, this section leverages the growth properties of the CF given by Theorems 2 and 3 and the two Corollaries 5 and 7 to define near-anomalies in a practical setting.

► **Definition 8** (eCF growth ratio). Given $\mathbf{x} \in \mathbb{X}$, the eCF growth ratio $\mathcal{G}_{n_2/n_1}(\mathbf{x})$ is defined for $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$ as follows:

$$\mathcal{G}_{n_2/n_1}(\mathbf{x}) = \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}} \quad (17)$$

The following definition defines *near-anomalies* thanks to the eCF growth ratio.

► **Definition 9** (Near-anomaly definition). Given the eCF growth ratio for $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, a data point $\mathbf{x} \in \mathbb{X}$ is defined as a near-anomaly if $\mathcal{G}_{n_2/n_1}(\mathbf{x}) > \tau_{na}$, where $\tau_{na} \in \mathbb{R}$ is a given threshold.

This definition proposes to define near-anomalies as data points that are inliers but whose eCF shows a growth significantly superior to core inliers. Interestingly, in the following section, we derive a method to set automatically the near-anomaly threshold τ_{na} .

3.4 Automatic determination of the near-anomaly threshold

The results provided in this section allow to derive a method to set automatically the near-anomaly threshold.

From Corollaries 5 and 7, the eCF growth ratio $\mathcal{G}_{n_2/n_1}(\mathbf{x})$ is expected to remain relatively low for inliers ($\mathbf{x} \in \mathbb{X}$, where $\mathbb{X}$ corresponds to the support of μ_N) while being much higher for outliers/anomalies ($\mathbf{x} \notin \mathbb{X}$). Near-anomalies are inliers but they lie in a range in which the eCF growth ratio is significantly superior to that of core inliers. Let us define as $\mathbb{X}_c \subset \mathbb{X}$ the set of core inliers and by $\mathbb{X}_{na} \subset \mathbb{X}$ the set of near-anomalies that we want to identify. For this purpose, there is a theoretical result that can be useful. The following lemma indeed gives the theoretical mean of the eCF on the support of the distribution.

► **Lemma 10** (Theoretical mean value of eCF in the support). *The theoretical mean value of the eCF over the support of μ_N is $s_d(n)$.*

Proof.

$$\begin{aligned}
 \int \Lambda_n^{\mu_N}(\mathbf{x})^{-1} d\mu_N(\mathbf{x}) &= \int K_n^{\mu_N}(\mathbf{x}, \mathbf{x}) d\mu_N(\mathbf{x}) \quad (\text{from Equation (4)}) \\
 &= \sum_{i=1}^{s_d(n)} \int p_i(\mathbf{x})^2 d\mu_N(\mathbf{x}) \quad (\text{from Equation (2)}) \\
 &= \sum_{i=1}^{s_d(n)} 1 \quad (\text{as } (p_i)_{i=1}^{s_d(n)} \text{ is an orthonormal basis of } \mathbb{R}_n[\mathbf{x}]) \\
 &= s_d(n)
 \end{aligned} \tag{18}$$

◀

To leverage this result, we propose to use the Chebyshev inequality [10] to identify the inliers low and high eCF growth ratio ranges. According to this inequality, for random distributions, at least $100(\frac{k^2-1}{k^2})\%$ of the samples are in the range given by the following interval:

$$[\bar{x} - k \times \sigma; \bar{x} + k \times \sigma], \tag{19}$$

where $\bar{x}$ and σ are the mean and the standard deviation of the distribution, respectively. We further assume that the interval upper bound $(\bar{x} + k \times \sigma)$ can be used as a threshold to identify high values and hence near-anomalies, which would amount to no more than $100(\frac{1}{k^2})\%$ of the samples.

Let us apply Chebyshev inequality to our problem, which translates in the following definition.

► **Definition 11** (Core inliers and near-anomaly samples). *Let us define $\mathbb{E}[\mathcal{G}_{n_2/n_1}]$ and $\sigma(\mathcal{G}_{n_2/n_1})$ as the mean and the standard deviation of the eCF growth ratio $\mathcal{G}_{n_2/n_1}$, core inliers and near-anomaly samples are characterized as follows:*

$$\forall \mathbf{x} \in \mathbb{X}_c \Leftrightarrow \mathcal{G}_{n_2/n_1}(\mathbf{x}) \in]0; \mathbb{E}[\mathcal{G}_{n_2/n_1}] + k \times \sigma(\mathcal{G}_{n_2/n_1})], \tag{20}$$

$$\forall \mathbf{x} \in \mathbb{X}_{na} \Leftrightarrow \mathcal{G}_{n_2/n_1}(\mathbf{x}) > \mathbb{E}[\mathcal{G}_{n_2/n_1}] + k \times \sigma(\mathcal{G}_{n_2/n_1}). \tag{21}$$

From Definition 11, it comes that the threshold τ_{na} that we want to determine is:

$$\tau_{na} = \mathbb{E}[\mathcal{G}_{n_2/n_1}] + k \times \sigma(\mathcal{G}_{n_2/n_1}). \tag{22}$$

In consequence, our problem comes down to determining $\mathbb{E}[\mathcal{G}_{n_2/n_1}]$ and $\sigma(\mathcal{G}_{n_2/n_1})$. The main result below provides a way to compute τ_{na} thanks to Lemma 10, which provides the theoretical mean value of eCF in the support, and some empirical computations.

