Risk-Based Thresholding for Reliable Anomaly Detection in Concentrated Solar Power Plants

Yorick Estievenart[1], Sukanya Patra[1], and Souhaib Ben Taieb[2] (✉)

[1] University of Mons, Mons, Belgium
yorick.estievenart@gmail.com,sukanya.patra@umons.ac.be
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates souhaib.bentaieb@mbzuai.ac.ae

Abstract. Efficient and reliable operation of Concentrated Solar Power (CSP) plants is essential for meeting the growing demand for sustainable energy. However, high-temperature solar receivers face severe operational risks, such as freezing, deformation, and corrosion, resulting in costly downtime and maintenance. To monitor CSP plants, cameras mounted on solar receivers record infrared images at irregular intervals ranging from one to five minutes throughout the day. Anomalous images can be detected by thresholding an anomaly score, where the threshold is chosen to optimize metrics such as the F1-score on a validation set. This work proposes a framework, using risk control, for generating more reliable decision thresholds with finite-sample coverage guarantees on any chosen risk function. Our framework also incorporates an abstention mechanism, allowing high-risk predictions to be deferred to domain experts. Second, we propose a density forecasting method to estimate the likelihood of an observed image given a sequence of previously observed images, using this likelihood as its anomaly score. Third, we analyze the deployment results of our framework across multiple training scenarios over several months for two CSP plants. This analysis provides valuable insights to our industry partner for optimizing maintenance operations. Finally, given the confidential nature of our dataset, we provide an extended simulated dataset[3], leveraging recent advancements in generative modeling to create diverse thermal images that simulate multiple CSP plants. Our code is publicly available[4].

Keywords: Deep Image Anomaly Detection, Risk Control, Irregular Time-series, Non-stationarity, Concentrated Solar Power Plants, Density Estimation, Reliable Decision Thresholds

1 Introduction

The global transition toward greener and more sustainable renewable energy sources is hindered by two critical challenges: (i) on-demand generation and (ii)

[3] https://tinyurl.com/macmnjyt
[4] https://github.com/yoest/reliable-ad-csp

dispatchability. Concentrated Solar Power (CSP) plants offer a promising solution, leveraging thermal energy storage to provide electricity even when sunlight is unavailable. Among the various CSP configurations, central tower-based plants are the most prevalent, using an array of mirrors to concentrate sunlight onto a receiver, where a heat transfer medium absorbs and stores the energy. However, the extreme operating temperatures make these systems highly susceptible to failures such as metal fatigue and tube blockages, directly impacting their efficiency, reliability, and operational lifespan. To mitigate these risks, thermal imaging from infrared cameras is used to monitor CSP plants. Nonetheless, the sheer volume and complexity of thermal image data render manual monitoring impractical, necessitating the development of an automated, data-driven Predictive Maintenance (PdM) pipeline. This problem naturally aligns with anomaly detection (AD), where the goal is to identify abnormal behaviours.

Despite significant progress in both deep and shallow AD research [20,32], existing image- and video-based approaches fall short in addressing the problem of detecting anomalous behaviours of operational CSP plants due to three key challenges. First, the lack of interpretability of the anomaly scores hinders decision-making in high-stakes applications without an appropriate thresholding strategy [28]. Traditional approaches rely on performance metrics such as F1-score or GMean to determine thresholds depending on the available labelled samples. These methods do not guarantee that the results will remain consistent in a deployment setting. Moreover, they assume that all CSP plants define risk similarly and follow the same operational strategies. In reality, this often differs (e.g., deploying a maintenance team may be preferable to replacing a tower component). Second, deep learning-based AD models are often perceived as unreliable [28] due to the uncertainty in predictions stemming from their inability to properly estimate the decision boundary, particularly when training data is limited. Thus, practitioners are hesitant to use the predictions even when the associated uncertainty is minimal, severely limiting their adoption in real-world applications. Also, unlike classical image- and video-based AD data, CSP plant monitoring involves thermal images without semantic content, lacks a fixed frame rate, and exhibits significant non-stationarity and temporal dependencies due to pronounced daily seasonal patterns. As a result, conventional image- and video-based anomaly detection methods are inappropriate. A recent forecasting-based AD method, `ForecastAD` [24], attempts to address these challenges by measuring per-pixel errors between predicted and observed thermal images. However, reconstruction-based AD methods suffer a critical flaw: models trained on normal data can inadvertently reconstruct and misclassify anomalous images as normal [10,22], leading to unreliable detection.

To overcome these limitations, we propose a principled, robust AD framework tailored for CSP plant monitoring. First, we introduce a risk-controlling thresholding strategy for anomaly scores that satisfies finite-sample performance guarantees on any chosen risk function (e.g., false positive rate or F1-score)—a critical requirement for reliable predictive maintenance (PdM) in industrial settings. To enhance trust and adoption, we integrate a machine-learning-with-

abstention framework [27] with adaptive thresholds that account for the overlap between normal and anomalous score distributions. This approach defers high-risk predictions to domain experts, ensuring human intervention when uncertainty is high. Furthermore, we propose an AD method based on density forecasting, `DensityAD`, which leverages conditional normalizing flows to model the likelihood of an observed sample being normal, given past thermal images and timestamps. This approach mitigates the limitations of reconstruction-based methods and enables likelihood-based thresholding for more effective anomaly detection. Our key contributions are:

- We propose a framework for computing reliable anomaly detection thresholds with finite-sample performance guarantees for any chosen risk function. The framework includes an abstention mechanism that defers decisions to domain experts under high uncertainty.
- We develop an unsupervised AD method that computes anomaly scores using density forecasting by estimating the conditional likelihood of an observed infrared image given a sequence of previously observed images.
- We conduct an extensive deployment analysis of our framework across multiple real-world scenarios over several months, using data from two CSP plants. This analysis provides valuable insights to our industry partner for maintenance operations.
- We release a simulated dataset by leveraging recent advancements in generative modelling to create diverse infrared images that emulate real-world data from CSP plants.

Our work not only advances the state of anomaly detection in renewable energy systems but also serves as an important milestone for future research in robust, data-driven PdM strategies for critical infrastructure monitoring.

2 Anomaly detection in thermal images from CSP plants

In the following, we describe our AD use case and the associated dataset.

