Autoencoder-Based Model of Nuclear Power Plant Core Temperature for Blockage Event Detection

Rémi Cogranne

Autoencoder-Based Model of Nuclear Power Plant Core Temperature for Blockage Event Detection

Rémi Cogranne
Troyes University of Technology
Troyes, France
remi.cogranne@utt.fr

ABSTRACT

This paper addresses the critical necessity of continuously monitoring systems of the temperature of fuel rod assemblies within a nuclear power plant and presents a quick and accurate detection of total and instantaneous blockages. Our research works propose an advanced and original autoencoder-based methodology for adaptive modelling of normal operational temperatures. This model allows incorporating detailed expertise on the specific malfunctions targeted, in order to detect and identify small heat increases indicative of blockages in the flow of a nuclear core coolant. The proposed online detection method ensures a reliable outcome in terms of false alarm probability as well as detection delay which are both of the utmost importance. Experimental results, utilizing real actual temperature measurements from the Superphénix power station, demonstrate the model's accuracy and the relevance of the proposed detection method.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Information extraction**; *Probabilistic reasoning*; *Control methods*; • **Hardware** → **Temperature monitoring**; *Safety critical systems*; Temperature control; • **Theory of computation** → *Unsupervised learning and clustering*.

KEYWORDS

Autoencoders, VAE, Detection, Coolant, Temperature, Real-Time, Monitoring

ACM Reference Format:
Rémi Cogranne. 2025. Autoencoder-Based Model of Nuclear Power Plant Core Temperature for Blockage Event Detection. In *Proceedings of ACM International Conference on Computing and Artificial Intelligence (ACM ICCAI '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Monitoring critical infrastructure is a top-priority concern due to the severe consequences of potential incidents. Offices and norms for critical national infrastructure protection exist in a vast majority of countries, such as the EU and US critical infrastructure protection (EUCIP and CIP) programs. Among critical infrastructure, nuclear power plants are the most important and sensitive, given the significant impact of any malfunction. This is especially true for new reactor types under development, such as the Sodium-cooled Fast Reactor (SFR): one of six models selected by the Generation IV International Forum (GIF) for future nuclear power plants [32, 35]. Despite rigorous development processes, safety and reliability can still be challenged by very unlikely events, such as core cooling incidents. One of such critical incidents is the Total and Instantaneous Blockage (TIB), which is extremely unlikely to happen, yet, if undetected, it can lead to the melting of several reactor rods. This paper focusses on the early detection of TIB.

Change detection in complex systems, including thermal incidents, has been extensively studied in various fields, such as finance, econometrics, and manufacturing. In this paper, we address the general problem of abrupt change detection within a complex critical infrastructure in a real-time context. More precisely, real-time monitoring of the nuclear reactor's core temperature is essential for detecting overheating and identifying various incidents, some of which are less obvious. In this operational context, the highest priority is to ensure the reliability of the detection system. This requires a detector with well-established statistical properties, guaranteeing a low false alarm rate and a very quick detection with the lowest possible delay.

In this context, the present paper proposes an effective use of autoencoders to model the temperature of nuclear cores, capturing complex temporal dependencies with precision. It enables the accuracy of blockage by integrating statistical detection theory. We believe that the combination of these two approaches offers a significant contribution to possible applications in a wide range of problems.

The present paper is organized as follows. Section 2 briefly presents the specific thermal anomaly event it is aimed at detecting and briefly introduces the main difficulties and also recalls the current art in the field of change detection and especially using artificial intelligence. Then Section 3 states the problem of the quickest detection of coolant flow blockage, considering the very weak heat anomaly it has at the location of the temperature probes. The original method we have designed is presented and explained in detail in the Section 4. Extensive numerical evaluation over a real dataset of the temperature of a nuclear power plant is presented in section 5 . Finally, Section 6 concludes the present paper by briefly recalling the theoretical findings and numerical validation and also draws possible future works.

ACM ICCAI '25, March 28–31, 2025, Kyoto, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 STATEMENT AND POSITION OF THE PROBLEM

This section details the Total and Instantaneous Blockage (TIB) event that the proposed method aims at detecting. We especially highlight the underlying challenges and introduce the notation used throughout the paper.

2.1 Total and Instantaneous Blockage

Figure 1 illustrates the structure of a Sodium-cooled Fast Reactor (SFR). The core contains the nuclear fuel, in the form of rods, and a coolant, which cools the core and exchanges heat with the alternator to produce electricity. The rods are arranged in a honeycomb structure, each placed in a steel pipe through which the coolant flows. The temperature of the rod assembly is monitored to detect overheating. However, direct measurement of the rods is impractical. Instead, the temperature of the output coolant is measured, providing an indirect measure of the rod temperatures.

As the name suggests, a total and instantaneous blockage occurs when the flow of the coolant is completely blocked all of a sudden. In this case, the actual temperature of the central increases very quickly yet the measured temperature of this rod remains roughly constant. The flows of the coolant being completely stopped, the measured temperatures do not correspond any longer with the actual rod temperature. However, the temperatures (actual and measured) increased due to thermal conductivity but only very lightly and rather slowly. This can cause nuclear core melting and an ensuing major accident if emergency systems are not activated immediately. A notable example of TIB occurred on October 5, 1966, at the FERMI-1 reactor in Michigan, USA. To quote the report [2]: "*during power ascension, zirconium plates at the bottom of the reactor vessel became loose and blocked sodium coolant flow to some fuel subassemblies. Two subassemblies started to melt and the reactor was manually shut down*" ; this event was the subject of the book [12].

It is therefore hoped that such an event will be the subject of a specific detection procedure [1, 23, 24]. However, this is a challenging problem, mostly because of the following difficulties:

- The temperature of a nuclear power plant is very difficult to model with accuracy. Indeed, the temperature is largely affected by the so-called control rods that absorb neutrons in order to control the rate of fission of the nuclear fuel. The temperature changes due to power adjustments and, more generally, depends on many factors that are out of control.
- The impact of total blockage on temperatures constitutes a very weak signal, at the beginning, again due to the absence of coolant flow.
- The detection method must be extremely reliable with respect to the following two detection criteria: (1) the detection delay must be below a certain limit after which the rods may start melting down, and (2) the false alarm rate must be kept extremely low, as detection leads to an emergency shutdown procedure, which is particularly costly in all respects.