► **Theorem 12** (Practical near-anomaly threshold). *Consider $n_1, n_2 \in \mathbb{N}^*$, $n_1 < n_2$, $k \in \mathbb{N}^*$, and $m, \Gamma \in \mathbb{R}^{+*}$, then a practical near-anomaly threshold $\tilde{\tau}_{na}$ is given by:*

$$\tilde{\tau}_{na} = \frac{s_d(n_2)}{m} + k \times \Gamma \quad (23)$$

$$\text{where } s_d(n_2) = \binom{n_2 + d}{n_2}, \quad (24)$$

$$m = \min \left\{ (\Lambda_{n_1}^{\mu_N}(\mathbf{x}))^{-1}, \mathbf{x} \in \mathbb{X} \right\}, \quad (25)$$

$$\Gamma = \sqrt{\text{var} \left\{ \mathcal{G}_{n_2/n_1}(\mathbf{x}), \mathbf{x} \in \mathbb{X} \right\}}. \quad (26)$$

Proof. From (22), we have $\tau_{na} = \mathbb{E}[\mathcal{G}_{n_2/n_1}] + k \times \sigma(\mathcal{G}_{n_2/n_1})$ and we have:

$$\mathbb{E}[\mathcal{G}_{n_2/n_1}] = \int \mathcal{G}_{n_2/n_1}(\mathbf{x}) d\mu_N(\mathbf{x}) = \int \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}} d\mu_N(\mathbf{x})$$

As $\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}$ is lower bounded on μ_N , there exists a constant $m \in \mathbb{R}^{+*}$ such that $\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1} \geq m$ [7]. So,

$$\int \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}} d\mu_N(\mathbf{x}) \leq \int \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{m} d\mu_N(\mathbf{x}) = \frac{1}{m} \int \Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1} d\mu_N(\mathbf{x})$$

and from Lemma 10:

$$\int \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}} d\mu_N(\mathbf{x}) \leq \frac{s_d(n_2)}{m} \Leftrightarrow \tau_{na} \leq \frac{s_d(n_2)}{m} + k \times \sigma(\mathcal{G}_{n_2/n_1}).$$

Because there is no theoretical result providing m and $\sigma(\mathcal{G}_{n_2/n_1})$, we propose to compute them empirically from the data set $\mathbb{X}$ at hand³. Finally,

$$\tau_{na} \leq \frac{s_d(n_2)}{\min \left\{ (\Lambda_{n_1}^{\mu_N}(\mathbf{x}))^{-1}, \mathbf{x} \in \mathbb{X} \right\}} + k \times \sqrt{\text{var} \left\{ \mathcal{G}_{n_2/n_1}(\mathbf{x}), \mathbf{x} \in \mathbb{X} \right\}} = \tilde{\tau}_{na}. \quad \blacktriangleleft$$

$\tilde{\tau}_{na}$ is more restrictive than τ_{na} but, as will be seen in the experiments, it provides a good practical threshold to identify near-anomalies. Algorithm 1 proposes the pseudo-code to compute the $\tilde{\tau}_{na}$ threshold.

3.5 Scaling up the Christoffel function

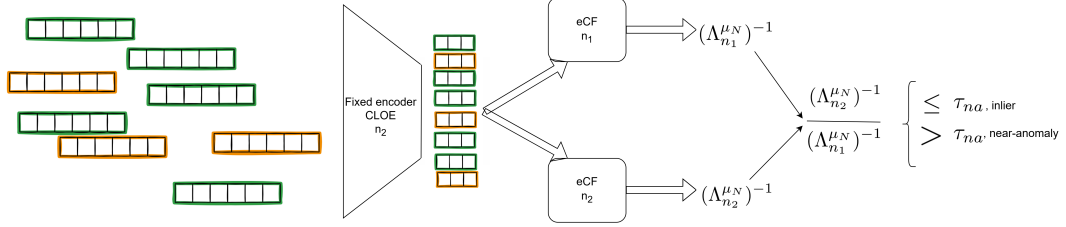
Even for moderate degrees n , the inverse of the Christoffel function is very expensive to compute if not impossible for feature dimensions $d > 8$ because of the necessity to invert the moment matrix $\mathbf{M}_n(\mu_N)$ of size $s_d(n) \times s_d(n)$ in equation (7).