Use-case. Concentrated Solar Power (CSP) plants are designed to harness solar energy for large-scale electricity generation while addressing two major challenges commonly associated with renewable energy sources – on-demand generation and dispatchability. Among the four primary CSP technologies currently in use, namely, Solar Tower, Parabolic Trough, Linear Fresnel, and Dish-Stirling systems, this study specifically focuses on the operational aspects of Solar Tower-based CSP plants. These plants comprise two critical components: the Thermal Solar Receiver and the Steam Generator. Positioned atop a central tower, the Solar Receiver functions as a solar furnace, absorbing concentrated sunlight reflected by an array of heliostats—movable mirrors strategically arranged on the ground around the tower. A high-capacity heat transfer medium, such as molten salts, circulates through vertical heat exchanger tubes configured as panels within the receiver, absorbing the thermal energy from the concentrated

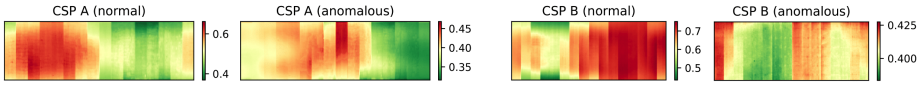


Fig. 1: Example of thermal images from CSP A and B.

sunlight. This heated transfer medium is subsequently stored in a Thermal Energy Storage (TES) system. It is later used to generate superheated steam, which drives the Steam Generator to produce electricity. Thus, the incorporation of TES enables the on-demand power generation capability of CSP plants and positions them as viable alternatives to conventional fossil fuel-based power plants. Despite their advantages, CSP plants encounter significant operational challenges operating in extremely high temperatures. These challenges include blockage or deformation of heat exchanger tubes, metal fatigue, and corrosion, all of which can impact plant efficiency and reliability. Therefore, continuous monitoring and real-time failure detection are crucial to ensuring uninterrupted power generation and preventing costly system failures. In this study, we focus on detecting failures and anomalous behaviours in the Thermal Solar Receiver.

Dataset. As previously discussed, the receiver consists of vertical heat exchanger tubes arranged in panels through which the heat transfer medium flows. During normal operation, the temperature of this medium increases as it moves through the tubes, absorbing heat from concentrated sunlight. It results in a surface temperature gradient along the flow direction, which is captured by infrared (IR) cameras. In this study, our goal is to identify anomalous behaviours of the solar receiver by monitoring these temperature gradients. The *solar receiver dataset* used in this study consists of sequences of IR images taken at irregular intervals ranging from one to five minutes throughout the day, with each sequence corresponding to an operational day of the CSP plant. Notably, the dataset lacks ground truth labels, as domain experts do not have prior knowledge of all possible failure types, and anomalies are inherently unknown apriori. Each operational day at the CSP plant comprises three distinct phases: (i) *preheating*, to prevent molten salt from freezing, (ii) *filling/draining*, during which salt circulates at the start and is drained at the end of the operation, and (iii) the *power* phase, where the salt absorbs thermal energy for power generation. Each phase exhibits a distinct surface temperature profile, which must be accounted for in modelling to ensure reliable AD. For example, low surface temperatures are expected during *preheating*, but the same behaviour during the *power* phase may signal a failure.

Building on prior work [24], we expanded the *solar receiver dataset* to include data from two distinct CSP plants, referred to as A and B, for anonymity. Specifically, we have access to 16343 samples from CSP A and 15181 from CSP B. Although the dataset exhibits similar key characteristics—such as non-stationarity, irregular sampling, and temporal dependence—certain differences exist across the plants. Notably, the thermal image resolutions differ, with CSP A capturing images of size 184×608 pixels, while CSP B captures images of size 196×528

pixels. Furthermore, CSP B exhibits an inversion in the thermal flow direction (left to right), whereas CSP A follow a right-to-left flow pattern. Examples of thermal images for both CSP plants can be seen in Figure 1. To maintain confidentiality, all thermal images have been normalized before analysis.

3 Background

Notations. We consider an unsupervised AD setting, where the training dataset, denoted as $\mathcal{D}_N = \{x_i\}_{i=1}^n$, consists of n *unlabeled* samples. Each sample $x_i = (y_i, t_i) \in \mathcal{X}$ is a tuple, where $\mathcal{X} = \mathbb{R}_+^d \times \mathbb{R}_+$. The first component, $y_i \in \mathcal{Y}$, represents a thermal image of dimension $d = H \times W$, where H and W denote the height and width, respectively, with $\mathcal{Y} = \mathbb{R}_+^d$. The second component, $t_i \in \mathbb{R}_+$, corresponds to the timestamp at which the thermal image y_i was captured. Following prior works [31], we assume that the training dataset $\mathcal{D}_N$ predominantly contains normal samples. Additionally, we introduce another *labelled* dataset, $\mathcal{D}_R = \big\{(x_i, z_i)\big\}_{i=1}^{n_R}$, consisting of n_R labeled pairs, where $n_R \ll n$. Each label $z_i \in \{0, 1\}$ indicates whether the corresponding sample is normal ($z_i = 0$) or anomalous ($z_i = 1$). Furthermore, the dataset $\mathcal{D}_R$ is partitioned into three disjoint subsets: validation ($\mathcal{D}_V$), calibration ($\mathcal{D}_C$), and test ($\mathcal{D}_T$).

Unsupervised AD. The goal of unsupervised AD is to estimate an anomaly score function $s(\cdot) : \mathcal{X} \to \mathbb{R}$ using $\mathcal{D}_N$, such that normal samples receive lower scores. A label (0 for normal or 1 for anomalous) is then assigned to a new test sample $x \in \mathcal{X}$ by thresholding its anomaly score:

$$\hat{z} = h(x) = \begin{cases} 0, & \text{if } s(x) \leq \lambda, \\ 1, & \text{if } s(x) > \lambda, \end{cases} \tag{1}$$

where $h : \mathcal{X} \to \{0, 1\}$ is the labelling function and $\lambda \in \mathbb{R}$ is a threshold to be determined, whose optimal value depends on the proportion of anomalies in the test set [26,29]. However, since the true proportion is unknown in practice, existing methods rely on test performance metrics to select a threshold $\lambda \in \Lambda$ from a set of feasible thresholds $\Lambda \subset \mathbb{R}$. Commonly adopted approaches include:

- **F1-score [2]**. The threshold λ_{F} yields the highest F1-score:

$$\lambda_{\mathrm{F}} = \arg\max_{\lambda \in \Lambda} \ \text{F1-Score}\big(\mathcal{H}_V\big), \tag{2}$$

  where $\mathcal{H}_V = \{(h(x), z) \mid (x, z) \in \mathcal{D}_V\}$ and F1-Score computes the harmonic mean of precision and recall.
- **G-Mean [17].** The threshold λ_{G} maximizes the G-Mean:

$$\lambda_{\mathrm{G}} = \arg\max_{\lambda \in \Lambda} \ \text{G-Mean}\big(\mathcal{H}_V\big), \tag{3}$$

  where G-Mean computes the geometric mean of precision and recall.

