Unsupervised Multimodal Learning for Fault Diagnosis and Prognosis – Application to Radiotherapy Systems

Kélian Poujade[1]
Université de Toulouse, Oncopole Claudius Regaud, Institut Universitaire du Cancer de Toulouse (IUCT), France
Université de Toulouse, CNRS, INSERM, Centre de Recherches en Cancérologie de Toulouse (CRCT), France

Louise Travé-Massuyès
Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS-CNRS), Université de Toulouse, CNRS, France

Jérémy Pirard
Airbus, Toulouse, France

Laure Vieillevigne
Université de Toulouse, Oncopole Claudius Regaud, Institut Universitaire du Cancer de Toulouse (IUCT), France
Université de Toulouse, CNRS, INSERM, Centre de Recherches en Cancérologie de Toulouse (CRCT), France

Abstract

Modern complex systems, such as radiotherapy machines, require robust strategies for fault detection, diagnosis, and prognosis to ensure operational continuity and patient safety. While data-driven methods have gained traction, few studies address diagnostic and prognostic tasks using multimodal operational data under unsupervised or semi-supervised learning settings. This gap is particularly critical given the scarcity of labeled failure data in real-world environments. This work aims to design a unified approach for fault detection, diagnosis, and prognosis using multimodal data in the absence of complete labeling. To this end, autoencoders (AEs) are employed due to their suitability for unsupervised and self-supervised learning, flexibility in handling heterogeneous data, and ability to construct latent representations optimized for various downstream tasks. A specific implementation based on a Long Short-Term Memory β-Variational Autoencoder (LSTM-β-VAE) was developed to detect anomalies in machine logs. This framework is applied to TomoTherapy® systems – a highly complex and under-explored use case within the radiotherapy domain. Initial results demonstrate strong anomaly detection performance on both a public benchmark dataset (HDFS) and a proprietary dataset derived from real-world TomoTherapy® machine faults. Beyond methodology, the paper includes a concise literature review of multimodal learning and data-driven diagnosis and prognosis with a focus on AEs. Based on this review, key research directions are identified for the continuation of the thesis, especially the integration of explainable AI as a means to enhance diagnosis capabilities in the absence of labeled faults.

2012 ACM Subject Classification Computing methodologies → Unsupervised learning

Keywords and phrases Artificial Intelligence, Diagnosis, Prognosis, Radiotherapy machines

Digital Object Identifier 10.4230/OASIcs.DX.2025.16

Category PhD Panel

[1] Corresponding author

36th International Conference on Principles of Diagnosis and Resilient Systems (DX 2025).
Editors: Marcos Quinones-Grueiro, Gautam Biswas, and Ingo Pill; Article No. 16; pp. 16:1–16:17

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Funding This research was supported by Accuray Inc as part of the doctoral funding. It has also benefited from the AI Interdisciplinary Institute ANITI funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002.

Acknowledgements This work is the result of a collaborative tripartite agreement involving IUCT-Oncopole, Airbus, and Accuray Inc.

1 Introduction

In complex modern systems, fault detection, diagnosis, and prognosis are critical components for maintaining system integrity and performance. As these systems grow in complexity, traditional rule-based maintenance approaches are increasingly being supplemented – or even replaced – by data-driven methods. These techniques hold promise not only for timely fault detection, but also for enabling robust diagnostic and prognostic capabilities. Despite significant advances, few studies have investigated comprehensive data-driven diagnosis and prognosis frameworks that leverage unlabeled multimodal operational data such as machine logs and time-series sensor measurements. Addressing this limitation is essential, especially in real-world settings, where labeled failure data is scarce and system behavior is often stochastic.

A first step in this direction is fault detection through anomaly detection. We have already explored the use of machine logs for this purpose. As presented in Section 2, a deep learning approach based on β-Variational Autoencoders (β-VAE) was developed to identify abnormal sequences in machine-generated log data. This method demonstrated the feasibility of using logs as a data source to detect anomalies in the absence of explicit fault labels.

Building upon this foundation, the next phases involve exploring data-driven diagnosis and prognosis methodologies, particularly under semi-supervised or unsupervised learning paradigms. Such approaches are more appropriate for the nature of real-world data, where anomalies may be poorly labeled or entirely unlabeled. This progression is illustrated through an original case study focusing on TomoTherapy® systems used in cancer radiotherapy.

In radiotherapy, equipment reliability is critical to ensuring effective and uninterrupted patient treatment. Unplanned faults in radiotherapy systems can lead to treatment delays, rescheduling, and workflow disruptions, all of which can compromise treatment efficacy and negatively affect patient outcomes. The continuity of treatment has a well-documented impact on clinical results, with studies such as [18] demonstrating that extended overall treatment times can negatively impact local tumor control and influence survival rates.
Among the various technologies employed in radiotherapy, TomoTherapy® machines (Accuray, Madison, WI) (Figure 1a) stand out for their high complexity and versatility. These systems integrate advanced image-guided radiotherapy (IGRT) with intensity-modulated radiation therapy (IMRT) in a helical delivery mode [35]. Treatment delivery involves a continuously rotating gantry and a synchronized, translating treatment couch, paired with a binary multi-leaf collimator (MLC) consisting of 64 pneumatically actuated leaves (Figure 1b). This configuration enables fine-tuned modulation, making the system particularly suitable for anatomically extensive or complex treatments such as total body irradiation, craniospinal irradiation, and re-irradiation [54]. Within this system, the MLC is a key subsystem due to its direct role in beam shaping and dose modulation. However, its mechanical and electronic complexity, relying on high-frequency actuation, makes it susceptible to faults. A particularly vulnerable component is the bumper pack – a mechanical absorber designed to cushion the rapid opening and closing movements of the leaves. Faults in the bumper pack can arise

from progressive wear or sudden rupture and may compromise both system performance and treatment accuracy. This work focuses on the MLC subsystem, with the aim of developing an approach that can be readily generalized to other subsystems.

Although preventive maintenance is standard practice in radiotherapy, including for MLC components [2], there is growing interest in transitioning toward predictive maintenance paradigms. Predictive maintenance aims to anticipate equipment failures using real-time or near-real-time operational data, enabling interventions before breakdowns occur. Such an approach can minimize unplanned interruptions, optimize spare parts logistics, and most importantly, safeguard the continuity of care for patients.

