Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26228-cd

Probabilistic Ensemble of Statistical and Machine Learning Techniques for the Detection of Abnormal Conditions in Safety-Critical Systems: Application to the Simulator of an Advanced Nuclear Reactor

Nicola Pedroni

Department of Energy, Politecnico di Torino, Italy. E-mail: nicola.pedroni@polito.it

The paper proposes a framework for the robust identification of anomalies (hereafter also called outliers) in the behavior of dynamic safety-critical systems, which is based on an ensemble of statistical and machine learning techniques, relying on three diverse principles: (i) proximity-based algorithms, defining outliers as those data points lying in low-density regions or considerably far from their neighbors; (ii) statistical models, capturing the underlying probabilistic structure of the data and identify the outliers as those observations characterized by the smallest likelihood; (iii) (deep) learning-based techniques, trained to reconstruct multivariate data and spot out abnormal samples as those characterized by the largest reconstruction error. Each algorithm of the ensemble is tailored to compute anomaly scores for all the available observations and quantify their "degree of outlyingness". These sets of anomaly scores are finally combined within an original probabilistic framework (based on cumulative distribution functions) to get a unique, robust and reliable ranking. The developed approach is tested on transients representing the time evolution of several physical parameters (fluid temperatures) taken from the simulator of a Generation-IV nuclear reactor. The results show that the ensemble improves the anomaly detection performance with respect to the individual detectors, with accuracy, precision and recall values of 97.24%, 96.94% and 98.77%, respectively.

Keywords: Anomaly detection, Advanced Nuclear Reactor, Machine Learning, Probabilistic Ensemble, Outlyingness scores, Cumulative distribution, Proximity-based algorithms, Deep learning, Statistical models.

1. Introduction

In the operation and control of safety-critical systems, operators must promptly recognize abnormal system conditions and identify their underlying causes by analyzing the temporal evolution of continuously monitored signals. This capability is essential for the early detection of behavioral deviations (hereafter also called "anomalies", "abnormalities" or – from a statistical perspective – "outliers") that may otherwise escalate into severe accident scenarios. However, this task is inherently challenging due to the typically high *dimensionality* and large *quantity* of transient monitoring signals and to the *complex interdependencies* among system variables. In addition, in real engineering applications the choice of an optimal anomaly detection approach is difficult, due to several theoretical, technical and practical issues (Aggarwal, 2017; Rollon de Pinedo et al., 2021): (a) the lack of a *rigorous* and *quantitative definition* of "anomalies"; (b) the lack of a *framework* based on *solid mathematical grounds* for their detection; (c) how to ensure that the chosen measure (or measures) that quantify this notions can be universally accepted as such; (d) due to the intrinsic *low probability* of occurrence of an anomaly (rare by definition), it is not trivial to evaluate the detection capabilities of any given algorithm over a wide variety of situations; (e) all available detection methods are *sensitive* to *hyperparameter* settings and fine tuning; (f) their performance can depend on the "physical origin" of the anomaly, and on the nature and spatial distribution of the data points (i.e., depending on the type, sparsity and homogeneity of the dataset). Due to such difficulties, in practice, *no single* detection algorithm can universally apply to *all* datasets. Also, since each anomaly detection algorithm presents *peculiar strengths* and *weaknesses* and relies on its *own definition* of how outliers differ from normal points, selecting a detector often requires some level of intuition or knowledge on the nature of the dataset (Ouyang et al., 2021).

In this respect, the present paper proposes a framework for the robust identification of anomalies in the dynamic behavior of safety-critical systems. The proposed approach is based on an *ensemble* of seven statistical and machine learning techniques grounded in three *diverse principles*: (i) *proximity-based methods*, which characterize outliers as observations located in low-density regions or at large distances from their neighbors (Local Outlier Factor, Isolation Forest); (ii) *statistical models*, which capture the underlying probabilistic structure of the data and identify outliers as observations associated with low likelihood values (Finite Mixture Models, Sliced Normal Distributions); and (iii) *(deep) learning-based techniques*, which are trained learn the inherent patterns and distribution of regular system behavior. On this basis, some algorithms reconstruct multivariate data (from the system) and detect abnormal samples as those showing the largest reconstruction errors (Autoencoders, Auto-Associative Kernel Regression). Other methods (One-Class Support Vector Machines) construct a tight-fitting hyperplane that encapsulates all the normal data points, and flag as anomalous any new data falling outside this established boundary. Each method in the ensemble independently assigns an *anomaly score* to all available observations, thereby quantifying their degree of *outlyingness*. These scores are subsequently aggregated within an original probabilistic framework based on *cumulative distribution functions* to produce a single, robust, and reliable anomaly ranking.

