Early warning in Molten Salt Fast Reactors based on a data-driven method for the online incident detection and diagnosis

Nuclear Engineering and Design 446 (2026) 114581

Contents lists available at ScienceDirect

Nuclear Engineering and Design

journal homepage: www.elsevier.com/locate/nucengdes

Early warning in Molten Salt Fast Reactors based on a data-driven method for the online incident detection and diagnosis☆

N. Abrate [a,*], N. Pedroni [a], N. Caruso [a], S. Dulla [a], S. Lorenzi [b]

[a] NEMO Group - Energy Department, Politecnico di Torino, Turin, Italy
[b] Nuclear Reactor Group - Energy Department, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Molten Salt Fast Reactors
Plant online monitoring
Incident detection and diagnosis
Early warning
Singular Value Decomposition
kNN algorithm
Incident precursors identification

ABSTRACT

This paper presents an innovative online incident detection and classification method, which aims at improving the safety, reliability and availability of Molten Salt Fast Reactor (MSFR) power plant, focusing on scenarios characterized by deviations from normal operational conditions.

The first part of the paper is devoted to describing and discussing the proposed online data-driven incident detection and classification methodology (based on adaptive Singular Value Decomposition-SVD and kNN algorithm), which aims at identifying abnormal plant conditions thanks to a continuous monitoring of some measurable parameters and variables (e.g., the molten salt temperatures in the secondary circuit). The developed incident detection algorithm is trained on a set of simulated scenarios featured by deviations of the main MSFR plant parameters from their nominal values. The data-driven model is then assessed considering increasingly complex incident classification rules and tasks, showing satisfactory performances in detecting and classifying plant anomalies (with an accuracy ranging between 89 % and 99 %). Finally, a fault diagnosis framework is proposed to carry out probabilistic inference on the most likely root causes (or precursors) – e.g., combinations of physical parameter values and component failures – that lead the system to the detected abnormal states.

1. Introduction

The Molten Salt Fast Reactor (MSFR) is a Generation IV reactor concept (Locatelli et al., 2013), conceived during a series of research projects (EVOL, SAMOFAR, SAMOSAFER…) financially supported by the European Union in the last fifteen years (Gerardin et al., 2017). The MSFR offers a set of advantages, compared to commercial light water reactors (LWRs), encompassing a larger utilisation fraction, an increased operational flexibility allowing to perform load following, the closure of the fuel cycle, which implies a consistent reduction of the long-lived radioactive waste and make the system proliferation-resistant, and enhanced safety features.

The MSFR employs a liquid core, circulating in the primary containment loop, made of a chloride- or fluoride-based molten salt containing fissile material. The molten salt, which operates nearly at atmospheric pressure and has a high boiling temperature, acts as fuel and heat transfer fluid at the same time, transferring the fission heat to the secondary loop with the 16 heat exchangers surrounding the core structure (Allibert et al., 2017).

The fluid core enhances the overall safety features of the plant, but poses a set of challenges for its operation, for instance the reduction of the importance of the delayed neutrons in the reactor dynamics, which is reflected by a smaller value of the effective delayed neutron fraction and, as a consequence, of a faster reactor dynamics compared to LWRs (Nicolino et al., 2008). This aspect is exacerbated by the fast spectrum characterising the reactor, which offers several advantages concerning the reactor sustainability but reduces the neutron mean generation lifetime (Bell and Glasstone, 1970).

Despite its fast dynamics, the large negative reactivity coefficient permits to control the power excursions in the system by relying only on the power extraction system (Fiorina, 2014; Laureau et al., 2022). The strength of the reactor reactivity feedback and the continuous adjustment of the salt composition eliminate the necessity of adopting control rods for the regulation of the reactor power and for the long-term reactivity control in the present design (Allibert et al., 2017).

These unique design features require *ad hoc* safety studies for demonstrating the inherent safety capabilities of the MSFR technology in both normal and off-normal conditions (Fiorina, 2014) and suitable

☆ This article is part of a special issue entitled: 'Molten Salt' published in Nuclear Engineering and Design.

* Corresponding author.
E-mail address: nicolo.abrate@polito.it (N. Abrate).

https://doi.org/10.1016/j.nucengdes.2025.114581
Received 31 January 2025; Received in revised form 17 October 2025; Accepted 31 October 2025
Available online 19 November 2025

operational strategies for maximising the power plant availability and reliability.

The objective of this work, which has been performed in the framework of the EU SAMOSAFER (SimulAtion MOdels and Safety Assessment for Fluid-fuel Energy Reactors) project as a follow-up to the EU projects EVOL and SAMOFAR, is the development of a data-driven fault detection and classification method for monitoring *online* the plant status and *promptly* identifying possible anomalies, by exploiting some measurable, safety–critical physical parameters and variables of the system (e.g., the molten salt temperatures at different positions of the cooling loops).

Data-driven methods are becoming increasingly popular for the safety analyses of critical systems and infrastructures, due to their accuracy and computational efficiency (Puppo et al., 2021). The data-driven algorithms for fault detection and diagnosis, whose effectiveness for the operation of nuclear power plants (NPPs) was reported in (Min et al., 2018), can be roughly divided into three broad categories, depending on how the data is generated, deployed and processed (Dai and Gao, 2013):

1. model-based (data-driven) methods, in which: i) only a small amount of data is typically available and used to detect and diagnose faults/incidents in the system; and ii) a physical–mathematical model of the plant is constructed from first principles or identified through system identification techniques. The physical models are exploited to produce an estimate (i.e., a prediction) of the system response (in normal and/or abnormal conditions) that is compared to the actual behaviour of the real monitored system (i.e., to the measured data), in order to spot out and classify possible anomalies.
2. signal-based methods, which exploit the relationship between faults/incidents and the time evolutions of some plant signals (when available) and require the association of one output signal (typically the most critical monitored plant parameter/variable) to the fault, according to an *a priori* human judgement and physical understanding of the system. Since the faults within the plant usually have an influence on the selected/identified critical output variable, there is no need for the construction of a detailed input–output physical model of the dynamic system of interest. This is beneficial for complex processes or systems, where accurate input–output models are often unavailable and/or their parameters are hard to estimate and calibrate.
3. history data-driven methods, which exploit historical data collected from systems similar to the one under investigation to train empirical (Machine Learning-ML and/or Artificial Intelligence-AI) models of the system behaviour. In other words, history data-driven methods are able to extract the necessary information and knowledge (about the *implicit* relationships and dependences between faults/incidents and the time evolution of some critical plant parameters/variables) from a large amount of process data. This represents the preferable choice when: i) a process or system is too complex to be modelled mathematically in the required detail; and/or ii) obtaining trustworthy expert knowledge is difficult; and/or iii) the signal patterns are not available straightforwardly and the corresponding signal analysis do not allow for an unambiguous detection and diagnosis.

Several works deal with the issues of fault detection and diagnosis in nuclear engineering applications. In (Nguyen et al., 2020) a framework based on quantitative *model-based* diagnosis, statistical change detection and probabilistic reasoning (i.e., Bayesian networks) is developed for fault detection and localization in a single-phase heat exchanger, showing high detection sensitivity and allowing noise and measurement uncertainty to be incorporated. In (Tolo et al., 2018), a robust online approach is applied to the case of Loss of Coolant Accidents (LOCAs), which combines Artificial Neural Networks (ANNs) and Bayesian statistics for providing the fault diagnosis and its confidence bounds. Similarly, *machine learning*, *artificial intelligence* and *deep learning* techniques (e.g., Deep Belief Networks-DBNs, Convolutional Neural Networks-CNNs, Long Short-Term Memory-LSTM networks, Transformers, Graph Convolutional Networks-GCN, Generative Adversarial Networks-GANs, Attention-based Networks-ANs and Stacked Sparse AutoEncoders-SSAEs) are employed in (Luo, 2024; Wang et al., 2024; Yang et al., 2024; Dai et al., 2023; Liu et al., 2024; Yang et al., 2022; Saeed et al., 2019; Mandal et al., 2017) to identify faults in several types of NPP components (e.g., thermocouples, pumps, pipes, sensors, controllers, drive circuits and bearings, respectively). Also, in (Farber and Cole, 2020), data-driven modelling and control-theoretic estimation techniques are combined to detect LOCAs and estimate their magnitudes in real-time. First, simulated process data for a variety of nominal operating conditions is collected using a generic pressurized water reactor simulator. Then, that data is used to train an ANN regression model that captures the nonlinear plant dynamics. Finally, the ANN regression model is used within a particle filter to detect the onset and estimate the magnitude of the leak. In (Choi and Lee, 2020), the Missforest imputation technique and gated recurrent unit with decay (GRUD) are used to improve the modelling capabilities of Recurrent Neural Networks (RNNs) in the development of a sensor fault-tolerant accident diagnosis approach for NPPs. In (Yong-kuo et al., 2018), a hybrid of Principal Component Analysis (PCA) (Li et al., 2018), signed directed graph (SDG), and Elman Neural Network (ENN) is proposed for fault detection, fault isolation, and severity estimation, respectively. The successful performance of the hybrid approach is verified on main steam line breaks and LOCAs simulated by the Personal Computer Transient Analyzer (PCTRAN) software. Similarly, in (Wang, 2021), the multi-stage hybrid combination of Least Squares Support Vector Machine (LS-SVM) and optimized Gaussian Process Regression (GPR) methods is introduced to efficiently identify system-level failures and their severities in a simulated Pressurized Water Reactor (PWR), with the specific objectives of facilitating the recognition of the real-time failure types and extracting both the constraint relationships and fault regularities of system-level parameters. In (Baraldi et al., 2015) different data-driven techniques (i.e., Auto-Associative Kernel Regression-AAKR, Fuzzy Similarity-FS, and Elman Recurrent Neural Network-RNN) are investigated and compared for reconstructing the system time-varying signals with the aim of fault identification. Another family of approaches is represented by (supervised or unsupervised) *clustering* (i.e., the intelligent grouping of different types and classes of system behaviours based on a similarity criterion), possibly aided by feature selection or extraction algorithms (to identify only the parameters or features that are relevant to the efficient characterization of the patterns of system evolution). For example, in (Al-Dahidi et al., 2018), a framework based on Spectral Clustering embedding an unsupervised *K*-Means algorithm is proposed for incrementally learning and reconciliating different (failure) clusters independently obtained for individual steam turbines coming from a heterogeneous fleet of NPPs. In (Wang et al., 2021), Kernel Principal Component Analysis (KPCA) and similarity clustering are combined and applied to a full scope PWR simulator. First, KPCA is used for anomaly detection (to distinguish actual faults form abnormal sensor readings) and for feature extraction (to analyse fault types and degrees); then, SVM carries out fault diagnosis by similarity clustering on the extracted KPCA features. In (de Pinedo, 2021), multivariate probabilistic clustering by Gaussian Mixture Models (GMMs) is applied on features (e.g., h-mode depth and dynamic time warping) extracted from transient signals describing an Intermediate Break Loss of Coolant Accident (IBLOCA) in a 900 MW French PWR. In (Hu et al., 2016), an adaptive diagnostic model relying on semi-supervised feature selection and SVM classifiers is proposed for industrial applications featured by evolving environments. A thorough review of data-driven methods can be retrieved in (Feng et al., 2013), while a critical review of signal-based techniques can be found in (Venkatasubramanian et al., 2003). Finally, recent and detailed reviews dedicated to fault detection and diagnosis in nuclear science and engineering can be found in (Huang et al., 2023; Qi et al., 2023; Ramezani et al., 2022; Ayodeji et al., 2020; Gomez-

Fernandez et al., 2020): the interested reader is referred to the mentioned publications for further details.

Despite such extensive literature, only a few papers deal with the issues of fault detection and diagnosis in MSFRs. In (Zhou and Hou, 2022), PCA is used for detecting faults. After an anomaly is detected, Reconstruction-based Contribution (RBC) analysis is adopted to diagnose the (output) signal that causes the anomaly. In (Jiang et al., 2023), the control rod drive mechanism (CRDM) of a liquid fuel thorium MSFR is taken as the research object, and a fault diagnosis system is proposed based on knowledge graph and Bayesian inference. Unlike data-driven or physics-based methods, the knowledge-based fault diagnosis approaches place no requirement on complete operational data or precise mathematical models: thus, they are very effective and have explicit interpretations. First, unstructured data (including design specification, operation and maintenance manual, alarm list, and other forms of expert experience) is used to build a fault event ontology model to label the entities and relationships involved in the corpus of CRDM fault events. Then, a three-layer robustly optimized bidirectional encoder representation from transformers (RBT3) pre-training approach combined with a text convolutional neural network (TextCNN) is introduced to facilitate the application of the constructed CRDM fault diagnosis graph database for fault query: in particular, the RBT3-TextCNN model searches the database, extracts the entities and recognizes the fault query intent simultaneously. Finally, Bayesian Networks combined with the variable elimination algorithm are used to develop an intelligent and reliable fault diagnosis and root cause identification system, due to their capability of describing causality in uncertain problems and realizing inference and prediction. In (Zhou et al., 2023), four Machine Learning (ML) algorithms (i.e., Recurrent Neural Networks-RNNs, Support Vector Machines-SVMs, Decision Trees-DTs and k-Nearest Neighbour-KNN classifiers) are tested and compared in the transient identification of a liquid-fuelled MSFR.

In this work, we propose a flexible but relatively simple data-driven approach (relying on an input–output dataset produced by a physical–mathematical computer model of an MSFR) that could be effectively employed to provide an *early warning* to the plant operators as soon as *anomalies* are detected in the time behaviour of the main plant parameters (also inferring potential *root causes* of the anomalies), which could help controlling their evolution and preventing more severe outcomes. Specifically, this approach allows to perform prompt, online: i) fault *detection* (based on the real-time monitoring of physical variables and parameters that are safety–critical for the reactor, e.g., the fuel temperatures); ii) fault *classification* (i.e., the assignment of the detected transients to a pre-labelled class of faults or failure modes); and iii) fault *diagnosis* (in the form of a probabilistic inference on the most likely root causes or precursors – e.g., combination of parameter values and component failures – that generated the detected transients). This family of methods has been preferred both to signal-based algorithms, because the direct input–output pairing for the MSFR plant under analysis is not straightforward, and to history-based data-driven techniques, since there is no operational experience with this unique type of NPP. The specific fault detection and classification methodology is based on a combination of (time-dependent) Singular Value Decomposition (SVD) and the k-Nearest Neighbours (kNN) algorithm, which are two traditional techniques for dimensionality reduction and pattern classification, respectively. The SVD-kNN approach is trained on a set of input–output signals, representing the failures of some plant components (e.g., the circulating pumps) and the corresponding system response (i.e., the safety–critical physical variables and parameters), respectively; once trained, the method allows to detect and classify new (unseen) scenarios.

The paper is organised as follows. In the next section, the structure of the SVD-kNN algorithm is presented and discussed. Then, in Section 3 the main features of the MSFR NPP are examined together with its simulator (Section 3.1), with the objective of identifying a set of parameters describing the plant status and one operational mode of the system for testing the algorithm (Section 3.2). In Section 3.3 the NPP numerical simulator is used to perform a thorough exploration of the plant parameter space with the main objective of generating an artificial database of transients for SVD-kNN training (Sections 3.4 and 3.5). In Sections 4 and 5, the main results obtained by the SVD-kNN algorithm are reported and discussed, focusing on the model training, testing and evaluation, respectively. Finally, in Section 6 we draw some conclusions and give some perspectives on possible extensions to the methodology.

2. The SVD-kNN algorithm for online incident detection and classification

In this section, the main steps constituting the algorithm for fault detection and classification are briefly outlined. Fig. 2.1 provides a graphical representation of the methodology, which is composed of three main steps.

The first step consists of a database generation, where the model of the MSFR plant is sampled for simulating a set of input and output time-dependent signals that are then divided into a training and a test set. According to the behaviour of the output signals with respect to some prescribed safety thresholds and margins, discussed in Section 3 each scenario is classified and labelled according to different taxonomies, i.e. categorisation rules (e.g., normal versus abnormal plant conditions, with possibly different and specific sub-classes of anomalous behaviours, depending on the peculiar time evolutions of the various observed physical variables).

The Singular Value Decomposition method (Lass and Volkwein, 2014) is the employed for extracting the most significant features from the signals, reducing the dimensionality of the dataset and, thus, the computational cost associated to their classification.

The reduced patterns of the training dataset are then used in the second step (i.e., the training phase) as input–output examples to optimally build a kNN classifier, i.e., a tool able to receive as input new monitored signals from the plant simulator and to return as output an exhaustive set of *probabilities* that the specific (new) transient falls into the available predefined classes (i.e. the kNN assigns a transient to a class with a given *probability* or *confidence level*). Since the classifier is not able to define new classes by itself (i.e., it is a *supervised* classifier), it is very important to generate and provide a training dataset that covers as thoroughly as possible the entire plant state-space. Hence, the larger is the number of classes, the larger the training dataset should be.

The founding principle of this algorithm is the simple assumption that data featured by similar characteristics are likely to belong to the same class, i.e. the hypothesis of feature similarity. With respect to the application presented in this work, if the simulated signals of each plant scenario exhibit a similar time evolution, they are likely to belong to the same failure class. Therefore, it is also likely that they are caused by a similar combination of input parameters, i.e. the component failures.

The choice of the kNN classifier (Soucy and Mineau, 2001; Mareboyana, 2018) is driven by a compromise between the algorithm simplicity and its robustness, despite it can be computationally expensive for large training datasets, since each prediction requires the evaluation of the distance between the example of interest (i.e., the *new* set of *test* transients to be classified) and all the training samples (i.e., the training database).