The two firsts problems can only be addressed using a very accurate model of the temperature of all rods under normal operating conditions. In addition, such a model must be computationally simple, as the system must be monitored in real time; complex physical models are therefore impossible. What's more, some parameters

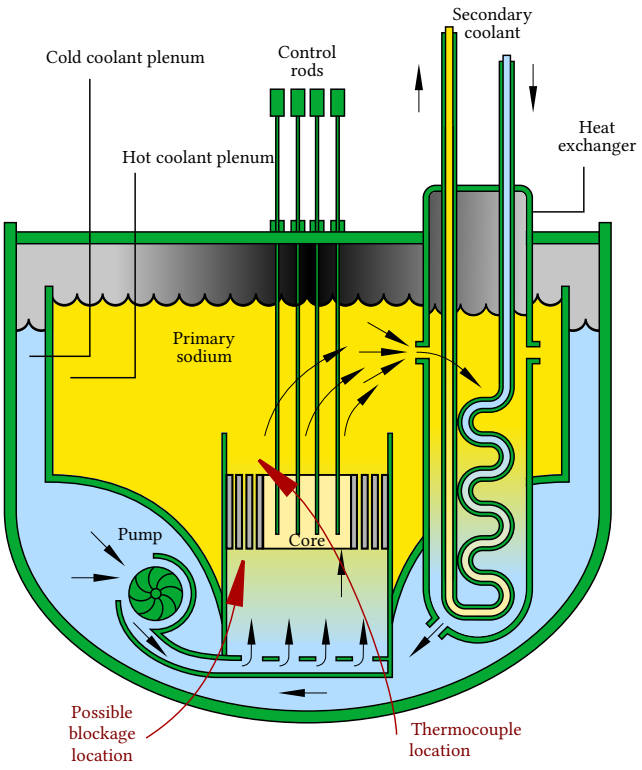


Figure 1: Illustration of the overall design of an SFR nuclear reactor, here using a pool architecture emphasizing the location of thermocouples with respect to the possible blockage. Source: based on an image in the public domain from www.Gen-4.org.

(outside temperature and humidity, water temperature, etc.) are not precisely known.

The last point is more technical but the usual sequential detection system focusses on the average detection delay. For instance, the well-known Sequential Probability Ratio Test (SPRT) [42] and the CUSUM [29] are, under mile condition, optimal for minimizing the worst-case average delay before detection. However, in our case, we want to maximize the detection under a strict maximal delay.

The latter problem is presented in more detail in Section 3.1 while the former is addressed in Sections 3.2 and 3.3.

2.2 Detection in Time Series With Deep Learning

Since the present paper aims at exploiting the latest artificial intelligence methods to detect anomalies in time series, we shall give a brief overview of the state of the art in this field. Interested readers can find comprehensive surveys on deep learning for time series in [3, 15], more specific reviews on transformers in time series in [43] and on autoencoders for anomaly detection in [25, 33].

Transformers have been effectively applied to anomaly detection in time series [22, 43], taking advantage of their ability to model

temporal dependencies and improve detection quality [44]. Several studies, including TranAD [41] and TransAnomaly [46], have combined transformers with neural generative models used as an efficient data augmentation means, hence improving performance. TranAD [41] uses adversarial learning to amplify reconstruction errors, addressing transformers' tendency to miss small anomalies, while TransAnomaly [46] uses autoencoders with the aim of reducing training costs. More generally, the reconstruction error has been widely used for anomaly detection and transformers have been used to approximated the inspected data hence generating reconstruction error, see for instance [45]. AnomalyTrans [44] associates Transformers with a Gaussian a priori to distinguish better anomalies, optimizing the model using a minimax strategy to improve association divergence. Last but not least, transformers have been extensively used in time-series for forcasting and this can also be used for anomaly detection with the underlying idea that when an anomaly occurs the forecasting is less accurate. The interested reader is reader, for example, to [17] and the references therein.

On the other hand, autoencoders (AEs) and variational autoencoders (VAEs) [18] have been widely applied for anomaly detection due to their ability to learn representation of complex signals [4]. In supervised anomaly detection, AE are generally trained first to learn the patterns of normal data, then self-refined on combined normal and abnormal data to distinguish between them by minimizing the reconstruction error. After training, test data are classified as abnormal if the reconstruction error exceeds a predefined threshold [30]. This approach combines the feature learning capabilities of AEs with the discriminating power of supervised classifiers, improving the accuracy of anomaly detection in real-world applications such as fraud detection [9], network security [20], and fault detection in industrial processes [8].

In unsupervised tasks, AEs are trained solely on normal data with the goal of minimizing reconstruction error with respect to input data. When abnormal data passes through the network, the reconstruction error is higher. A threshold is set based on this error to classify the data as abnormal [4]. The versatility of EAs and their ability to adapt to various types of data make this method efficient in various applications, especially when the dataset is highly imbalanced. This includes, for example, cybersecurity to identify network intrusions [19], manufacturing to spot defects [31] and finance to detect fraud [16]. Data contribute to their widespread use in unsupervised anomaly detection scenarios, enhancing system security and reliability.

3 STATISTICAL DETECTION IN NUISANCE PARAMETERS

In order to understand the contribution of the present paper, this section presents how we addressed the three main difficulties presented in Section 2.1. First, Section 3.1 recalls primers on the quickest detection problem and explains the originality of our work and the proposed sequential method. Next, the general methodology for dealing with nuisance parameters is presented in Section 3.2. Last, but not least, the model of the abnormal temperature due to a blockage, that is the signal it is aimed at detecting, is presented in the Section 3.3.

3.1 Quickest Detection Problem

The temperature of nuclear rods must be monitored in real time to detect blockages. This setting, where data are analysed one-by-one as they are received, is referred to as sequential detection [36].

Let $T_{i;x,y}$ denote the temperature of the rod at location $(x, y) \in \mathcal{X}$ at time $i \in [1, \dots, I]$. The temperature surface of all p rods at time i is denoted by $\mathbf{T}_i \in \mathbb{R}^p$. The sequential problem can be stated as follows:

$$\mathcal{H}_0 \qquad \mathbf{T}i \sim \mathcal{P}_{\boldsymbol{\theta}_i} \qquad \forall i \leq I \tag{1}$$

$$\mathcal{H}_{1,\nu} \quad \mathbf{T}i \sim \begin{cases} \mathcal{P}_{\boldsymbol{\theta}_i} & \forall i < \nu \\ \mathcal{P}_{\boldsymbol{\theta}_i} + \mathbf{a}_{i-\nu} & \forall i \in \{\nu, \dots I\}, \end{cases} \tag{2}$$

where $\mathcal{P}_{\boldsymbol{\theta}_i}$ is a known distribution parametrized by $\boldsymbol{\theta}_i$ describing the regular rod's temperature, and ν is the (unknown) time of change point. The quickest detection means that upon occurrence, hypothesis $\mathcal{H}_{1,\nu}$ must be detected with minimal delay $I - \nu$. The distribution parameter $\mathbf{a}_n$ represents the temperature drift after the TIB occurs and, of course, it changes over time.

We can already note that the detection problem is complex because hypothesis $\mathcal{H}_{1,\nu}$ is composite in the sense that at each time i there are many possible possibilities for the changepoint hence the "different alternative hypothesis" for the occurrence of the blockage.