To use the proposed method in high-dimensional settings, we propose to use the Christoffel-based Loss for Autoencoder (CLOE) [2]. CLOE consists of a method that applies the eCF to high-dimensional data for semi-supervised anomaly detection. It allows to reduce the dimension of high-dimensional datasets in a latent space specifically regularized with the CF score in the loss function of the autoencoder (AE).

³ To avoid division by extremely small values, a 99% percentile can be estimated from the validation set.

■ **Algorithm 1** Computation of $\tilde{\tau}_{na}$.

Input: $n_2 > n_1 > 0$,
Input: $k > 0$,
Input: Trained $(\Lambda_{n_1}^{\mu_N})^{-1}$ and $(\Lambda_{n_2}^{\mu_N})^{-1}$,
Input: Validation dataset $X_v = \{x_i\}_{i \in (1, |X_v|)}$
Output: The threshold $\tilde{\tau}_{na}$
 $d \leftarrow$ dimension of x_1
 $\mathcal{G}_{n_2/n_1} \leftarrow \left\{ \frac{\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}}{\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}}, x \in X_v \right\}$ ▷ Equation 17
 $\bar{x} \leftarrow s_d(n_2) = \binom{n_2+d}{n_2}$ ▷ Equation 24
 $\min \leftarrow \min \left\{ \Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}, x \in X_v \right\}$ ▷ Equation 25
 $\sigma \leftarrow \sqrt{\text{Var}(\mathcal{G}_{n_2/n_1})}$ ▷ Equation 26
 $\tilde{\tau}_{na} \leftarrow \frac{\bar{x}}{\min_{n_1}} + k \times \sigma$ ▷ Theorem 12
return $\tilde{\tau}_{na}$



■ **Figure 4** With normal data as input, the trained encoder from CLOE delivers low dimension encoded data in the latent space. Both near-anomaly samples, in orange, and core inlier samples, in green, are present. Then two eCF are trained, one with parameter n_1 and the other with n_2 , $n_1 < n_2$. Finally, the ratio of the two eCF scores is compared to the near-anomaly threshold, τ_{na} (in practice $\tilde{\tau}_{na}$), to decide if the data is core inlier or near-anomaly.

CLOE includes a pretraining stage and a joint training stage. The pretraining stage is performed using only the Mean Square Error (MSE) between the input and the output of the AE as the loss function. Then, the joint training stage combines the MSE with the CF-based loss to structure the latent space. For this latter step, the only hyperparameter required is the degree n for the CF computation, which is set to $n = n_{CLOE}$. The trained encoder reduces the data dimension to a lower dimension in the latent space (8 dimensions are proposed in [2]). The latent space data is then used to train the eCF. The threshold for outlier detection is chosen as the maximum value of the eCF.

In this work, CLOE is used as follows. CLOE is trained with normal data, the parameter n_{CLOE} being set to n_2 , which provides a low dimension latent space characterized by the set of moments up to the order n_2 . After this dimension reduction phase, the obtained low dimension latent space data is used to compute the eCF for the two degrees $n = n_1$ and $n = n_2$, $n_1 < n_2$, i.e., to compute $\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}$ and $\Lambda_{n_2}^{\mu_N}(\mathbf{x})^{-1}$, and the CANARI method can be applied to identify near-anomalies. This is illustrated in Figure 4.

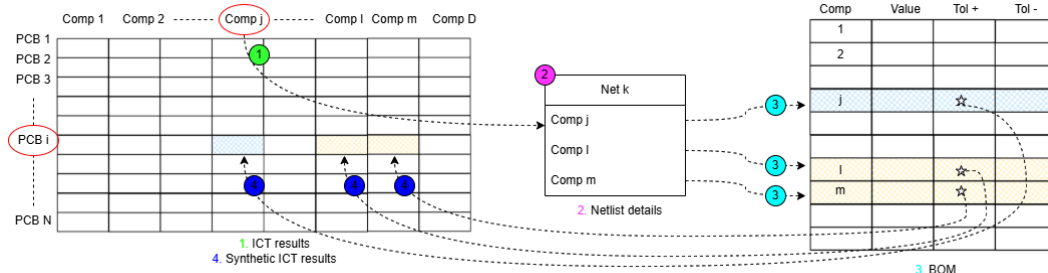
4 Experiments and results

4.1 Datasets

CANARI is applied to real production data acquired in industrial plants. In these plants, PCBs are produced in an almost fully automated way. First, the PCBs are assembled in the front-end line. Components are placed and welded during this step. At the end of this

■ **Table 1** Dimension and size of each dataset.