The developed approach is tested on transients representing the time evolution of several physical parameters (fluid temperatures) taken from the simulator of a Molten Salt Fast Reactor (MSFR), a Generation IV concept conceived during a series of research projects (EVOL, SAMOFAR, SAMOSAFER…) financially supported by the European Union in the last fifteen years (Tripodo et al., 2019).

2. Case study: Generation-IV nuclear reactor

A Molten Salt Fast Reactor (MSFR) design is selected, which corresponds to a 3000 MWth reactor housed in a cylindrical vessel approximately 2.25 m in both diameter and height. The vessel, fabricated from a nickel-based alloy, contains about 18 m^3 of molten salt, which serves simultaneously as fuel and coolant. The reactor operates at near-atmospheric pressure with an average temperature of 750 °C. Driven by the primary circulation pumps, the molten salt flows upward through the central core region and downward through circumferentially arranged heat exchangers. The fuel salt composition is $LiF–ThF_4–^{233}U$, which has a melting temperature of approximately 585 °C (858 K). The intermediate (secondary) circuit employs $NaF–NaBF_4$ as the working fluid, with a melting temperature of 384 °C, while helium is used as the heat transfer medium in the gas circuit dedicated to power conversion.

2.1. *The MSFR power plant simulator*

This study is based on publicly available time-series data generated by a power plant simulator developed within the SAMOFAR project (Abrate et al., 2026). The simulator is a control-oriented plant dynamics tool implemented in the open-source, object-oriented Modelica language (Tripodo et al., 2019). The principal modeling assumptions underlying the simulator are as follows: (i) thermal-hydraulic phenomena are described using a one-dimensional formulation with volume-averaged variables; (ii) the transport of delayed neutron precursors is treated within a one-dimensional framework; and (iii) neutron kinetics are represented using the point kinetics approximation. The latter assumption is not expected to significantly affect the accuracy of the results, as the core is homogeneous and optically small due to the relatively large mean free path characteristic of fast neutrons.

2.2. *Main plant and controlled parameters*

The identification of a suitable set of Main Plant Parameters (MPPs) for the MSFR is a key prerequisite for the development of an effective anomaly detection algorithm. Each MPP can be assigned to one or more *functional categories*, noting that such classification may depend on the plant's operational state. *Controlled parameters* are quantities that are maintained within predefined thresholds through control and protection strategies implemented via dedicated functions. *Control parameters* are quantities that can be directly manipulated by the control system and whose variations influence one or more controlled parameters. *Monitored parameters* (possibly including control and controlled variables) are quantities that can be

measured either directly or indirectly and provide a real-time representation of the reactor state. Building upon previous investigations within the SAMOFAR and SAMOSAFER projects (Tripodo et al., 2019), abnormal scenarios are assumed to be primarily driven by variations in the *control parameters*, as these quantities are directly or indirectly caused by anomalies affecting plant components. Assuming operation under nominal "reactor in power" conditions - namely, with thermal power between 20% and 100% of the rated value and the balance of plant connected to the electrical grid - the following control parameters are considered: (i) the mass flow rate of the fuel (primary) circuit, $\dot{g}_{FC}$; (ii) the mass flow rate of the intermediate (secondary) circuit, $\dot{g}_{IC}$; (iii) the mass flow rate of the gas in the energy conversion (gas) circuit, $\dot{g}_{GC}$. The primary criterion for selecting these control parameters is their direct measurability and controllability through the plant circulation pumps. Owing to this coupling, the pumps constitute safety-critical components and abnormal conditions can be represented by perturbations of the corresponding control variables. Since the mass flow rates significantly affect the operating temperatures of the fuel and intermediate circuits, these temperatures are selected as *monitored (controlled) parameters* for training the anomaly detection ensemble, whose objective is the identification of potential pump failures during "reactor in power" operation. A representation of the three MSFR circuits considered, together with the corresponding observed parameters, is provided in Figure 1.

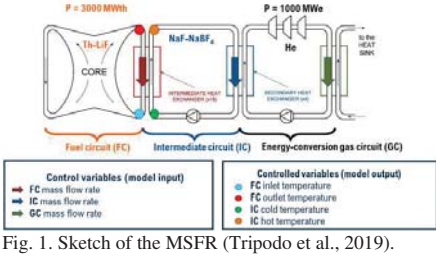

Fig. 1. Sketch of the MSFR (Tripodo et al., 2019).

The admissible temperature range is defined as follows: the peak temperature must remain below the maximum allowable temperature for containment structures, T_{max} = 1373 K, while the minimum temperature must exceed the highest freezing temperature of the primary and secondary salts, T_{min} = 858 K.