A sequential detection scheme is a stopping rule S on $\mathbf{T}1, \mathbf{T}2, \dots, \mathbf{T}_i$ that indicates when to take a final decision. The stopping time t_S is the first time i at which the stopping rule S decides that the change point ν has been reached. False alarms occur if $t_S \leq \nu$, and the detection delay $t_S - \nu$ should be minimized.

Two criteria are widely used for sequential detection; On the one hand, a popular choice is the so-called Lorden criterion [21] which focusses on the worst possible detection delay

$$\sup_{\nu \in \mathbb{N}} \sup \mathbb{E}\left[t_S - \nu\right]. \tag{3}$$

Here $\mathbb{E}$ stands for the expectation operator. This quantity represents the worst-case expected value for the detection delay over all possible change-point locations ν and all observations of temperature distribution parameters $\boldsymbol{\theta}_i$. Focussing only on this worst possible detection delay is clearly a pessimistic approach.

Alternatively, a Bayesian approach was proposed in [34] considering that the change point ν is drawn from a known statistical distribution $\boldsymbol{\pi} : \pi_k = \mathbb{P}[\nu = k], k \in \mathbb{N}$. In this case, it has been proposed to minimize the average detection delay :

$$\mathbb{E}_\pi\left[t_S - \nu\right]. \tag{4}$$

In our case, neither the worst detection delay nor the average detection delay are a relevant criterion. Indeed, the change point in our problem is a deterministic but unknown value, and the goal is to minimize the probability that the change point is detected with a prescribed strict maximal delay. This constraint is dictated by the operational context: if the TIB is not detected within 6 seconds, the core begins to melt, causing irreversible damage. Conversely, it does not matter whether the change point is detected after one or five seconds, as long as it is identified before the maximum delay of six seconds, allowing for an emergency shutdown of the fission reaction without any damage.

In this operational context, our goal is to find a stopping rule such

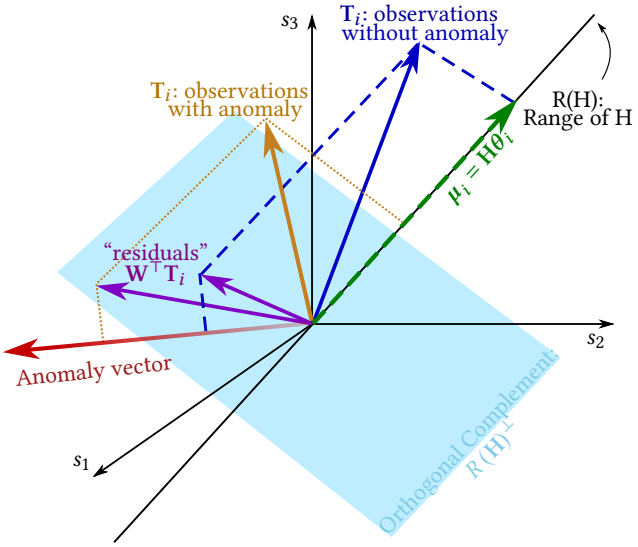


Figure 2: Illustration of the methodology to deal with linear nuisance parameters: the non-anomalous background lies within $R(H)$ which is spanned by the green vector ; the orthogonal complement is the blue surface onto which the observations are projected to subtract the normal temperature and obtain residuals (in purple). The statistical test consists in projecting the residual onto the red vector that defines the anomaly it is aimed at detecting.

that the worst probability of detecting a change point ν with maximal delay M is maximized. The power function of the stopping rule S is:

$$\beta_S^{(M)} = \inf_{\nu \in \mathbb{N}} \mathbb{P}[t_S < \nu + M | t_S \geq \nu]. \tag{5}$$

Additionally, the largest probability of false alarm over a run length of L observations is:

$$\alpha_S^{(L)} = \sup_{i \in \mathbb{N}} \mathbb{P}[t_S \in i - L + 1, \ldots, i | \nu > i]. \tag{6}$$

To this end we will use a sliding window likelihood ratio test (SW-LRT) which has been shown to be optimal for the worst possible power function (5), over all possible stopping time ν, under a prescribed mean time to false alarm (6), see details in [14, 39, 40] and the references therein.
Before presenting in more detail this sliding window likelihood ratio test detection procedure, we need to explain how to deal with the nuisance parameters that normal temperatures are.

3.2 Dealing with Nuisance Parameters

As already explained, the regular temperature of the rods, without any anomaly, is a nuisance parameters in the sense that they have no interest for the detection of an anomaly but it must be considered carefully in order to detect an anomaly from this "non-anomalous background".

Dealing with such nuisance parameters has always been an important research topic in the field of statistical detection, see for instance [10, 11] and the references therein.

A usual simple approach that is convenient as a basis for explanation is when the nuisance parameters is linear. More precisely, this model assumes that under normal operating conditions, the temperature at time index i of all the rods can be modelled as:

$$\mathbf{T}i \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I}_p\right), \tag{7}$$

where $\boldsymbol{\mu}_i$ is the expected value of all the rods' temperature at index time i, σ^2 is the variance of all the rods, and $\mathbf{I}_p$ is the identity matrix of size $p \times p$ with p the number of rods.
The linear nuisance parameters model consists in assuming that the average values of the temperature can be represented with a small set of basis vectors:

$$\mathbf{T}_i \sim \mathcal{N}\left(\mathbf{H}\boldsymbol{\theta}_i, \sigma^2 \mathbf{I}_p\right), \tag{8}$$

where $\mathbf{H}$ is a known full-column rank matrix of size $p \times q$ and $\boldsymbol{\theta} \in \mathbb{R}^q$ is a vector representing the q-dimension nuisance parameter. This model has been widely used in signal processing because it is simple and can be accurate enough in several applications. Under this model, which is depicted in the Figure 2, everything that falls within the column space spanned by $\mathbf{H}$, denoted $R(\mathbf{H})$ is considered as the nuisance parameters. Therefore, the rejection of the nuisance parameter can be carried out by simply projecting the vector of measured temperature $\mathbf{T}_i$ onto the orthogonal complement $R(\mathbf{H})^\perp$ of the column space $R(\mathbf{H})$. This projection is defined with the orthonormal matrix $\mathbf{W} = (w_1, \ldots, w_{p-q})$, where w_i are the eigenvectors of the projection matrix $\mathbf{P}_\mathbf{H}^\perp = \mathbf{I}_p - \mathbf{H}\left(\mathbf{H}^\top\mathbf{H}\right)^{-1}\mathbf{H}^\top$ corresponding to eigenvalues 1. The matrix $\mathbf{W}$ verifies the following properties:

$$\mathbf{W}^\top\mathbf{H} = \mathbf{0}, , \quad \mathbf{W}\mathbf{W}^\top = \mathbf{P}_\mathbf{H}^\perp, \quad \mathbf{W}^\top\mathbf{W} = \mathbf{I}_{p-q}. \tag{9}$$

The rejection of a linear nuisance parameter is represented in the Figure 2, it can be simply carried out as $\mathbf{W}^\top\mathbf{T}_i$ such that:

$$\mathbf{W}^\top\mathbf{T}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_{p-q}\right), \tag{10}$$

However, it is also well known that it is very limited in the case of a complex "non-anomalous background", such as the temperatures of the rods are [28].
Recently it was proposed to use a locally adaptive model which uses the last measurements in order to adjust the linear model ; such an approach has been used, for instance, for industrial inspection systems in [37, 38], and for network traffic monitoring in [5, 13, 26, 27].