Dataset	D1	D2	D3	D4	D5	D6	D7	D8
Dimension	26	821	80	691	703	714	1139	18
Nb samples	57183	56334	59947	25827	10091	8421	26729	41151



■ **Figure 5** 1. A PCB product (PCB i) and one of its component (Comp j) are randomly chosen (red circle). 2. Netlist is checked to find the net (k) to which comp j belongs. All the components of the net k (j , l and m) are modified. 3. The value of these components and their upper and lower tolerances Tol+ and Tol-, respectively, are obtained from the Bill Of Materials (BOM), from which the upper and lower limits, l^+ and l^- respectively, are computed. 4. l^+ or l^- is randomly chosen. The corresponding ICT results of the concerned components are modified to appear below or above their limit.

line, ICT is performed. This test consists of injecting current into the PCB at many points on the circuit (between 10 and 2000, depending on the size of the PCB) and measuring the voltage and the current intensity to check the conformance of the PCB. Test engineers define limits for each test, and the PCB continues the manufacturing process if all its ICT results are within the limits. Otherwise, the PCB is removed from the production line. The second line is the back-end line where the PCB is encapsulated in an appropriate box. Finally, the PCB can be delivered to the customer.

The ICT results constitute tabular data with dimensions ranging from 10 to 1200, depending on the number of test points for each PCB. 8 different PCBs produced in 2023 were considered for this study, resulting in eight datasets D1 to D8. The ICT results may be PASS or FAIL. Only products that pass the ICT constitute the datasets. The dimensions of these products vary between 18 and 1139. The details of the dimension and the size for each dataset are indicated in Table 1. Each dataset is divided into 3 different sets: the training set, composed of 1150 samples, the validation set, composed of 4600 samples, and the testing set composed of the rest of the samples.

To evaluate the method, synthetic anomalies were injected into the datasets, as illustrated in Figure 5. Starting from real datasets and known ICT threshold values and upper and lower tolerances Tol+ and Tol-, respectively, from the Bill Of Materials (BOM), 10% of the samples in each test dataset were modified. First, a component is randomly chosen. Then the upper limit l^+ or lower limit l^- is randomly chosen. The ICT result for this component is modified to be close to the limit l . The order of magnitude o of the ICT result is computed with :

$$o = 10^{\lfloor \log(l^+ - l^-) \rfloor}. \quad (27)$$

The final value of the modified component is $l \pm o \times \tau_{synth}$, where τ_{synth} is the disturbed magnitude of the near anomaly samples. Finally, knowing the design of the PCB, all the components connected by the first net in the netlist, where the component belongs, are modified according to the same upper or lower decision. With this process, synthetic datasets are created taking into account the connection of the different components on the PCB and the ICT.

4.2 Baselines and Metrics

CLOE with double threshold is considered as the first baseline. The original threshold is considered to compute the new threshold to detect near-anomalies. 5%, 20% and 50% of the original threshold are used as the double threshold to detect near-anomaly samples.

Then, CANARI with different thresholds is used as the second baseline. The new thresholds are the ratio of the theoretical means of the two eCFs, τ_{mean} , and the empirical percentile 99th of the ratio, τ_{upper} .

F1-score and Matthews Correlation Coefficient (MCC) [8] are the two metrics used to evaluate CANARI. F1-score, given in Definition 13, is computed with the recall and precision metrics. Its lower value is 0 and its best value is 1.

► **Definition 13 (F1-score).** *Let TP be the true positives, TN the true negatives, FP the false positives, and FN the false negatives. The metric F1-score is defined as:*

$$F1_{-score} = \frac{2TP}{2TP + FP + FN} = 2 \frac{precision \cdot recall}{precision + recall}. \quad (28)$$

MCC, given in Definition 14, computes the Pearson product moment correlation coefficient in a contingency matrix method. Its worst value is -1, and its best value is 1, 0 corresponds to a random classification. This metric is stronger against unbalanced datasets than F1-score and accuracy [3].

► **Definition 14 (MCC).** *Let TP be the true positives, TN the true negatives, FP the false positives, and FN the false negatives. The metric MCC is defined as:*

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (29)$$

The testing set is composed of two classes: core inliers and synthetic near-anomalies. Note that the core inliers testing dataset may also be corrupted by real unknown near-anomalies because it is based on real world data, so comparing the metrics for baselines and CANARI only makes sense on our datasets.

4.3 Results

4.3.1 Observation of the growth of the eCF with n

The goal of this subsection is to visually assess the eCF growth patterns for fail, pass, and near-anomaly samples. Pretraining and joint-training steps of CLOE are performed for $n_{CLOE} = 4$. Then the eCF is computed for n varying from 1 to 5. Figure 6 (a) illustrates this experiment with PCBs that fail the ICT represented by the red graphs (only 2 red graphs because only 2 such PCBs) and PCBs that pass the ICT represented by green graphs. The testing set includes fail products in this experiment. The fail samples (in red in the

■ **Table 2** Correspondence between the parameter k and the proportion of samples outside the upper bound according to the Chebyshev inequality [10].

k	1	$\sqrt{2}$	1.5	2	$2\sqrt{2}$	3	4	5	10
Maximum percentage	100%	50%	44.5%	25%	12.5%	11.11%	6.25%	4%	1%

■ **Table 3** F1-score for the 8 datasets D1 to D8 for our method and the baselines method.