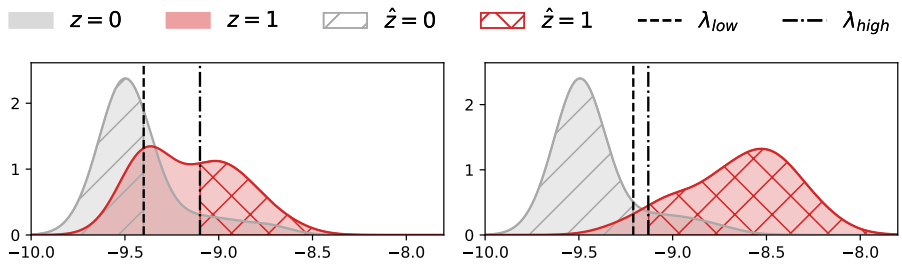


Fig. 2: Illustration of thresholds for AD with abstention under high (left) and low (right) overlap in anomaly score distributions of normal and anomalous samples.

– **Z-score**. Let $\mathcal{S}_V$ be the set of anomaly scores for normal samples in $\mathcal{D}_V$, defined as $\mathcal{S}_V = \{s(x) \mid (x, 0) \in \mathcal{D}_V\}$ with the corresponding mean and standard deviation denoted by μ_{S_V} and σ_{S_V}, respectively. The threshold λ_z is set k standard deviations above μ_{S_V}. Unlike Eq. 1, where the threshold is applied directly to $s(x)$, here it is applied to the z-scores, defined as $s_z(x) = \left|\frac{s(x)-\mu_{S_V}}{\sigma_{S_V}}\right|$.

For a comprehensive discussion on existing methods for selecting λ, we refer to [26]. A key limitation of these approaches is that they do not account for uncertainty when the anomaly score distributions of normal and anomalous samples overlap, as illustrated in Figure 2. However, given the high-risk nature of AD applications, it is essential to abstain from assigning labels under high uncertainty. This allows domain experts to intervene, reducing the risk of incorrect classifications and ensuring more reliable decision-making.

Unsupervised AD with abstention. To enable abstention from labeling under high uncertainty, we augment the labeling function with an abstention label (Ⓡ) and introduce two thresholds (λ^l and λ^h), reformulating $h(x)$ as follows:

$$\hat{z} = h(x) = \begin{cases} 0, & \text{if } x \in \hat{C}_{\text{nor}}, \quad \hat{C}_{\text{nor}} = \{x' \in \mathcal{X} \mid s(x') \leq \lambda^l\}, \\ 1, & \text{if } x \in \hat{C}_{\text{ano}}, \quad \hat{C}_{\text{ano}} = \{x' \in \mathcal{X} \mid s(x') \geq \lambda^h\}, \\ \text{Ⓡ}, & \text{if } x \in \hat{C}_{\text{abs}}, \quad \hat{C}_{\text{abs}} = \{x' \in \mathcal{X} \mid \lambda^l < s(x') < \lambda^h\}, \end{cases} \tag{4}$$

where $\hat{C}_{\text{nor}}$ is the *normal prediction region*, $\hat{C}_{\text{ano}}$ is the *anomalous prediction region*, and $\hat{C}_{\text{abs}}$ is the *abstention region*, where the model refrains from making a decision.

Figure 2 illustrates two examples of this decision-making process. The parameters λ^l and λ^h define the normal and anomalous prediction regions while also regulating the abstention region, thereby controlling the abstention rate.

Ideally, the pair of thresholds (λ^l, λ^h) should adapt to the anomaly score distribution, effectively capturing the overlap between normal and anomalous scores.

A trivial yet uninformative approach is to set $\lambda^l = -\infty$ and $\lambda^h = +\infty$, which results in abstaining from prediction for all samples. We aim to propose a principled method for selecting a reliable pair of thresholds.

4 Reliable Decision Thresholds for AD

Let us define a Risk-Controlling Prediction Set (RCPS) $\widehat{C}_\lambda$ for a given threshold $\lambda \in \Lambda \subset \mathbb{R}$ as follows:

Definition 1 (RCPS [6]). *Let $\lambda \in \Lambda$ be a random variable and $R(\cdot) : 2^{\mathcal{X}} \to \mathbb{R}$ a risk function. The set $\hat{C}_\lambda$ is defined as an (α, δ)-risk-controlling prediction set if it satisfies the condition $\mathbb{P}(R(\hat{C}_\lambda) \leq \alpha) \geq 1 - \delta$, where $\alpha \in [0, 1]$ is the risk tolerance and $\delta \in [0, 1]$ is the error level.*

One method for constructing an RCPS is *conformal risk control*, an extension of conformal prediction (CP) [3] designed to control the expected value of a risk function, assuming it is monotonically non-increasing with respect to a single threshold λ. However, this approach is limited to a single-parameter setting, as in (1), and relies on a restrictive assumption about the risk function.

To overcome these limitations, we propose leveraging the *Learn then Test* (LTT) procedure [4]. We consider the unsupervised AD problem with abstention, as defined in (4). Our objective is to determine a pair of reliable thresholds (λ^l, λ^h) that define a RCPS $\hat{C}_{(\lambda^l, \lambda^h)} = \hat{C}_{\text{nor}} \cup \hat{C}_{\text{ano}}$ with finite-sample coverage guarantees for any given risk function $R(\cdot) : 2^{\mathcal{X}} \to \mathbb{R}$ (e.g., the false positive rate). Additionally, we seek to adapt the abstention rate based on the complexity of the risk function.