In this paper, we propose an original adaptive model based on deep learning. Specifically, autoencoders (AEs) are well-known models with latent variables that are particularly effective for manifold learning, which involves representing data as a complex subspace. It consists of a deep learning unsupervised model trained to minimize the reconstruction error, which is given is the case of the usual binary cross entropy (BCE):

$$Loss = \sum_p \left[\mathbf{T}_{i,p} \log\left(\widehat{\mathbf{T}}_{i,p}\right) + (1 - \mathbf{T}_{i,p}) \log\left(1 - \widehat{\mathbf{T}}_{i,p}\right)\right], \tag{11}$$

where $\widehat{\mathbf{T}}_i = AE(\mathbf{T}_i)$ represents the reconstructed temperature by the Autoencoder AE.

In our case, we leveraged the knowledge on the monitoring system and especially the statistical properties of the noise defined as

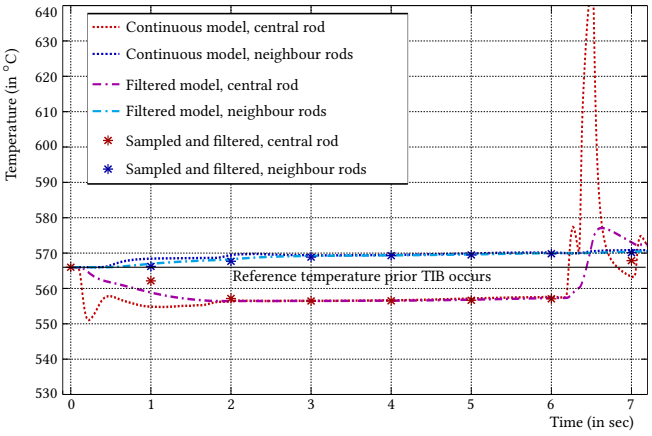


Figure 3: Illustration of the model of TIB contribution on both the central rod and its neighbours.

an uncorrelated Gaussian random field (7). To this end we proposed to use a Variational Autoencoder (VAE) in which the latent variable $\mathbf{z}$ is drawn from a random distribution, whose parameters depend on the input $\mathbf{T}_i$, and add to the loss function it is aimed at minimizing a constraint on the minimal distance between input and output distribution. Typically, with the usual ELBO (Expectation Lower BOund) this loss function is defined as

$$Loss = \mathbb{E}_{\mathbf{z}}\left[\log\left(p_\eta(\mathbf{T}_i|\mathbf{z})\right)\right] + D_{KL}\left(q(\mathbf{z}|\mathbf{T}_i)||p_\eta(\mathbf{z})\right), \qquad (12)$$

where, simply stated, $p_\eta(x)$ is the probability distribution of data point x by the model η and the expectation is computed over the latent variables z, drawn from an auxiliary distribution $q_\eta(zjx)$. The first term corresponds to the reconstruction error and the second ter the Kullbach-Leibler divergence used as a distance between input and output distribution, which, in our case, is known with fairly good accuracy.

After training a simple, yet efficient, VAE with a short sequence of input temperature, in order to learn the temperature model in time as well as in space, we subtract the reconstructed temperature. This is an original alternative to deal with complex nuisance parameters exploiting, on the one hand, powerful unsupervised learning methods and, on the other hand, leveraging the known statistical model of observation within the framework of hypothesis testing theory.

The methodology is illustrated in the Figure 2 except that with our original approach the residuals are obtained but subtracting the temperature estimated by the Variational Autoencoders instead of using a fixed linear projector.

3.3 Modelling Blockage Impact on Temperatures

In order to model the impact of a blockage-induced anomaly (TIB) on temperatures, we used numerical data measured at a Sodium Fast Reactor (SFR) station. The temperature measurements were obtained using K-type thermocouples, whose transfer function is well documented. The impulse response function of the thermocouples is given by:

$$h(t) = \frac{1}{\tau}\exp\left(\frac{-t}{\tau}\right), \qquad (13)$$

where τ is the time constant of the thermocouple, approximately 0.5-1 seconds in our setting. We used a time constant of $\tau = 1$ second for our simulations.

Due to the impracticality of creating a real TIB to measure its impact on rod temperatures, we relied on numerical simulations using the model of the ASTRID reactor. The simulations accounted for relevant thermodynamic parameters, such as materials, assembly geometry, and sodium cooling flow. However, the simulated temperatures did not exactly match the measured temperatures due to differences in sampling frequency and thermocouple response.

To transform the simulated temperatures into realistic observable temperatures, we applied the thermocouple response function, defined in (13) and subsampled the data by averaging temperatures over one second. This resulted in a model for the anomalous temperature due to the TIB, denoted as $\mathbf{a}$. Figure 3 illustrates the temperature after TIB occurs for both the rod at which the blockage happens and its neighbours.

The figure shows that the temperature at neighbours increases at a rather constant pace of about 0.6°C per second, while the temperature of the central rod drops suddenly after TIB, followed by a sudden increase of about 150°C after about 7 seconds, due to local vaporization of sodium. The difference between the simulated and observable temperatures is evident, especially for the central rod. We assumed a spatially invariant model of TIB impact, but the proposed methodology can be adapted to more accurate models by modifying the vector $\mathbf{a}$ according to inspected rod position and temperature.

All in all, with this model for the anomaly temperature it is aimed at detecting, the global decision function is given by:

$$\Lambda(\mathbf{T}_i) = \left(\mathbf{T}_i - \widehat{\mathbf{T}}_i\right)^\top \mathbf{a}, \qquad (14)$$

where $\widehat{\mathbf{T}}_i$ is the regular temperature estimated by the variational encoder model and $\mathbf{a}$ is set to the maximal admissible detection delay. That is, we focus on detection with a delay of the longest admissible duration to maximize the probability of detection before the time detection delay is exceeded, see details in [14, 39, 40] and the references therein.