2.3. *Exploration of the MSFR parameter space and database generation*

Each system-level simulation is performed under *free-dynamics* conditions, with the objective of analyzing the evolution of the MPPs driven solely by the intrinsic feedback mechanisms of the MSFR. Simulations are initiated from steady-state nominal operating conditions. At t = 50 s, the mass flow rates in the fuel (FC), intermediate (IC) and gas (GC) circuits ($\dot{g}_{FC}$, $\dot{g}_{IC}$ and $\dot{g}_{GC}$) are perturbed, following an exponential decay from their nominal values to circuit-specific reduced values. Each scenario is simulated over a total duration of 800 s (with discretization step of 1 s, thus generating N_{time} = 800 data points for each transient). Each mass flow rate is varied over the interval [−90%, 0%] with respect to its nominal value, which encompasses both moderate and severe changes in the monitored parameters, corresponding to plant states that cannot be classified as minor deviations from nominal operation. With these settings, a dataset is generated by sampling the input parameter space, executing the MSFR simulator, and extracting the time-dependent monitored output variables of interest: a uniform sampling strategy is adopted to ensure adequate coverage of the parameter space. In general, a simulation is labeled as *safe* (or *normal*), if the system successfully fulfills its operational objective - namely, power generation within the prescribed safety margins - over the entire duration of the transient and *all* the salt temperatures remain within the prescribed safety limits $[858\ K, 1373\ K]$. Conversely, a scenario is classified as *unsafe* (or *abnormal*, *anomalous*) if, at *any time* during the transient, *at least one* of the four monitored temperatures $\boldsymbol{T}(t) = \{T_i(t)\colon i = 1, 2, 3, 4;\ t = 1, 2, \ldots, N_{time}\} = \{T_i(t)\colon i = FC_{cold}, FC_{hot}, IC_{cold}, IC_{hot};\ t = 1, 2, \ldots, N_{time}\}$ violates the admissible range: $T_i(t) \notin [T_{min}, T_{max}] = [858\ K, 1373\ K]$. In the present study, N_{train} = 3487 normal (safe) transients are generated to train the anomaly detection model, $D_{train} = \{\boldsymbol{T}_k^{train}(t)\colon k = 1, 2, \ldots, N_{train};\ t = 1, 2, \ldots, N_{time}\} = \{T_{i,k}^{train}(t)\colon i = 1, 2, 3, 4;\ k = 1, 2, \ldots, N_{train};\ t = 1, 2, \ldots, N_{time}\}$. Instead, N_{test} = 1748 scenarios (equally split between normal and anomalous)

are selected for testing the algorithm, $D_{test} = \{\boldsymbol{T}_k^{test}(t)\text{: } k = 1, 2, \ldots, N_{test}\}$.

3. The Method: Probabilistic Ensemble for the Detection of Abnormal Conditions

A *data-driven* approach is proposed (relying on the input-output dataset produced by the *physical-mathematical* computer *model* of the MSFR of Section 2) to effectively detect *anomalies* (pump failures) in the reactor behavior based on the time evolution of the four monitored salt temperatures $\boldsymbol{T}(t) = \{T_i(t)\text{: } i = 1, 2, 3, 4\} = \{T_i(t)\text{: } i = FC_{cold}, FC_{hot}, IC_{cold}, IC_{hot}\}$. As for many fault detection algorithms, the idea is to compare the *real* behavior of the system (e.g., the measured signal values) *during operation* with the one estimated (e.g., reconstructed) by a *model* as if the system were in *normal* operating conditions. Such a model is here built using the training dataset D_{train} containing $N_{train} = 3478$ "safe" scenarios (Section 2.3). The test set D_{test} serves the purpose of the "real" measured (observed) behavior of the system during operation. The "discrepancy" between *observed* and *modeled* (normal) values reveal the presence of anomalous conditions.

Three steps are foreseen: (1) *reduce* the *dimensionality* of the training time series; (2) *build* the model of the *normal* system behavior using D_{train}; (3) *test* the algorithm using D_{test}.

With respect to Step (1), the training time series $\{T_{i,k}^{train}(t)\text{: } i = 1, 2, 3, 4; k = 1, 2, \ldots, N_{train}; t = 0, 1, 2, \ldots, N_{time}\}$ are transformed into a lower dimensional representation, say $\{s_{i,k}^{train}(p_i)\text{: } i = 1, 2, 3, 4; k = 1, 2, \ldots, N_{train}; p_i = 1, 2, \ldots, N_{s,i}^{train}\}$, using properly selected functions $\boldsymbol{S}_i(\cdot)$: $R^{N_{time}} \rightarrow R^{N_{s,i}^{train}}$ ($N_{s,i}^{train} << N_{time}$). In this paper, Stacked Sparse AutoEncoders (SSAEs) (Zhao et al., 2019) are employed for dimensionality reduction, due to their proven ability to deal with complex and nonlinear data (see Section 3.1). The idea underlying this pre-processing step is to perform the subsequent anomaly detection task in the "reduced" space (defined by the SSAE transformation functions $\boldsymbol{S}_i$) rather than in the (dynamic multivariate) time domain. This is required by some algorithms of the ensemble, which cannot handle time series and/or severely suffer of the curse of dimensionality (see Sections 3.1-3.3).