Our LTT procedure for reliable threshold selection. We propose an extension of the LTT procedure, denoted as `xLTT`, which generalizes the framework to consider a pair of thresholds (λ^l, λ^h) instead of a single threshold λ. The procedure begins by defining a set of paired threshold values, $\Lambda = \{(\lambda^l_{(a)}, \lambda^h_{(b)}) \mid a, b \in \{1, \ldots, m\},\ \lambda^l_{(a)} \leq \lambda^h_{(b)}\}$. Next, we define the null hypotheses $\mathcal{H}_j : \hat{R}_{n_C}(\hat{C}_{(\lambda^l_j, \lambda^h_j)}) > \alpha$ for each $(\lambda^l_j, \lambda^h_j) \in \Lambda$, $j \in \{1, \ldots, |\Lambda|\}$ and $\alpha \in [0, 1]$, where $\hat{R}_{n_C}(\cdot) : 2^{\mathcal{X}} \to \mathbb{R}$ is an empirical risk function computed on the calibration set $\mathcal{D}_C$. Accepting $\mathcal{H}_j$ indicates that $(\lambda^l_j, \lambda^h_j)$ does not control the risk. To decide whether to accept or reject $\mathcal{H}_j$ and thus verify whether the risk is controlled for a given pair $(\lambda^l_j, \lambda^h_j)$, we compute a valid p-value p_j for every $\mathcal{H}_j$ using α. This is achieved via a concentration inequality (e.g., the Hoeffding-Bentkus inequality [6]). Based on the set of p-values $P = \{p_j\}_{j \in \{1, \ldots, |\Lambda|\}}$, we then select the threshold pairs for which the risk is controlled. Since multiple comparisons increase the likelihood of false positives, a correction function $\mathcal{A} : P \to P'$ with $P' \subseteq P$ is required to maintain the desired risk control. For example, we define the set $\mathcal{O} = \mathcal{A}(P) \subset \Lambda$ using Bonferroni correction as $\mathcal{A}(P) = \{(\lambda^l_j, \lambda^h_j) \mid p_j \leq \frac{\delta}{|\Lambda|}, p_j \in P\}$. If $\mathcal{O} = \emptyset$, we set $\mathcal{O} = \{(-\infty, \infty)\}$. Finally, any pair $(\lambda^l, \lambda^h) \in \mathcal{O}$ ensures that $\hat{C}_{(\lambda^l, \lambda^h)}$ forms a risk-controlling prediction

set. This method enables the use of any risk function in a post-hoc manner (i.e., without requiring retraining of a given anomaly detector), making it particularly valuable for AD in CSP plants with diverse and evolving requirements.

Optimal threshold selection for AD. Now that we have obtained the set $\mathcal{O}$ of threshold pairs that control the risk, our next objective is to (1) avoid trivial selections where $\lambda^l = -\infty$ and $\lambda^h = \infty$, and (2) minimize false positives and false negatives while keeping the abstention rate as low as possible. Let $\mathcal{I}_1 = \{i \mid z_i = 1\}$ and $\mathcal{I}_0 = \{i \mid z_i = 0\}$ be the set of indices for anomalous and normal points, respectively. $\hat{z}_i$ are the predicted labels computed using (4), with $i = 1, \ldots, |\mathcal{D}_V|$. We propose selecting the optimal thresholds λ^l_* and λ^h_* by computing:

$$\lambda^l_*, \lambda^h_* = \arg\min_{\lambda^l, \lambda^h \in \mathcal{O}} \underbrace{\frac{|\{i \in \mathcal{I}_1 \mid \hat{z}_i = 0\}|}{|\mathcal{I}_1|}}_{\text{False Negative Rate (FNR)}} + \underbrace{\frac{|\{i \in \mathcal{I}_0 \mid \hat{z}_i = 1\}|}{|\mathcal{I}_0|}}_{\text{False Positive Rate (FPR)}} + \underbrace{\frac{|\{i \mid \hat{z}_i = \varnothing\}|}{|\mathcal{D}_V|}}_{\text{Abstention Rate}}.$$

Density-Based Anomaly Score Functions. Recent work [23] examined the intrinsic connection between anomaly detection and conformal prediction, demonstrating how insights from each field can mutually enhance the other. Building on this perspective, we leverage recent advancements in CP [13] to develop novel anomaly score functions $s(\cdot)$[5] for the labeling function in (4). These score functions are further integrated with the reliable threshold selection procedure `xLTT`.

Our framework is based upon an invertible, conditional generative model (e.g., normalizing flows) $\hat{g} : \mathcal{V} \times \mathcal{C} \times \mathbb{R}_+ \to \mathcal{Y}$, where $\mathcal{V}$ is a latent variable with a known distribution and $\mathcal{C}$ is the space of the conditioning variable. We defer the discussion of the exact model used to Section 5. Formally, $\hat{g}(\hat{g}^{-1}(y; c, t); c, t) = y$ for any $c \in \mathcal{C}$, $y \in \mathcal{Y}$ and $t \in \mathbb{R}_+$. The invertibility allows us to compute the exact density $\hat{f}(y \mid c, t)$ via the change of variables formula. For a test observation $x = (y, t)$, and given $\hat{g}$, we consider the following two approaches:

- `DR-xLTT`**.** The negative log-likelihood is the score function:

$$s_{\mathrm{DR}}(x; c) = -\log(\hat{f}(y \mid c, t)). \tag{5}$$

- `L-xLTT`**.** The second approach is based on an invertible model, following the L-CP method introduced in [13]. Unlike the output space $\mathcal{Y}$, we expect the latent space $\mathcal{V}$ to be more structured, where normal samples are ideally clustered near the origin. Consequently, in `L-xLTT`, we frame the decision-making process as a one-class classification problem in the latent space. Assuming the latent variable follows a standard normal distribution, we use the ℓ_2 distance of the latent representation from the origin as the anomaly score for a test point x:

$$s_{\mathrm{L}}(x; c) = \|v\|, \quad \text{where } v = \hat{g}^{-1}(y; c, t). \tag{6}$$

[5] Hereafter, the score function incorporates contextual information c.

5 Density-based AD Model

The most recent AD model for CSP plants, `ForecastAD`, is a reconstruction-based AD methods. However, prior research has shown that anomalies, despite significantly different from normal data, can often be reconstructed in practice [10]. For instance, in a bimodal distribution, the distance between the two peaks is greater than the distance between a peak and the local minimum separating them. In such cases, when a prediction aligns with one of the peaks, observations near the local minimum exhibit lower reconstruction errors and thus are incorrectly deemed more likely [22]. Figure 3 presents examples of IR images that are well reconstructed but are anomalous and exhibit empirically low density.

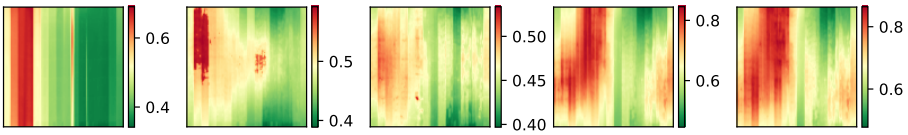


Fig. 3: Well-reconstructed anomalous thermal images with empirically low density.