4 AUTOENCODERS-BASED MODEL FOR DEALING WITH NUISANCE PARAMETER

We have designed an original mode of a variational autoencoder (VAE) whose architecture is depicted in Figure 4. First of all, one of the main originality is that the model accepts as input a series of 16 temperature measurements $\mathbf{T}_{i-31}, \ldots, \mathbf{T}_i$. The first convolution layer is made of 8 kernels of size 16×9 : the idea is on allows time modelling of the temperature thanks to this layer. However, the loss function is measured only the very last temperature $\mathbf{T}_i$.

Then the proposed variational autoencoder is makde of three convolution layers followed by two connection layers for modelling the temperature surface of the core of an experimental nuclear power plant. The decoder is made of the reciprocal three convolution layers to obtain a set of 8 tempertures of rods which are finally pooled in a last convolution layer with kernel size 1×1. This proposed model therefore output from a time series of the last 16 temperature a probabilistic representation of the temperature surface.

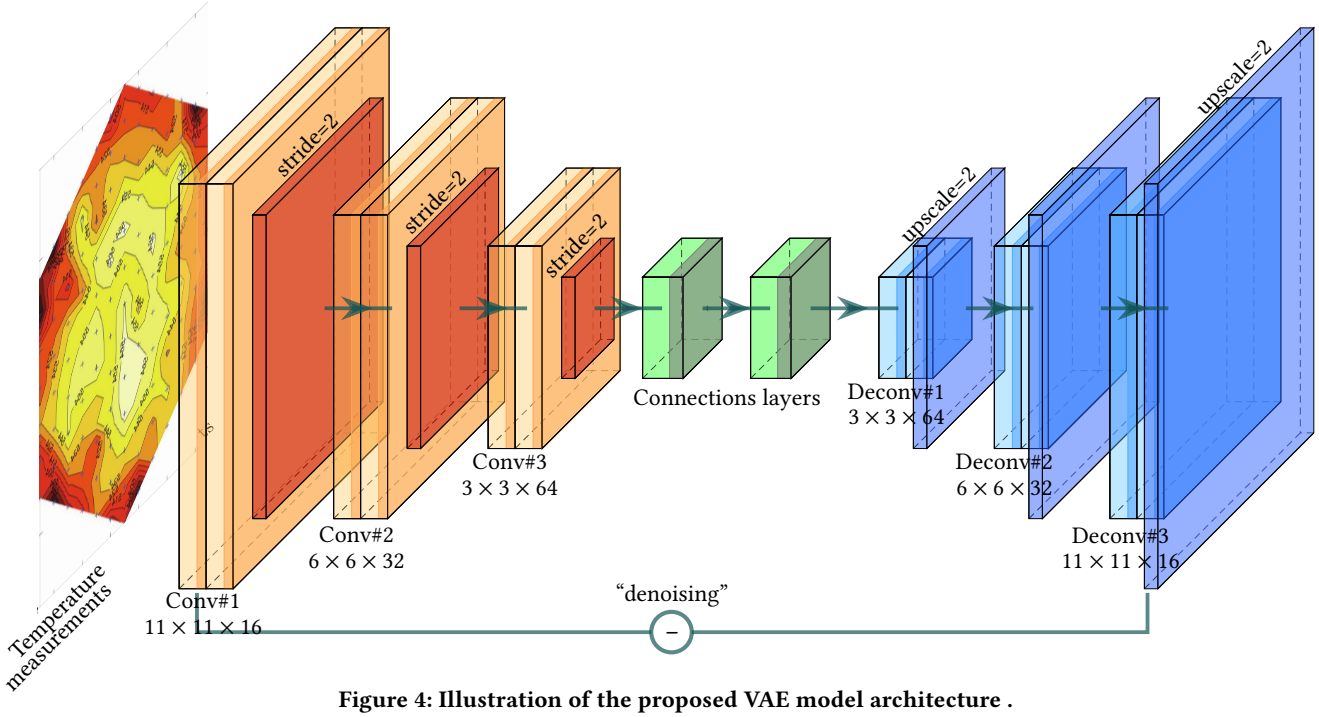


Figure 4: Illustration of the proposed VAE model architecture .

More precisely the the architecture of the encoder is the following, see also the Figure 4. After the initial convolution which generates 8 output of the same size as the original temperature measurement, that is 11×11. Then the first convolutional layer takes the 2D input temperatures and applies $3x3$ convolution with 16 filters, followed by a GELU. We applied a "same" padding of size 1 prior to the convolution and then using a stride size of 2 to reduce the size to 6×6 padding prior to the The output of this layer is a feature map with 16 which is given to the second layer with the same architecture: 3×3 convolution with 32 follwed by GELU activation and strided with stride 2. A last convolution with the same architectire except with 64 filters is applied. We thus eventually end up with 64 layers of size 2×2.

This resullts of the encoder is given to a first connection layer: the input is flattened and passed through a fully connected layer with 128 units, followed by a GELU activation function. A second connection layer with the same architecture of a function connected model with 128 units, followed by a GELU activation function. This constitutes our final latent space.

The VAE learns a probabilistic representation of the temperature surface via the encoder: its outputs a mean and variance for each dimension of the latent space, which are used to sample a latent vector. The decoder takes this latent vector and generates a reconstructed temperature surface.

The decoder consists of the reciprocal architecture as the encoder : it is made of three time the same convolution layer with kernel of size 3×3 together with outper padding of size 1 to increase the dimension. The number of filters is halved at each layer from 64 to 32, 16 and eventually 8 which are merged with a GELU activation function and pooling ; empirically the average pooling gave the best only at this layer, all other pooling are carried out using the maximal value.

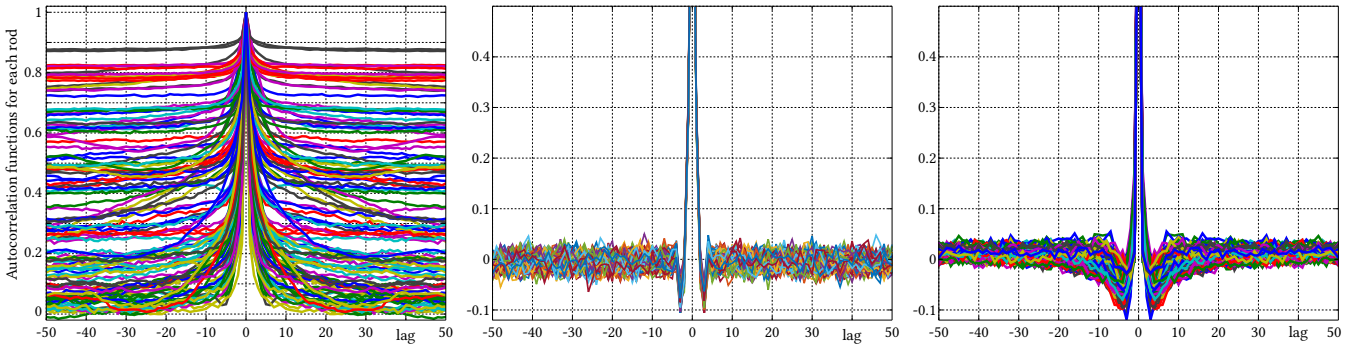
In the present paper we obtained best results using a loss function which is the sum of the mean squared error (the latest temperatures and the reconstructed values) and the Kullbach-Leibler divergence used to regularize the latent space and encourage the model to learn a meaningful representation of the temperature surface.