To date, research in radiotherapy has focused primarily on performance monitoring techniques. Studies have analyzed trajectory log files from systems such as standard photon linacs (TrueBeam® Varian Medical Systems, Palo Alto, CA) and proton therapy machines, employing methods like threshold-based monitoring [56] and Statistical Process Control (SPC) [1] . Some studies have investigated predictive methods for system fault anticipation, typically relying on data from routine quality control (QC) or assurance (QA) tests [34, 16]. This approach limits real-time detection of incipient faults and restricts the resolution of predictive models. To our knowledge, no studies have explored predictive strategies for monitoring radiotherapy systems using continuously generated operational data.
This study aims to address this gap by developing a predictive framework tailored to TomoTherapy® machines. It proposes to go beyond fault detection by leveraging operational logs and sensor data for fault diagnosis and prognosis in a data-driven manner. Ultimately, this approach aims to improve equipment reliability and support more consistent and effective radiotherapy delivery, thereby enhancing the quality of patient care.

Figure 1 (a) Schema of TomoTherapy® system (model TomoHD™). (b) Schema of main subsystems composing the TomoTherapy® machines.

2 Conducted research

We first explored anomaly detection on machine log data. These logs, which chronicle the states, events, and procedures of the system, represent a valuable but underutilized resource for modeling system behavior and identifying early signs of malfunction. To address the challenges posed by limited labeled data, the approach relies on Autoencoders (AEs) and semi-supervised learning strategy, enabling the model to learn normal patterns without requiring exhaustive labels. A key objective is to construct a latent space that not only

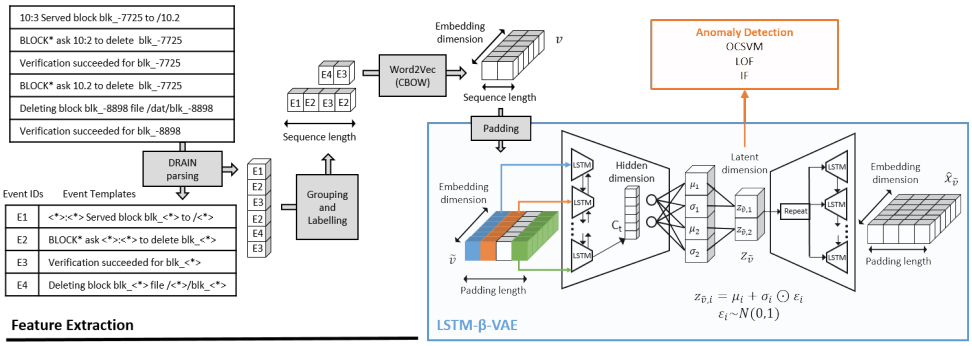


Figure 2 Overview of the proposed pipeline. The first stage includes DRAIN parsing, grouping, labeling and Word2Vec to extract features from log files. Obtained embedded sequences are padded to obtain the input $\tilde{v} \in \tilde{V}$ that feed an LSTM-β-VAE. While reconstructing the input at its output $\hat{x}_{\tilde{v}}$, this later learns compressed latent representations $z_{\tilde{v}} \in Z_{\tilde{v}}$ useful for anomaly detection. Different algorithms are then applied on the learned latent space $Z_{\tilde{v}} = \{z_{\tilde{v}}\}$.

captures the essential structure of normal operations, but also improves interpretability and facilitates the distinction between normal and anomalous behaviors. This framework aims to support more transparent and generalizable anomaly detection in complex systems. A preprint version of this work has been deposited on HAL [42] and the associated code is available online [2].

2.1 Proposed pipeline

We developed a semi-supervised pipeline for anomaly detection in log sequences. It assumes the availability of a subset of logs representing normal system behavior, which is used to train representation learning models. The pipeline consists of three main stages: log preprocessing and feature extraction, latent space learning via a Long Short-Term Memory-based β-Variational Autoencoder (LSTM-β-VAE), and anomaly detection on learned latent space using traditional machine learning algorithms enhanced with conformal prediction. The proposed pipeline is illustrated in Fig. 2. The full pipeline is designed to generalize well across different datasets, relying on minimal tuning and emphasizing the interpretability and robustness of the learned representations.

2.1.1 Log Preprocessing and Feature Extraction

Raw log files are first parsed using the DRAIN algorithm [22], which groups similar log lines into structured templates, allowing them to be represented as sequences of discrete event identifiers. These log lines are then grouped into sequences – either based on procedure windows or sliding time windows – and each sequence is labeled as either normal or faulty. The event identifiers from DRAIN parsing are then embedded into dense vector representations using the Word2Vec model (CBOW variant) [37]. Each log sequence is thus transformed into a sequence of embedding vectors $v \in V$, which are then padded or truncated to a fixed length to obtain $\tilde{v} \in \tilde{V}$. This process ensures compatibility with the subsequent neural architecture.

[2] `https://gitlab.laas.fr/addram/anomaly-detection-in-log-data.git`

2.1.2 Learning Latent Representations

To capture both semantic and sequential characteristics of log sequences, a latent space is learned using a LSTM-β-VAE trained using normal labeled data. This model encodes each embedded sequence $\tilde{v} \in \tilde{V}$ into a low-dimensional latent vector $z_{\tilde{v}} \in Z_{\tilde{v}}$ that captures the key patterns and structure of normal behavior, while reconstructing the input sequence from this compressed representation. The β term in the objective function controls the balance between reconstruction accuracy and latent space regularization [23], allowing better separation of anomalies from normal samples.

2.1.3 Anomaly Detection in the Latent Space

Then, traditional anomaly detection algorithms $\mathcal{A}$ such as One-Class SVM (OCSVM)[48], Isolation Forest (IF)[27], and Local Outlier Factor (LOF)[8] are applied on the learned latent representations $z_{\tilde{v}} \in Z_{\tilde{v}}$. These algorithms are learned in a semi-supervised manner using the latent representations of the normal labeled data used for the LSTM-β-VAE training. To provide statistical guarantees on model predictions, Conformal Anomaly Detection (CAD) – a technique derived from the Conformal Prediction framework [5] – is used. This method allows to define a threshold on anomaly scores returned by the algorithms $\mathcal{A}$ to correct the predictions and guarantee a statistical confidence in the results. Each confidence level is associated to a threshold value.