With respect to Step (2), the model of the MSFR in normal conditions is here represented by an *ensemble* of several statistical and artificial intelligence techniques, relying on *diverse* principles. In all generality, consider a group of N^{AD} (different) anomaly detection algorithms, namely AD_1, AD_2, …, AD_l, …, $AD_{N^{AD}}$, and a (multidimensional) test datum $\boldsymbol{x}^{test}$ to classify as normal or anomalous. Notice that, depending on the *nature* of the detector (see Sections 3.1-3.3), the generic test datum $\boldsymbol{x}^{test}$ may be represented by: (i) the four time-dependent monitored salt temperatures $\boldsymbol{T}^{test} = [T_1^{test}(t), T_2^{test}(t), T_3^{test}(t), T_4^{test}(t)]$, $t = 1, 2, \ldots, N_{time}$; or (ii) the projection of the four temperature time series on the SSAE reduced feature space $\boldsymbol{s}^{test} = [s_1^{test}(p_1), s_2^{test}(p_2), s_3^{test}(p_3), s_4^{test}(p_4)]$, $p_i = 1, 2, \ldots, N_{s,i}^{test}$. Each algorithm AD_l, $l = 1, 2, \ldots, N^{AD}$, produces a metric $m(\boldsymbol{x}^{test})^{AD_l}$ that is used to quantitatively evaluate the degree of abnormality of (each) $\boldsymbol{x}^{test}$ (usually, the larger $m(\boldsymbol{x}^{test})^{AD_l}$ the larger the dissimilarity of $\boldsymbol{x}^{test}$ with respect to the bulk of the normal observations). Due to the possibly broad and diverse nature and taxonomy of the anomalies, the performance of the available detection methods largely depends on the "*physical origin*" of the anomaly itself, on the *nature* and *spatial distribution* of the data points (i.e., on the type, sparsity and homogeneity of the dataset). In practice, *no single* detection algorithm can universally apply to *all* datasets. In this view, adopting several *diverse* detection algorithms within an *ensemble* framework would produce more *robust* and *reliable* results. The underlying idea is that some algorithms will perform well on a particular subset of points (i.e., the abnormalities that are the most "obvious" from their respective standpoints), whereas other algorithms will do better on other subsets of points. By so doing, the ensemble combination is often able to obtain better results because of its ability to combine the outputs of multiple algorithms and to "correct" errors committed by the single, diverse members. For most approaches, however, the anomaly score $m(\boldsymbol{x}^{test})^{AD_l}$ (when available) is not easily interpretable. The scores provided by varying methods differ widely in their *scale*, *range* and *meaning*. Also, for many methods, the scaling of occurring values of the anomaly score even differs within the same method from dataset to dataset. In many cases, even within one dataset,

an identical score for two different objects can denote substantially different degrees of outlierness, depending on different *local data distributions*. Most importantly, opposite decisions on the outlierness of single objects may be equally meaningful since there is *no* generally *valid definition* of what constitutes an outlier in the first place. For different applications, a different selection of anomaly detection approaches may be meaningful. Thus, these methods cannot usually translate their scoring result into a label: there is no clear decision boundary that is generally sufficient to designate any data point as an anomaly. Obviously, this makes the *interpretation*, *comparison* and eventually *aggregation* of different detection models a very difficult task.