5 NUMERICAL RESULTS, SIMULATION AND EVALUATION

5.1 Common Core of All Experiments

The primary dataset analysed in this study originated from the French Commission of Atomic and Alternative Energies (CEA) and was obtained from the experimental French Phenix Nuclear Reactor. The dataset consisted of synchronized one-week time stamps from February 16^{th} to February 22^{nd} and from March 2^{nd} to March 09^{th} 2009, encompassing approximately $1,201,600$ temperature observations per rod, across 120 rods. Due to the reactor's construction date and experimental goals, a sampling frequency of 1 Hz was adopted to avoid excessive data accumulation over the years.

The proposed Variational Autoencoder model was implemented using the PyTorch deep learning framework and the FastAI library. We randomly split the samples into training (40%), validation (10%), and testing (50%) sets. This choice was motivated by the need for a large testing dataset to evaluate low false-alarm rates. During our simulation, the statistical test was carried out using a small subset of 3x3 temperature values (a central rod subject to blockage and its neighbours). We maximized the number of simulations by computing the detection statistics on all possible subsets of rods and for all possible measurement times. The statistical test was



(a) Autocorrelation function of all rods' residuals using the polynomial "fixed" model.

(b) Autocorrelation function of all rods' residuals using the proposed adaptive model.

(c) Autocorrelation function of all rods' residuals using the polynomial with ARMAX model [28].

Figure 5: Comparison of the autocorrelation functions, plotted as a function of the lag, obtained with different models.

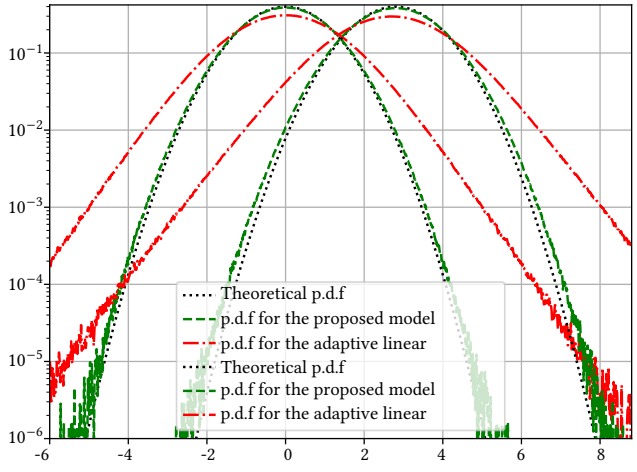


Figure 6: Theoretical and empirical distributions of the detection statistics (14)**. The results obtained with the proposed model are much closer to the theoretical results than those obtained with an adaptive linear model as proposed in [5, 38].**

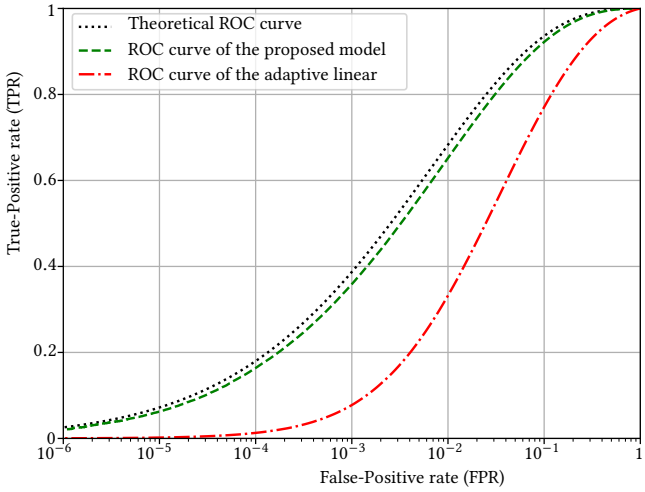


Figure 7: Performance of the detectors via a ROC curve plotting the true-positive rate (TPR) as a function of the false-positive rate (FPR). The theoretical results are based on measurements noise variance and the BTI signal computed as in Section 3.3.

applied randomly to all possible blockage locations at all possible times, resulting in approximately 35 million test data points under each hypothesis.

5.2 Assessement of Proposed Model Accuracy

To evaluate the proposed VAE's performance, we examined its ability to learn temperature representations and the characteristics of the resulting residual noise. Figure 5 presents the normalized autocorrelation function of all residual rod temperatures. On the left-hand side, the residuals obtained with a simple linear model based on a polynomial approximation of the temperature surface are shown. The residuals are heavily biased due to the model's lack of accuracy. In contrast, the results obtained using a simple Autoregressive-Moving Average model with exogenous inputs (ARMAX) model, as proposed in [28], where the input is the power produced by the core, show much better performance. The noise residuals are only loosely autocorrelated in time, demonstrating a great capability to remove the content. However, this model is not accurate enough, as will be seen in the next section 5.3. The autocorrelation obtained by subtracting the temperature reconstructed by the proposed VAE model is shown in the middle. The autocorrelation of the noise residuals is very small and quickly decreases to zero. Obtaining these results is challenging, and it is encouraging because it indicates that the proposed model can learn a sparse representation of the temperature, such that errors do not affect subsequent samples. This result is a necessary but insufficient condition for the statistical test, based on the projection of residuals onto the anomaly vector, to perform accurately. Indeed, the model can remove a large part of the anomaly if it is not sharp enough. In contrast, results such as those presented on the left-hand side, using a simple polynomial approximation, result in non-zero mean test statistics and biased detection statistics, which corrupt the decision.

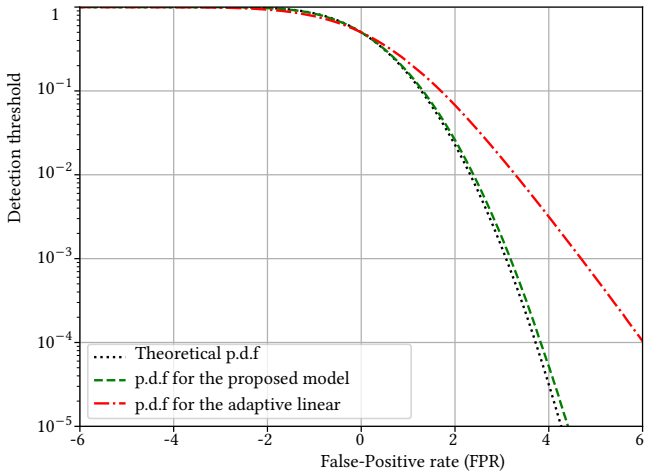


Figure 8: False-positive rate (FPR) as a function of the detection threshold ; comparison between the theoretical values, the empirical results obtained with the proposed hybrid autoencoders/transformers model and using a linear adaptive mode as proposed in [5, 38].