In synthesis, in the third step (i.e., the testing phase) a new signal, say x , is classified by a plurality vote of its neighbours, with the observation x being assigned to the class most common among its k nearest neighbours, where k is a positive, typically small, integer. The classification process carried out in this work is practically actuated by minimizing the expected misclassification cost, defined as

$$\widehat{y} = arg\left[\min_{y=1,\ldots,n_c} \sum_{j=1}^{m} \widehat{P}(j|x)C(y|j)\right],$$

where $\widehat{y}$ is the predicted classification, n_c is the total number of available

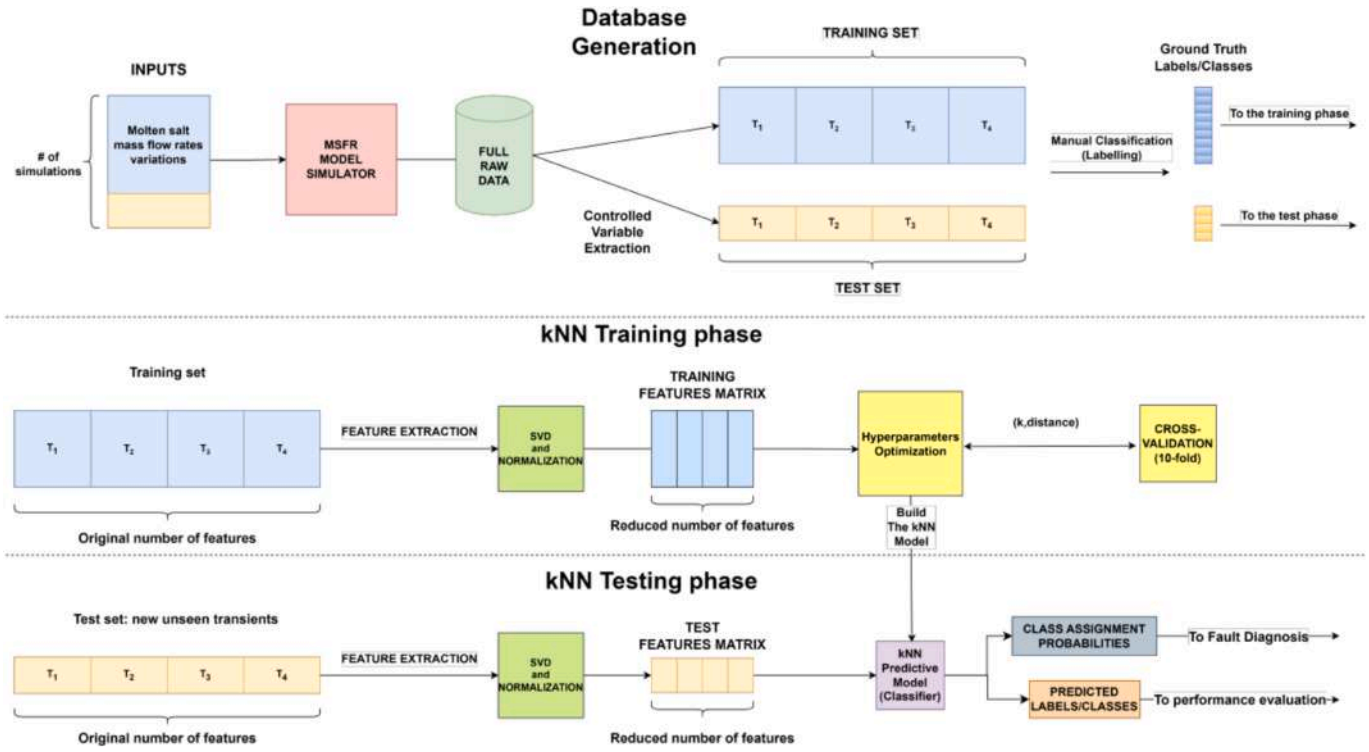


Fig. 2.1. Conceptual scheme of the data-driven fault detection and classification algorithm.

classes, $C(y|j)$ is the cost of classifying a new observation x as y, provided that its real class is j, and $\widehat{P}(j|x)$ is the posterior probability of class j for the new test observation x. This last term is computed as follows,

$$\widehat{P}(j|x) = \frac{\sum_{i \in n_b} W(i)\, \delta_{Y(X(i))=j}}{\sum_{i \in n_b} W(i)},$$

where $W(i)$ is the weight assigned to the i-th element of the k training data samples that belongs to the "neighbourhood" n_b of the new observation x and $Y(X(i))$ is the "real" class of the i-th training sample $X(i)$. The term $\delta_{Y(X(i))=j}$ is equal to 1 when $Y(X(i)) = j$ or 0 otherwise.

The posterior probabilities provided by the classifier should be interpreted as the membership probabilities of scenario x to class j. These probabilities are mutually exclusive and exhaustive. The class featuring the largest membership probability is assigned to the signal. It should be noted that in general there are no minimum threshold values to be exceeded by these probabilities for the class to be assigned (for example, a crisp assignment of a transient to a given class may be decided in the presence of a posterior probability equal to 0.97, as well as equal to 0.48). However, such thresholds may be defined *ad hoc* and imposed by the analyst, in case he/she wants to build a desired level of confidence and robustness in the classification process (by way of example, the analyst may consider a transient reliably assigned to a given class, only if the corresponding posterior probability exceeds, e.g., 0.7).

For simplicity, the classification cost $C(y|j)$ is unitary when $j \neq y$ and vanishes when $j = y$, meaning that the cost for a correct classification is 0 and the cost for an incorrect classification is 1. Moreover, all the training data belonging to the neighbourhood n_b of the new observation x have the same weight in the classification process. Clearly, the computational cost of the classifier is proportional to the number of classes n_c and to the number of features, i.e., the dimensionality, of the observable vector $X(i)$.

In this work, we rely on the *fitcknn* built-in function in MATLAB, which automatically adopts acceleration strategies depending on the dataset size and dimensionality, such as KD-tree structures for efficient nearest-neighbour searches or optimized exhaustive methods for high-dimensional data. Moreover, the search process can be parallelized. These features substantially reduce the time required for both (off-line) model training and (online) class prediction, favouring the applicability of the proposed approach in real-time monitoring frameworks. In case of even larger datasets, further acceleration could be achieved through approximate nearest-neighbour methods with negligible loss in accuracy (Muja and Lowe, 2014; Dong et al., 2011). Finally, it is also worth reminding that the most expensive step (from a computational viewpoint), i.e., the training of the classification model, is performed off-line. Instead, its online deployment (i.e., the class prediction phase by the trained classifier) is expected to be much faster, also in case of larger datasets.

The training dataset and the membership probabilities provided by the kNN-SVD algorithm are finally combined to perform a preliminary Fault Diagnosis, which consists in a probabilistic inference on the most likely causes that led the system to the detected failure. In this work, the inference is performed exploiting the Theorem of the Total Probability, which allows to retrieve the probability distribution of the physical perturbation $\vec{P}$ inducing the failure,

$$f\left(\vec{P}\right) = \sum_{j=1}^{N} f\left(\vec{P}\middle|C_j\right) g\left(\tau\left(\vec{P}\right) \in C_j\right)$$

as the weighted sum, over the N classes, of the distribution of the physical perturbation yielding the j-th class, $f\left(\vec{P}\middle|C_j\right)$, times the membership probability of each transient $\tau\left(\vec{P}\right)$ belonging to each class C_j, namely $g\left(\tau\left(\vec{P}\right) \in C_j\right)$. From the posterior probability distribution $f\left(\vec{P}\right)$, the most likely causes of the system failure can be inferred.

Some final considerations are in order with respect to the features of the adopted classifier and its construction process. The kNN classifier is a supervised, non-parametric algorithm, i.e., it does not make any assumptions about the underlying distribution featuring the data.

However, the algorithm performances heavily depend on the distance metric used to estimate the distances of the test points from the class centroids and on the hyperparameter k, which indicates the number of nearest neighbours used to define a class. Since an *a priori* optimal prediction for these parameters is barely impossible on a theoretical ground, their values are optimally assigned through minimisation of the classification error on a set of test samples.

The hyperparameter optimization routine is performed by a cross-validation approach. In general, a small value of k is associated with a data overfitting, whereas a too large value of k could smooth out the prediction, since very wide neighbourhoods would be considered.

For ensuring an adequate exploration of the k optimal values, in this paper the optimisation process considers values ranging between 1 and 40. The optimization routine also allows to choose the optimal distance metric that best performs in combination with the selected value of k. For each combination of k and distance metrics, one kNN classifier is trained and cross-validated, and the associated miss-classification error is computed. The cross-validation process is accomplished by partitioning the training dataset into a training fold and a testing fold. The classifier is trained with each of the N sets obtained by taking $N - 1$ samples and using the left one as a testing point (namely, Leave-One-Out Cross-Validation, LOO-CV).

It should be acknowledged that, in principle, also other classifiers could be embedded in the algorithm, as alternatives to the kNN or to support its performances. For instance, the class membership probabilities may be obtained by applying a majority voting procedure involving all the classifiers adopted.

3. The molten salt fast reactor nuclear power plant

The current design of the MSFR is a 3000 MWth reactor consisting of a cylindrical vessel with diameter and height of about 2.25 m. The vessel, made of a nickel-based alloy, is filled with roughly 18 m^3 of liquid salt, which acts both as fuel and as a coolant. The molten salt is operated at a pressure near the ambient one with a mean temperature of 750 °C and flows, thanks to the primary circulating pumps, in the primary circuit in the upward direction through the central core zone and in the downward direction through the heat exchangers, located circumferentially around the core. Between the core and the heat exchangers, a container filled with a fertile blanket containing a thorium-based salt is present to increase the system breeding. A dedicated heat removal system is foreseen to remove the power generated in the fertile blanket. Fig. 3.1 shows a sketch of the core and of the full containment building.

The MSFR can be operated with different fuel compositions, thanks to its online fuel control and flexible fuel processing: its initial fissile inventory can be composed of enriched natural uranium or transuranic elements, like plutonium and other minor actinides. The fission fragments produced in the salt are then removed in a salt treatment unit, with the main purpose of controlling thermophysical properties. The fuel salt composition considered for this work is LiF-ThF_4-^{233}U, whose melting temperature is around 585 °C (858 K).

The fluid flowing in the Intermediate (secondary) Circuit (IC) is NaF-$NaFB_4$, with a melting temperature of 384 °C, while the heat transfer fluid in the Gas Circuit (GC), devoted to the power conversion, is helium.

3.1. The MSFR power plant simulator

In this work, we rely on a power plant simulator developed during the SAMOFAR project for investigating time-dependent response of the system. The simulator is a control-oriented plant- dynamics tool, developed using the open-source, object-oriented Modelica language (Nguyen et al., 2020). Fig. 3.2 shows the schemes of the primary and energy conversion circuits of the Modelica system-level model.

In order to explore the MSFR plant behaviour at a system-level, including the energy conversion system, the power plant simulator developed in the open-source, object-oriented Modelica language at Politecnico di Milano (PoliMi) in the frame of the SAMOFAR project is adopted, to simulate the various system responses to perturbations in the nominal mass flow rates (Laureau, 2017; Tripodo et al., 2019). The power plant simulator has been benchmarked with other system-level tools like the PANDAS code, developed at the Centre national de la recherche scientifique (CNRS), and a version of TRACE modified at the Paul Scherrer Institute (PSI), providing satisfactory results (Laureau, 2017). The main modelling assumptions of the PoliMI power plant simulator are the following:

- thermal-hydraulics phenomena are modelled with a one-dimensional approach, adopting volume-averaged quantities,
- the delayed neutron precursors motion is treated with a one-dimensional approach,
- the neutron kinetics is approximated with the point kinetics model. This assumption is not expected to impact significantly the accuracy of the calculations, since the core is homogeneous and optically small due to the typically large mean free path featuring fast neutrons.

The Modelica-based simulator does not fully reproduce the spatial resolution and measurement characteristics of an actual instrumentation and control (I&C) system. In a real MSFR plant, sensor signals (e.g., wall-mounted thermocouples, ex-core neutron detectors, …) would be subject to additional uncertainties, time delays, and noise, and their deployment would be constrained by technological and economic considerations. In practice, the number, location and type of detectors are typically determined by a reasonable trade-off between cost, fabricability and diagnostic power (i.e., the capability to maximize the accident identification and classification rate). Therefore, the present paper should be regarded as an idealized framework aimed at testing the

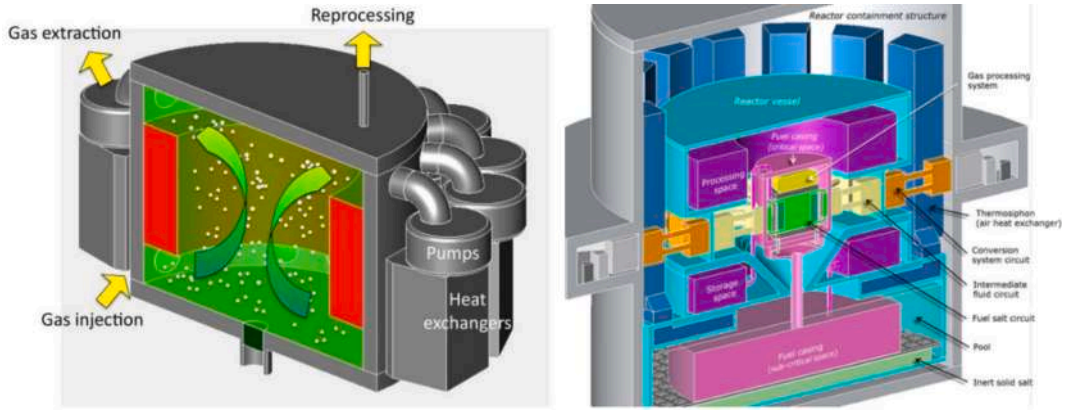


Fig. 3.1. Sketch of the MSFR reactor (left) and of the MSFR reactor containment structure (left). The pictures are taken from [3.

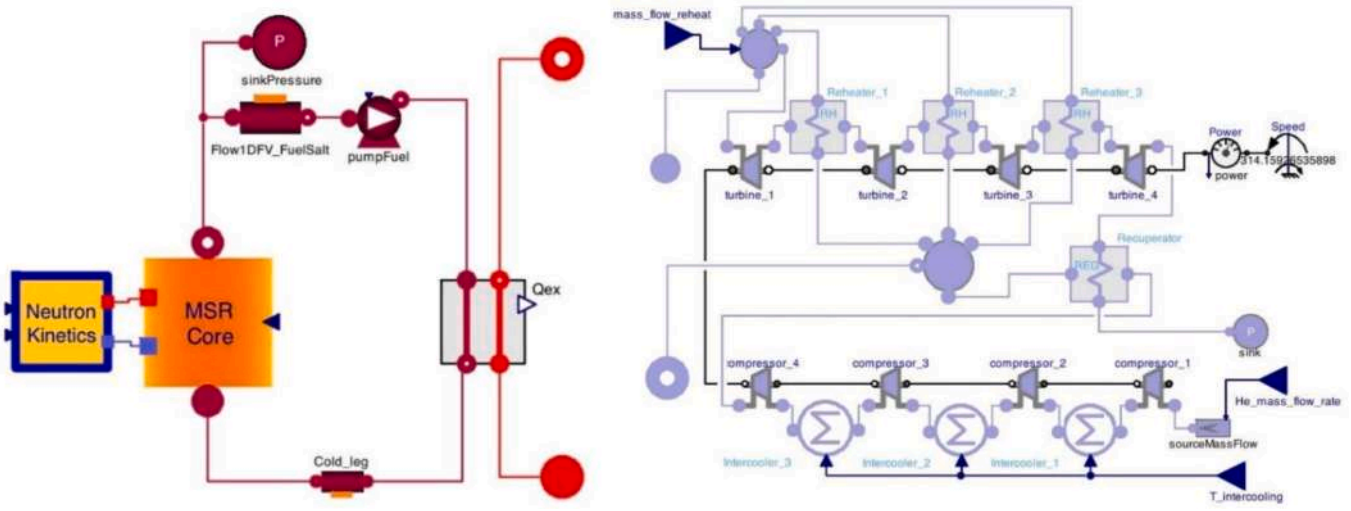


Fig. 3.2. Models of the fuel circuit (left) and energy conversion circuit (left) (Dong et al., 2011).

feasibility and robustness of the proposed diagnostic methodology under controlled, noise-free conditions. The extension of this work to more realistic monitoring configurations, including measurement noise and variance, and sensor manufacturability and cost is envisaged as a natural follow-up of this study, also building upon the insights available in the literature on the topic. Specifically, several studies have previously addressed the issues of measurement noise and signal uncertainty through filtering and feature-extraction techniques, such as moving-average preprocessing and PCA (Roma et al., 2021; Zhu et al., 2021), or by using deep autoencoders to automatically perform denoising while at the same time reducing dimensionality (Roma et al., 2022; Zhu and Yin, 2025). Recurrent Neural Network-based surrogate models have also shown notable resilience to signal degradation (Chevalier-Jabet and Verma, 2024).

Despite these limitations, the model still retains an acceptable accuracy with respect to the main physical phenomena, allowing to perform a computationally efficient analysis of the plant, also thanks to the adoption of an adaptive time step algorithm.

3.2. Identification of the main plant parameters and controlled parameters

The identification of an appropriate set of main plant parameters for the MSFR is crucial for the development of a fault/incident detection and classification algorithm. Any main plant parameter (MPP) can be identified with one or more of the following classes, keeping in mind that the membership to a certain class depends on the operational state of the plant:

- *Controlled parameters*: any parameter that can be regulated by means of control and protection strategies, which are realised in practice with some control and protection functions, aiming at ensuring that the parameter stays within the control threshold. This definition implies the possibility to measure the controlled parameters in time and to act on its value by means of a control system, which acts on other variables called control variables.
- *Control parameters*: any parameter that can be directly regulated by means of a control system and whose variations may act on one or more controlled parameters. To result effective during the reactor operations, the regulation of these parameters should occur at short term, except when the regulations regard the long-term operations. Due to these features, these parameters are crucial for the development of an automatic control scheme and are strongly related to the evolution of possible incidents. Failures involving some control systems may directly affect the value of the control parameters.
- *Monitored parameters*: any parameter that can be directly or indirectly measured. These parameters are crucial for delivering a real-time picture of the reactor state. As such, they can be employed to assess whether the plant is currently working in normal or abnormal operating conditions. If the monitored parameters exceed the protection thresholds defined for a specific operational stage, the protection system is triggered. For minimising spurious interventions, an ~~optimal~~ appropriate subset of monitored parameters should be identified to get reliable and robust information about the plant status, properly accounting for the safety margins. In addition to the set of parameters, the effectiveness of the monitoring system should also be equipped with, for example, redundancy, majority voting and feature selection strategies. Moreover, the selection of these parameters should also encompass the detection time, including the dead time, the noise of the measurements, the elaboration of the signals, and the response of the actuator, accounting for the various time scales of the incidental or accidental situation as well. These features potentially reduce the signals' correlation with the (time-dependent) accidental evolutions, thus making them more difficult to identify and diagnose. In addition, the presence of noise may increase the rate of false positives, with a detrimental effect on the discriminative capability and diagnostic effectiveness of the classifier (Chevalier-Jabet and Verma, 2024). These extensions and improvements could leverage on (and be addressed by) modern techniques available in the literature, such as de-noising and pre-classification approaches, as mentioned above (Roma et al., 2021; Zhu et al., 2021; Roma et al., 2022; Zhu et al., 2025; Chevalier-Jabet and Verma, 2024).

Starting from previous studies carried out in the SAMOFAR and SAMOSAFER projects (Cammi, et al., 2018; Boisseau et al., 2022), we select a subset of the main plant parameters for the fault detection and classification.

For a given plant, featured by a set of components and working in a certain operational state, any deviation in a main plant parameter may be ascribed, directly or indirectly, to a loss of functionality in one or more components. For developing the fault detection and classification algorithm, we assume that the incidental scenarios are driven by the control parameters, since they are directly affected by anomalies in the components of the plant.