To address these issues, a *normalization* method is employed for a range of different models to become independent from the specific data distribution in a given dataset, as well as a *statistically sound* motivation for mapping the scores into the range [0, 1], readily interpretable as the *probability* of a given database object of being an anomaly (Kriegel et al., 2011). The *unified scaling* and *better comparability* of different methods (within a common *probabilistic framework*) can facilitate the construction of the ensemble. The basic idea underlying the method is the following: since we do not have a direct, unique and unambiguous way to interpret the anomaly metrics $m(\boldsymbol{x}^{test})^{AD_l}$, l = 1, 2, …, N^{AD}, we evaluate – for *each* detector – how "unusual" the score of a point is, using all the detector's output scores as *one-dimensional* projections of the original *multi-dimensional* data (i.e., $\boldsymbol{x}^{test} = \boldsymbol{T}^{test}$ or $\boldsymbol{s}^{test}$). By so doing, we do *not* draw any assumptions on the distribution of the original data, but we work only on the distribution of the anomaly scores produced. Also, we are not forced to grasp (analytically or empirically) the distribution of the anomaly scores $m(\boldsymbol{x}^{test})^{AD_l}$, l = 1, 2, …, N^{AD}, in the (very high-dimensional) original space of the data points to be classified. Any (cumulative) distribution function can be used for this purpose, depending on how well it fits to the distribution of the anomaly scores derived. In this work, we employ a nonparametric bootstrapped Kernel Density Estimation approach with reflection as boundary correction method and optimal bandwidth to fit the available scores and avoid making any assumption on their (cumulative) distribution. The training datapoints $\{\boldsymbol{T}_k^{train}(t)$: k = 1, 2, …, N_{train}; t = 1, 2, …, $N_{tiime}\}$ (or their "reduced" counterparts $\{\boldsymbol{s}_k^{train}$: k = 1, 2, …, $N_{train}\}$), representing the behavior of the system in normal (safe) conditions, are used to fit the empirical cumulative distribution functions (CDFs) $\hat{F}^{AD_l}(\cdot)$ of the anomaly metrics $m\left(\boldsymbol{T}_k^{train}\right)^{AD_l}$ (or $m\left(\boldsymbol{s}_k^{train}\right)^{AD_l}$) produced by the single detectors AD_l, l = 1, 2, …, N^{AD}. The resulting group of empirical CDFs $\hat{F}^{AD_l}(\cdot)$, l = 1, 2, …, N^{AD}, represents a *nonparametric probabilistic model* of the normal behavior of the system, where: (i) the (output) time series are first mapped to the one-dimensional anomaly metrics produced by the single detectors; (ii) the heterogeneous distributions of such metrics are then *normalized* and made *homogeneous* within a *statistical* framework based on CDFs.

Finally, with respect to Step (3), for a new (test) datum $\boldsymbol{x}^{test}$ the normalized version $\sigma(\boldsymbol{x}^{test})^{AD_l}$ of the anomaly score $m(\boldsymbol{x}^{test})^{AD_l}$ produced by detector AD_l, l = 1, 2, …, N^{AD}, is obtained as the corresponding empirical CDF value $\hat{F}^{AD_l}(m(\boldsymbol{x}^{test})^{AD_l})$. The multiple, diverse normalized probabilistic scores $\sigma(\boldsymbol{x}^{test})^{AD_l}$ are then combined (by means of a proper *aggregation* function $C(\cdot)$) to obtain a single score $\sigma(\boldsymbol{x}^{test}) = C(\sigma(\boldsymbol{x}^{test})^{AD_1}, \sigma(\boldsymbol{x}^{test})^{AD_2}, \ldots \sigma(\boldsymbol{x}^{test})^{AD_l}, \ldots, \sigma(\boldsymbol{x}^{test})^{AD_{N^{AD}}})$. Several aggregation functions $C(\cdot)$ have been proposed in the literature: see (Aggarwal, 2013; Ouyang et al., 2021). In this work, maximization (i.e., simple combination taking the maximum scores over the diverse detectors) is found to provide the most satisfactory results: $\sigma(\boldsymbol{x}^{test}) = max(\sigma(\boldsymbol{x}^{test})^{AD_1}, \sigma(\boldsymbol{x}^{test})^{AD_2}, \ldots \sigma(\boldsymbol{x}^{test})^{AD_l}, \ldots, \sigma(\boldsymbol{x}^{test})^{AD_{N^{AD}}})$. This normalized score can be interpreted as the *probability* of being an anomaly. In this framework, $\boldsymbol{x}^{test}$ is finally identified as an anomaly, if $\sigma(\boldsymbol{x}^{test})$ exceeds a threshold value α (here, 0.999) selected: (i) by the analyst according to the "level of risk" he/she is willing to accept in the detection process; or (ii) by a rigorous procedure aimed at finding over a validation set the optimal trade-off between (conflicting) figures of merit (e.g., accuracy, false positive and false negative rates).

The seven algorithms adopted to build the ensemble are briefly summarized hereafter for

space limitations, classified as follows: learning-based (Section 3.1), proximity-based (Section 3.2) and statistical (Section 3.3) ones.

3.1. *Learning-based techniques*

These models are trained to learn multivariate data representing the system *regular* behavior: on this basis, they screen abnormalities from normal points. (Stacked Sparse) Autoencoder (SSAE), Auto-Associative Kernel Regression (AAKR) and One-Class Support Vector Machine (OCSVM) algorithms are here considered.

An AE is a type of Artificial Neural Network (ANN) that is designed to learn new *low-dimensional* latent features of the data by trying to reconstruct the input data. It is composed of an *encoder* network, that maps N_{time}-dimensional time series $\{T_{i,k}^{train}(t)$: $i = 1, 2, 3, 4$; $k = 1, 2, \ldots, N_{train}$; $t = 1, 2, \ldots, N_{time}\}$ into a lower-dimensional (namely, hidden) feature space; and a *decoder* network that transforms the hidden representation back into $\hat{T}_{i,k}^{train}(t)$ (i.e., the *reconstruction* of $T_{i,k}^{train}(t)$) (Zhao et al., 2019).