To overcome such limitations of reconstruction-based approaches, we introduce `DensityAD`, an invertible generative model that directly estimates the density of thermal images given the contextual information from past images. `DensityAD` operates in two main steps: (i) concatenating the K preceding images and their timestamps as a context vector c, and (ii) leveraging this context to estimate the density of the current observation $x = (y, t)$, i.e. $f(y \mid c, t)$.

Context encoding. Building on [24], given a test observation $x_i = (y_i, t_i)$, we construct a rich contextual representation c_i for AD by encoding both spatial and temporal information from the preceding K images. First, at each time step t_{i-k}, where $k = 1, \cdots, K$, the corresponding image y_{i-k} is mapped into a lower-dimensional latent space. Specifically, we define an image encoder $\phi_e(\cdot; W_e) : \mathcal{Y} \to \mathcal{V}'$, which transforms images from the high-dimensional input space $\mathcal{Y}$ into a lower dimensional latent space $\mathcal{V}' = \mathbb{R}^{d'}$, where $d' \ll d$. Then, to capture temporal dependencies, we consider two temporal features: the inter-arrival time $\tau_{i-k} = t_{i-k} - t_{i-(k+1)}$, which represents the time elapsed since the previous observation, and the relative time since the start of operation $\gamma_{i-k} = t_{i-k} - t_0$, which situates the observation within the broader operational cycle. These temporal attributes are encoded using a sinusoidal function $\psi(\cdot)$. The final embedding for each data point (y_{i-k}, t_{i-k}) is then constructed by concatenating the temporal encodings with the image embedding as $\hat{c}_{i-k} = \phi_e(y_{i-k}; W_e) \oplus \psi_\tau(\tau_{i-k}) \oplus \psi_\gamma(\gamma_{i-k})$. Lastly, to generate the fixed-dimensional context vector c_i at time step t_i, the embeddings of the past K

images are aggregated using a deep sequence model.

Conditional Normalizing Flow. The conditional PDF $f(y_i \mid c_i, t_i)$ of the current image y_i, given context c_i at timestep t_i, is estimated using a conditional normalizing flow, specifically GLOW [21]. The invertibility property of normalizing flows [30,16] enables exact likelihood computation, which is essential for the threshold selection methods discussed in Section 4. To model $f(y_i \mid c_i, t_i)$, we apply conditional invertible transformations g, mapping y_i to a latent variable v_i as $v_i = g(y_i; c_i, t_i)$. The conditional log-likelihood is then computed using the change-of-variables formula. For further details, we refer the reader to [21].

6 Experiments

Here, we compare the performance of `DensityAD` against existing baselines and assess the efficacy of our proposed decision thresholds for risk-controlled AD.

6.1 Experimental Setup

Dataset. We use data from two CSP plants, denoted as A and B. The validation set also serves as a calibration set. For the first data point of each day, both τ and γ are initialized to a small positive value, $\epsilon =$ 1e-5.
Baselines. In our evaluation, we compare the performance of `DensityAD` against deep image-based AD methods, specifically CFlow [14] and DRÆM [37]. To extend the comparison to AD approaches that incorporate historical sequences of observations, similar to `DensityAD`, we include a spatiotemporal autoencoder (STAE) architecture [15,12,34] and TimeSformer [9], a transformer-based video classification framework, as baselines, along with `ForecastAD` [24].
Experimental details. To prevent numerical instability during training, images are resized to 64×64, and we employ 3 flows per block across 5 blocks. The model is trained using the Adam optimizer with a learning rate of 0.0001 and a weight decay of 0.00001. Early stopping is applied based on the validation AUPR[6], maintaining a fixed balance between normal and anomalous samples in the validation set during training to mitigate the impact of dataset imbalance. The baseline models are trained following their published training setups. We used TimeSformer as the encoder in an encoder-decoder architecture, using the decoder from `ForecastAD`, and trained with a mean squared error loss. For the decision thresholds, we use $\alpha = \delta = 0.1$. We conduct an ablation study in Section 2 of the supplementary material on the context length K and the importance of time embeddings τ and γ. Based on the analysis, we opt for the sequence length $K = 30$ and only consider τ in `DensityAD` for modelling the temporal dynamics.
Evaluation metrics. We evaluate `DensityAD` using two primary metrics: the AUROC[7] and the AUPR. Additionally, we assess the proposed thresholding

[6] Area Under the Precision-Recall Curve (AUPR)
[7] Area Under the Receiver Operating Characteristic Curve (AUROC)

Table 1: AUROC and AUPR performances of `DensityAD` against baseline methods. Style: best in **bold**, and second best underlined.

CSP	Model	AUROC (%) ↑	AUPR (%) ↑
A	CFlow [14]	76.46 ± 0.92	70.32 ± 1.20
	DRÆM [37]	81.55 ± 1.9	74.8 ± 2.79
	STAE	89.47 ± 1.59	87.38 ± 2.4
	TimeSformer [9]	87.8 ± 2.46	83.36 ± 3.15
	`ForecastAD` [24]	86.28 ± 1.74	87.57 ± 1.38
	`DensityAD`	**94.25 ± 0.2**	**93.88 ± 0.48**
B	CFlow [14]	55.8 ± 5.47	57.56 ± 4.85
	DRÆM [37]	78.82 ± 5.72	71.75 ± 8.56
	STAE	89.9 ± 1.18	88.98 ± 1.68
	TimeSformer [9]	88.59 ± 2.14	89.84 ± 1.29
	`ForecastAD` [24]	81.76 ± 0.7	82.88 ± 1.39
	`DensityAD`	**91.93 ± 0.52**	**90.66 ± 0.46**

methods by reporting the risk, along with the F1-score and the corresponding abstention rate for two controlled risk measures relevant to our context: the FPR and the F1-score. These choices are not fixed—any risk function can be selected to meet the specific requirements of a CSP plant. We also report these risk measures for existing threshold selection methods. For all experiments, we present the mean over three runs along with one standard error.

6.2 Results and Discussion

AD models. Table 1 presents the performance of `DensityAD` for both CSP plants. The results indicate that `DensityAD` consistently outperforms all baseline methods on both datasets. While STAE, `ForecastAD`, and TimeSformer perform well, they still fall short of the performance achieved by our `DensityAD`.

Anomaly scores. Figure 4 shows the distributions of normal and anomalous scores for test samples on CSP A, using the proposed scores, introduced in Section 4 (i.e., s_{DR} and s_{L}) and the reconstruction score s_{REC} from `ForecastAD`. In this example, s_{DR} and s_{REC} scores effectively distinguish normal from anomalous samples, as shown by the overlapping area (OA) between both distributions.