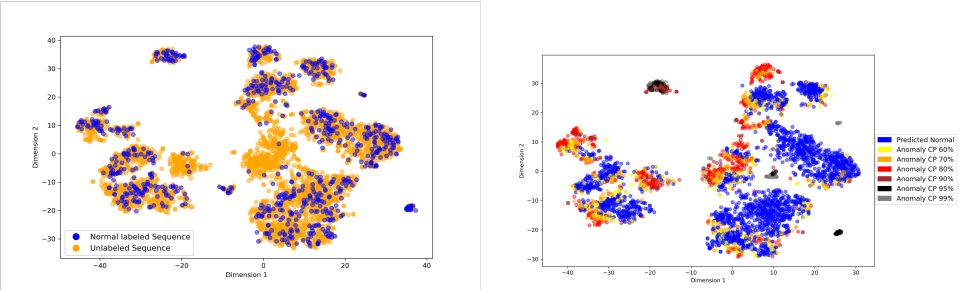
2.2 Evaluation

The evaluation of the proposed pipeline was conducted using two datasets: a publicly available benchmark (HDFS) [66] and a proprietary dataset derived from the logs of TomoTherapy® machines (TOMO). The TOMO dataset focuses on MLC faults linked to the bumper pack, gathering data from 20 fault cases across 15 different machines. Each fault case is referenced as FC(MM/DD/YYYY), identified by the date of the bumper pack repair.

One of the objectives of this study was to investigate the use of Mass-Volume (MV) and Excess-Mass (EM) scores [12, 20] – metrics specifically designed for unlabeled datasets – as a means to optimize the model's parameters and evaluate its performance in the absence of ground truth labels. These metrics evaluate the distance between the level sets of an anomaly scoring function and those of the underlying data distribution. They offer a principled way to assess a model's ability to capture the structure of the data and produce a meaningful anomaly score with statistical rigor.

For the HDFS dataset, both EM and MV scores were computed and compared with conventional classification metrics, including the area under the receiver operating characteristic curve (ROC), the area under the precision-recall curve (PR), and the F1 score. We first demonstrated that using the EM and MV scores to optimize the model and evaluate anomaly detection algorithms in the learned latent space yielded results comparable to those obtained with conventional metrics such as ROC and PR.

Then, for the TOMO dataset – where only a limited number of log sequences were labeled as normal (specifically, those recorded after system repairs) – only the EM and MV scores were employed. These metrics were used to assess the model's performance on a subset of normal labeled sequences unseen during model training.



(a) (b)

Figure 3 T-SNE visualization of log sequences from FC(06/15/2023) using their learned LSTM-β-VAE latent representations. (a) Colored by normal label (blue) and unlabeled (orange). (b) Colored by conformal predictions with different confidence levels based on learned OCSVM scoring function.

2.3 Results

Building on these findings, the EM and MV scores were then used to guide model optimization on the TOMO dataset and to assess the performance of various machine learning algorithms applied in the latent space. To complement this quantitative evaluation, several visualizations of the test data latent representations related to the fault case FC(06/15/2023) were produced using t-distributed stochastic neighbor embedding (t-SNE) algorithm [53], which projects high-dimensional representations into a two-dimensional space for interpretability.

Fig. 3a shows the latent space learned by the LSTM-β-VAE, where a clear separation appears between labeled normal sequences (blue) and a central group of unlabeled sequences (orange), indicating behaviors differing from the normal patterns used during training.

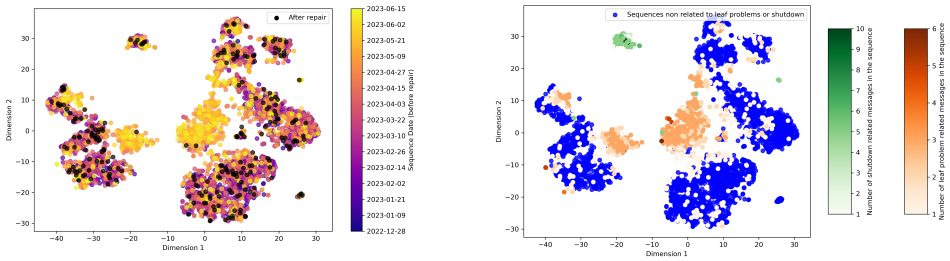
Fig. 3b presents the conformal prediction results of OCSVM, the best-performing algorithm based on EM and MV metrics. The color-coded conformal predictions can be seen as level sets of the OCSVM's scoring function. This figure highlights sequences in the center – isolated from normal ones 3a – as anomalies, mainly with 80% statistical confidence. Additional isolated groups are flagged as abnormal with 90–95% confidence, despite containing sequences labeled as normal.

To contextualize these detections, Fig. 4a maps the detected anomalies to their temporal position relative to subsystem maintenance. Central sequences flagged as abnormal predominantly occurred in the month before the identified fault, with none appearing post-repair, confirming the relevance of the detections.

Finally, Fig. 4b links the detected anomalous groups to specific log message content. Messages related to MLC issues (e.g., overtravel, position, bounce) (orange in Fig. 4b), known indicators of faults, are found in the central anomaly cluster. Another high-confidence anomaly group, detected in black in Fig. 3b and overlapping normal data in Fig. 3a, corresponds to machine shutdowns (green in Fig. 4b), reinforcing the consistency of the detection results with expert knowledge.

3 Thesis roadmap

So far, an approach has been developed to detect anomalous log sequences from the perspective of the system under study. This method leverages AEs to learn, in a semi-supervised manner, a latent space optimized for anomaly detection while preserving interpretability. The next phase of the project aims to enrich this framework by integrating additional data sources,



(a) (b)

Figure 4 T-SNE visualization of log sequences from FC(06/15/2023) using their learned LSTM-β-VAE latent representations. (a) Colored by temporality. (b) Colored by the number of known MLC-problem related messages.

such as time-series sensor measurements, to move toward fault diagnosis and prognosis using multimodal data and AI models. In this context, AEs remain central due to their key advantages: they naturally support multimodal learning, enable unsupervised or semi-supervised training, and allow for the construction of meaningful latent representations tailored to the different mentioned tasks.