To improve the representational and modelling power, multiple pretrained AEs can be stacked to form a multiple-hidden-layer ANN, called SSAE. Since training non-linear AEs with multiple hidden layers is difficult, a breakthrough approach is adopted, which consists of a *pre-training* (consecutive training of multiple basic AEs that are finally stacked) and a *fine-tuning* phase (the SSAE obtained from the pre-training is fine-tuned using the classical ANN backpropagation of error derivatives). Initially, the first basic AE is trained using the time series $T_{i,k}^{train}(t)$; then, the corresponding extracted features $\boldsymbol{s}_{i,k,1}^{train}$ (of dimension $N_{1,i}^{train}$) are used as training input vectors for the next (second) basic AE, which transforms $\boldsymbol{s}_{i,k,1}^{train}$ into $\boldsymbol{s}_{i,k,2}^{train}$ (of dimension $N_{2,i}^{train} < N_{1,i}^{train}$). This step is repeated up to the last (S-th) AE, which is trained to deliver the hidden features $\boldsymbol{s}_{i,k,S}^{train}$ (of dimension $N_{s,i}^{train}$). Then, the SSAE is built by stacking all the basic SAEs. The *encoder* part of the SSAE thereby trained is used to carry out the dimensionality reduction from $T_{i,k}^{train}(t)$ to the features $\boldsymbol{s}_{i,k,S}^{train}$ of dimension $N_{s,i}^{train}$ (referred to as $\boldsymbol{s}_{i,k}^{train}$ for the sake of compact notation); the *decoder* translates $\boldsymbol{s}_{i,k}^{train}$ back to the reconstruction $\hat{T}_{i,k}^{train}(t)$. The SSAE model is trained exclusively on data representing normal operating conditions, so it learns to reconstruct such data with low Root Mean Square Errors (RMSEs), used as anomaly metrics $m\left(\boldsymbol{T}_k^{train}\right)^{SSAE} = RMSE_k^{train}$, $k = 1, 2, \ldots, N_{train}$. When presented with an anomalous (test) sample $\boldsymbol{T}^{test}$ that deviates from the learned normal patterns, the SSAE fails to reconstruct it accurately, resulting in a significantly higher RMSE ($m(\boldsymbol{T}^{test})^{SSAE} = RMSE^{test}$).

Similarly, AAKR is an unsupervised, nonparametric approach that, given a training dataset of normal observations (time series), reconstructs a new (test) sample as a kernel-weighted average (where kernels – here Gaussian – are employed to quantify the *similarity* between samples). As for SSAEs, $m(\boldsymbol{T}^{test})^{AAKR} = RMSE^{test}$ (Jung et al., 2025).

Instead, OCSVM learns an *optimal* decision boundary (i.e., a multivariate function or hyperplane) that encloses the region of the parameter space corresponding to normal data. After training, OCSVM assigns anomaly scores $m(\boldsymbol{T}^{test})^{OCSVM}$ related to the *distance* of a point $\boldsymbol{T}^{test}$ from the learned hyperplane: points with higher anomaly scores are considered more likely to be anomalies (Scholkopf et al., 2001).

3.2. *Proximity-based algorithms*

They define anomalies as those data points lying in *low-density regions* or considerably *far* from their *neighbors*. In this paper, the Local Outlier Factor (LOF) and the Isolation Forest (IF) techniques have been adopted. These algorithms are fed with the projection of the training (time series) data onto the reduced SSAE feature space, i.e., $\{\boldsymbol{s}_k^{train}$: $k = 1, 2, \ldots, N_{train}\}$.

The LOF algorithm identifies anomalies by measuring the *local deviation* of a given data point with respect to its neighbors (Breunig et al., 2000). In synthesis, LOF is based on a concept of a *local (reachability) density*, where "locality" is given by the analysis of k nearest neighbors, whose distance is used to estimate the density. Then, the $LOF(\boldsymbol{s}^{test})$ for datum $\boldsymbol{s}^{test}$ (used as anomaly metric $m(\boldsymbol{s}^{test})^{LOF}$) is computed as the ratio between the average local reachability density of the k neighbors of $\boldsymbol{s}^{test}$ and the test datum's own local reachability density.

Isolation Forest (IF) (Liu et al., 2008) detects outliers by *isolating anomalies* from normal

points through a fast and approximate data *density* estimation carried out by an ensemble of binary trees, each one trained for a subset of the available data, sampled without replacement: points with a low-density estimate are considered anomalies. IF *recursively* splits the parameter space using lines that are parallel to the standard basis of the parameter space and assigns higher anomaly scores $m(\boldsymbol{s}^{test})^{IF}$ to data points that need fewer splits to be isolated: actually, the premise of the IF algorithm is that anomalous data points are easier to separate from the rest of the sample.