Method	D1	D2	D3	D4	D5	D6	D7	D8	Mean
CANARI	0.76$\pm$0.1	0.49$\pm$0.06	0.64$\pm$0.15	0.45$\pm$0.04	0.41$\pm$0.06	0.50$\pm$0.06	0.50$\pm$0.25	0.72$\pm$0.08	0.56
CANARI- τ_{mean}	0.16 $\pm$ 2.10 $\cdot$ 10 $^{-6}$	0.21 $\pm$ 0.002	0.16 $\pm$ 10 $^{-6}$	0.31 $\pm$ 0.01	0.20 $\pm$ 0.004	0.22 $\pm$ 0.003	0.03 $\pm$ 0.003	0.07 $\pm$ 4.10 $^{-6}$	0.17
CANARI- τ_{upper}	0.25 $\pm$ 0.0004	0.21 $\pm$ 0.0009	0.22 $\pm$ 0.001	0.14 $\pm$ 0.01	0.07 $\pm$ 0.006	0.17 $\pm$ 0.02	0.49 $\pm$ 0.06	0.45 $\pm$ 0.001	0.25
CLOE $_{5\%}$	0.16 $\pm$ 0.02	0.06 $\pm$ 0.006	0.04 $\pm$ 0.0004	0.17 $\pm$ 0.01	0.15 $\pm$ 0.01	0.17 $\pm$ 0.01	0.34 $\pm$ 0.03	0.34 $\pm$ 0.04	0.18
CLOE $_{20\%}$	0.26 $\pm$ 0.02	0.09 $\pm$ 0.002	0.08 $\pm$ 0.002	0.28 $\pm$ 0.02	0.16 $\pm$ 0.01	0.24 $\pm$ 0.01	0.37 $\pm$ 0.04	0.39 $\pm$ 0.03	0.23
CLOE $_{50\%}$	0.45 $\pm$ 0.01	0.19 $\pm$ 0.005	0.27 $\pm$ 0.02	0.39 $\pm$ 0.004	0.25 $\pm$ 0.003	0.31 $\pm$ 0.02	0.41 $\pm$ 0.05	0.45 $\pm$ 0.03	0.34

image) exhibit the greatest increase, even for low values of n . Then, between $n = 3$ and $n = 4$, the graphs of some pass samples show a change in growth pattern, these products are the near-anomaly products we are looking for. The other pass samples have a polynomial growth from $n = 1$ to $n = 5$.

With this experiment, the change in the growth pattern with the parameter n is visible. Part of the distribution is clearly different from the rest of the distribution.

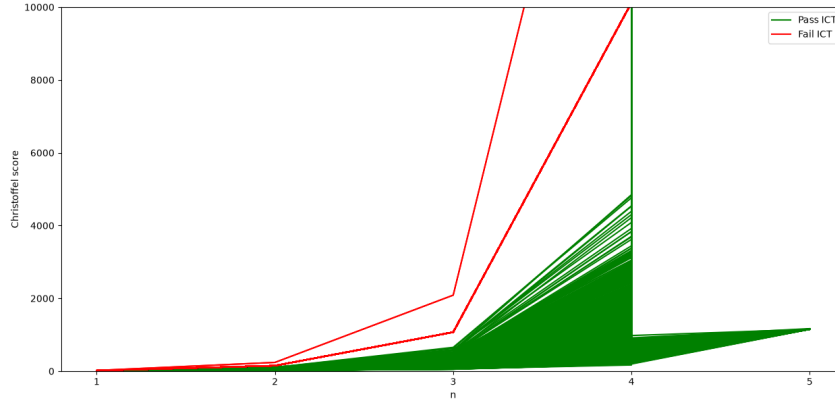
4.3.2 Evaluation of the proposed method

In this section, and for computational reasons, n_1 is fixed to 1 and n_2 is fixed to 4. Different values for the parameter k are tested, namely $k \in \{1, \sqrt{2}, 1.5, 2, 2\sqrt{2}, 3, 4, 5, 10\}$. According to the Chebyshev inequality [10], this values correspond to an upper bound on the proportion of samples falling below the defined threshold τ_{na} , as summarized in Table 2. A slightly smaller proportion falls below $\tilde{\tau}_{na}$. The value of τ_{synth} is fixed to 10^{-2} to create the disturbed magnitude of the near anomaly samples.