Assuming that the reactor is working in the "*reactor in power*" nominal condition, i.e. with the thermal power ranging from 20 % to 100 %

and with the Balance of Plant connected to the electrical grid, the following control parameters are considered:

- the mass flow rate of the Fuel (primary) Circuit (FC), $\dot{m}_{FC}$
- the mass flow rate of the Intermediate (secondary) Circuit (IC), $\dot{m}_{IC}$
- the mass flow rate of the gas flowing in the energy conversion (Gas) Circuit (GC), $\dot{m}_{GC}$. The re-heat gas flow rate is also considered, although it is assumed proportional to the gas flow rate.

The main criterion for the selection of these controlled parameters is the possibility to measure them and to control them directly with the circulation pumps of the plant. Due to this direct relationship, the pumps are safety–critical components and anomalies in their behaviour can be simulated by perturbing these control variables. On top of this, previous simulations of the MSFR power plant evidenced that these parameters strongly affect the behaviour of the plant (Tripodo et al., 2019).

Since the various flow rates have a strong impact on the operating temperatures of the FC and IC, these temperatures are considered the monitored parameters for training the kNN classifier, whose goal is the detection of possible pump failures occurring in the "reactor in power" mode. The lower and upper limits in the fluid temperatures are defined according to previous analyses (Allibert et al., 2017): the peak temperature should be lower than the maximum acceptable temperature for the containment structures, $T_{max} = 1373$ K, while the minimum temperature should be above the maximum between the freezing temperatures of the primary and secondary salt, i.e., $T_{min} = 858$ K,

In the following, the plant is assumed to be equipped with a proper measurement system (consisting of multiple thermos-couples and other instruments based, for instance, on ultrasonic velocity measurement) capable of providing the real-time value of the fluid temperature averaged on the pipe cross section.

A schematic representation of the three circuits of the MSFR considered and the corresponding observed parameters are given in Fig. 3.3.

3.3. Exploration of the MSFR parameter state

Each system-level simulation is carried out in free dynamics, aiming at studying the evolution of the main plant parameters only in presence of the inherent feedback mechanisms of the MSFR, starting from the steady state, nominal operation. At $t = 50$ s, where t indicates the time, different combinations of the mass flow rates in the FC, IC and GC are assumed to follow an exponential decay from their nominal value to a perturbed value, which is different in each circuit. The exponential reduction is featured by a time constant that simulates the inertia of the pumps' flywheel. Each scenario is simulated for 800 s.

The choice of simulating all the failures at the same time may not appear natural, since the common loss of these physically separated and independent pumps may occur, in practice, only with the loss of electrical power. However, this assumption turns out to be very useful to demonstrate the fault detection and classification capabilities of the algorithm in a challenging situation, i.e., when the deviations of the various monitored parameters occur at the same time. On top of this, this choice is convenient for reducing the input parameter space from 6 dimensions (i.e., the magnitude of the three mass flow reductions and the time instants of their occurrence) to 3 (i.e., only the magnitude of the three mass flow reductions). This choice allows reducing the number of simulations needed to thoroughly map the plant behaviour, thus simplifying the development and the testing of the algorithm.

Before running the simulations, a set of preliminary calculations have been carried out for performing a sensitivity study on the maximum discretisation time step Δt adopted in the Modelica solver, with the objective of optimising this parameter in view of the large number of simulations needed.

Fig. 3.4 shows that the evolution of the four output monitored parameters, namely the inlet and outlet core temperatures and the inlet and outlet temperatures of the secondary salt, is rather insensitive to the discretisation time step, due to the adaptive time step algorithm implemented in the Modelica code.

Fig. 3.5 shows the reduction in the disk space occupied by the Modelica output file and in the computational time required by the code computed with respect to the reference case with $\Delta t = 0.01$ s considering different discretisation time steps. Based on these results, the optimal maximum time step was set to 1 s, as a reasonable compromise between the storage occupation, the computational time and the accuracy of the results. With these simulation settings, each simulation took about 90 s on a commercial desktop computer to be completed.

To define a meaningful range of variation for the mass flow rates, a set of simulations was run to examine the MSFR response, in the wake of the work performed in (Allibert et al., 2017). Each mass flow rate was varied in the interval [0, −100 %], keeping the other flow rates at their nominal values. For each simulation, the value of the controlled variables was extracted at the same time instant, $t = 800$ s, which was taken large enough to ensure that all the outputs reached their steady state.

In the following, see Fig. 3.6, an example of scenario employed to characterise the system response is briefly analysed. In this specific case, the GC flow rate and, consequently, the re-heat flow rate drop exponentially to 10 % of their nominal values.

The reduction of the gas mass flow in the energy conversion circuit causes a strong reduction in the heat extracted from the IC, which causes

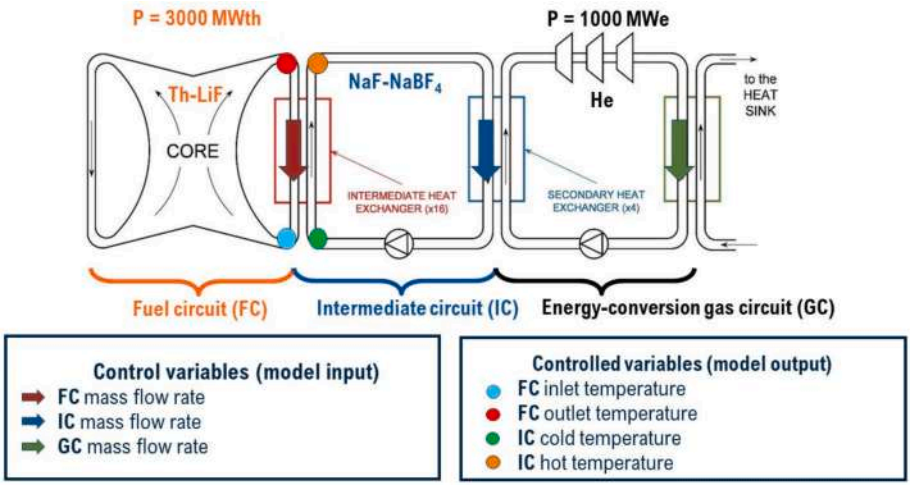


Fig. 3.3. Definition of control and controlled (monitored) parameters for the algorithm development (the plant sketch was taken from (Tripodo et al., 2019).

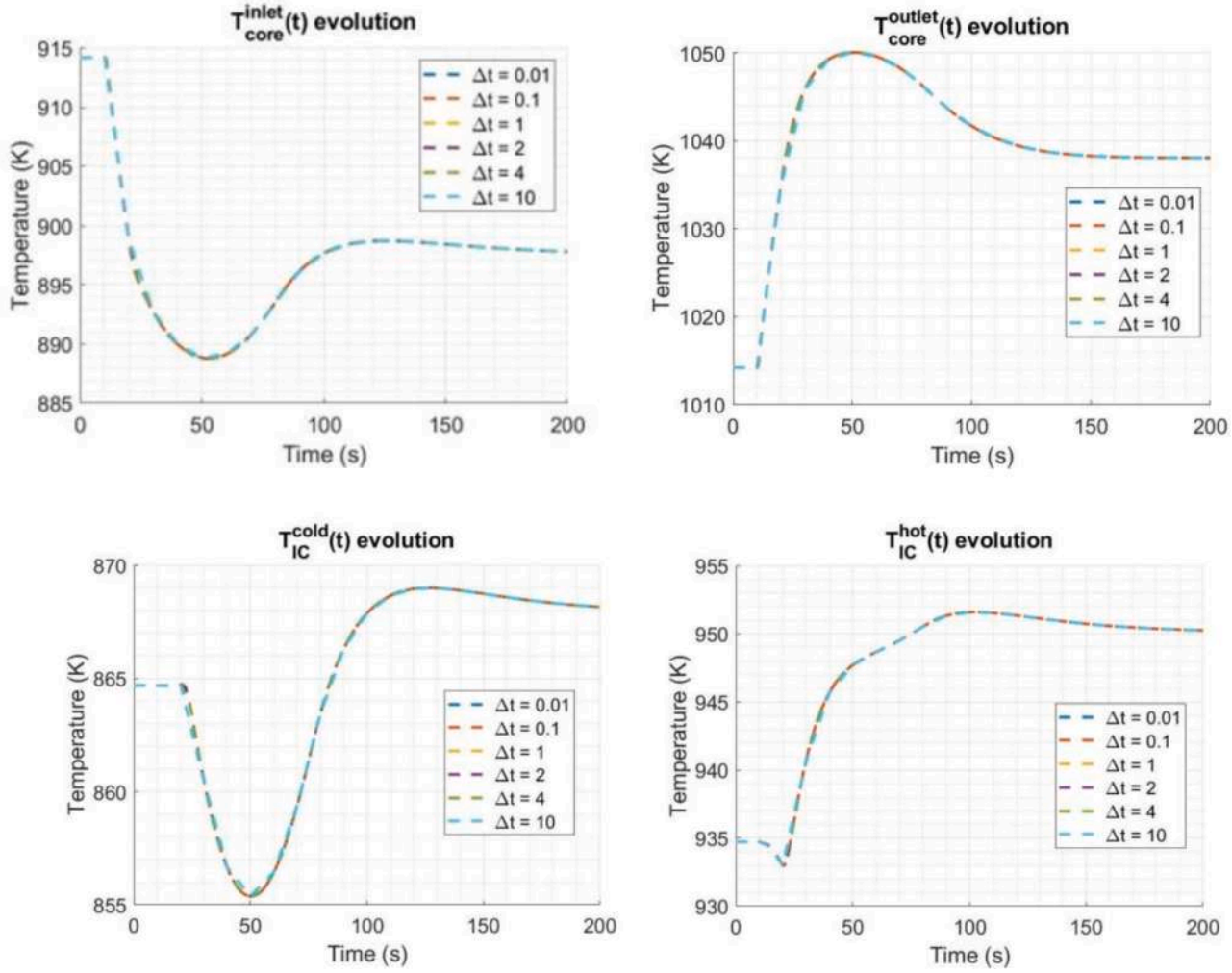


Fig. 3.4. Time evolution of the monitored parameters computed with different maximum time steps.

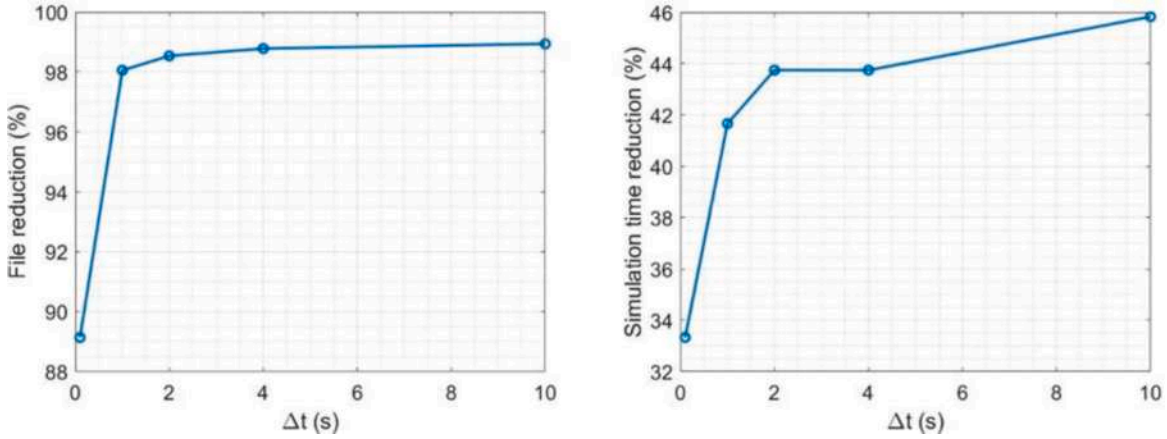


Fig. 3.5. Memory storage and computational time trend as a function of the time step adopted.

an increase in the IC salt inlet and outlet temperatures that in turn reduce the heat extraction capabilities of the Intermediate Heat eXchanger (HX). The ultimate consequence is the increase in the core average temperature, which induces a sharp decrease in the core thermal power. After a relatively short time, in virtue of the strong thermal feedback, the system reaches a new equilibrium condition, featured by a higher inlet core temperature and a lower temperature difference between inlet and outlet, which is a consequence of the lower thermal power produced in the core. The complete time evolution of the IC and FC temperatures and output power can be observed in Fig. 3.7.

This preliminary simulation campaign is useful both to investigate the sensitivity of the monitored parameters to the input variations and to test the numerical limits of the simulator, which is not designed to deal with strong variations in the plant parameters that could bring the

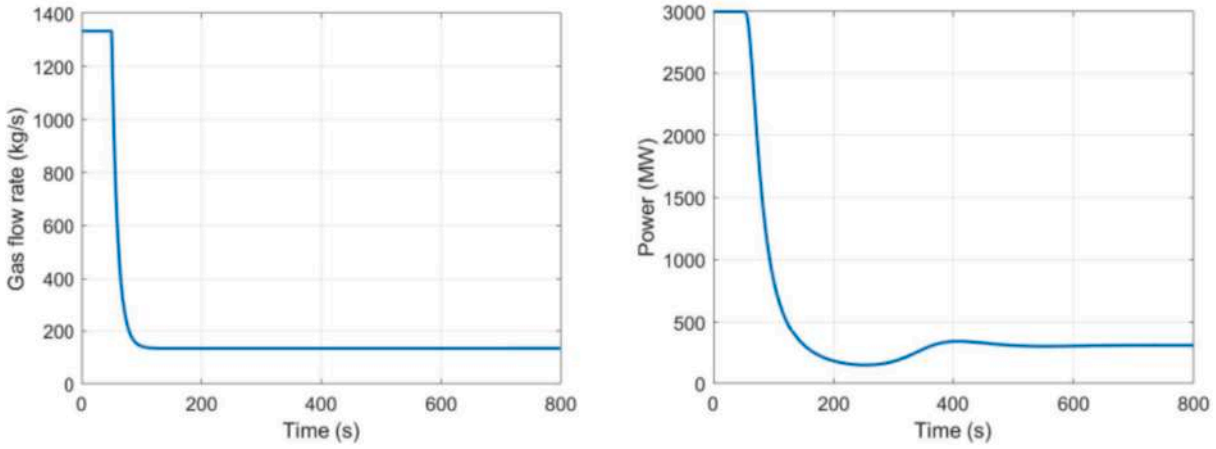


Fig. 3.6. Exponential reduction in the gas flow rate (left) and the consequent thermal power reduction (right.

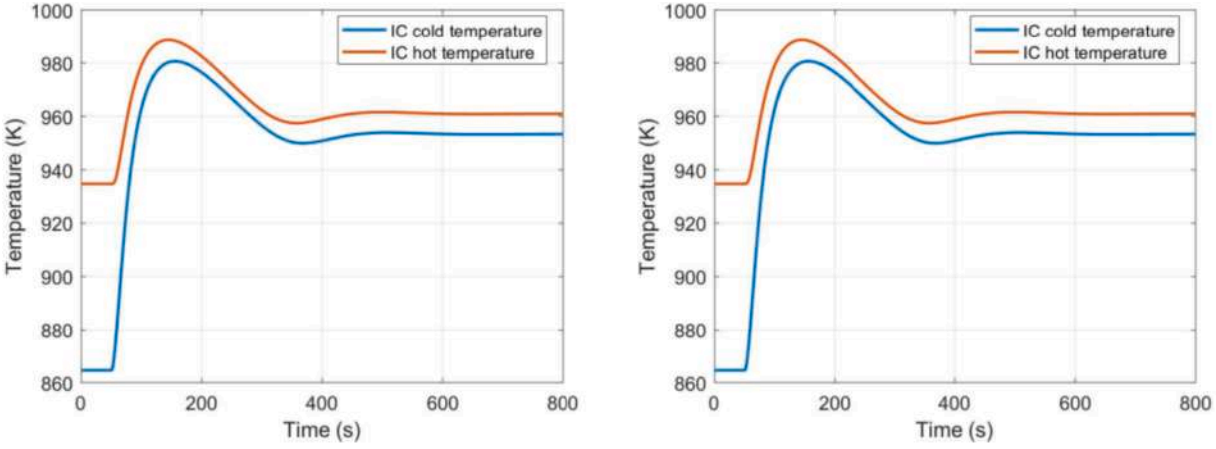


Fig. 3.7. Evolution of the IC temperatures (left) and FC temperatures (right) following an exponential reduction of the gas mass flow rate.

reactor in serious accidental conditions, e.g., salt freezing or the fluid temperatures above the maximum allowable temperature for the structures. The study confirms the numerical limitations in the Modelica power plant simulator, which may not be able to complete the simulation to the requested final simulation time ($t = 800$ s), if the mass flow reductions are significant. This analysis suggests that relative variations ranging between −90 % and 0 % can induce significant variations in the monitored parameters and, thus, a change in the plant status that cannot be ascribed to "slight deviations". Despite the mass flow reductions sampled from the range [−90 %, 0 %] may still cause the numerical failure of the simulator, the range is judged conservative, as it includes also more important deviations from nominal conditions.

3.4. Database generation

With the simulation settings and the input parameter range defined in Section 3.3, the training dataset is constructed by sampling the input parameter space, running the MSFR power plant simulator, extracting the time-dependent monitored output parameters of interest and labelling them according to the classification rules detailed in Section 3.6.

It is worth emphasizing that at present, no MSFR power plant is in operation worldwide. Consequently, the amount of experimental or operational data available for such systems — particularly under off-normal or accidental conditions — is extremely limited. In this respect, the use of simulation-driven, physics-enhanced machine learning (ML) approaches (like the one proposed in this work) represents a pragmatic and effective way to mitigate the inherent lack of real measurements and plant-specific data. By relying on high-fidelity system models to synthetically reproduce a relatively broad spectrum of (artificial) transient and failure scenarios, the (data-driven) classifier can be trained to recognize physical trends that are consistent with reactor dynamics, even in the absence of real measured data (Destino et al., 2021). The resulting diagnostic model could then be *hybridized* with experimental information, within an effective physics-enhanced ML framework: in particular, it could be progressively refined and validated, as soon as relevant operational data become available, bridging the gap between (pure) model-based analysis and (pure) data-driven learning (Cicirello, 2909; Lye et al., 2025).

Since the computational cost of the system-level model is quite limited, about 90 s per transient simulation on a commercial laptop, and the input parameter space has only 3 dimensions, i.e., the three mass flow rates $\dot{m}_{FC}$, $\dot{m}_{IC}$ and $\dot{m}_{GC}$, a simple uniform sampling scheme is adopted for ensuring an adequate mapping of the space. Each sampled triplet is then employed to define the exponential reduction in the mass flow rates in the Modelica solver, assuming the system to be in free dynamics. The choice of ignoring the control systems is justified by the fact that the design of the automatic control scheme requires information on the natural plant behaviour. One of the intended applications of the fault detection and classification algorithm is to provide the system designers with suggestions on the control actions. Hence, the data-driven classifier should be first trained to recognise possible deviations from the nominal conditions in free dynamics, to provide an early warning to the control systems. After the definition of suitable control strategies and the corresponding actuators, the fault detection and

classification algorithm could then be trained also to handle deviations from nominal controlled conditions, thus providing an indication of possible failures in the control system.