3.3. *Statistical models*

They capture the underlying probabilistic structure of the observations and identify the anomalies as those characterized by the *smallest likelihood*. Finite Mixture Models (FMMs) and Sliced Normal (SN) distributions are here adopted. Also these algorithms are fed with reduced training data $\{\boldsymbol{s}_k^{train}: k = 1, 2, \ldots, N_{train}\}$. FMMs are a flexible modeling tool for multivariate data, providing a formal approach to unsupervised learning. FMMs analyze a set of data; assume it is generated by a linear combination of known random models (Gaussian in this case), also called components; then, they infer the components' parameters to identify the distribution that originated the data. In this paper, the *iterative* SNOB algorithm is adopted to *automatically* determine the number of components, their relative weights and parameters simultaneously, while obtaining a trade-off between the model complexity and the goodness of fit (Rollon de Pinedo et al., 2022).
SN distributions are instead a generalization of Gaussian distributions where the quadratic argument of the exponential is replaced with a sum of squares polynomial. They can characterize the uncertainty in complex multivariate (multi-modal, non-symmetric and skewed) datasets in terms of semi-algebraic, nested *sets*, which allows *tightly enclosing* the (normal) data (Crespo et al., 2019). FMMs and SNs are used to fit a PDF $f_s(\boldsymbol{s})$ ($f_s^{FMM}(\boldsymbol{s})$ and $f_s^{SN}(\boldsymbol{s})$, respectively) on the training data $\{\boldsymbol{s}_k^{train}: k = 1, 2, \ldots, N_{train}\}$. In this case, the *normalized* anomaly score ($\sigma(\boldsymbol{s}^{test})^{FMM}$ and $\sigma(\boldsymbol{s}^{test})^{SN}$) for a test datum $\boldsymbol{s}^{test}$ can be *directly* computed as the *probability* of $\boldsymbol{s}^{test}$ being an anomaly, i.e., as the probability mass associated to the points that are more likely than $\boldsymbol{s}^{test}$: $\sigma(\boldsymbol{s}^{test}) = \int f_{\boldsymbol{u}}(\boldsymbol{u}) \mathbb{I}\{f_{\boldsymbol{u}}(\boldsymbol{u}) \geq f_{\boldsymbol{u}}(\boldsymbol{s}^{test})\} d\boldsymbol{u}$, where $\mathbb{I}\{f_{\boldsymbol{u}}(\boldsymbol{u}) \geq f_{\boldsymbol{u}}(\boldsymbol{s}^{test})\}$ is equal to 1, if $f_{\boldsymbol{u}}(\boldsymbol{u}) \geq f_{\boldsymbol{u}}(\boldsymbol{s}^{test})$, and equal to 0, otherwise.

4. Results

The performance of each detector and of the ensemble is assessed in terms of the following well-known quantitative indicators: *Accuracy*, *Precision* and *Recall*. In this view, each prediction of the detector can be categorized as: True Positive (TP) (the actual and the predicted transient category is "anomalous"); True Negative (TN) (the actual and the predicted transient category is "normal"); False Negative (FN) (the actual class is "anomalous", while the prediction is "safe"); False Positive (FP) (the actual class is "normal", while the prediction is "unsafe"). Using these definitions, the *accuracy*, i.e., the percentage of overall correct (TP and TN) predictions referred to the total number of predictions (instances), is calculated as $\frac{TP+TN}{TP+FP+TN+FP}$. Instead, *precision*, i.e., the proportion of true positive predictions among all positive predictions made by the model, is quantified as $\frac{TP}{TP+FP}$. Finally, *recall*, i.e., the percentage of correctly predicted "positive" outcomes referred to the total number of real "positive" instances (namely, the sum of true positives and false negatives), is defined as $\frac{TP}{TP+FN}$. The results are shown in Table 1. Referring to the single detectors, SSAE (learning-based) and IF (proximity-based) are the top performing ones with respect to all metrics, showing accuracy, precision and recall values ranging in 0.9510-0.9528, 0.9395-0.9451 and 0.9637-0.9645, respectively. With respect to the LOF algorithm (distance-based), IF shows a better capability to deal with high-dimensional spaces. Among the statistical (probability-based) models, SN distributions achieve outstanding performances (comparable to those of SSAE and IF) showcasing accuracy, precision and recall values of 0.9341, 0.9257 and 0.9438, respectively. SNs significantly outperform FMMs (by 4-12% over all the metrics), thanks to their ability to capture multivariate dependences in high-dimensional spaces (where FMMs fail by a large amount).