3.1 Exploration

3.1.1 Multimodal learning

Multimodal learning refers to the development of AI models capable of processing and integrating heterogeneous data sources – such as images, time series, text, or tabular data – within a unified framework. This paradigm is particularly relevant in radiotherapy systems monitoring, where combining event logs, sensor streams, and contextual metadata yields richer representations of machine behavior. The goal is to extract complementary information from each modality to improve tasks such as anomaly detection, diagnosis, and prognosis. A key challenge lies in aligning and fusing modalities with differing structures, semantics, and temporal characteristics. Central to this is representation learning, which encodes raw multimodal inputs into robust, task-relevant vector representations.

Early approaches, like Multimodal Deep Boltzmann Machines [50], illustrated the potential of probabilistic graphical models to jointly model diverse modalities via a shared latent layer. However, these models are computationally intensive and require complex variational inference [50]. More scalable alternatives have emerged with AEs, now central to unsupervised and self-supervised multimodal learning. In this setting, AEs use modality-specific encoders (e.g., CNNs, RNNs or transformers) fused into a shared latent space.

AEs are well-suited for self-supervised learning, where reconstruction or auxiliary tasks drive representation learning without labels. Robinet et al. (2024) [46] introduced DRIM, which uses dual encoders per modality. The DRIM-U variant minimizes reconstruction of modality-unique components, using a tailored loss inspired by supervised contrastive learning and an adversarial objective. Contrastive learning has emerged as a powerful unsupervised approach, with methods like contrastive predictive coding [39], which use InfoNCE loss to bring similar representations closer and push others apart. Geng et al. (2022) [19] proposed Multimodal Masked AEs (M3AE), which learn joint vision-language representations via masked token prediction, avoiding modality-specific encoders and contrastive learning. Feng et al. (2024) [17] added a modality-consistency detection task, where the network learns to identify tampered modalities, enhancing cross-modal coherence.

AEs also optimize latent spaces for downstream goals. DRIM separates shared representations – rich in patient-specific information – from unique ones by minimizing mutual information, improving interpretability and predictive utility [46]. Correlational Neural Networks (CorrNet) [10] maximize cross-modal correlation in the latent space. Yang et al. extended this idea to temporal data with CorrRNN [61].

Fusion strategies in AE-based architectures vary. As surveyed by Zhao et al. (2024) [64], fusion can occur at the raw-data, feature, or decision levels. DRIM combines shared and unique representations through attention-based fusion, handling missing modalities effectively [46]. Other techniques like Tensor Fusion Networks model high-order modality interactions, though with higher computational cost [63].

Overall, AE-based multimodal learning offers a flexible framework for heterogeneous data, especially when labels are limited. Key directions include incorporating temporal dynamics, improving robustness to missing modalities, and optimizing latent spaces for real-world tasks such as anomaly detection, diagnosis, and prognosis.

3.1.2 Diagnosis

Fault diagnosis is a core task in system monitoring and maintenance, ensuring safety, reliability, and efficiency. It refers to the reasoning process used to identify the nature and root cause of a failure based on observed symptoms from measurements, checks, or tests. Formally, it can be seen as an inference problem: determining a system's internal (possibly faulty) state from its external outputs. Diagnosis typically involves three stages: fault detection (whether a fault occurred), fault isolation (identifying the faulty component), and fault identification (characterizing the fault's type and severity). Methods vary by system complexity and data availability, and can be categorized into model-based, data-driven, or hybrid approaches combining physical knowledge with statistical or machine learning techniques.

3.1.2.1 Model-based diagnosis

Model-based approaches have long been the dominant paradigm in fault diagnosis [7, 40]. These approaches rely on constructing analytical or physical models that describe the system's behavior under nominal and faulty conditions. Such models are usually derived from first-principles knowledge, including conservation laws, differential equations, discrete event models, or logical constraints, and are validated through expert analysis and simulation. Model-based methods provide high interpretability, allow detection of specific fault patterns with precision, and generally perform well in systems where accurate models are available.

However, their effectiveness is often limited by the complexity of real-world systems. Modeling nonlinear dynamics, stochastic disturbances, time-varying behaviors, and interactions between subsystems remains a challenging task. See [41] for challenges referring to the DX approaches. Furthermore, developing and validating such models requires significant domain expertise and resources, which can be prohibitive in systems that evolve rapidly or lack comprehensive documentation.

3.1.2.2 Data-driven diagnosis

Data-driven fault diagnosis has gained momentum over the last decade due to increasing sensor data availability, advances in machine learning, and enhanced computational power. Unlike model-based methods, these approaches infer patterns or fault signatures directly from historical or real-time data. This enables fault detection in complex or poorly understood

systems and improves scalability. Reviews such as [11, 47] highlight the maturity and applicability of these methods across domains like HVAC systems and general industrial equipment, underlining their potential to automate diagnostics and reduce expert reliance.

Among these techniques, AEs are widely used for their ability to learn compact, informative representations from high-dimensional data. Often, they serve as feature extractors, with supervised classifiers trained on the latent space. For instance, Han Liu et al. (2018) [28] proposed a recurrent AE-based method using gated recurrent units (GRU-NP-DAEs), where each AE is trained on a specific fault class and the classification is determined by identifying the AE that minimizes the reconstruction error. Similarly, Lang Liu et al. (2024) [29] introduced a variable-wise stacked temporal AE (VW-STAE), in which a variable sensitivity analysis guides the classification, again relying on supervised training per fault type.

Other works improved AE architectures for better feature extraction. Shao et al. (2021) [49] used adaptive Morlet wavelets to capture nonlinearities. Yang et al. (2020) [60] combined sparse and denoising AEs with ensemble learning. Qiu et al. (2025) [43] proposed a multimodal fusion scheme using multiscale stacked denoising AEs for noise robustness. Zhao et al. (2024) [65] presented a semi-supervised Gaussian mixture VAE for few-shot learning, adapting to new fault classes via episodic training and a dynamic multimodal prior.

However, these methods often assume the availability of labeled data for all fault types – an unrealistic assumption in practice. Real-world systems frequently encounter rare or unknown faults, and collecting exhaustive labeled datasets is infeasible. Fully supervised AE-based methods may thus struggle to generalize.