Anomaly threshold selection. Figure 5 provides an overview of the threshold selection approaches. The results clearly show that the proposed methods effectively control risk for both risk functions, whereas existing methods do not offer such guarantees. The `DR-xLTT` methods demonstrate strong performance, balancing risk control with a high F1-score while maintaining a low abstention rate. Notably, they outperform approaches that select the maximum validation set value. Furthermore, these methods adapt to the complexity of the risk function, recognizing that controlling the F1-score presents greater predictive challenges than the FPR. They also fully adjust to user requirements, increasing the abstention rate when constraints are too stringent (e.g., attempting to control

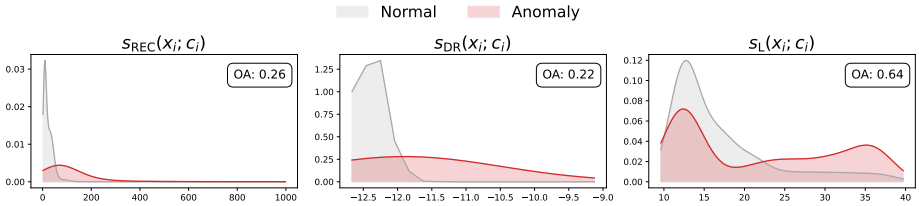


Fig. 4: Empirical score distributions of normal and anomalous test samples from CSP *A* for our proposed score functions and the one used by `ForecastAD`, with the overlapping area (OA) between both distributions in the top right corner.

Table 2: Total CPU time and memory used for training the models.

	`ForecastAD`		`DensityAD`	
	CSP A	CSP B	CSP A	CSP B
Training time (s)	4151 ± 327	2204 ± 334	3760 ± 836	3248 ± 411
Memory used (Gb)	1.63 ± 0.13	1.73 ± 0.05	1.73 ± 0.02	1.63 ± 0.02
Inference time (ms)	194 ± 6.73	177 ± 1.44	201.3 ± 17.41	178 ± 2.37

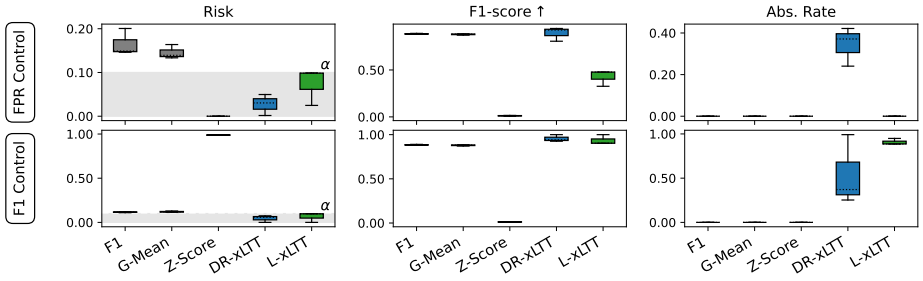
the F1-score with a weak underlying model).

Computation requirements. The models are trained using a single NVIDIA A100 GPU with 40 GB of memory, along with 8 CPU cores and 20 GB of RAM. Table 2 presents the training times (excluding the risk-control), memory usage and inference time for a single test point. As shown in Table 1, `DensityAD` performs better than `ForecastAD` while using similar resources.

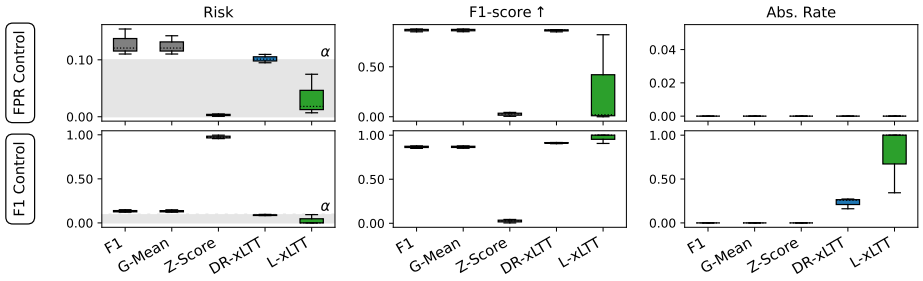
7 Deployment

We deployed our threshold selection methods using `DensityAD` on 5 and 6 months of anonymized data from CSP plants A and B, respectively. Figure 6 presents the thresholding results, where the FPR is used as the controlled risk. The results demonstrate that risk is effectively controlled in deployment, with `DR-xLTT` emerging as the most consistent method across both CSP plants. All methods maintain a low abstention rate, making them well-suited for deployment. Additionally, the deployment results align closely with those observed during testing. Performance fluctuations across months can be attributed to variations in the frequency and complexity of anomalies, with some months exhibiting a higher occurrence or more challenging cases.

Figure 7 evaluates the performance of `DensityAD` in deployment under three training configurations: training on CSP A, training on CSP B, and training on a combination of data from both CSP plants. As expected, deploying a model trained on a different tower results in a performance decline. Furthermore, training on data from both plants does not yield any performance improvement, sug-



(a) For CSP A



(b) For CSP B

Fig. 5: Risk control over FPR (top row) and F1-score (bottom row) for existing and proposed methods. The risk is FPR (top row) and 1−F1 (bottom row).

gesting that information from one tower does not generalize well to another. Although thermal flow patterns are similar across CSPs, anomaly definitions vary due to site-specific factors such as geographic location and operational context. This limits cross-site generalization, indicating the need for per-site models or fine-tuning. Future work could address this through domain adaptation techniques. Although originally not designed for this purpose, `DensityAD` offers a general framework that can be extended to multivariate time series anomaly detection. In this work, we focus on its application to anomaly detection in CSPs, where the anomaly score is computed and subsequently processed through a thresholding mechanism. Finally, the results suggest that the proposed method supports practical deployment by enabling control over operational risk. For instance, it allows organizations to meet predefined detection targets—such as identifying 90% of anomalies—thereby supporting compliance with regulatory or performance requirements.