5.3 Detection performance

We now move to the evaluation of the performance of the proposed overall detection method. Figure 7 presents the ROC (Receiver Operating Characteristic) curve, showing the true-positive rate, also referred to as the sensitivity of the test, as a function of the false-positive rate. In addition to the performance of the proposed detection methodology based on the original VAE, the performance of an advanced auto-adaptive model is reported. This model, as proposed in [5, 38], uses the principal components of the last temperature measurement to improve the polynomial model, whose residuals were presented in Figure 5. In other words, this model takes advantage of both the linear model and the ARMAX model, which models temperature over time. While this model performs quite well, it is clear that the proposed method for subtracting the residuals using the original Variational Autoencoders performs much better. Lastly, Figure 8 shows the distribution of the false-positive rate as a function of the decision threshold. Interestingly, the plot presents a comparison with the theoretical false-alarm probability if one had a model that could perfectly remove the normal temperature. This curve is calculated using the noise of all thermocouples, which is known with a fairly good accuracy. It is noticeable that the proposed detection method leveraging the VAE is quite close to the theoretical false-positive rate, indicating that the subtraction of the non-abnormal temperature is nearly optimal. This capability remains true even for very low false-positive rates of about 10^{-5}. In contrast, the false-positive rates obtained with the adaptive model are good but are much further from the theoretical values. More importantly, this simpler model is not relevant for very low false-positive rates, which is the operational context, as a false-positive rate of only 10^{-3} would be hardly usable in real-life applications.

6 CONCLUSION AND FUTURE WORKS

This paper investigates the detection of Total and Instantaneous Blockage (TIB) of coolant flow around a fuel rod in a nuclear power plant core. This problem is particularly challenging because, when the coolant flow is obstructed, the measured temperature does not accurately reflect the true temperature of the rod due to the placement of the instrumentation. Additionally, the problem is statistically complex for two primary reasons: first, the temperature of the rods is complex and it fluctuates over time; second, the usual sequential detection criterion is not relevant, as it is necessary to detect the blockage with a maximal fixed delay.

To model the "normal" temperature variations in a nuclear power plant core, a simple yet efficient variational autoencoder is proposed. Numerical results demonstrate the accuracy of this adaptive model and its effectiveness for TIB detection. The detection of non-stationary temperature changes is addressed using a fixed size sliding likelihood ratio test. This approach matches the maximal delay criterion and has the significant advantage of providing bounds on its statistical performance, which is crucial in the industrial operational context of nuclear power plants.

Numerical results using real data validate the effectiveness of the proposed two Fixed Length Windows method for sequential TIB detection. Theoretical results also confirmes the reliability and sharpness of the proposed original methodology.

An important possible barrier to the practical application of the method proposed in the present paper lies in the lack of explainability of deep learning in general [6, 7]. This could limit the acceptability of users, as well as the occurrence of false alarms may undermine trust in the decision of the proposed automatic surveillance system.

Our future work will focus on integrating contrastive learning into the detection process, more specifically by incorporating the projection onto the anomaly during the learning phase.

REFERENCES

[1] 2011. *Experimental Facilities for Sodium Fast Reactor Safety Studies.* Technical Report. Task Group on Advanced Reactors Experimental Facilities (TAREF), Committee on the Safety of Nuclear Installations (CSNI), Nuclear Energy Agency (NEA). 144 pages.

[2] Erika N Bailey. 2011. *Independent Confirmatory Survey Summary and Results for the Enrico Fermi Atomic Power Plant, Unit 1 Newport, Michigan (Final Report No. 2).* Technical Report. Oak Ridge Inst. for Science and Education (ORISE), Oak Ridge, TN (United States).

[3] Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, François-Xavier Aubet, Laurent Callot, and Tim Januschowski. 2022. Deep Learning for Time Series Forecasting: Tutorial and Literature Survey. *ACM Comput. Surv.* 55, 6, Article 121 (Dec. 2022), 36 pages. https://doi.org/10.1145/3533382

[4] Kamal Berahmand, Fatemeh Daneshfar, Elaheh Sadat Salehi, Yuefeng Li, and Yue Xu. 2024. Autoencoders and their applications in machine learning: a survey. *Artificial Intelligence Review* 57, 2 (2024), 28.

[5] Rémi Cogranne, Guillaume Doyen, Nisrine Ghadban, and Badis Hammi. 2018. Detecting Botclouds at Large Scale: A Decentralized and Robust Detection Method for Multi-Tenant Virtualized Environments. *IEEE Transactions on Network and Service Management* 15, 1 (2018), 68–82. https://doi.org/10.1109/TNSM.2017.2785628

[6] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[7] Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M Ali. 2022. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences* 615 (2022), 238–292.

[8] Yifei Ding, Jichao Zhuang, Peng Ding, and Minping Jia. 2022. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliability Engineering & System Safety* 218 (2022), 108126. https:

//doi.org/10.1016/j.ress.2021.108126

[9] Hosein Fanai and Hossein Abbasimehr. 2023. A novel combined approach based on deep Autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications* 217 (2023), 119562. https://doi.org/10.1016/j.eswa.2023.119562

[10] Mitra Fouladirad and Igor Nikiforov. 2005. Optimal statistical fault detection with nuisance parameters. *Automatica* 41, 7 (2005), 1157–1171.

[11] Paul M Frank. 1990. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results. *Automatica* 26, 3 (1990), 459–474.

[12] John Grant Fuller. 1975. We almost lost Detroit. (1975).

[13] Nisrine Ghadban, Rémi Cogranne, and Guillaume Doyen. 2017. A decentralized approach for adaptive workload estimation in virtualized environments. In *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. IEEE, 1186–1194.

[14] Blaise Kévin Guépié. 2013. *Sequential Detection of Transient Changes: application to water distribution network monitoring* in french. Ph. D. Dissertation. Troyes University of Technology (UTT).

[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33, 4 (2019), 917–963.

[16] Robert KL Kennedy, Zahra Salekshahrezaee, Flavio Villanustre, and Taghi M Khoshgoftaar. 2023. Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. *Journal of Big Data* 10, 1 (2023), 106.

[17] Jina Kim, Hyeongwon Kang, and Pilsung Kang. 2023. Time-series anomaly detection with stacked Transformer representations and 1D convolutional network. *Engineering Applications of Artificial Intelligence* 120 (2023), 105964.