To address this, semi- and unsupervised AE-based methods have emerged. These models learn representations of normal data patterns without relying on fault labels. For example, Amini and Zhu (2022) [4] introduced a source-aware AE that can operate with or without labels. Cacciarelli and Kulahci (2022) [9] proposed an orthogonal AE to decorrelate latent features, improving fault detection and interpretability. Ma et al. (2018) [33] developed a deep coupling AE for multimodal sensory data, learning a shared representation and applying late fusion for diagnosis.

These studies reflect growing interest in unsupervised learning frameworks for early fault detection in safety-critical systems. While supervised AEs remain popular, semi-supervised or unsupervised models better match real-world constraints, offering scalable, realistic solutions for modern diagnostic challenges.

3.1.2.3 Fault diagnosis and AI explainability

Another promising avenue for fault diagnosis using data-driven and unsupervised learning methods is to investigate the role of explainability in AI models. Explainability refers to the extent to which a model's internal mechanisms, decisions, and outputs can be interpreted and understood by humans. In fact, the explainability of the models can be viewed as a form of diagnosis. While fault detection methods typically indicate the presence of an abnormal state, explainability goes further by identifying the elements or patterns that contributed to this state – essentially answering why the system deviated from normal behavior. In this sense, providing an explanation can lead to diagnosing the cause of the fault. Therefore, designing an explainable fault detection model is inherently aligned with building a diagnosis system. Despite its importance, this connection has received limited attention in data-driven research. In such unsupervised contexts, explainability plays a crucial role, as it enables the interpretation – and thus the diagnosis – of detected anomalies in the absence of explicit ground truth. Exploring explainability as a diagnostic tool is therefore not only a promising direction, but also a necessary one to improve the understanding, trustworthiness, and practical applicability of data-driven fault detection systems under realistic constraints.

Explainability methods are typically categorized into two families: post-hoc methods, which seek to interpret already trained models, and intrinsic methods, which embed interpretability directly into the model architecture or training process. This distinction is particularly relevant for AEs, which are widely used in unsupervised tasks such as anomaly detection but often operate as black boxes.

Post-hoc explainability techniques are applied after the model is trained, often without modifying the model's structure. Among the most widely used post-hoc tools are Shapley Additive Explanations (SHAP)[32], which attribute contributions of input features to a model's predictions. SHAP has been extensively employed in the context of AEs to understand anomalies and latent representations. For instance, Antwarg et al. (2021) [6] applied Kernel SHAP to explain reconstruction-based anomalies by linking reconstruction errors to influential input features. Similarly, Xu et al. (2021) [59] used SHAP in a dynamic multimodal VAE (DMVAE) to provide both local and global feature attribution in a clinical prediction task. In genomics, Li et al. (2023) [26] introduced XA4C, an AE-based pipeline where SHAP values derived from XGBoost identify critical genes contributing to latent representations, supporting downstream biological interpretation.

Another line of work focuses on counterfactual explanations, which generate hypothetical examples to highlight what changes would be required to alter a model's output, which shares high similarity with the concept of *conflict* known in the logical diagnosis theory [44, 13]. Using contrastive supervision, Todo et al. (2023) [52] trained a VAE to disentangle class-relevant and class-irrelevant components in multivariate time series, enabling the generation of plausible counterfactuals by manipulating only the class-relevant latent subspace. Extending this concept, Haselhoff et al. (2024) [21] proposed the Gaussian discriminant VAE (GdVAE), a self-explainable generative model that integrates a class-conditional latent space with closed-form counterfactual generation, balancing interpretability and quality of explanations in vision tasks.

Gradient-based attention mechanisms have also been applied to AEs to enhance interpretability. Liu et al. (2020) [30] derived visual attention maps from VAE latent variables to localize anomalies in images, while Nguyen et al. (2019) [38] used gradient-based fingerprinting in an unsupervised VAE for network anomaly detection.

Another line of work involves surrogate models like LIME [45], which approximate AE behavior locally using interpretable models (e.g., decision trees). Wu and Wang (2021) [57] proposed a LIME-based framework with explainers for reconstruction, classification, and global behavior in fraud detection.

In contrast to post-hoc approaches, intrinsic explainability is built into the model architecture or training objective. One strategy is to enforce interpretable representations through structured constraints. For example, Di Clemente et al. (2025) [15] developed a physics-informed AE where latent codes directly correspond to astrophysical quantities such as mass and radius of neutron stars. By embedding domain knowledge and explicit constraints in the loss function, the model achieves physical interpretability of latent variables. Other studies integrate interpretability by combining AEs with inherently transparent models. Aguilar et al. (2022) [3] proposed a decision tree-based AE capable of handling categorical data without encoding, offering interpretable internal representations through the branching structure of the tree itself. In probabilistic frameworks, Bayesian autoencoders (BAEs) can enhance interpretability via uncertainty estimation. Yong and Brintrup (2022) [62] introduced coalitional Bayesian AEs, where explanations are derived from the mean and epistemic uncertainty of log-likelihood estimates, providing insight into model behavior under covariate shift without relying on additional explainer models. Other works focused on providing explanations based on reconstruction errors from AEs. Kieu et al. (2022) [24] used Robust Principal Component

Analysis (RPCA) combined with AEs to improve the explainability of outlier detection in time series, separating outliers from clean data. Martinez-Garcia et al. (2019) [36] proposed the entropy of the AE's reconstructed outputs as a form of explanation.

3.1.3 Prognosis

In prognosis, as in diagnosis, methodologies can broadly be divided into model-based and data-driven approaches, each offering distinct strategies for predicting the Remaining Useful Life (RUL) and anticipating system failures.

Model-based methods rely on physical laws and tools like Physics of Failure (PoF), Kalman filters, and finite element analysis to estimate RUL without historical data.

Data-driven methods leverage sensor data and failure history to predict degradation. They include stochastic models (e.g., Weibull distribution, Bayesian networks, Hidden Markov Models), statistical techniques (e.g., ARMA, ARIMA), and AI-based approaches.