The total number of training scenarios generated with the power plant simulator, sketched in Fig. 3.8, roughly amounts to 6000, while the number of test scenarios roughly amounts to 1750.

The inlet and outlet temperature signals are represented in Fig. 3.9 for the FC and IC. The figures show that all scenarios are in nominal conditions until t = 30 s, when the mass flow rates are suddenly perturbed. According to the value of the perturbation, the temperatures may evolve quite differently. The signals in green indicate the "safe" scenarios, i.e. the temperature stay within the safety limits, while the signals in orange, labelled as "deviations", overcome the minimum and/or maximum temperatures. The signals represented in red, labelled as "incidents", overcome the safety thresholds and do not reach the prescribed end simulation time at 800 s because of some numerical error in the simulator (e.g., when the salt temperature falls much below the freezing temperature). For these simulations, the signals are artificially extended from the time of occurrence of the numerical failure, t_{NF}, to the defined end simulation time. Since the simulated signals are meant to represent physical measurements, their values are modified such that they do not fall below 300 K (which would be unphysical) and they do not raise above 1500 K. Hence, any signal assuming a value exceeding these ultimate temperature limits is set to either 300 K or 1500 K.

During the operation of real nuclear power plants, different limits may be introduced, in practice, for the MPPs. For instance, proper "intermediate" thresholds may be considered for triggering the limitation and protection functions, which are actuated by the control and protection systems, respectively, to avoid reaching the extreme safety/structural limits mentioned above. Since the specific values of these settings have not been established yet for the MSFR NPP, we decided to distinguish the scenarios according to the safety limits defined by salt freezing and the structural integrity. This choice is also somehow consistent with the fact of considering the free dynamics evolution of the scenarios, which ignores the possible actions of the control systems.

3.5. Rules for the scenario classification

An appropriate set of output data should be provided to the kNN classifier, which is trained to distinguish small operational deviations from more serious, potentially severe, accidents. To this aim, it is required that the set of reference simulations performed to explore the plant states are pre-classified by labelling each scenario according to the evolution of the subset of output parameters chosen to monitor the state of the system.

This section outlines the criteria adopted for the classification of the scenarios. In general, a simulation is labelled as "safe" if the system accomplishes its mission, i.e., producing power within the safety margins of the plant, for the whole duration of the transient.

In any other scenario where the following condition occurs,

$$T_i(t) \notin [T_{min}, T_{max}] = [858\,K,\ 1373\,K],$$

the transient is indicated as "*unsafe*". In this case, the transients can be furtherly divided into the ones causing numerical failures in the simulator, the so-called "*deviations*", and the ones exceeding the safety limits but simulated up to the final simulation time, the so-called "*incidents*". The "*unsafe*" scenarios can be furtherly subdivided according to the specific physical phenomenon causing the anomaly, namely:

- *High Temperature in the Fuel Circuit* (HT-FC): the molten salt temperature exceeds T_{max} in the FC.
- *Low Temperature Fuel Circuit* (LT-FC): the molten salt temperature falls below T_{min} in the FC.
- *High Temperature Intermediate Circuit* (HT-IC): the molten salt temperature exceeds T_{max} in the IC.
- *Low Temperature Intermediate Circuit* (LT-IC): the molten salt temperature falls below T_{min} in the IC.

Clearly, in real operations also other scenarios can occur. For instance, additional monitored parameters may be observed, according to the level of detail of the plant design and the technological solutions employed for the monitoring.

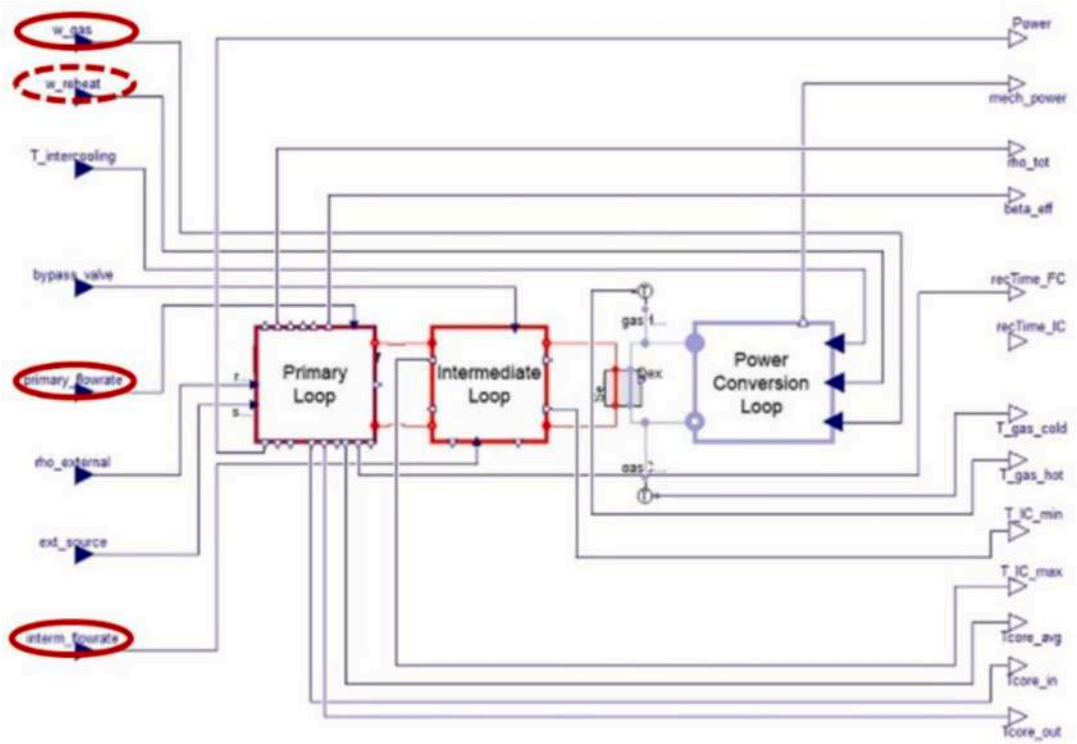


Fig. 3.8. Scheme of the Modelica power plant simulator. The red ellipses indicate the perturbed input parameters.

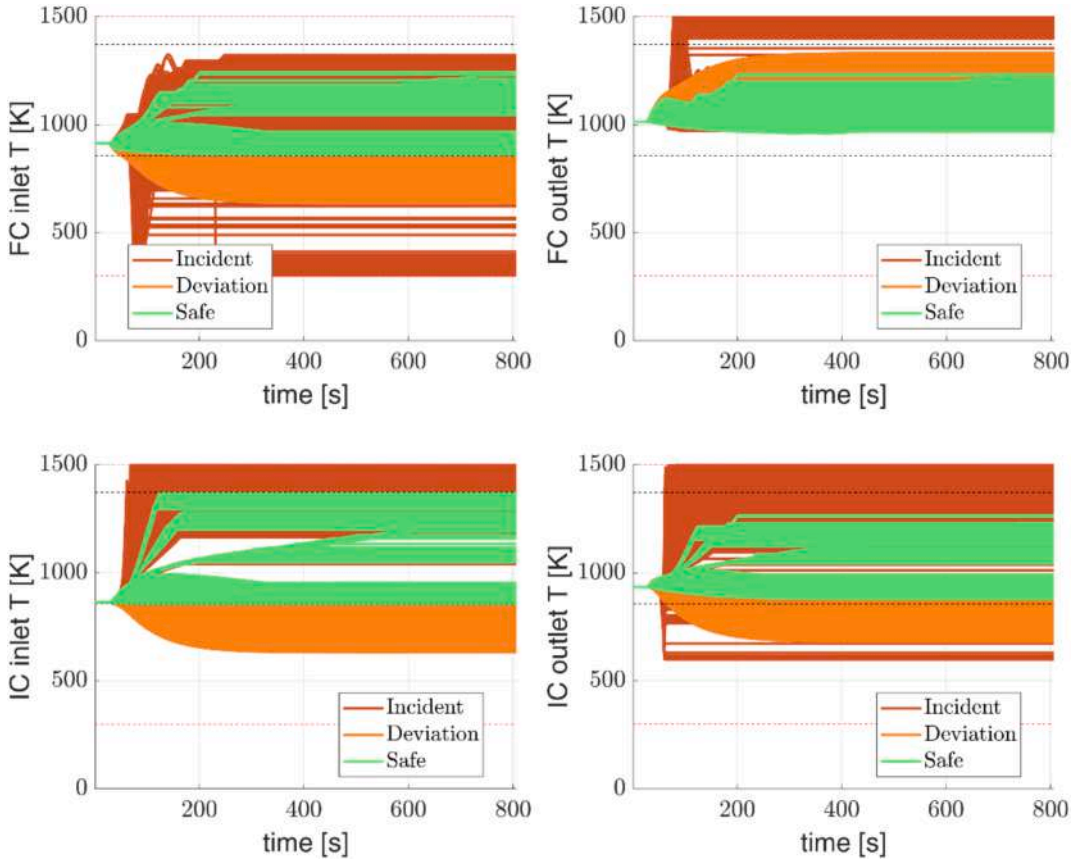


Fig. 3.9. Simulated inlet (left) and outlet (right) temperature signals for the FC (top) and IC (bottom).

The classification rules adopted for the supervised labelling of the training transients for the kNN classifier are summarised in Fig. 3.10. The first set of classes, addressed as C1 in the following, is very elementary and can be adopted for identifying the boundaries of the safe operational zone, regardless of the type of failure mode occurred. Hence, the SVD-kNN approach would work in *fault detection* mode. Differently from C1, C2 is not only oriented to the incident identification, but also to the *classification* of the *first* failure mode occurred during the transient. This choice is motivated by practical considerations: during the operation of the plant, the occurrence of one of the failure modes defined above could directly bring the plant to emergency shutdown or to a damaged state; from this perspective, the occurrence of successive multiple failures, which appear in the model, may not be relevant from a physical viewpoint during real plant operation.

Fig. 3.11 shows the distribution of the training scenarios according to the C1 classification with respect to the mass flow rate variation. These maps show that a well-defined "safe", conic-shaped region exists (green). This cone enlarges progressively when the FC and IC mass flow reductions occurs with a contemporary GC reduction. The physical explanation for this peculiar behaviour is that the amount of heat extracted by the energy conversion system diminishes with the reduction of the gas flow rate, helping both FC and IC to maintain the temperature difference within acceptable values, despite their mass flow reductions.

The map also provides useful insights on the distribution of the scenarios that suffer from some numerical issues. Specifically, they tend to cluster around the lower limit of the IC and GC mass flow reductions, which are associated with an almost complete loss of one or more circulation pumps.

The detailed type of physical failure detected in the "unsafe" scenarios can be inferred by inspection of the maps produced with the C2 classification. Most of these cases are classified as HT_{IC}, consistently with the mass flow reduction in this circuit. Most of the cases that are classified as "deviations" in C1 are classified here as LT_{IC}, i.e., the salt in the IC may reach the freezing temperature of the primary salt. The simultaneous inspection of C1 and C2 maps suggests a possible control action, aiming at reducing the GC mass flow rate when the LT_{IC} class is detected.

Fig. 3.10. Classification rules adopted in the training phase of the kNN algorithm.

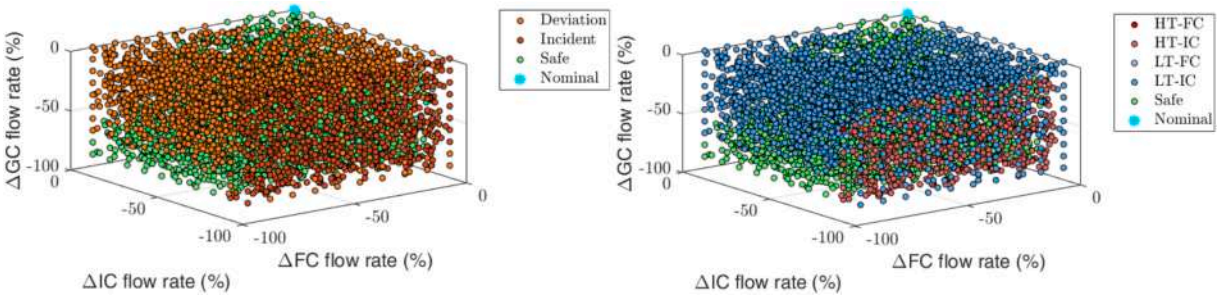


Fig. 3.11. Distribution of the training scenarios according to the C1 (left) and C2 (right) classification rules.

4. Model training

After the generation of the dataset, the SVD-kNN can be applied to build an efficient classifier. Since this application is intended to mimic a practical scenario, the classification is carried out by decomposing the temperature signals into incremental time windows, i.e. $\Delta t_i = (i+1)\,\Delta t$, with $i = 1, \cdots, n_T$ where Δt is the time step of the first window. This partitioning of the signals reflects the behaviour of a real acquisition system, which yields the measurements with a certain time delay and in a discrete time interval. Fig. 4.1 shows an example of the signal time binning for a time window width of 100 and 30 s, respectively.

The SVD-kNN algorithm is trained *iteratively* for each Δt_i, incrementing the length of the signal in time, with the objective of yielding a set of *time-dependent membership probabilities* which are supposed to become more accurate as time goes on. Exploiting this set of time-dependent probabilities, suitable control actions may be *promptly* carried out in a real plant operation.

In this section we show the results of the SVD-kNN application to the case of the MSFR plant. The first step of the training phase consists in the dimensionality reduction, which is achieved with the standard SVD algorithm. In this work we decided to apply the SVD independently to each temperature signal, which is cast in a suitable (m, n_T) matrix and then reduced to a set of basis functions, constituting a (p, nT) matrix, and a set of coefficients, cast in a (m, p) matrix. The effectiveness of the SVD for the data reconstruction after their reduction is visible in Fig. 4.2, where an example of the SVD decomposition of the temperature signals is reported, as well as the Percentage of Variance Explained (PVE), i.e., the information content in the reduced dataset, as a function of the number of basis functions employed in the data reconstruction.

It can be appreciated that a relatively low number of basis functions, around 7, is needed to achieve an acceptable value of the PVE for representing the full signal. The number of basis functions and coefficients clearly depends on the complexity of the signal. Since the whole signal is incrementally increased every Δt s, the number of basis functions changes according to the time instant considered. In order to ensure a consistent feature extraction, a PVE around 99.999 % was imposed for each SVD operation. In the case of real, measured signals, a larger number of basis functions should be expected, due to the presence of noise in the experimental measures.

After the reduction, the kNN classifier is trained with the signals from $t = 0$ to $t = \Delta t_i$. The hyperparameters of the kNN are optimised with a leave-one-out cross-validation process. Examples of the cross-validation errors at different time instants (for different numbers k of neighbours and different distance metrics) are reported in Fig. 4.3. For the first 30–40 s, the temperature signals are quite similar, so the classification error is quite large. As time goes on, the cross-validation error tends to stabilise between 3.5 % and 7 %. The analysis of the cross-validation error trend suggests that the optimal value of the hyperparameter k is always below 10. For this reason, only the points between 1 and 6 are reported. Concerning the choice of the metric, which clearly impacts the evaluation of the k nearest neighbours, the best option seems to be the "seuclidean", which indicates the standardized euclidean distance, while the "correlation" distance seems appropriate only when the transient is quite developed, i.e. after 50 s.

5. Model testing and evaluation

In this section, we present the main results concerning the testing phase of the time-dependent SVD-kNN algorithm. To evaluate the performances of the model, different figures of merits will be adopted in the following (including accuracy, recall and precision). The model will be tested also considering different thresholds for the assignment of the membership probabilities and adopting different time windows for the signal collection and discretisation.

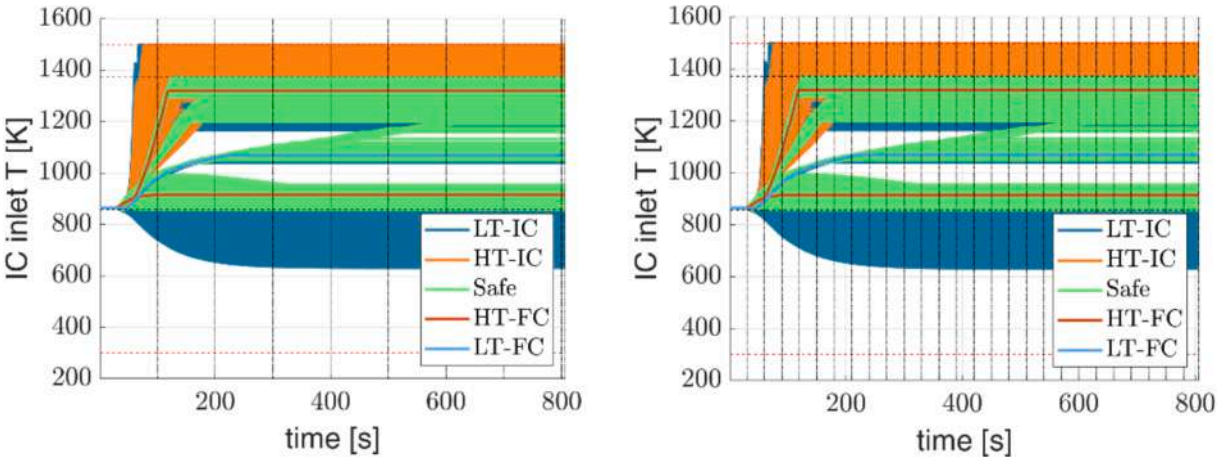


Fig. 4.1. Inlet temperature signals in the IC with a binning time of 100 s (left) and 30 s (right). The vertical dotted lines indicate the time windows, while the horizontal lines indicate the safety limits (in black) and the ultimate temperature limits (in red).

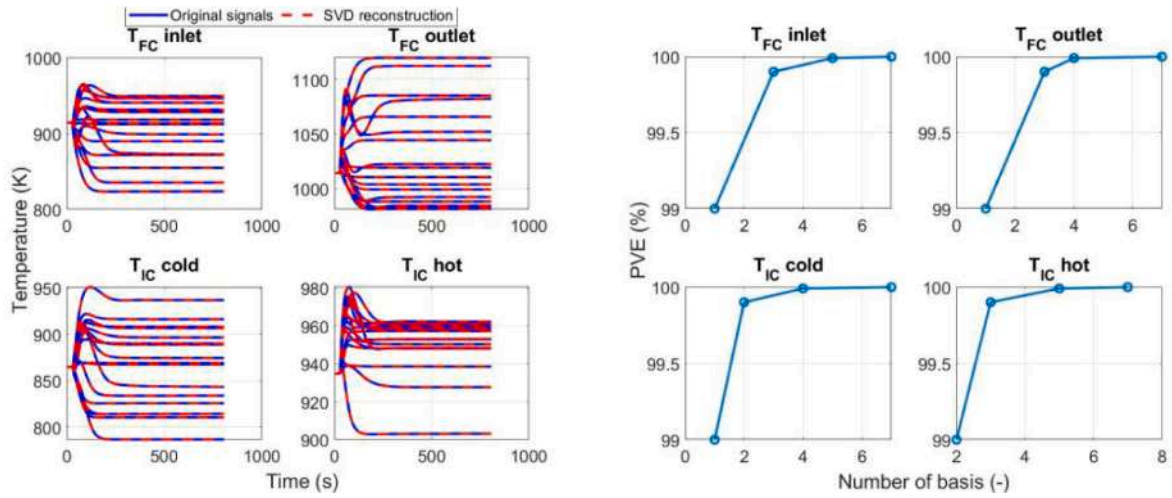


Fig. 4.2. Example of SVD reconstruction of the signals (left) and percentage of variance explained (right) according to the number of basis functions adopted in the reconstruction.