[18] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[19] Brian Lewandowski and Randy Paffenroth. 2022. Autoencoder feature residuals for network intrusion detection: Unsupervised pre-training for improved performance. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1334–1341.

[20] Ivandro O. Lopes, Deqing Zou, Ihsan H. Abdulqadder, Francis A. Ruambo, Bin Yuan, and Hai Jin. 2022. Effective network intrusion detection via representation learning: A Denoising AutoEncoder approach. *Computer Communications* 194 (2022), 55–65. https://doi.org/10.1016/j.comcom.2022.07.027

[21] G. Lorden. 1971. Procedures for Reacting to a Change in Distribution. *The Annals of Mathematical Statistics* 42, 6 (1971), 1897 – 1908. https://doi.org/10.1214/aoms/1177693055

[22] Mingrui Ma, Lansheng Han, and Chunjie Zhou. 2024. Research and application of Transformer based anomaly detection model: A literature review. *arXiv preprint arXiv:2402.08975* (2024).

[23] Serge Marguet. 2023. Reactor Accidents in the Early Days of Nuclear Power. In *A Brief History of Nuclear Reactor Accidents: From Leipzig to Fukushima*. Springer, 29–209.

[24] Sinuhe Martinez-Martinez, Nadhir Messai, Jean-Philippe Jeannot, and Danielle Nuzillard. 2015. Two neural network based strategies for the detection of a total instantaneous blockage of a sodium-cooled fast reactor. *Reliability Engineering & System Safety* 137 (2015), 50 – 57. https://doi.org/10.1016/j.ress.2014.12.003

[25] Harindra S. Mavikumbure, Chathurika S. Wickramasinghe, Daniel L. Marino, Victor Cobilean, and Milos Manic. 2022. Anomaly Detection in Critical-Infrastructures using Autoencoders: A Survey. In *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society*. 1–7. https://doi.org/10.1109/IECON49645.2022.9968505

[26] Tan Nguyen, Remi Cogranne, and Guillaume Doyen. 2015. An optimal statistical test for robust detection against interest flooding attacks in CCN. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. 252–260. https://doi.org/10.1109/INM.2015.7140299

[27] Tan Nguyen, Hoang-Long Mai, Rémi Cogranne, Guillaume Doyen, Wissam Mallouli, Luong Nguyen, Moustapha El Aoun, Edgardo Montes De Oca, and Olivier Festor. 2019. Reliable Detection of Interest Flooding Attack in Real Deployment of Named Data Networking. *IEEE Transactions on Information Forensics and Security* 14, 9 (Sept. 2019), 2470–2489. https://doi.org/10.1109/TIFS.2019.2899247

[28] Igor Nikiforov, Fouzi Harrou, Rémi Cogranne, Pierre Beauseroy, Edith Grall, Blaise Kevin Guépié, Lionel Fillatre, and Jean-Philippe Jeannot. 2020. Sequential detection of a total instantaneous blockage occurred in a single subassembly of a sodium-cooled fast reactor. *Nuclear Engineering and Design* 366 (2020), 110733. https://doi.org/10.1016/j.nucengdes.2020.110733

[29] ES Page. 1954. Continuous inspection schemes. *Biometrika* (1954), 100–115.

[30] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* 54, 2, Article 38 (March 2021), 38 pages. https://doi.org/10.1145/3439950

[31] Moschos Papananias, Thomas E McLeay, Mahdi Mahfouf, and Visakan Kadirkamanathan. 2023. A probabilistic framework for product health monitoring in multistage manufacturing using Unsupervised Artificial Neural Networks and Gaussian Processes. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 237, 9 (2023), 1295–1310. https://doi.org/10.1177/09544054221136510 arXiv:https://doi.org/10.1177/09544054221136510

[32] Igor L Pioro and Gilles H Rodriguez. 2023. Generation IV international forum (GIF). In *Handbook of Generation IV Nuclear Reactors*. Elsevier, 111–132.

[33] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. 2019. Anomaly Detection with Conditional Variational Autoencoders. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 1651–1657. https://doi.org/10.1109/ICMLA.2019.00270

[34] A. N. Shiryaev. 1963. On Optimum Methods in Quickest Detection Problems. *Theory of Probability & Its Applications* 8, 1 (1963), 22–46. https://doi.org/10.1137/1108002 arXiv:https://doi.org/10.1137/1108002

[35] Alexander Stanculescu. 2018. GIF R&D outlook for generation IV nuclear energy systems: 2018 update. In *Proceedings of the Generation IV International Forum, Paris, France*. 16–18.

[36] A. Tartakovsky, I. Nikiforov, and M. Basseville. 2014. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Taylor & Francis.

[37] Karim Tout, Rémi Cogranne, and Florent Retraint. 2016. Fully automatic detection of anomalies on wheels surface using an adaptive accurate model and hypothesis testing theory. In *2016 24th European Signal Processing Conference (EUSIPCO)*. 508–512. https://doi.org/10.1109/EUSIPCO.2016.7760300

[38] Karim Tout, Rémi Cogranne, and Florent Retraint. 2018. Statistical decision methods in the presence of linear nuisance parameters and despite imaging system heteroscedastic noise: Application to wheel surface inspection. *Signal Processing* 144 (2018), 430–443. https://doi.org/10.1016/j.sigpro.2017.10.030

[39] Karim Tout, Florent Retraint, and Rémi Cogranne. 2017. Wheels coating process monitoring in the presence of nuisance parameters using sequential change-point detection method. In *2017 25th European Signal Processing Conference (EUSIPCO)*. 196–200. https://doi.org/10.23919/EUSIPCO.2017.8081196

[40] Karim Tout, Florent Retraint, and Rémi Cogranne. 2018. Non-Stationary Process Monitoring for Change-Point Detection With Known Accuracy: Application to Wheels Coating Inspection. *IEEE Access* 6 (2018), 6709–6721. https://doi.org/10.1109/ACCESS.2018.2792838

[41] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. 2022. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284* (2022).

[42] A. Wald and J. Wolfowitz. 1948. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics* 19, 3 (1948), 326–339. http://www.jstor.org/stable/2235638

[43] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125* (2022).

[44] Jiehui Xu. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642* (2021).

[45] Fanyu Zeng, Mengdong Chen, Cheng Qian, Yanyang Wang, Yijun Zhou, and Wenzhong Tang. 2023. Multivariate time series anomaly detection with adversarial transformer architecture in the Internet of Things. *Future Generation Computer Systems* 144 (2023), 244–255.

[46] Hongwei Zhang, Yuanqing Xia, Tijin Yan, and Guiyang Liu. 2021. Unsupervised anomaly detection in multivariate time series through transformer-based variational autoencoder. In *2021 33rd Chinese Control and Decision Conference (CCDC)*. IEEE, 281–286.

submitted 05 Novembre 2024; revised 26 Novembre 2024