Among AI-based methods, similarity-based learning has gained prominence. Widodo et al. (2025) [55] demonstrated an approach for boiler prognosis using Support Vector Machines (SVMs), Random Forest Algorithms (RFAs), and Dynamic Time Warping (DTW) for RUL estimation, showing potential for real-world deployment in power plants.

Deep learning models, particularly LSTM networks, have been widely used to model temporal dependencies in degradation data. Liu et al. (2021) [31] introduced an elastic-net-regularized LSTM (E-LSTM) to mitigate overfitting and improve RUL prediction stability for rolling bearings. Wu et al. (2018) [58] utilized vanilla LSTM networks and dynamic differential technology to enhance RUL prediction under varying operational conditions and noise levels.

AE-based architectures have also been explored for prognostic tasks. Robinet et al. (2024) [46] proposed a method for survival prediction using disentangled representations from incomplete multimodal healthcare data, applying a discretized time model supervised by a specialized loss function for censored survival data as described by Kvamme and Borgan (2021) [25]. This approach models hazard probabilities over time intervals and learns individualized survival curves from multimodal inputs. De Pater and Mitici (2023) [14] designed an LSTM-AE with attention to develop health indicators for aircraft system in an unsupervised manner. These health indicators are further used to predict the RUL of the aircraft system using a similarity-based matching approach. In a more recent contribution, Tefera et al. (2025) [51] introduced a constraint-guided deep learning framework to generate physically consistent health indicators from bearing sensor data. The proposed AE model integrates domain constraints – monotonicity, bounded output, and energy-consistency – into the training process via a custom optimization scheme. Compared to baseline models, this approach enhances trendability, robustness, and consistency, yielding interpretable degradation profiles aligned with physical expectations.

Overall, the landscape of prognosis methodologies continues to evolve, with hybrid approaches combining physics-based insight and data-driven learning offering powerful solutions for anticipating failures and optimizing maintenance in complex systems.

4 Schedule

The proposed research roadmap is structured into three interconnected phases. Each phase explores a fundamental capability – fault detection, diagnosis, and prognosis – through the lens of multimodal and unsupervised (or semi-supervised) learning. A central prerequisite for each stage is the curation of a clean and well-structured dataset focused on the MLC subsystem, enabling a controlled yet realistic environment for experimentation.

4.1 Phase 1 – Multimodal Representation Learning for Anomaly Detection

Phase 1 aims to construct joint representations of operational logs and time-series sensor data through unsupervised deep learning. Inspired by recent advances such as DRIM [46] and Correlational Neural Networks [10], this step will evaluate various fusion strategies – including dual encoder architectures, disentangled shared/unique representations, and correlation-maximizing latent spaces.

The goal is to design an embedding space where normal and abnormal behaviors can be effectively separated, even in the absence of fault labels. Special attention will be given to robustness in the presence of missing modalities and asynchronous data.

Milestones:

- Construction of a labeled and time-aligned multimodal dataset focused on MLC.
- Leveraging and extending existing multimodal AE approaches (e.g., DRIM, CorrNet) to construct a latent space that better suit the characteristics of radiotherapy system data.
- Testing different self-supervised learning strategies: reconstruction [46], masked token prediction [19] and modality-consistency detection task [17].

4.2 Phase 2 – Explainable Fault Diagnosis via Latent Representations

Phase 2 focuses on leveraging the learned multimodal latent space to perform fault diagnosis in an unsupervised or weakly supervised setting. A key hypothesis is that explainability can serve as a proxy for diagnosis, especially when ground truth labels are scarce. Building on the work of Todo et al. [52] and the logical theory of conflict-based diagnosis [44, 13], counterfactual explanation techniques will be explored as a means of identifying latent dimensions or input factors contributing to anomalies.

In parallel, post-hoc tools such as SHAP will be employed to generate interpretable attributions on both the model outputs and the latent encoding. The interplay between these explanations and traditional diagnostic tasks (fault isolation, severity ranking) will be investigated.

Milestones:

- Adaptation of counterfactual explanations to the latent space of multimodal AEs.
- SHAP-based analysis of log and sensor contributions to fault signatures.
- Evaluation of the diagnostic capability of selected explainable methods.

4.3 Phase 3 – Prognosis and Remaining Useful Life Estimation

Phase 3, and the final stage, addresses long-term prediction of subsystem degradation. Two complementary directions will be explored: (1) survival analysis with multimodal latent embeddings, following Robinet et al. [46]; (2) unsupervised health indicator construction with physical constraints, following Tefera et al. [51].

In the first direction, discrete-time hazard models will be used to estimate individualized survival curves from latent variables, integrating sensor and log-derived features. In the second, attention will be paid to embedding physical priors (e.g., monotonicity, boundedness) directly into AE training, to produce interpretable and consistent degradation profiles.

Milestones:

- Derivation of latent health indicators from log and sensor embeddings.
- Modeling of hazard probabilities from multimodal data (DRIM-like survival modeling).
- Training of constraint-aware AEs to enforce physically consistent degradation behavior.

Comparison across Phases: Throughout all phases, systematic comparisons will be conducted between semi-supervised learning (enabled by partially labeled TOMO data) and fully unsupervised alternatives, to assess scalability and realism in operational settings.

5 Conclusion

This work presents a semi-supervised learning framework for fault detection in complex systems using log data. The method, based on a LSTM-β-VAE, demonstrated effective anomaly detection on both benchmark and real-world datasets, leveraging an optimized latent space combined with conformal prediction. The case study on TomoTherapy® machines highlights the practical relevance of this approach in a safety-critical healthcare setting.

Looking forward, the research will expand to incorporate time-series sensor data alongside log sequences, moving toward a multimodal diagnostic and prognostic framework. Upcoming phases will explore multimodal fusion strategies, counterfactual and SHAP-based explainability techniques for unsupervised diagnosis, and survival modeling for prognosis. These directions aim to produce interpretable, generalizable models that support fault isolation and remaining useful life estimation under realistic operational constraints.

References

1 Charles M Able, Alan H Baydush, Callistus Nguyen, Jacob Gersh, Alois Ndlovu, Igor Rebo, Jeremy Booth, Mario Perez, Benjamin Sintay, and Michael T Munley. A model for preemptive maintenance of medical linear accelerators—predictive maintenance. *Radiation Oncology*, 11:1–9, 2016.