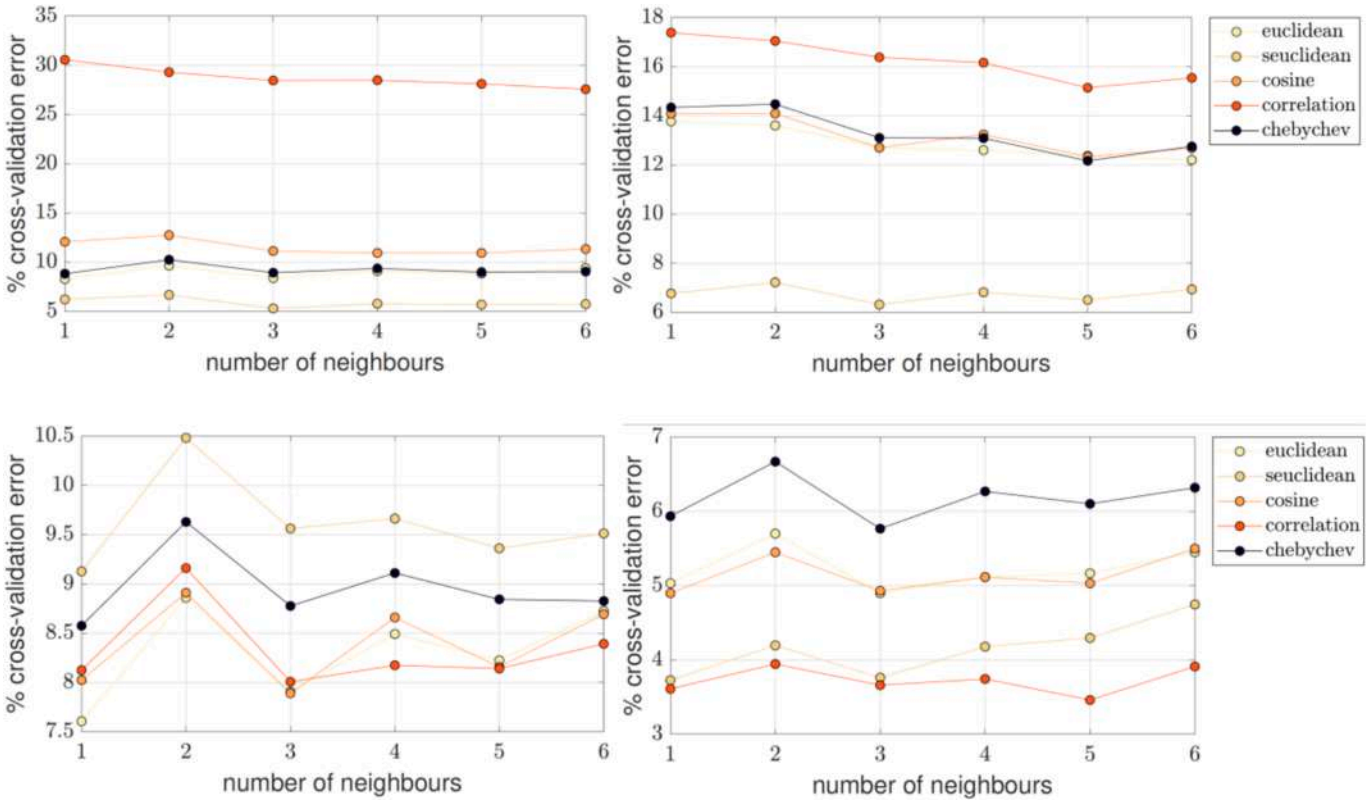
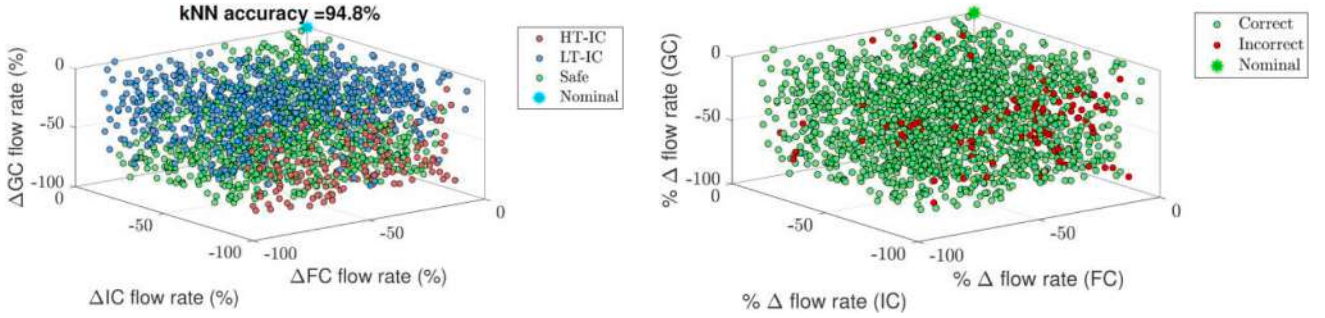


Fig. 4.3. Percentage cross-validation error for the signal classification at 32 s (top left), 50 s (top right), 90 s (bottom left) and 300 s (bottom right).

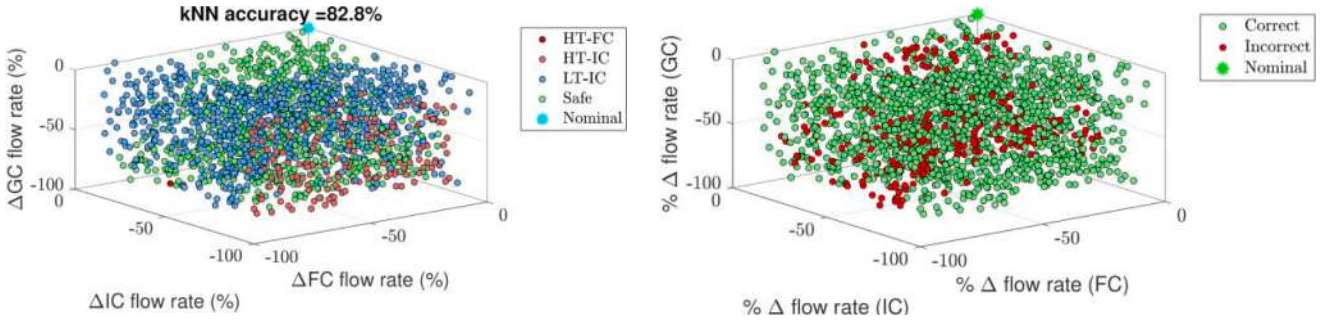
5.1. Prediction maps

The so-called "prediction maps" present the spatial distribution of the predicted classes in the simulations input space, similar to Fig. 3.11. These figures also report the *global accuracy* of the classifier, which is evaluated as the percentage of overall correct predictions referred to the total number of predictions (of safe and unsafe transients). In order to appreciate the distribution of the scenarios that are not classified correctly, we also report the so-called "misclassified maps" for highlighting the regions of the input space where the model performs well and poorly in terms of classification accuracy.

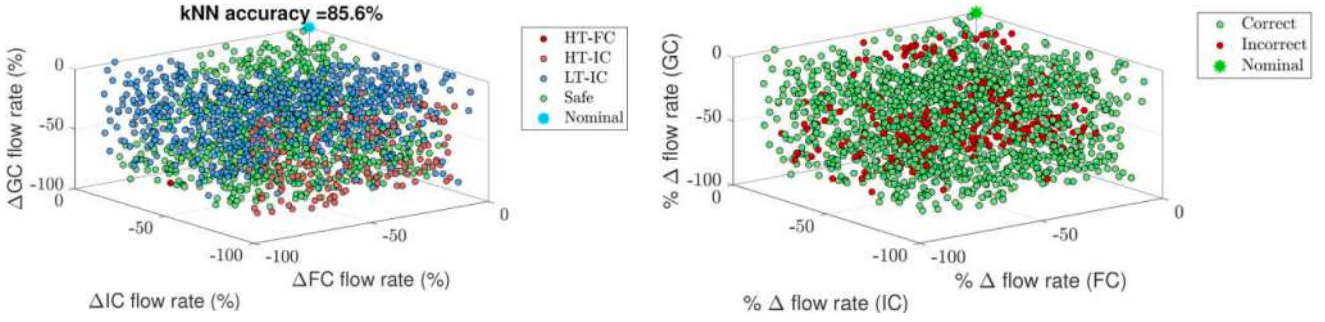
Fig. 5.1 shows the predicted distribution maps on the left and the misclassification maps on the right at some time instants. As one may expect, the accuracy is very large (94.8 %) at the beginning of the transients, when all scenarios are very close to the nominal state, but start departing towards other (possibly unsafe) states in peculiar ways, easily recognized by the classifier; then, it slightly deteriorates (82.8–85.6 %) as new signals (types and classes) are processed, due to the different evolutions featuring each scenario; finally, the overall accuracy increases again (93.2 %) when the (almost complete) transients are fully developed and made available to the classifier (i.e., at $t = 790$ s) The misclassification maps help to appreciate that the distribution of the scenarios that are not classified correctly also evolve in time: in particular, these maps indicate that the misclassified scenarios are roughly



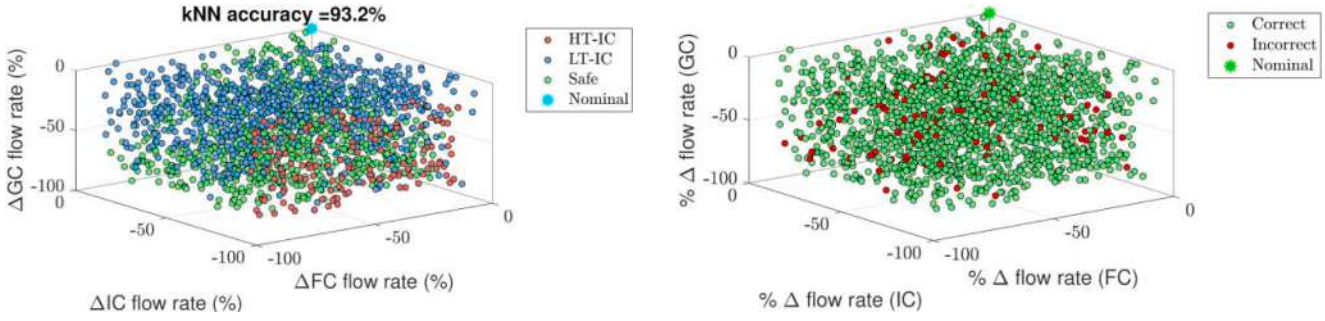
a) Maps at t = 34 s.



b) Maps at t = 150 s.



c) Maps at t = 402 s.



d) Maps at t = 790 s.

Fig. 5.1. Predicted distribution maps (left) and misclassification maps (right) for different time instants considering a time window of 2 s.

uniformly distributed in the parameter space, suggesting that the classifier is apparently unbiased.

5.2. Confusion matrix

The information conveyed by Fig. 5.1 is clearly not sufficient for judging appropriately the performances of the classifier. Therefore, we employ the "confusion matrix" charts to shed some light on the behaviour of the SVD-kNN algorithm. These charts introduce additional figures of merit for the assessment of the classification performances, i.e. the class-wise precision and recall. For a given class, each prediction of the classifier can be categorized as:

- True Positive (TP): the actual and the predicted class are the same.
- False Negative (FN): if the actual class is "positive" (in this case, "unsafe scenario"), the prediction is "negative" (in this case, "safe scenario").
- False Positive (FP): if the actual class is "negative" (in this case, "safe scenario"), the prediction is "positive" (in this case, "unsafe scenario").

A False Positive may be interpreted as a conservative classification from the point of view of system *safety*, but it is an undesirable feature from the point of view of the system *availability*. Conversely, the False Negative is dangerous for the safety of the plant. Exploiting these definitions, different Figures of Merit (FOMs) can be defined:

- *accuracy*, i.e. the percentage of correct true positive and true negative predictions referred to the total number of predictions (instances),

$$\frac{TP+TN}{TP+FP+TN+FP}.$$

In case some classes are more populated than others, a high accuracy may not be sufficient to fully qualify the model performances. For instance, in case the number of physical failures was very small, the accuracy could be close to unity even in case the classifier was not able to identify these cases, whose missing detection could have serious consequences for the plant safe operation.

- *recall*, i.e. the percentage of correctly predicted "positive" outcomes referred to the total number of "positive" instances, namely the sum of true positives and false negatives,

$$\frac{TP}{TP+FN}.$$

As mentioned, a FN for the "failure" state does not constitute an issue from the availability point of view, thus the recall may be useful for assessing the performances of the classifier from the plant safety perspective. Ideally, the higher the recall, the lower is the number of scenarios that can undermine the plant safety and that are not properly detected.

- *precision*, i.e. the percentage of true positive outcomes referred to the total number of "positive" predictions, i.e., true and false "positive",

$$\frac{TP}{TP+FP}.$$

Since FP events refer to a "failure" state, the higher is the precision, the lower is the number of spurious interventions of the safety system, increasing the plant availability.

Fig. 5.2 shows the confusion matrix charts evaluated at the time instants considered in Fig. 5.1. Each confusion matrix reports a square table where the rows indicate the "true" classes and the columns the predicted classes. The entries of the matrix represent the numbers of scenarios associated with each class with respect to the "true" one: thus, the confusion matrix of an ideal classifier should be diagonal. The rectangular matrices on the right and at the bottom of the square matrix represent the class-wise recall and precision, respectively. Therefore, moving along the *i*-th row it is possible to see the true positive outcomes on the *i*-th column and the false negatives (i.e., the outcomes which have been assigned to other classes instead of being assigned to the "correct" one) on the remaining columns. The first column of the right rectangular matrix is the recall of the *i*-th class, while the second column is the False Negative rate,

$$\frac{FN}{TP+FN} = 1 - \frac{TP}{TP+FN}.$$

Moving along the columns, it is possible to see the number of true positive outcomes on the *j*-th row and the number of false positives (i.e., the outcomes which have been assigned to the *j*-th class instead of being assigned to the other ones) on the remaining rows. Thus, the first row of the bottom rectangular matrix indicates the precision of the *j*-th class, while the second row indicates the False Positive Rate,

$$\frac{FP}{TP+FP} = 1 - \frac{TP}{TP+FP}.$$

With respect to classification rule C1 of Section 3.5 (i.e., simple fault/deviation/incident detection and identification of the boundaries of the safe operational zone, *regardless of the type of failure mode occurred*), the

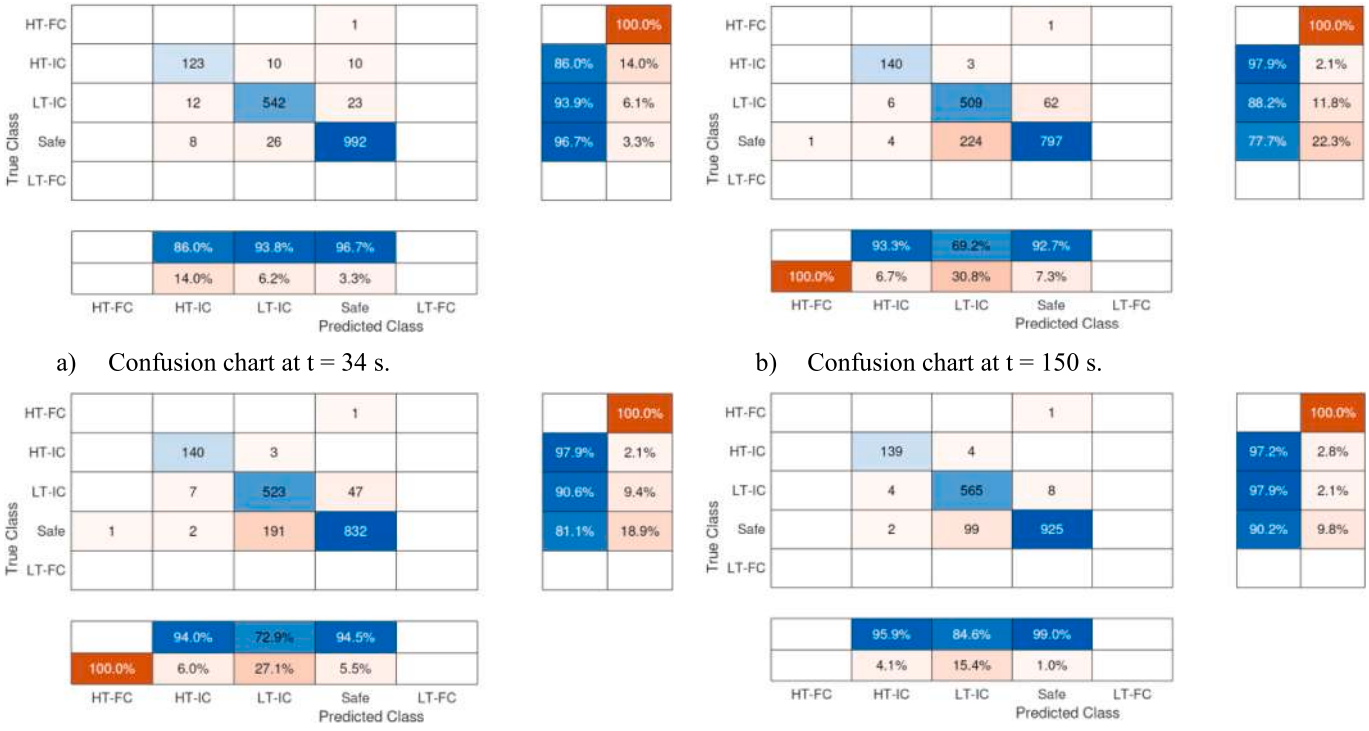


Fig. 5.2. Confusion charts for different instants of time considering a time window of 2 s.

capability of the classifier of discriminating between normal (i.e., "Safe") and abnormal (i.e., "HT-IC", "LT-IC", "HT-FC" and "LT-FC") behaviour is quite appreciable. Actually, the overall accuracy associated with such discrimination task remains high along the transient, i.e., 96.1 %, 83.3 %, 86.2 % and 93.7 % at $t = 34$, 150, 402 and 790 s, respectively.