Table 1. Accuracy, Precision and Recall values for the seven diverse detectors and for the probabilistic ensemble

	Learning-based models			Statistical (probabilistic) models		Proximity-based models		ENSEMBLE (Max.)
	SSAE	**AAKR**	**OCSVM**	**FMM**	**SN**	**LOF**	**IF**	
Accuracy	0.9528	0.9108	0.9114	0.8465	0.9341	0.8901	0.9510	0.9724
Precision	0.9415	0.9085	0.8939	0.8093	0.9257	0.8609	0.9395	0.9694
Recall	0.9645	0.9136	0.9331	0.9048	0.9438	0.9293	0.9637	0.9877

The probabilistic (CDF-based) ensemble with maximization aggregation function is superior to *all* the baseline detectors: the accuracy, precision and recall are improved by about 2.06% (resp., 14.87%), 2.96% (resp., 19.78%) and 2.41% (resp., 9.16%) with respect to the best (resp., worst) detector, respectively.

4. Conclusions

The paper has proposed a framework for the identification of anomalies in the behavior of dynamic safety-critical systems, which is based on an ensemble of seven diverse statistical and machine learning techniques, combined within a nonparametric probabilistic framework. An application to the detection of failure scenarios in an innovative (Gen-IV) nuclear reactor has shown that the ensemble: (i) has better performances than *all* the individual detectors; (ii) has a high capability to detect the failures, i.e., to improve system *safety* (the recall is 0.9877); and (iii) keeps the false positive rate (i.e., the number of unnecessary stops of the plant) to a comparatively low value (the "imprecision" is 1 – 0.9604 = 0.0306, which is important from the point of view of system *availability*). Future work is devoted to: (a) apply the model to larger-sized systems and possibly real (non-simulated) data; (b) apply the framework to other type of functional data (e.g., space-dependent quantities in the nuclear reactor core); (c) «inform» empirical (data-driven) detection models with physical models; (d) implement *adaptive* ensembles, where detectors are *iteratively* and *sequentially* added (or removed) based on the optimization of proper figures of merit (e.g., ensemble predictive capability).

References

Abrate, N., N. Pedroni, N. Caruso, S. Dulla, S. Lorenzi (2026). Early warning in Molten Salt Fast Reactors based on a data-driven method for the online incident detection and diagnosis. *Nucl. Eng. Des.* 446, 114581.

Aggarwal, C. (2013). Outlier ensembles: position paper. *SIGKDD Explor. Newsl.* 14(2), 49–58.

Aggarwal, C. (2017). *Outlier Analysis* (2nd ed.). Springer: Berlin/Heidelberg, Germany.

Breunig, M.M., H.-P. Kriegel, R.T. Ng, J. Sander (2000). LOF: identifying density-based local outliers. *ACM SIGMOD Record* 29, 93–104.

Crespo, L.G., B.K. Colbert, S.P. Kenny, D.P. Giesy (2019). On the quantification of aleatory and epistemic uncertainty using Sliced-Normal distributions. *Syst. Control Lett.* 134, 104560.

Jung, S., J. Kim, S. Kim (2025). A Hybrid Fault Detection Method of Independent Component Analysis and Auto-Associative Kernel Regression for Process Monitoring in Power Plant. *IEEE Access* 13, 39135-39151.

Kriegel, H.P., P. Kröger, E. Schubert, A. Zimek (2011). Interpreting and Unifying Outlier Scores. *Proceedings of the 11th SIAM International Conference on Data Mining*, Mesa, AZ, 13–24.

Liu, F.T., K.M. Ting, Z.H. Zhou (2008). Isolation Forest. *8th IEEE Int. Conference on Data Mining*, Pisa, Italy, 15-19/12/2008, 413-422.

Ouyang, B., Y. Song, Y. Li, G. Sant, M. Bauchy (2021). EBOD: An ensemble-based outlier detection algorithm for noisy datasets. *Knowledge-Based Systems* 231, 107400.

Rollón de Pinedo, A., M. Couplet, B. Iooss, N. Marie, A. Marrel, et al. (2021). Functional Outlier Detection by Means of h-Mode Depth and Dynamic Time Warping. *Appl. Sci.* 11, 11475.

Scholkopf, B., J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson (2001). Estimating the support of a high-dimensional distribution. *Neural Computation* 13(7), 1443–1471.

Tripodo, C., A. Di Ronco, S. Lorenzi, A. Cammi (2019). Development of a control-oriented power plant simulator for the molten salt fast reactor. *EPJ Nucl. Sci. Technol.* 5, paper 13, 24 pages.

Zhao, R., R. Yan, Z. Chen, K. Mao, P. Wang, R.X. Gao (2019). Deep learning and its applications to machine health monitoring. *Mech Syst Signal Process*, 115, 213–237.