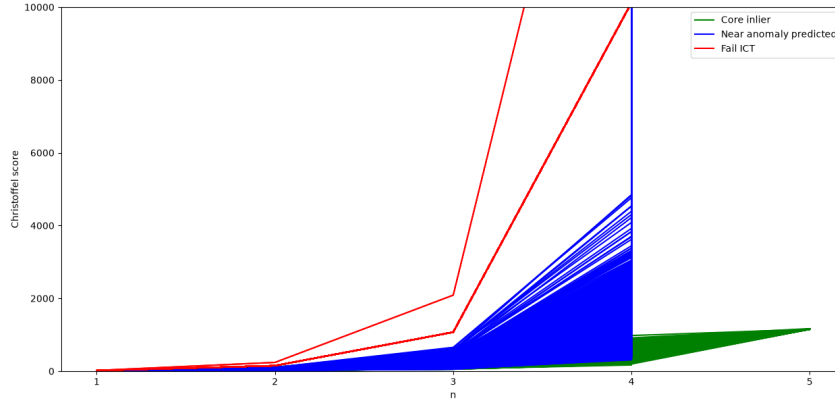
Figure 6 (b) shows the growth of the eCF with n varying from 1 to 5. The blue graphs correspond to the near-anomaly samples detected by CANARI while the green ones correspond to the samples detected as core inliers. We can observe that the detected near-anomaly samples correspond to samples whose growth lies between that of fail PCB samples and core inlier PCB samples. This observation confirms that CANARI detects samples with a growth between fail samples and core inlier samples.

Then synthetic datasets were created for PCBs as described in Section 4.1. Five runs are performed with different random seeds. The random seed is the same for each dataset and each method to ensure reproducibility across the same run. The mean and the standard deviation across these five runs are displayed in Table 3 for the F1-score metric and in Table 4 for the MCC metric.

The performance obtained by CANARI on the datasets D1 to D8 is better than double thresholded CLOE with different percentage thresholds for both metrics. For datasets D1 to D8, CANARI improves the F1-score by 65.8% and the MCC by 34.3% comparing to the CLOE methods with percentage. Classic CANARI obtains better performances than CANARI- τ_{mean} and CANARI- τ_{upper} . This shows that the proposed threshold $\tilde{\tau}_{na}$ can actually detect near anomaly samples.



(a) Pass ICT in green, and Fail ICT in red.



(b) Predicted core inliers (green) and near anomalies (blue) and real fail samples (red).

Figure 6 The figure highlights the evolution of the eCF as n varies from 1 to 5. (a) displays the true label, PASS or FAIL, with the samples that succeeded ICT in green and the samples that failed ICT in red. In this dataset, only two products have failed the ICT, hence only two red graphs. (b) displays the predicted label, with core inliers in green and near anomalies in blue. In red, we find the two Fail ICT samples kept to compare their graph growth. For $n = 5$, core inliers have slightly different scores around 1500, hence the “strange” value concentration due to the scale. Between $n = 4$ and $n = 5$, the near anomaly sample scores increase so sharply that their values are out of the figure for $n = 5$, producing a “strange” blue line that appears to be vertical at $n = 4$ in the figure. Red graphs, i.e., those of Fail ICT samples, exhibit a faster growth than other graphs. The difference between the growth of the green and blue graphs, i.e., core inliers and near anomalies respectively, really occurs in the interval between $n = 3$ to $n = 4$. Over the interval from $n = 1$ to $n = 5$, we can hence distinguish a polynomial growth for the green graphs, i.e., those of core inliers, compared to an exponential growth for the red graphs, i.e., those of Fail ICT samples, the near anomaly graphs growth standing in between.

■ **Table 4** MCC for the 8 datasets D1 to D8 for our method and the baselines method.

Method	D1	D2	D3	D4	D5	D6	D7	D8	Mean
CANARI	0.76$\pm$0.09	0.48$\pm$0.07	0.61$\pm$0.17	0.47$\pm$0.04	0.39$\pm$0.08	0.49$\pm$0.04	0.50$\pm$0.25	0.73$\pm$0.08	0.55
CANARI- τ_{mean}	0.09 $\pm$ 10 ⁻⁵	0.15 $\pm$ 0.003	0.09 $\pm$ 9.10 ⁻⁵	0.28 $\pm$ 0.01	0.17 $\pm$ 0.006	0.19 $\pm$ 0.005	0.06 $\pm$ 0.001	0.07 $\pm$ 9.10 ⁻⁶	0.14
CANARI- τ_{upper}	0.37 $\pm$ 0.003	0.32 $\pm$ 0.001	0.32 $\pm$ 0.002	0.21 $\pm$ 0.02	0.11 $\pm$ 0.01	0.23 $\pm$ 0.03	0.48 $\pm$ 0.05	0.53 $\pm$ 0.0006	0.32
CLOE _{5%}	0.27 $\pm$ 0.01	0.16 $\pm$ 0.001	0.14 $\pm$ 0.001	0.27 $\pm$ 0.01	0.22 $\pm$ 0.02	0.28 $\pm$ 0.01	0.40 $\pm$ 0.04	0.43 $\pm$ 0.03	0.27
CLOE _{20%}	0.36 $\pm$ 0.02	0.20 $\pm$ 0.003	0.19 $\pm$ 0.004	0.37 $\pm$ 0.01	0.24 $\pm$ 0.02	0.34 $\pm$ 0.01	0.10 $\pm$ 0.04	0.47 $\pm$ 0.02	0.28
CLOE _{50%}	0.52 $\pm$ 0.008	0.30 $\pm$ 0.003	0.37 $\pm$ 0.02	0.45 $\pm$ 0.004	0.32 $\pm$ 0.005	0.39 $\pm$ 0.01	0.43 $\pm$ 0.05	0.52 $\pm$ 0.02	0.41

4.4 Ablation and sensitivity studies

An ablation study is performed to evaluate the robustness of CANARI. This section relies exclusively on MCC as the evaluation metric. First, Table 5 reports a sensitivity analysis of parameter k across datasets. The optimal values for k across the different datasets are $\sqrt{2}$, 1.5 and 2. Accordingly, the parameter k should be selected among these three values.