Concerning the classification rule C2 (i.e., failure mode classification and diagnosis) and analysing the results from the point of view of plant safety, the recall values associated with the "unsafe" scenarios HT-IC and LT-IC are 86.0 % and 93.9 % ($t = 34$ s), 97.9 % and 88.2 % ($t = 150$ s), 97.9 % and 90.6 % ($t = 402$ s), 97.2 % and 97.9 % ($t = 790$ s), respectively. Thus, the performance of the classifier in classifying failure scenarios remains satisfactorily high along the entire time window considered. Analysing the results from the point of view of plant availability, after the first 30–40 s, the precision values associated with the HT-IC and "safe" states become very high: in particular, they reach 86.0 % and 96.7 % ($t = 34$ s), 93.3 % and 92.7 % ($t = 150$ s), 94.0 % and 94.5 % ($t = 402$ s), 95.9 % and 99.0 % ($t = 790$ s), respectively. Instead, the precision related to the LT-IC scenario may not be considered fully satisfactory with respect to the objective of maximizing plant availability. The precision values range from 69.2 % to 93.8 %, which is mainly due to the fact that many scenarios predicted as "LT-IC" are in facts "safe".

Conversely, the assignment of some false positives to HT-IC instead of LT-IC (e.g., at $t = 34$ s) may be still considered a satisfactory performance from the perspective of the fault/deviation/incident detection rate (i.e., the classifier is still able to correctly discriminate "unsafe" and "safe" scenarios). However, it may represent an issue with respect to the fault diagnosis and system control tasks, since the protection and mitigation actions introduced to bring the plant back to its nominal operational state may be very different for these two scenarios. Hence, various different practical implications should be carefully taken into account when judging the overall performances of the classifier. Similar considerations can be drawn for the recall. For instance, the recall values associated to HT-IC and LT-IC scenarios are quite large over the entire time window considered (ranging between 86.0 % and 97.9 %), while the one featuring the "safe" scenarios is comparatively low at some time instants during the transients (e.g., it equals 77.7 % and 81.1 % at $t =$ 150 and 402 s, respectively): this may be considered *conservative* from the point of view of plant safety, but it can be detrimental for the plant availability. Nevertheless, the values of both recall and precision are generally quite large, suggesting that the kNN is appropriate for the detection and classification of the plant status (and the corresponding faults/deviations/incidents).

Fig. 5.3 shows the instantaneous and mean values of the three figures of merit computed using two time windows of 2 s (left) and 4 s (right) for the signal scoring. The FOMs are evaluated by classifying the scenarios without imposing any constraint on the minimum value of the membership probabilities provided by the kNN classifier. The mean value is defined as the time-averaged FOM over a certain time T,

$$\bar{f}(T) = \frac{1}{T}\int_0^T dt f(t).$$

After about 200 s, the time-averaged FOMs are very close to their asymptotic values. The asymptotic averaged accuracy and recall raise up to about 85 %, while the asymptotic averaged precision is about 75 %. These values are satisfactory from the perspective of the safety of the plant, despite the rate of FP, which is about 15 %, could reduce the availability of the plant.

5.3. Impact of the confidence level

In this section, we assess the performances of the model both in terms of the impact of the confidence level required to accept a classification and of the time required for obtaining a correct detection.

In this respect, Fig. 5.4, which shows the distribution of the detection times featuring correct classifications, provides an interesting information: the algorithm is able to correctly classify online most of the cases very *promptly* (at $t = 30$–40 s, i.e., within 3–10 s from the beginning of the transient itself, with a very high peak located at $t = 30$–35 s). This is of paramount importance to ensure *early warning* in critical situations for the MSFR, which is the main objective of the proposed approach.

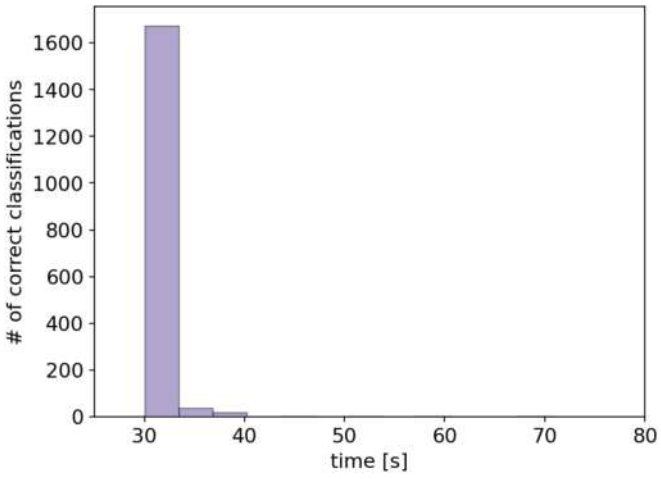


Fig. 5.4. Distribution of the detection time for correct classifications using the signals scored every 2 s.

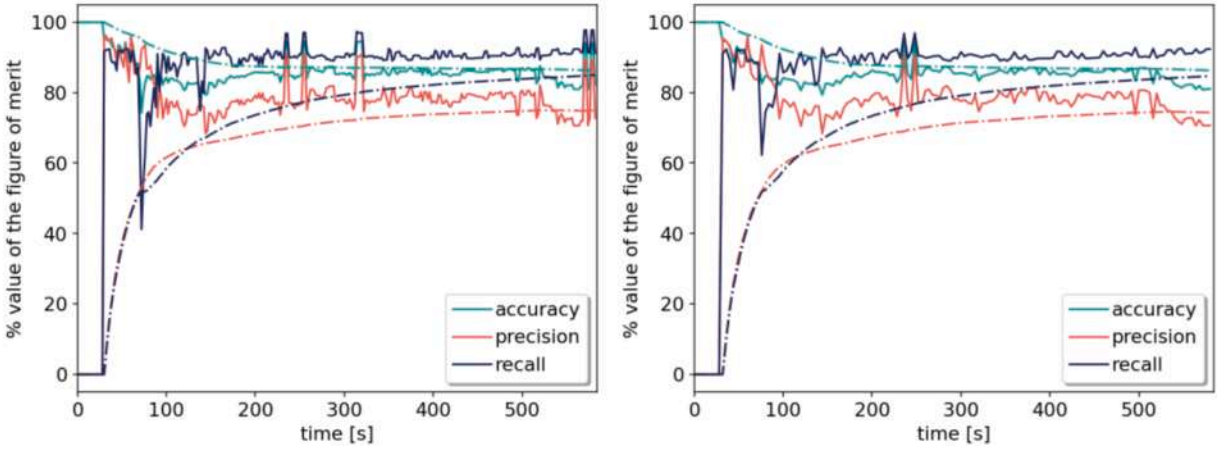


Fig. 5.3. Time-dependent behaviour of the accuracy, precision and recall using $\Delta t = 2$ s (left) and $\Delta t = 4$ s (right) until the steady state is reached.

As discussed in Section 2, the k-NN classifies the events with respect to the maximum of the membership probabilities, say P_{max}, disregarding its absolute value. However, in order to guarantee *robustness* in the fault/deviation/incident detection and diagnosis and to ensure that the answer provided by the classifier contains a certain level of "confidence", some thresholds (p_{thresh}) can be introduced in the classification process. In this respect, the transients can be assigned to a given class only when $P_{max} \geq P_{thresh}$. Figs. 5.5–5.8 show the number of correctly classified scenarios and the corresponding FOMs computed using only the data characterised by a predefined minimum level of "confidence" (p_{thresh}) in each time window. The figures on the left are realised by using the signals scored every 2 s, while the figures on the left refer to the signals scored every 4 s. As expected, the number of correctly classified scenarios grows in time, and it is inversely proportional to the threshold probability, reaching its nominal value after about 15 s from the operational perturbation. It is also important to note that the performance of the classifier is excellent even when very high levels of confidence are requested: for example, even for $p_{thresh} = 0.9$, the percentage of correct classifications remains above 90 % and 80 % for the two different time windows considered, i.e., 2 s and 4 s, respectively.

Conversely, the FOMs, which are evaluated with an increasingly larger pool, tend to decrease reaching asymptotic values between 95 % and 97 %. These values seem quite independent with respect to the threshold probability used to perform the classification, suggesting that the membership probabilities yielded by the SVD-kNN algorithm become soon quite reliable. The time variation in the FOMs is more evident for higher probability thresholds.

Another possible approach to increase the reliability and the trustworthiness of the kNN-SVD consists in accepting the proposed classification only if the class assigned by the algorithm does not change for a certain number of time steps: in other words, for the transient to be assigned to a given class, the SVD-kNN algorithm should "confirm" the answer for a predefined number of validation time steps. The results of this acceptance criterion are illustrated in Fig. 5.9, which reports the number of scenarios that are correctly classified as a function of the corresponding detection and classification time, for different number of validation time steps (from 3 to 9). As expected, increasing the number of "validation" time steps has the overall effect of shifting the distributions towards larger detection times and of reducing the height of the peaks. The figure also suggests that the shift towards the right in the distribution is also proportional to the time step considered for scoring the input signals (i.e., 2 and 4 s).

Fig. 5.10 shows the distribution of the detection times obtained by combining the two acceptance criteria, i.e., threshold probabilities (or minimum level of confidence for the classification) and number of required validation timesteps. In order to ease the comparison between the different cases, each histogram shows six batches of bars per detection time bin, one for each combination of threshold probabilities and validation timesteps. Quite reasonably, the peaks tend to assume lower height for large numbers of validation time steps and higher threshold probabilities. However, it is very important to notice that the distribution *shapes* and the *location* of the corresponding peaks are quite robust (i.e., not very sensitive) to the different combinations of threshold probabilities and number of validation time steps. For example, for signals scored every 2 s (left) the detection time is still peaked around 40–50 s for almost all the combinations tested (i.e., the correct detection and classification of the transients occurs after about 10–20 s from the system perturbation). Instead, for signals scored every 4 s (right), many peaks are located between 45 and 70 s (i.e., the correct detection and classification of the transients occurs after about 15–30 s from the perturbation).

5.4. Performances of the fault diagnosis model

In this section, the performances of the fault diagnosis model are evaluated assuming that the kNN-SVD works prescribing both a minimum probability threshold p_{thresh} of 75 % and a number of minimum timesteps $n_{T,\ thresh}$ equal to 5 and considering the signals scored every 2 s. Fig. 5.11 and Fig. 5.12 show the empirical distributions of the three mass flow rates perturbations when the scenario "Safe" and "LT-IC" are obtained, respectively.

By exploiting these data, which are part of the training database, and the membership probabilities, the Total Probability equation yields the distributions of the physical perturbations referring to the specific scenarios under analysis. Fig. 5.13 shows an example of the empirical, posterior distribution obtained for a testing scenario featured by a significant perturbation in both the IC and GC flow rates. It can be appreciated how the mode of the distributions falls in correspondence of the actual values of the physical perturbations, reported in the figure.

6. Comparison of the proposed approach with state-of-the-art methods applied to MSFRs

To better position the proposed methodology within the current state of the art, Table 1 its performance metrics with those of other Machine Learning-based approaches recently developed for fault detection and classification tasks in Molten Salt Reactor (MSR) systems (see Section 1). The approach developed in this work (based on a time-dependent kNN

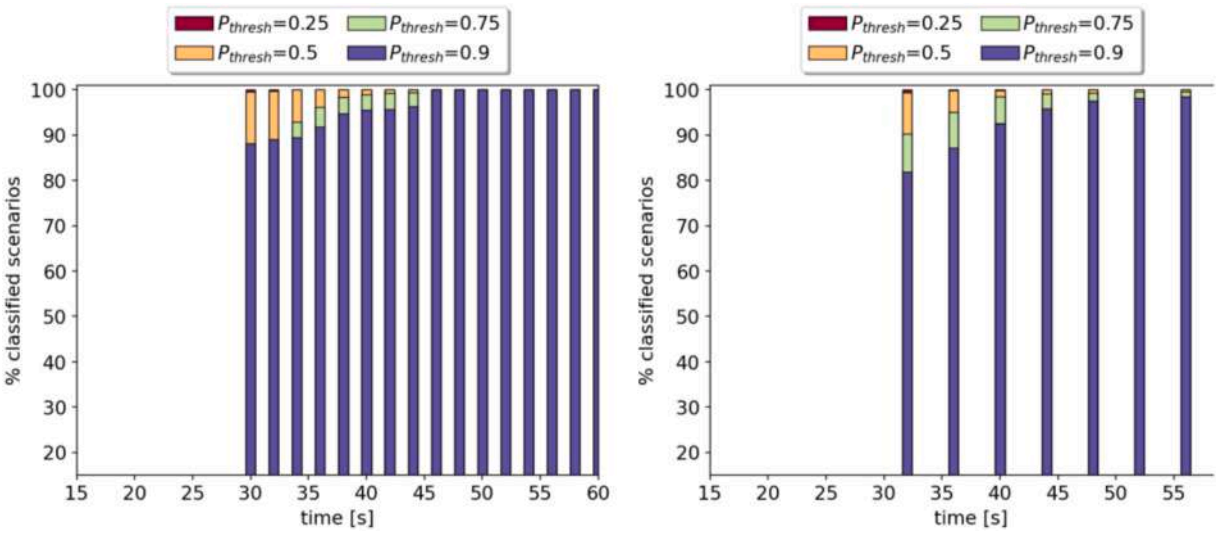


Fig. 5.5. Percentage of classified scenarios (left) for a given number of classified events (right) using the signals scored 2 s (left) and 4 s (right).

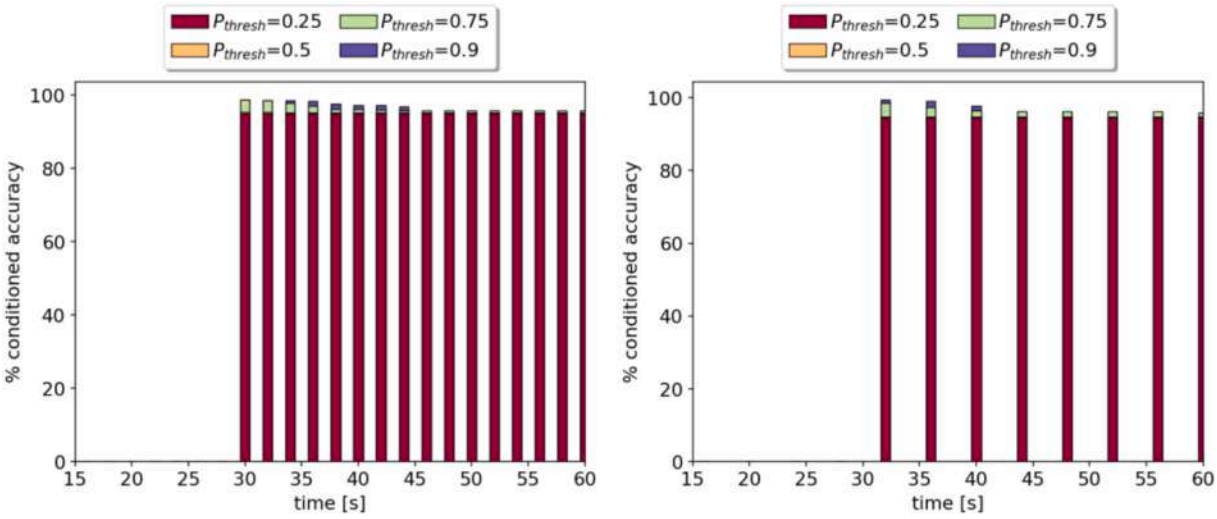


Fig. 5.6. Percentage accuracy for a given number of classified events (right) using the signals scored every 2 s (left) and 4 s (right).

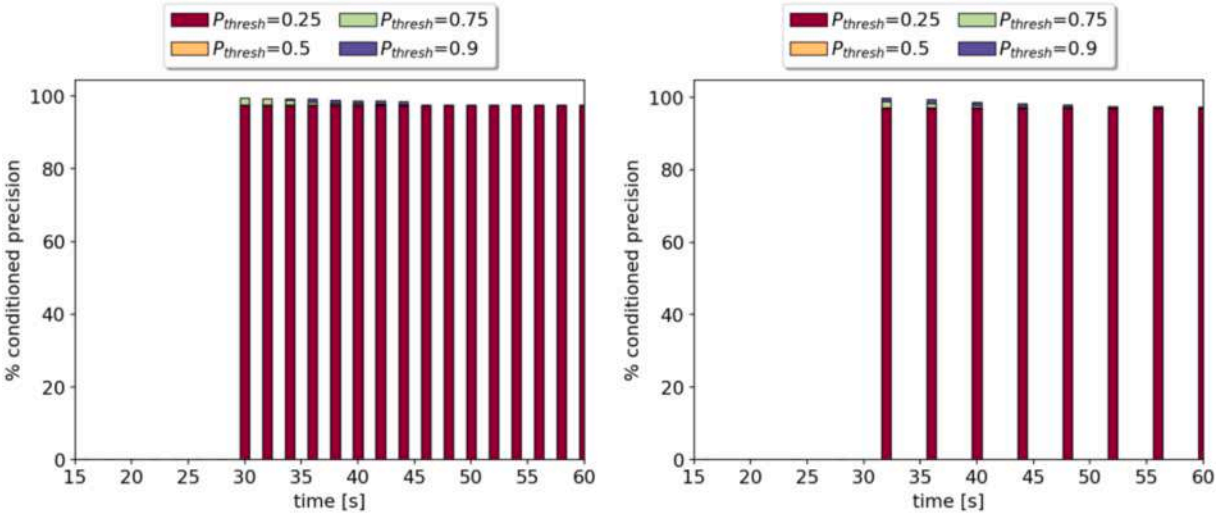


Fig. 5.7. Percentage precision for a given number of classified events (right) using the signals scored every 2 s (left) and 4 s (right).