8 Simulated dataset

Building on the methodology described in [24], we construct a simulated dataset to facilitate the reproducibility and validation of our results. `DensityAD` enables

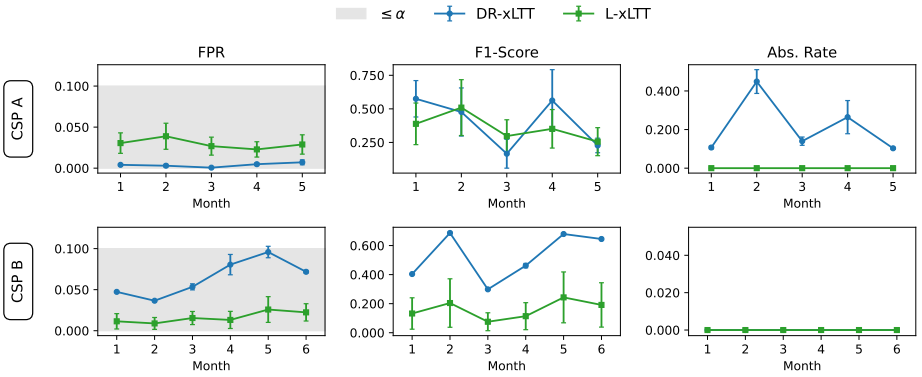


Fig. 6: FPR control for the proposed approaches in a deployment setting over multiple months, for the two CSP plants.

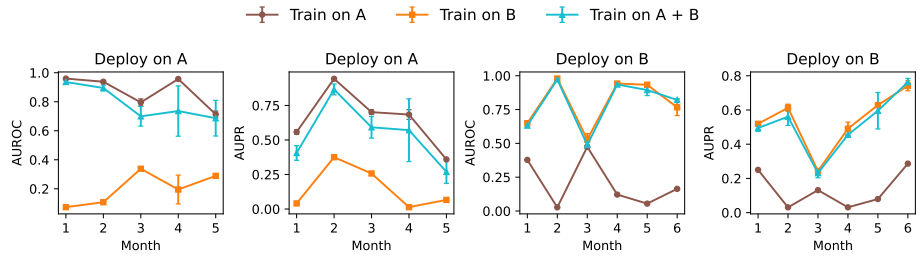


Fig. 7: AUROC and AUPR for the two CSP plants over multiple months using three different training settings (i.e. training on A, B, and $A + B$).

exact likelihood computation while also allowing sampling from the learned distribution. Leveraging this capability, we generate high-quality samples using our proposed density-based model. The dataset simulates two distinct CSP setups, providing a valuable resource for advancing anomaly detection research in CSP plants. Further details are provided in the supplement.

Due to transformations applied during anonymization, we assessed the reliability of the generated images by comparing the average daily temperature profiles of both CSP plants. As shown in Figure 8, temperature levels during the critical period (08:00 to 20:00) closely match between the simulated and original datasets. Temperatures outside this interval, which are notably lower, were regarded as trivial outliers and excluded from the simulated data.

9 Related work

Unsupervised AD. Based on the assumption of a "clean" training dataset, i.e., containing only normal samples, unsupervised AD approaches have been

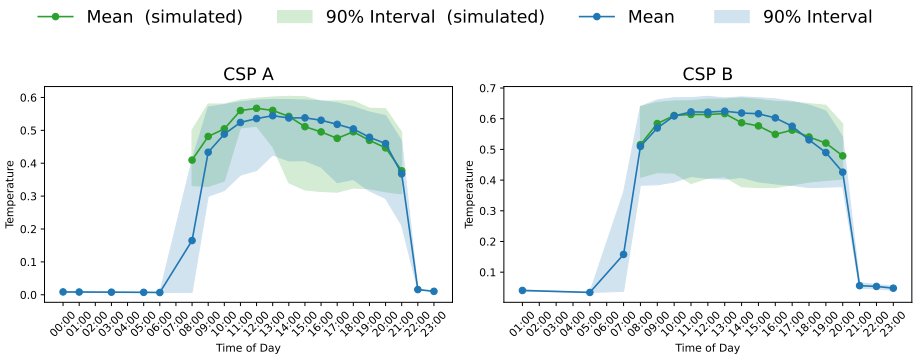


Fig. 8: Mean daily temperature for the original and simulated datasets. The shaded area represents the 90% temperature interval.

proposed with the aim of training models that learn a "compact" representation of the normal behaviour. Then, anomalies are identified as deviations from this learned normality. Existing methods can be broadly categorized into four families [32]. First, both deep and shallow *one-class classifiers* [35,33] learn a decision boundary around normal samples with classical methods such as support vector data description. Second, *feature embedding-based* methods store or learn normal data representations using pre-trained models [31,19] or student-teacher networks [38,8,25]. Third, *reconstruction-based* methods use encoder-decoder architectures to map normal samples into a lower-dimensional bottleneck and reconstruct them with high fidelity. Lastly, *density-based* methods estimate the probability distribution of normal samples under the *concentration assumption*, where anomalies are expected to be in low-density regions. For a comprehensive survey, we refer readers to [20,32].

Beyond image-based AD, prior research also investigated AD in videos (VAD) [11,18,1], using historical sequences of observations to identify deviations. However, our setting differs in two key ways: (1) our IR images lack the semantic content of typical video frames, and (2) our solar dataset is captured at irregular intervals, while videos are captured at a fixed frame rate. Although [24] introduced a forecast-based AD approach for CSP plants, it lacks a reliable selection of AD threshold. Moreover, their study is limited to a single CSP plant, whereas our setting involves multiple plants, introducing additional heterogeinity.

Anomaly detection thresholds. To assign labels, AD methods typically threshold anomaly scores. Commonly used decision thresholds particularly relevant to our use case involve optimizing performance metrics, such as the F1-score [2], G-Mean [17], or the area under the Precision-Recall Curve (PRC), on the validation set over a range of possible thresholds. Another class of methods builds on conformal prediction (CP) [36], a distribution-free framework for constructing prediction sets (e.g., those defined in our decision-making process) providing a

finite-sample coverage guarantee. [7] introduces a method for computing conditionally valid conformal p-values for nonparametric outlier detection, framing the problem within a multiple hypothesis testing context. A key extension of CP, named *conformal risk control* [5], shifts the guarantee from coverage to managing any monotonically non-increasing risk function. The *Learn then Test* procedure [4] further allows us to extend this concept to any risk function, irrespective of its monotonicity, to generate risk-controlled prediction sets [6].

10 Conclusion

We introduced a principled and robust framework for anomaly detection (AD) designed to monitor CSP plants using infrared imagery captured at irregular intervals throughout the day. Our approach labels images as normal or anomalous by first assigning an anomaly score using a model trained on an unlabeled image dataset, followed by a thresholding procedure. To address the challenges of unsupervised AD for CSP plants, our contributions are fourfold. First, we proposed a framework for computing reliable anomaly detection thresholds with finite-sample risk coverage guarantees for any chosen risk function while allowing deferral to domain experts under high uncertainty. Second, to compute more robust anomaly scores for an observed image, we developed a density forecasting method that estimates its likelihood conditional on a sequence of previously observed images. Third, we conducted an extensive real-world deployment analysis over several months across two operational CSP plants, providing valuable insights for industrial maintenance. Lastly, we released a simulated dataset leveraging recent advancements in generative modeling, facilitating data-driven predictive maintenance (PdM) for critical infrastructure. By enhancing the reliability of renewable energy systems, our work supports the broader adoption of sustainable energy solutions for a greener future.