Second, Figure 7 compares different pairs of parameters n_1 and n_2 , under the constraint $n_1 < n_2$. This heatmap shows the mean MCC associated with each possible combination of n_1 and n_2 across all datasets. The best combination is the one used in CANARI, namely $n_1 = 1$ and $n_2 = 4$.

Then, Table 6 compares CLOE with the kernelized inverse Christoffel function (KIC) [1] and the univariate Christoffel function [5] as approaches for eCF computation in high dimensional datasets before applying CANARI⁴. To compute the threshold $\tilde{\tau}_{na}$, the best value of k is selected separately for each method and dataset. For KIC [1] and UCF [5] and for some datasets, the minimum value or the 99% percentile of the eCF $\Lambda_{n_1}^{\mu_N}(\mathbf{x})^{-1}$ can be very low, in the order of 10⁻⁶⁸, and the ratio between the two Christoffel functions cannot be computed. These datasets are therefore not included in the present analysis. CLOE consistently achieves the best results across all datasets. Only the KIC method yields similar performance on the last dataset.

Finally, Table 7 presents a comparison of several values of τ_{synth} , corresponding to the disturbance magnitude used for generating the synthetic near-anomaly dataset. Except for one dataset with $\tau_{synth} = 0.1$, CANARI is able to detect synthetic near-anomaly samples generated with amplitudes between 0.1 and 10⁻⁴, which illustrates the robustness of the method that does not need to define a strict near-anomaly threshold for the synthetic near anomalies.

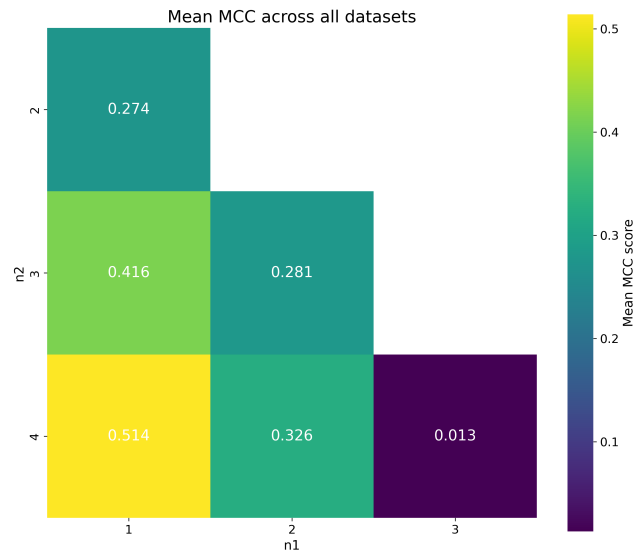
5 Conclusion and Future Work

This paper presents CANARI and introduces the near-anomaly concept. Near-anomalies are samples that, while not yet anomalous, lie close to the boundary of the distribution and are likely to transition into anomalies in the near future. CANARI is an unsupervised near-anomaly detection method based on the eCF that leverages theoretical properties of the CF distinguishing its growth with the parameter n for data in the support and data outside the support. CANARI is constructed with two eCFs with two different values for the parameter n from which the growth can be assessed. A threshold based on the Chebyshev

⁴ The implementation used for these two methods is proposed in the framework ChRISToffel Anomaly Locator available at <https://github.com/fgrivet/CRISTAL>

■ **Table 5** MCC for the 8 datasets D1 to D8 for different values of the parameter k .

k	D1	D2	D3	D4	D5	D6	D7	D8
1	0.743	0.342	0.539	0.418	0.293	0.442	0.318	0.677
$\sqrt{2}$	0.750	0.446	0.610	0.445	0.351	0.494	0.415	0.733
1.5	0.759	0.435	0.609	0.441	0.344	0.487	0.402	0.725
2	0.712	0.478	0.606	0.469	0.387	0.493	0.472	0.687
$2\sqrt{2}$	0.640	0.461	0.565	0.462	0.326	0.450	0.496	0.602
3	0.632	0.451	0.551	0.443	0.325	0.448	0.501	0.601
4	0.536	0.398	0.458	0.428	0.257	0.437	0.496	0.575
5	0.479	0.343	0.386	0.386	0.257	0.425	0.476	0.552
10	0.053	0.138	0.088	0.051	0.022	0.108	0.434	0.151



■ **Figure 7** Heatmap of the mean of the MCC on all the datasets for different values of n_1 and n_2 . The best combination is obtained for $n_1=1$ and $n_2=4$, which are the values that are used in CANARI.

inequality and on theoretical properties of the eCF is proposed to detect near-anomaly samples. In high-dimensional settings, CANARI can be coupled with CLOE [2] to reduce the data dimension.