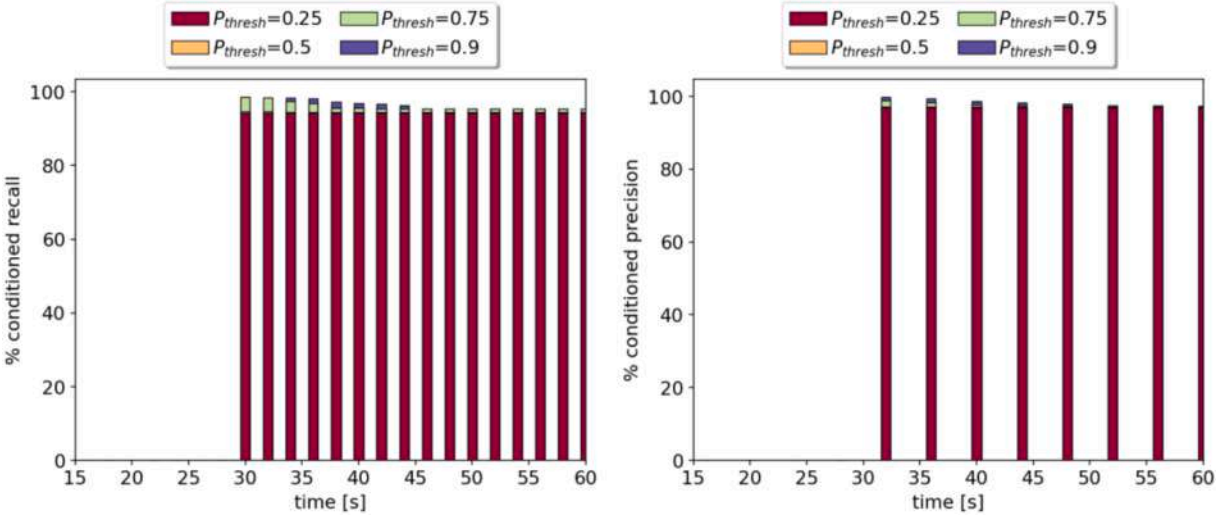


Fig. 5.8. Percentage recall for a given number of classified events (right) using the signals scored every 2 s (left) and 4 s (right).

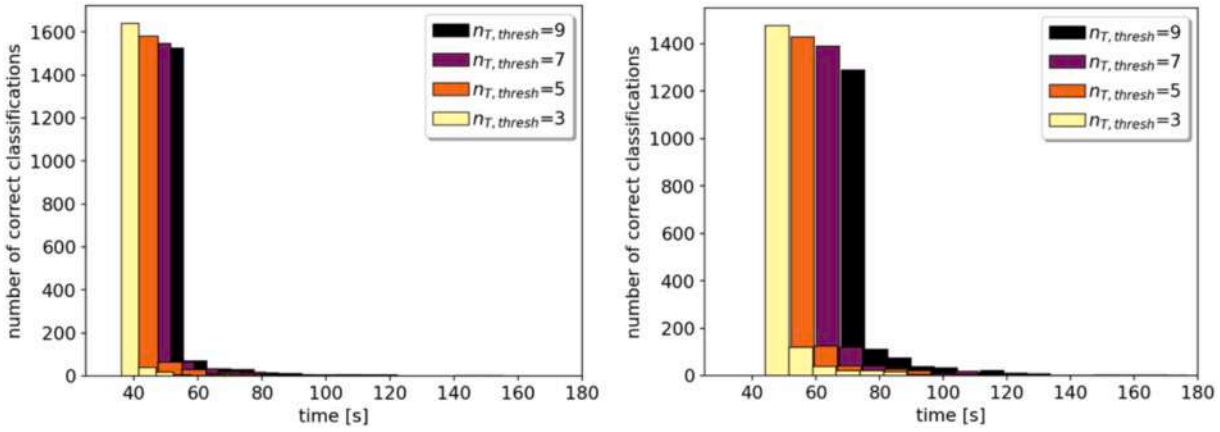


Fig. 5.9. Distribution of the detection time for the correct classifications considering different numbers of validation time steps using the signals scored every 2 s (left) and 4 s (right).

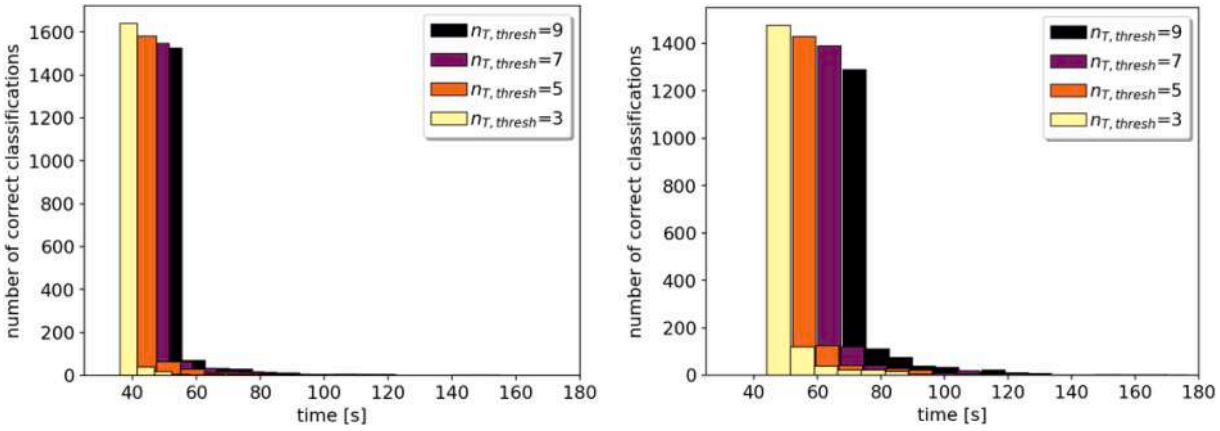


Fig. 5.10. Distribution of the detection time for the correct classifications considering two thresholds (0.25 and 0.75) for the assignment probabilities and different numbers of validation time steps using the signals scored every 2 s (left) and 4 s (right).

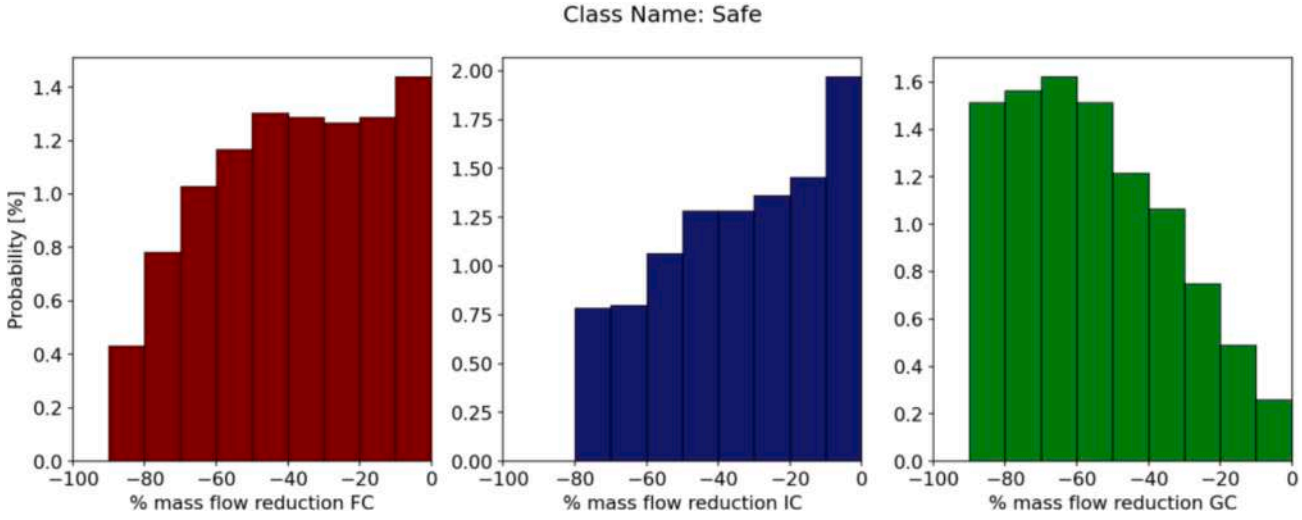


Fig. 5.11. Empirical distribution of the input parameters when the class is "Safe".

classifier coupled with Singular Value Decomposition-SVD for dimensionality reduction) performs comparably to (or even better than) a broad spectrum of methodologies of literature. In this respect, it is worth noting that such benchmark approaches range from relatively simple and classical algorithms, e.g., Principal Component Analysis (PCA) (Zhou and Hou, 2022), Decision Trees (DTs), Random Forest (RF) and Support Vector Machines (SVMs) (Prantikos et al., 2023), to much more sophisticated architectures, including ensembles of ML tools (Prantikos et al., 2023), Neural Networks (NNs) (Zhou et al., 2023), Bayesian Knowledge Graphs (Jiang et al., 2023), and Data-Driven rReduced

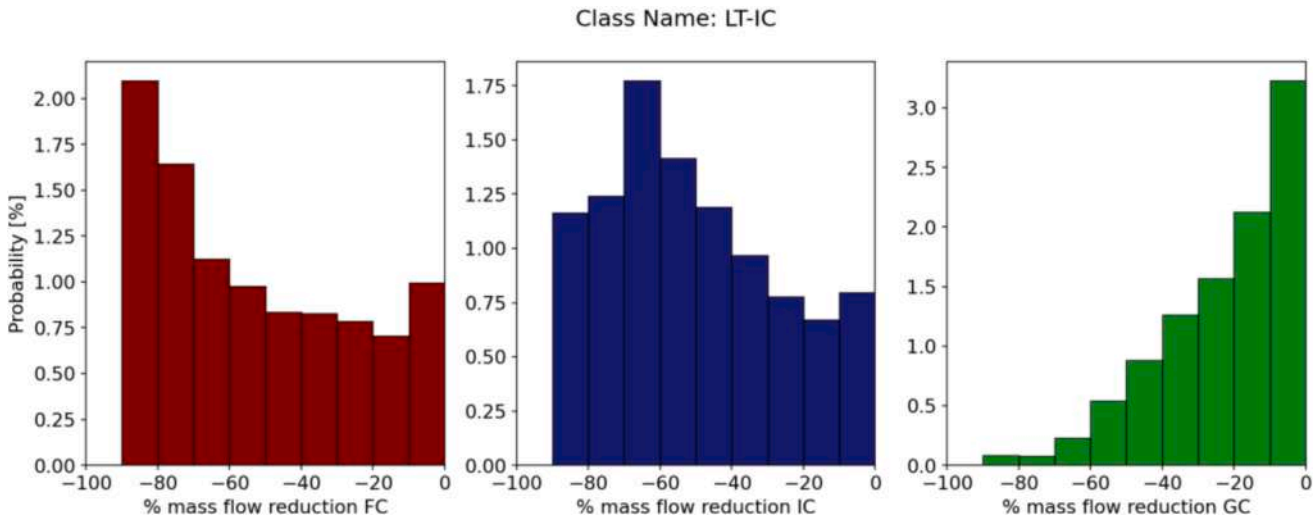


Fig. 5.12. Empirical distribution of the input parameters when the class is "LT-IC".

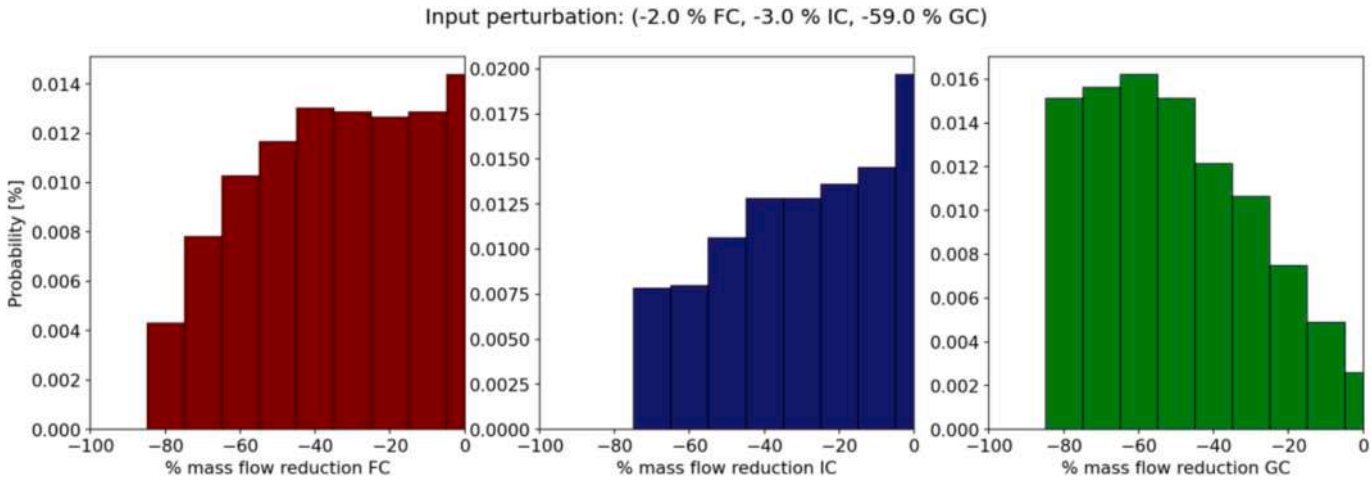


Fig. 5.13. Empirical distribution of the input parameters for a scenario belonging to the testing dataset.

oOrder mModelling (DDROM) relying on the Tikhonov-regularised Generalised Empirical Interpolation Method (TR-GEIM) and the Parameterised-Background Data-Weak formulation (PBDW) for model reduction, and on data-driven Gaussian Process Regression (GPR) for retrieving the missing information (or correct the wrong information) of faulty sensors (Riva et al., 2025). These very satisfactory performances represent a strong statement in favour of the proposed framework, which addresses an important gap in the safety analysis and monitoring of Generation IV reactor concepts, particularly MSFRs, by proposing an *effective* and *computationally efficient real-time* tool capable of detection, classification and probabilistic root-cause diagnosis (inference). A final word of caution is in order with respect to the information presented in Table 1. Although it certainly allows to position our approach within the state of the art, it must be acknowledged that a direct and precise comparison of the performances is not easy due to the possibly different plants/systems/components, operational phases and algorithmic parameters' configurations considered in the other works of literature.

7. Conclusions and future perspectives

The paper shows the development of a fault/deviation/incident detection method that could support the deployment of the MSFR technology, with the ~~ultimate~~ goal of providing early warning and fault classification of anomalies potentially arising during the operation, thus ultimately ensuring the safe and reliable operation of the MSFR power plant. The research work presented mainly focuses on operational scenarios involving some deviations from normal operating conditions.

After presenting and defining a set of suitable control and controlled variables for the plant online monitoring, i.e. the main circuit mass flow rates and the inlet and outlet temperatures of the primary and secondary circuits, respectively, a set of simulations representative of some possible abnormal states of the MSFR power plant has been performed exploiting a Modelica-based power plant simulator, as a physically consistent surrogate in the absence of experimental and operational data. The results of these simulations have been then discretised into time windows, to emulate the response of a real acquisition system, though neglecting the noise featuring real-life signals. The signals received at each Δt_i have been then employed to optimally train a kNN data-driven classifier, coupled with a time-dependent SVD algorithm for dimensionality reduction, considering different classification rules. The performances of the classifier turned out to be very satisfactory in terms of: i) detection and classification capabilities (measured by different figures of merit like accuracy, recall, precision and time of correct fault detection and classification); and ii) reduced computational cost, allowing in principle to exploit measurable system parameters and variables for a continuous, online plant monitoring. In particular, it has been shown that: i) the time-dependent accuracy in discriminating between normal and abnormal behaviour (fault detection mode) ranges

Table 1
Comparison of the proposed approach with fault detection and classification methods recently developed for Molten Salt Reactors.

Reference	Method/classifier	System/case study	Reported performance	Remarks
This work	Time-dependent kNN classifier coupled with SVD for dimensionality reduction	MSFR power plant (synthetic transient data)	Detection accuracy: 83.3–96.1 %; Classification accuracy: 82.8–94.8 %;Recall (unsafe cases) : 86.0–97.9 %; Precision: 69.2–93.8 %; Correct detection within 3–10 s from transient onset	Robust and computationally efficient; based on physically consistent simulated data; noise-free signals;flexible and reliable across classification rules and time-window widths
W. Zhou and J. Hou (2022) (Zhou and Hou, 2022)	PCA and contribution analysis	Fault isolation in MSR systems	Classification accuracy ≈ 90 %; effective feature contribution analysis	Focused on dimensionality reduction and interpretability; limited dynamic response handling
X. J. Jiang, et al. (2023) (Jiang et al., 2023)	Knowledge graph and Bayesian inference	Control rod drive mechanism (CRDM) fault diagnosis	Qualitative reasoning; probabilistic inference with > 90 % correct fault identification	Strong knowledge-based reasoning;limited quantitative validation
T. Zhou et al. (2023) (Zhou et al., 2023)	Neural Networks (various architectures tested)	MSR transient identification	Accuracy up to 95 %; good generalization on synthetic data	Includes multiple fault classes and transient types; computationally heavier than kNN-type classifiers
S. Riva, et al. (2025) (Riva et al., 2025)	Data-Driven Reduced Order Modelling (DDROM) and malfunctioning sensor recovery	MSR (loop-type) – Reduced Order Model (TR-GEIM and PBDW)	Accuracy ≈ 85–95 % for sensor recovery and fault identification	Emphasizes robustness against missing/noisy data;combines ROM (TR-GEIM and PBDW) with ML (in particular, Gaussian Process Regression)
K. Prantikos, et al. (2023a) (Prantikos et al., 2023)	Decision Tree, Random Forest (RF), SVM, kNN, Neural Networks	MSR heat exchanger channel plugging detection and localization	Decision Tree and RF best: accuracy > 95 % on synthetic data with noise	Focused on single-component diagnostics; includes robustness tests to sensor noise
K. Prantikos, et al. (2023b) (Prantikos et al., 2023)	Ensemble of ML classifiers	MSR heat exchanger, maintenance prediction	Accuracy ≈ 90–97 % (depending on feature set)	Comprehensive comparison of ML classifiers under uncertainty

between 83.3 % and 96.1 %, whereas the overall accuracy in fault classification varies between 82.8 % and 94.8 % along the transients; ii) the time-dependent recall values (i.e., the class-wise performance) associated to the "unsafe" scenarios (HT-IC and LT-IC) vary between 86.0 % and 97.9 %, which is a very promising result in order to ensure plant safety; iii) the time-dependent precision is very high (consistently above 90 %) for the "HT-IC" and "safe" states, whereas it may require improvements in some time windows (69.2–93.8 %) for the "LT-IC" state, due to some false positives: this may be conservative for the plant safety, but detrimental for the plant availability; iv) most of the correct detections and classifications are performed by the kNN very promptly, within 3–10 s from the beginning of the transients themselves (i.e., from the system perturbation). Most important, a thorough sensitivity study has shown that the performance of the SVD-kNN algorithm is *reliable*, *robust* and *not very dependent* on the classification rules adopted, on the width of the discretisation time windows, on the classification confidence imposed and on the number of validation time steps required to confirm the assignment. Also, an accurate literature review has shown that the approach developed in this work performs comparably to (or even better than) a broad spectrum of state-of-the-art methodologies recently proposed for fault detection and classification tasks in MSFR systems (ranging from relatively simple and classical tools, like Principal Component Analysis, Decision Trees, Random Forest and Support Vector Machines, to more sophisticated algorithmic schemes, including ensembles of ML models, Neural Networks and hybrid Data-Driven Reduced Order Models). All these aspects represent strong statements in favour of the flexibility and applicability of the proposed approach to realistic problems.