2 Christina Elizabeth Agnew, Sergio Esteve, Glenn Whitten, and William Little. Reducing treatment machine downtime with a preventative mlc maintenance procedure. *Physica Medica*, 85:1–7, 2021.

3 Diana Laura Aguilar, Miguel Angel Medina-Pérez, Octavio Loyola-Gonzalez, Kim-Kwang Raymond Choo, and Edoardo Bucheli-Susarrey. Towards an interpretable autoencoder: A decision-tree-based autoencoder and its application in anomaly detection. *IEEE transactions on dependable and secure computing*, 20(2):1048–1059, 2022. doi:10.1109/TDSC.2022.3148331.

4 Nima Amini and Qinqin Zhu. Fault detection and diagnosis with a novel source-aware autoencoder and deep residual neural network. *Neurocomputing*, 488:618–633, 2022. doi:10.1016/J.NEUCOM.2021.11.067.

5 Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. arXiv:2107.07511.

6 Liat Antwarg, Ronnie Mindlin Miller, Bracha Shapira, and Lior Rokach. Explaining anomalies detected by autoencoders using shapley additive explanations. *Expert systems with applications*, 186:115736, 2021. doi:10.1016/J.ESWA.2021.115736.

7 Pietro Baroni, Gianfranco Lamperti, Paolo Pogliano, and Marina Zanella. Diagnosis of large active systems. *Artificial Intelligence*, 110(1):135–183, 1999. doi:10.1016/S0004-3702(99)00019-3.

8 Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000. doi:10.1145/342009.335388.

9 Davide Cacciarelli and Murat Kulahci. A novel fault detection and diagnosis approach based on orthogonal autoencoders. *Computers & Chemical Engineering*, 163:107853, 2022. doi:10.1016/J.COMPCHEMENG.2022.107853.

10 Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. Correlational neural networks. *Neural computation*, 28(2):257–285, 2016. doi:10.1162/NECO_A_00801.

11 Zhelun Chen, Zheng O'Neill, Jin Wen, Ojas Pradhan, Tao Yang, Xing Lu, Guanjing Lin, Shohei Miyata, Seungjae Lee, Chou Shen, et al. A review of data-driven fault detection and diagnostics for building hvac systems. *Applied Energy*, 339:121030, 2023.

12 Stéphan Clémençon and Jérémie Jakubowicz. Scoring anomalies: a m-estimation formulation. In *Artificial Intelligence and Statistics*, pages 659–667. PMLR, 2013. URL: http://proceedings.mlr.press/v31/clemencon13a.html.

13 Johan De Kleer and Brian C Williams. Diagnosing multiple faults. *Artificial intelligence*, 32(1):97–130, 1987. doi:10.1016/0004-3702(87)90063-4.

14 Ingeborg de Pater and Mihaela Mitici. Developing health indicators and rul prognostics for systems with few failure instances and varying operating conditions using a lstm autoencoder. *Engineering Applications of Artificial Intelligence*, 117:105582, 2023. doi:10.1016/J.ENGAPPAI.2022.105582.

15 Francesco Di Clemente, Matteo Scialpi, and Michał Bejger. Explainable autoencoder for neutron star dense matter parameter estimation. *Machine Learning: Science and Technology*, 2025.

16 Tai Dou, Benjamin Clasie, Nicolas Depauw, Tim Shen, Robert Brett, Hsiao-Ming Lu, Jacob B Flanz, and Kyung-Wook Jee. A deep lstm autoencoder-based framework for predictive maintenance of a proton radiotherapy delivery system. *Artificial Intelligence in Medicine*, 132:102387, 2022. doi:10.1016/J.ARTMED.2022.102387.

17 Wenjun Feng, Xin Wang, Donglin Cao, and Dazhen Lin. An autoencoder-based self-supervised learning for multimodal sentiment analysis. *Information Sciences*, 675:120682, 2024. doi:10.1016/J.INS.2024.120682.

18 José A González Ferreira, Javier Jaén Olasolo, Ignacio Azinovic, and Branislav Jeremic. Effect of radiotherapy delay in overall treatment time on local control and survival in head and neck cancer: review of the literature. *Reports of Practical Oncology and Radiotherapy*, 20(5):328–339, 2015.

19 Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022. doi:10.48550/arXiv.2205.14204.

20 Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016. arXiv:1607.01152.

21 Anselm Haselhoff, Kevin Trelenberg, Fabian Küppers, and Jonas Schneider. The gaussian discriminant variational autoencoder (gdvae): A self-explainable model with counterfactual explanations. In *European Conference on Computer Vision*, pages 305–322. Springer, 2024. doi:10.1007/978-3-031-73668-1_18.

22 Pinjia He, Jieming Zhu, Zibin Zheng, and Michael R Lyu. Drain: An online log parsing approach with fixed depth tree. In *2017 IEEE international conference on web services (ICWS)*, pages 33–40. IEEE, 2017. doi:10.1109/ICWS.2017.13.

23 Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

24 Tung Kieu, Bin Yang, Chenjuan Guo, Christian S Jensen, Yan Zhao, Feiteng Huang, and Kai Zheng. Robust and explainable autoencoders for unsupervised time series outlier detection. In *2022 IEEE 38th International conference on data engineering (ICDE)*, pages 3038–3050. IEEE, 2022. doi:10.1109/ICDE53745.2022.00273.

25 Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, 27(4):710–736, 2021.

26 Qing Li, Yang Yu, Pathum Kossinna, Theodore Lun, Wenyuan Liao, and Qingrun Zhang. Xa4c: explainable representation learning via autoencoders revealing critical genes. *PLOS Computational Biology*, 19(10):e1011476, 2023. doi:10.1371/JOURNAL.PCBI.1011476.

27 Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. doi:10.1109/ICDM.2008.17.

28 Han Liu, Jianzhong Zhou, Yang Zheng, Wei Jiang, and Yuncheng Zhang. Fault diagnosis of rolling bearings with recurrent neural network-based autoencoders. *ISA transactions*, 77:167–178, 2018.

29 Lang Liu, Ying Zheng, and Shaojun Liang. Variable-wise stacked temporal autoencoder for intelligent fault diagnosis of industrial systems. *IEEE Transactions on Industrial Informatics*, 2024.