Acknowledgments. This research is funded by the project "Federated Learning and Augmented Reality for Advanced Control Centers." Special thanks to Thibault GEORGES and Adrien FARINELLE from John Cockerill for their assistance in analyzing the dataset and identifying related abnormal behaviors. We also acknowledge Victor DHEUR for his valuable feedback on the risk control approach.

References

1. Abdalla, M., Javed, S., Radi, M.A., Ulhaq, A., Werghi, N.: Video anomaly detection in 10 years: A survey and outlook. arXiv preprint arXiv:2405.19387 (2024)
2. Akcay, S., Ameln, D., Vaidya, A., Lakshmanan, B., Ahuja, N., Genc, U.: Anomalib: A deep learning library for anomaly detection. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 1706–1710. IEEE (2022)
3. Angelopoulos, A.N., Bates, S.: Conformal prediction: A gentle introduction. Found. Trends® Mach. Learn. **16**(4), 494–591 (2023)

4. Angelopoulos, A.N., Bates, S., Candès, E.J., Jordan, M.I., Lei, L.: Learn then test: Calibrating predictive algorithms to achieve risk control. The Annals of Applied Statistics **19**(2), 1641 – 1662 (2025). https://doi.org/10.1214/24-AOAS1998, https://doi.org/10.1214/24-AOAS1998
5. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L., Schuster, T.: Conformal risk control. arXiv [stat.ME] (Aug 2022)
6. Bates, S., Angelopoulos, A., Lei, L., Malik, J., Jordan, M.: Distribution-free, risk-controlling prediction sets. Journal of the ACM (JACM) **68**(6), 1–34 (2021)
7. Bates, S., Candès, E., Lei, L., Romano, Y., Sesia, M.: Testing for outliers with conformal p-values. The Annals of Statistics **51**(1), 149–178 (2023)
8. Batzner, K., Heckler, L., König, R.: EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 128–138 (1 2024)
9. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? **2**(3), 4 (2021)
10. Bouman, R., Heskes, T.: Autoencoders for anomaly detection are unreliable. arXiv [cs.LG] (Jan 2025)
11. Chandrakala, S., Deepak, K., Revathy, G.: Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis. Artificial Intelligence Review **56**(4), 3319–3368 (2023)
12. Deepak, K., Chandrakala, S., Mohan, C.K.: Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing **15**(1), 215–222 (2021)
13. Dheur, V., Fontana, M., Estievenart, Y., Desobry, N., Taieb, S.B.: A unified comparative study with generalized conformity scores for multi-output conformal regression. arXiv [stat.ML] (Jan 2025)
14. Gudovskiy, D., Ishizaka, S., Kozuka, K.: CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows. Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022 pp. 1819–1828 (7 2021). https://doi.org/10.1109/WACV51458.2022.00188
15. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 733–742 (2016)
16. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems **31** (2018)
17. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the 14th International Conference on Machine Learning pp. 179–186 (1997)
18. Le, V.T., Kim, Y.G.: Attention-based residual autoencoder for video anomaly detection. Applied Intelligence **53**(3), 3240–3254 (2023)
19. Lee, S., Lee, S., Song, B.C.: CFA: Coupled-hypersphere-based Feature Adaptation for Target-Oriented Anomaly Localization. IEEE Access **10**, 78446–78454 (6 2022). https://doi.org/10.1109/ACCESS.2022.3193699
20. Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., Jin, Y.: Deep industrial image anomaly detection: A survey. Machine Intelligence Research **21**(1), 104–135 (2024)
21. Lu, Y., Huang, B.: Structured output learning with conditional generative flows. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5005–5012 (2020)

22. Moore, A., Morelli, D.: ConDENSE: Conditional density estimation for time series anomaly detection. J. Artif. Intell. Res. **79**, 801–824 (Mar 2024)
23. Novello, P., Dalmau, J., Andéol, L.: Exploring the link between out-of-distribution detection and conformal prediction with illustrations of its benefits (2025)
24. Patra, S., Sournac, N., Taieb, S.B.: Detecting abnormal operations in concentrated solar power plants from irregular sequences of thermal images. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 5578–5589. KDD '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3637528.3671623
25. Patra, S., Taieb, S.B.: Revisiting deep feature reconstruction for logical and structural industrial anomaly detection. Transactions on Machine Learning Research (2024)
26. Perini, L., Bürkner, P.C., Klami, A.: Estimating the contamination factor's distribution in unsupervised anomaly detection pp. 27668–27679 (2023)
27. Perini, L., Davis, J.: Unsupervised anomaly detection with rejection. Advances in Neural Information Processing Systems **36**, 69673–69691 (2023)
28. Perini, L., Vercruyssen, V., Davis, J.: Quantifying the confidence of anomaly detectors in their example-wise predictions. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 227–243. Springer (2020)
29. Perini, L., Vercruyssen, V., Davis, J.: Transferring the Contamination Factor between Anomaly Detection Domains by Shape Similarity. Proceedings of the AAAI Conference on Artificial Intelligence **36**(4), 4128–4136 (6 2022). https://doi.org/10.1609/AAAI.V36I4.20331, https://ojs.aaai.org/index.php/AAAI/article/view/20331
30. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
31. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)
32. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)
33. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep One-Class Classification. In: International Conference on Machine Learning. pp. 4393–4402. PMLR (2018)
34. Sudhakaran, S., Lanz, O.: Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). pp. 1–6. IEEE (2017)
35. Tax, D.M., Duin, R.P.: Support vector domain description. Pattern Recognition Letters **20**(11-13), 1191–1199 (11 1999). https://doi.org/10.1016/S0167-8655(99)00087-2
36. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer, New York, NY, 2005 edn. (Mar 2005)
37. Zavrtanik, V., Kristan, M., Skočaj, D.: DRÆM - A discriminatively trained reconstruction embedding for surface anomaly detection. Proceedings of the IEEE International Conference on Computer Vision pp. 8310–8319 (8 2021). https://doi.org/10.1109/ICCV48922.2021.00822
38. Zhang, J., Suganuma, M., Okatani, T.: Contextual affinity distillation for image anomaly detection pp. 149–158 (2024)