In spite of the satisfactory results obtained, many aspects of the proposed methodology are worthy of further developments. First of all, although the reduced computational cost associated with the present power plant simulator allows the adoption of traditional a brute-force sampling, more efficient techniques could be employed for the intelligent exploration of the MSFR state space (e.g., *adaptive* sampling methods combined with fast-running reduced order models or meta-models), especially in view of adding more control and controlled parameters to the analyses.

Also, although the method assigns class probabilities (which still allows to build a desired level of confidence in the classification process), it does *not* quantify several sources of uncertainty possibly affecting: (i) the underlying (possibly noisy) *measurements* and the model *input parameters*; (ii) the system *model output estimates* (due to internal model errors, discrepancies, approximations and lack of knowledge of the analyst on some relevant phenomena); and (iii) the *predictions* of the (kNN) *classification algorithm* itself. In this respect, the robustness of the proposed methodology could be improved by consistently and rigorously quantifying such uncertainties by means of, e.g., (nested) non-parametric bootstrapped kNN ensembles (Villa Medina and Boqué, 2009), bagged Bayesian approaches (Sugahara and Aomi, 2022), or combinations of kNN classifiers with generalised uncertainty representation frameworks (e.g., fuzzy or imprecise probability theories for uncertain class assignments) (Salem et al., 2022).

In addition, the model should be extended and improved to consider, e.g., that the signals coming from a real data acquisition system do not represent bulk properties of the fluids, but rather they are localised, featured by different time delays and affected by noise. These characteristics potentially weaken the signals' correlation with (time-dependent) accidental behaviors, thus making them more difficult to detect and classify. Also, the presence of significant noise may increase the fraction of false positives, with a detrimental effect on the discriminative capability and diagnostic power of the classifier (Chevalier-Jabet and Verma, 2024). These extensions and improvements could leverage on modern techniques available in the literature, such as de-noising and pre-classification approaches (Roma et al., 2021; Zhu et al., 2021; Roma et al., 2022; Zhu et al., 2025; Chevalier-Jabet and Verma, 2024).

Another possible development could be represented by the adoption of additional classifiers, which could be combined with the kNN within an *ensemble*, to increase the accuracy and robustness of the fault/deviation/incident detection and classification process, for instance with the adoption of majority voting. Finally, the fault/deviation/incident detection phase could be decoupled from the classification and diagnosis step. In particular, anomaly detection methods (based, e.g., on Deep Learning algorithms, Long Short-Term Memory Networks, Recurrent Neural Networks, Auto-Associative Neural Networks and AutoEncoders, or on statistical techniques including clustering, Finite Mixture Models, Isolations Forests, Local Outlier Factor, One-Class Support Vector Machines) could be specifically developed to decide whether the anomaly is actually due to an incidental condition, or it is just a spurious

operational fluctuation, in order to properly trigger (or not) the intervention of the fault/deviation/incident classification algorithm.

Finally, the proposed method should also be tested within a controlled-dynamics framework. Since the controller inherently acts to mitigate operational deviations (by trying to keep the critical physical variables within nominal ranges), its intervention is expected to systematically reduce the observable magnitude of anomalies, thus potentially decreasing the classification accuracy. However, this effect could be partially compensated by extending the training dataset to include closed-loop simulations, allowing the classifier to learn the modified system response under control actions. In this respect, if the classifier can promptly detect and classify the anomalous deviations despite the mitigative action of the controller, then the joint system can still be considered effective for early warning and diagnostic purposes. In perspective, a robust integration of control and diagnostic layers could be achieved through adaptive or hybrid schemes, in which the classifier continuously updates its decisions based on feedback signals or residual trends.

CRediT authorship contribution statement

N. Abrate: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **N. Pedroni:** Writing – original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **N. Caruso:** Software, Investigation, Data curation. **S. Dulla:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **S. Lorenzi:** Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project has received funding from the Euratom research and training programme 2014–2018 under grant agreement No. 847527. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and/or views expressed therein lies entirely with the authors.

The authors would like to thank their French partners from CNRS, Framatome and CEA for their precious feedback on this work during the SAMOSAFER project. The authors would also like to thank the reviewers for their precious suggestions, which have contributed to improve the quality of the paper.

Data availability

The complete datasets and scripts employed to pre- and post-process the calculations presented in this paper are available in the open access repository 4TU Research Data (DOI https://doi.org/10.4121/0ae20eee-97a6-4634-9f57-eb1887018fc2).

References

Al-Dahidi, S., Di Maio, F., Baraldi, P., Zio, E., Seraoui, R., 2018. A framework for reconciliating data clusters from a fleet of nuclear power plants turbines for fault diagnosis. Appl. Soft Comput. J. 69, 213–231. https://doi.org/10.1016/j.asoc.2018.04.044.

Allibert, M., et al., 2017. «SAMOFAR European Project D1.1 Description of initial reference design and identification of safety aspects», fasc. 661891.

Ayodeji, A., Liu, Y., Chao, N., Yang, L., 2020. A new perspective towards the development of robust data-driven intrusion detection for industrial control systems. Nucl. Eng. Technol. 52 (12), 2687–2698.

Baraldi, P., Di Maio, F., Genini, D., Zio, E., 2015. Comparison of data-driven reconstruction methods for fault detection. IEEE Trans. Reliab. 64 (3), 852–860. https://doi.org/10.1109/TR.2015.2436384.

Bell, G.I., Glasstone, S., 1970. Nuclear Reactor Theory. New York.

Boisseau, T., et al., 2022. SAMOSAFER D.6.2: List and description of the plant operational states with the corresponding safety margins, fasc. 847527.

Cammi, A., et al., 2018. SAMOFAR D1.4 Safety issues of normal operation conditions, fasc. November, pp. 1–69.

Chevalier-Jabet, K., Verma, L., Kremer, F., 2024. Using a surrogate model for the detection of defective PWR fuel rods. Ann. Nucl. Energy 209, 110779. https://doi.org/10.1016/j.anucene.2024.110779.

Choi, J., Lee, S.J., 2020. A sensor fault-tolerant accident diagnosis system. Sens. Switz. 20 (20), 1–17. https://doi.org/10.3390/s20205839.

Cicirello, A., 2024. Physics-enhanced machine learning: a position paper for dynamical systems investigations. J. Phys. Conf. Ser. 2909 (1), 012034. https://doi.org/10.1088/1742-6596/2909/1/012034.

Dai, Y., et al., 2023. An intelligent fault diagnosis method for imbalanced nuclear power plant data based on generative adversarial networks. J. Electr. Eng. Technol. 18 (4), 3237–3252.

Dai, X., Gao, Z., 2013. From model, signal to knowledge: a data-driven perspective of fault detection and diagnosis. IEEE Trans. Ind. Inform. 9, fasc. 4, 2226–2238. https://doi.org/10.1109/TII.2013.2243743.

de Pinedo, Á., et al., 2021. Functional outlier detection by means of h-mode depth and dynamic time warping. Appl. Sci. 11 (23), 11475.

Destino, V., Pedroni, N., Bonifetto, R., Di Maio, F., Savoldi, L., Zio, E., 2021. Metamodeling and on-line clustering for loss-of-flow accident precursors identification in a superconducting magnet cryogenic cooling circuit. Energies 14 (17), 5552. https://doi.org/10.3390/en14175552.

Dong, W., Moses, C., Li, K., 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In: Proceedings of the 20th international conference on World wide web, in WWW '11. New York, NY, USA: Association for Computing Machinery, mar., pp. 577–586. doi: 10.1145/1963405.1963487.

Farber, J.A., Cole, D.G., 2020. Detecting loss-of-coolant accidents without accident-specific data. Prog. Nucl. Energy 128, fasc. July, 103469. https://doi.org/10.1016/j.pnucene.2020.103469.

Feng, Z., Liang, M., Chu, F., 2013. Recent advances in time–frequency analysis methods for machinery fault diagnosis: a review with application examples. Mech. Syst. Sig. Process. 38 (1), 165–205. https://doi.org/10.1016/j.ymssp.2013.01.017.

Fiorina, C., et al., 2014. Modelling and analysis of the MSFR transient behaviour. Ann. Nucl. Energy 64, 485–498. https://doi.org/10.1016/j.anucene.2013.08.003.

Gerardin, D., Allibert, M., Heuer, D., Laureau, A., Merle-Lucotte, E., Seuvre, C., 2017. Design evolutions of molten salt fast reactor. Int. Conf. Fast React. Relat. Fuel Cycles 1–10.

Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W.-K., Yang, H., 2020. Status of research and development of learning-based approaches in nuclear science and engineering: a review. Nucl. Eng. Des. 359, 110479.

Hu, Y., Baraldi, P., Di Maio, F., Zio, E., 2017. A systematic semi-supervised self-adaptable fault diagnostics approach in an evolving environment. Mech. Syst. Signal Process., 88, fasc. February 2016, pp. 413–427, doi: 10.1016/j.ymssp.2016.11.004.

Huang, Q., et al., 2023. A review of the application of artificial intelligence to nuclear reactors: where we are and what's next. Heliyon 9 (3).

Jiang, X.-J., Zhou, W., Hou, J., 2023. Construction of fault diagnosis system for control rod drive mechanism based on knowledge graph and Bayesian inference. Nucl. Sci. Tech. 34 (2), 21.

Lass, O., Volkwein, S., 2014. Adaptive POD basis computation for parametrized nonlinear systems using optimal snapshot location. Comput. Optim. Appl. 58 (3), 645–677. https://doi.org/10.1007/s10589-014-9646-z.

Laureau, A., Bellè, A., Allibert, M., Heuer, D., Merle, E., Pautz, A., 2022. Unmoderated molten salt reactors design optimisation for power stability. Ann. Nucl. Energy 177, 109265. https://doi.org/10.1016/j.anucene.2022.109265.

Laureau, A., et al., 2017. D1 . 3 Development of a power plant simulator, fasc. August 2015.

Li, W., Peng, M., Wang, Q., 2018. Fault detectability analysis in PCA method during condition monitoring of sensors in a nuclear power plant. Ann. Nucl. Energy 119, 342–351. https://doi.org/10.1016/j.anucene.2018.05.024.

Liu, S., Zhou, X., Yu, J., Wang, Y., Xu, T., Wang, H., 2024. Graph attention network-based model for multiple fault detection and identification of sensors in nuclear power plant. Nucl. Eng. Des. 419, 112949.

Locatelli, G., Mancini, M., Todeschini, N., 2013. Generation IV nuclear reactors: current status and future prospects. Energy Policy 61, 1503–1520. https://doi.org/10.1016/j.enpol.2013.06.101.

Luo, H., et al., 2024. New RMC energy deposition treatment and its application in multi-physics simulation new RMC energy deposition treatment and its application in multi-physics simulation. Nucl. Sci. Eng. 1–21. https://doi.org/10.1080/00295639.2024.2316955.

Lye, A., Ong, T.K.C., Xiao, S., Chung, K.Y., 2025. Physics-enhanced machine learning for probabilistic risk assessment in nuclear safety: an overview, recent developments, and perspectives. Ann. Nucl. Energy 222, 111562. https://doi.org/10.1016/j.anucene.2025.111562.

Mandal, S., Santhi, B., Sridhar, S., Vinolia, K., Swaminathan, P., 2017. Nuclear power plant thermocouple sensor-fault detection and classification using deep learning and generalized likelihood ratio test. IEEE Trans. Nucl. Sci. 64 (6), 1526–1534. https://doi.org/10.1109/TNS.2017.2697919.

Mareboyana, M., Le Moigne, J., 2018. Super-resolution of remote sensing images using edge-directed radial basis functions, p. 35, doi: 10.1117/12.2303732.

Min, J.H., Kim, D.W., Park, C.Y., 2018. Demonstration of the validity of the early warning in online monitoring system for nuclear power plants. Nucl. Eng. Des., 349, fasc. November, pp. 56–62, 2019, doi: 10.1016/j.nucengdes.2019.04.028.

Muja, M., Lowe, D.G., 2014. Scalable nearest neighbor algorithms for high dimensional data. IEEE Trans. Pattern Anal. Mach. Intell., 36(11), pp. 2227–2240, doi: 10.1109/TPAMI.2014.2321376.

Nguyen, T.N., Downar, T., Vilim, R., 2020. A probabilistic model-based diagnostic framework for nuclear engineering systems. Ann. Nucl. Energy 149, 107767. https://doi.org/10.1016/j.anucene.2020.107767.

Nicolino, C., Lapenta, G., Dulla, S., Ravetto, P., 2008. Coupled dynamics in the physics of molten salt reactors. Ann. Nucl. Energy 35, fasc. 2, 314–322. https://doi.org/10.1016/j.anucene.2007.06.015.

Prantikos, K., Lee, T., Tsoukalas, L.H., Heifetz, A., 2023. Conceptual machine learning-based strategy for molten salt heat exchanger channel plugging detection and localization. In: Proceedings of the 2023 American Nuclear Society Annual Meeting, Indianapolis, IN, USA, pp. 11–14.

Prantikos, K., Lee, T., Heifetz, A., 2023. Machine learning classification of molten salt heat exchanger channel plugging using synthetic data. Argonne National Laboratory (ANL), Argonne, IL (United States).

Puppo, L., Pedroni, N., Maio, F.D., Bersano, A., Bertani, C., Zio, E., 2021. A framework based on finite mixture models and adaptive kriging for characterizing non-smooth and multimodal failure regions in a nuclear passive safety system. Reliab. Eng. Syst. Saf. 216, 107963. https://doi.org/10.1016/j.ress.2021.107963.

Qi, B., Liang, J., Tong, J., 2023. Fault diagnosis techniques for nuclear power plants: a review from the artificial intelligence perspective. Energies 16 (4), 1850.

Ramezani, I., Moshkbar-Bakhshayesh, K., Vosoughi, N., Ghofrani, M.B., 2022. Applications of soft computing in nuclear power plants: a review. Prog. Nucl. Energy 149, 104253.

Riva, S., Introini, C., Zio, E., Cammi, A., 2025. Data-driven reduced order modelling with malfunctioning sensors recovery applied to the Molten Salt Reactor case. EPJ Nucl. Sci. Technol. 11, 55. https://doi.org/10.1051/epjn/2025054.

Roma, G., et al., 2021. A Bayesian framework of inverse uncertainty quantification with principal component analysis and Kriging for the reliability analysis of passive safety systems. Nucl. Eng. Des. 379, 111230. https://doi.org/10.1016/j.nucengdes.2021.111230.

Roma, G., et al., 2022. Passive safety systems analysis: A novel approach for inverse uncertainty quantification based on Stacked Sparse Autoencoders and Kriging metamodeling. Prog. Nucl. Energy 148, 104209. https://doi.org/10.1016/j.pnucene.2022.104209.

Saeed, H.A., Wang, H., Peng, M., Hussain, A., Nawaz, A., 2020. Online fault monitoring based on deep neural network & sliding window technique. Prog. Nucl. Energy, 121, fasc. August 2019, p. 103236, doi: 10.1016/j.pnucene.2019.103236.

Salem, H., Shams, M.Y., Elzeki, O.M., Abd Elfattah, M., Al-Amri, J.F., Elnazer, S., 2022. Fine-tuning fuzzy KNN classifier based on uncertainty membership for the medical diagnosis of diabetes. Appl. Sci. 12 (3), 950. https://doi.org/10.3390/app12030950.

Soucy, P., Mineau, G.W., 2001. A simple KNN algorithm for text categorization. Proc. - IEEE Int. Conf. Data Min. ICDM 647–648. https://doi.org/10.1109/icdm.2001.989592.

Sugahara, S., Aomi, I., Ueno, M., 2022. Bayesian network model averaging classifiers by subbagging. Entropy 24 (5), 743. https://doi.org/10.3390/e24050743.

Tolo, S., et al., 2018. Robust on-line diagnosis tool for the early accident detection in nuclear power plants. Reliab. Eng. Syst. Saf., 186, fasc. April, pp. 110–119, 2019, doi: 10.1016/j.ress.2019.02.015.

Tripodo, C., Di Ronco, A., Lorenzi, S., Cammi, A., 2019. Development of a control-oriented power plant simulator for the molten salt fast reactor. EPJ Nucl. Sci. Technol. 5, 13. https://doi.org/10.1051/epjn/2019029.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K., 2003. A review of process fault detection and diagnosis: Part III: Process history based methods. Comput. Chem. Eng. 27 (3), 327–346. https://doi.org/10.1016/S0098-1354(02)00162-X.

Villa Medina, J.L., Boqué, R., Ferré, J., 2009. Bagged k-nearest neighbours classification with uncertainty in the variables. Anal. Chim. Acta 646 (1), 62–68. https://doi.org/10.1016/j.aca.2009.05.016.

W. Zhou, Hou, J., 2022. Implementation of fault isolation for molten salt reactor using PCA and contribution analysis. Ann. Nucl. Energy 173, 109138.

Wang, Z., et al., 2021. A multi-stage hybrid fault diagnosis approach for operating conditions of nuclear power plant. Ann. Nucl. Energy 153, 108015. https://doi.org/10.1016/j.anucene.2020.108015.

Wang, H., Jun Peng, M., Yu, Y., Saeed, H., Ming Hao, C., Kuo Liu, Y., 2021. Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants. Ann. Nucl. Energy 150, 107786. https://doi.org/10.1016/j.anucene.2020.107786.

Wang, Y., Liu, J., Qian, G., 2024. Hierarchical FFT-LSTM-GCN based model for nuclear power plant fault diagnosis considering spatio-temporal features fusion. Prog. Nucl. Energy 171, 105178.

Yang, C., et al., 2024. Sparse convolutional autoencoder-based fault location for drive circuits in nuclear reactors. Qual. Reliab. Eng. Int. 40 (2), 819–837.

Yang, Z., Baraldi, P., Zio, E., 2022. A method for fault detection in multi-component systems based on sparse autoencoder-based deep neural networks. Reliab. Eng. Syst. Saf. 220, 108278.

Yong-kuo, L., Abiodun, A., Zhi-bin, W., Mao-pu, W., Min-jun, P., Wei-feng, Y., 2018. A cascade intelligent fault diagnostic technique for nuclear power plants. J. Nucl. Sci. Technol. 55 (3), 254–266. https://doi.org/10.1080/00223131.2017.1394228.

Zhou, T., Yu, K., Cheng, M., Li, R., Dai, Z., 2023. Development and validation of machine learning-based transient identification models in a liquid-fueled molten salt reactor system. Nucl. Eng. Des. 415, 112682.

Zhu, S., et al., 2021. A robust strategy for sensor fault detection in nuclear power plants based on principal component analysis. Ann. Nucl. Energy 164, 108621. https://doi.org/10.1016/j.anucene.2021.108621.

Zhu, S., Yin, W., Xia, H., 2025. Fault detection for nuclear power plant based on improved moving window and sparse autoencoder. Ann. Nucl. Energy 222. https://doi.org/10.1016/j.anucene.2025.111626, 111626.