To evaluate this method, industrial datasets composed of ICT from a PCB manufacturing process are considered. Synthetic datasets are created to simulate near-anomaly samples based on real industrial scenarios. Experiments confirm that CANARI detects near-anomaly samples more efficiently than a classical anomaly detector with double thresholds. Moreover, CANARI with the presented threshold achieves better performance than CANARI with other thresholds.

A sensitive analysis and an ablation study were conducted to assess the robustness of CANARI. The optimal values of parameter k are identified through this study. The selected combination of n_1 and n_2 corresponds to the one that yields the best average performance across all datasets. The variation in the amplitude used to generate the synthetic near-anomalies dataset confirms the stability of CANARI. Finally, CLOE is compared with two

■ **Table 6** MCC for the 3 datasets D1, D3, and D8 for different methods to reduce dimension before applying CANARI.

Method to reduce dimension	D1	D3	D8
$CANARI_{CLOE}$	0.76 ± 0.09	0.61 ± 0.17	0.73 ± 0.08
$CANARI_{KIC}$	0.56 ± 0.01	0.31 ± 0.09	0.69 ± 0.007
$CANARI_{UCF}$	0.31 ± 0.08	0.41 ± 0.11	0.29 ± 0.26

■ **Table 7** MCC for the 8 datasets D1 to D8 for different values of τ_{synth} .

τ_{synth}	D1	D2	D3	D4	D5	D6	D7	D8
0.1	0.61	0.33	0.69	0.34	0.40	0.45	0	0.57
10^{-2}	0.76	0.48	0.61	0.47	0.39	0.49	0.50	0.73
10^{-3}	0.62	0.34	0.73	0.34	0.48	0.47	0.59	0.58
10^{-4}	0.62	0.33	0.73	0.36	0.49	0.47	0.61	0.55

alternative methods for computing the eCF in high-dimensional dataset. The performance obtained by the two other approaches support the choice of CLOE made in this paper. Moreover, the two alternative methods fail to compute the ratio on five datasets because of very small eCF values.

The CANARI method can be adapted to any quality process with many tests with defined limits. These quality processes are commonly used in industry, where thresholds are defined on the basis of physical models. Such physical models fail to consider the production distribution and consequently reduce the ability to detect other problems. It can be used to anticipate future failures in products that initially look normal but can have some invisible defects and cause issues at a later step.

References

- 1 Armin Askari, Forest Yang, and Laurent El Ghaoui. Kernel-based outlier detection using the inverse christoffel function. *arXiv preprint arXiv:1806.06775*, 2018. [arXiv:1806.06775](#).
- 2 Léa Billet, Louise Travé-Massuyès, Elodie Chanthery, and Alexandre Gaffet. Cloe: Christoffel loss autoencoder for anomaly detection. *The 16th International Conference on Information Science and Technology*, 2026.
- 3 Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.
- 4 Kévin Ducharlet, Louise Travé-Massuyès, Jean-Bernard Lasserre, Marie-Véronique Le Lann, and Youssef Miloudi. Leveraging the christoffel function for outlier detection in data streams. *International Journal of Data Science and Analytics*, 20(3):1–17, 2024.
- 5 Florian Grivet, Didier Henrion, Jean-Bernard Lasserre, and Louise Travé-Massuyès. Scalable anomaly detection via a univariate christoffel function. *arXiv preprint arXiv:2606.12483*, 2026. [doi:10.48550/arXiv.2606.12483](#).
- 6 Jean B Lasserre and Edouard Pauwels. The empirical christoffel function with applications in data analysis. *Advances in Computational Mathematics*, 45(3):1439–1468, 2019. [doi:10.1007/S10444-019-09673-1](#).
- 7 Jean Bernard Lasserre, Edouard Pauwels, and Mihai Putinar. *The Christoffel–Darboux kernel for data analysis*, volume 38. Cambridge University Press, 2022.

- 8 Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
- 9 Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it until you make it: Towards accurate near-distribution novelty detection. In *The eleventh international conference on learning representations*, 2022.
- 10 Friedrich Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88–91, 1994.
- 11 Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi:10.1109/JPROC.2021.3052449.
- 12 Durgesh Samariya and Amit Thakkar. A comprehensive survey of anomaly detection algorithms. *Annals of Data Science*, 10(3):829–850, 2023.
- 13 Luc PJ Sträter, Mohammadreza Salehi, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Generalad: Anomaly detection across domains by attending to distorted features. In *European conference on computer vision*, pages 448–465. Springer, 2024.