30 Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8642–8651, 2020.

31 Zhao-Hua Liu, Xu-Dong Meng, Hua-Liang Wei, Liang Chen, Bi-Liang Lu, Zhen-Heng Wang, and Lei Chen. A regularized lstm method for predicting remaining useful life of rolling bearings. *International Journal of Automation and Computing*, 18:581–593, 2021. doi:10.1007/S11633-020-1276-6.

32 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

33 Meng Ma, Chuang Sun, and Xuefeng Chen. Deep coupling autoencoder for fault diagnosis with multimodal sensory data. *IEEE Transactions on Industrial Informatics*, 14(3):1137–1145, 2018. doi:10.1109/TII.2018.2793246.

34 Min Ma, Chenbin Liu, Ran Wei, Bin Liang, and Jianrong Dai. Predicting machine's performance record using the stacked long short-term memory (lstm) neural networks. *Journal of Applied Clinical Medical Physics*, 23(3):e13558, 2022.

35 T Rockwell Mackie, John Balog, Ken Ruchala, Dave Shepard, Stacy Aldridge, Ed Fitchard, Paul Reckwerdt, Gustavo Olivera, Todd McNutt, and Minesh Mehta. Tomotherapy. In *Seminars in Radiation Oncology*, volume 9, pages 108–117. Elsevier, 1999.

36 Miguel Martinez-Garcia, Yu Zhang, Jiafu Wan, and Jason Mcginty. Visually interpretable profile extraction with an autoencoder for health monitoring of industrial systems. In *2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 649–654. IEEE, 2019.

37 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

38 Quoc Phong Nguyen, Kar Wai Lim, Dinil Mon Divakaran, Kian Hsiang Low, and Mun Choon Chan. Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In *2019 IEEE Conference on Communications and Network Security (CNS)*, pages 91–99. IEEE, 2019. doi:10.1109/CNS.2019.8802833.

39 Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

40 Bernhard Peischl and Franz Wotawa. Model-based diagnosis or reasoning from first principles. *IEEE intelligent systems*, 18(3):32–37, 2005.

41 Ingo Pill and Johan De Kleer. Challenges for model-based diagnosis. In *35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024)*, pages 6–1. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/OASIcs.DX.2024.6.

42 Kélian Poujade, Louise Travé-Massuyès, Jérémy Pirard, and Laure Vieillevigne. B-variational autoencoder based anomaly detection in log data – application to radiotherapy systems. Preprint, HAL archive, 2025. URL: `https://hal.science/hal-05209127`.

43 Zhi Qiu, Shanfei Fan, Haibo Liang, and Jincai Liu. Multimodal fusion fault diagnosis method under noise interference. *Applied Acoustics*, 228:110301, 2025.

44 Raymond Reiter. A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95, 1987. `doi:10.1016/0004-3702(87)90062-2`.

45 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

46 Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. Drim: Learning disentangled representations from incomplete multimodal healthcare data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 163–173. Springer, 2024. `doi:10.1007/978-3-031-72384-1_16`.

47 Atma Ram Sahu, Sanjay Kumar Palei, and Aishwarya Mishra. Data-driven fault diagnosis approaches for industrial equipment: A review. *Expert Systems*, 41(2):e13360, 2024. `doi:10.1111/EXSY.13360`.

48 Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

49 Haidong Shao, Min Xia, Jiafu Wan, and Clarence W de Silva. Modified stacked autoencoder using adaptive morlet wavelet for intelligent fault diagnosis of rotating machinery. *IEEE/ASME Transactions on Mechatronics*, 27(1):24–33, 2021.

50 Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

51 Yonas Tefera, Quinten Van Baelen, Maarten Meire, Stijn Luca, and Peter Karsmakers. Constraint-guided learning of data-driven health indicator models: An application on the pronostia bearing dataset. *arXiv preprint arXiv:2503.09113*, 2025. `doi:10.48550/arXiv.2503.09113`.

52 William Todo, Merwann Selmani, Béatrice Laurent, and Jean-Michel Loubes. Counterfactual explanation for multivariate times series using a contrastive variational autoencoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. `doi:10.1109/ICASSP49357.2023.10095789`.

53 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

54 David C Westerly, Emilie Soisson, Quan Chen, Katherine Woch, Leah Schubert, Gustavo Olivera, and Thomas R Mackie. Treatment planning to improve delivery accuracy and patient throughput in helical tomotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 74(4):1290–1297, 2009.

55 Achmad Widodo, Toni Prahasto, Mochamad Soleh, and Herry Nugraha. Diagnostics and prognostics of boilers in power plant based on data-driven and machine learning. *International Journal of Prognostics and Health Management*, 16(1), 2025.

56 Binbin Wu, Pengpeng Zhang, Bill Tsirakis, David Kanchaveli, and Thomas LoSasso. Utilizing historical mlc performance data from trajectory logs and service reports to establish a proactive maintenance model for minimizing treatment disruptions. *Medical physics*, 46(2):475–483, 2019.

57 Tung-Yu Wu and You-Ting Wang. Locally interpretable one-class anomaly detection for credit card fraud detection. In *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 25–30. IEEE, 2021.

58 Yuting Wu, Mei Yuan, Shaopeng Dong, Li Lin, and Yingqi Liu. Remaining useful life estimation of engineered systems using vanilla lstm neural networks. *Neurocomputing*, 275:167–179, 2018. `doi:10.1016/J.NEUCOM.2017.05.063`.

59 Yiming Xu, Xiaohong Liu, Liyan Pan, Xiaojian Mao, Huiying Liang, Guangyu Wang, and Ting Chen. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1362–1373, 2021. doi:10.1109/JBHI.2021.3103271.

60 Jing Yang, Guo Xie, and Yanxi Yang. An improved ensemble fusion autoencoder model for fault diagnosis from imbalanced and incomplete data. *Control Engineering Practice*, 98:104358, 2020.

61 Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A Bernal, and Jiebo Luo. Deep multimodal representation learning from temporal data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5455, 2017.

62 Bang Xiang Yong and Alexandra Brintrup. Coalitional bayesian autoencoders: Towards explainable unsupervised deep learning with applications to condition monitoring under covariate shift. *Applied Soft Computing*, 123:108912, 2022. doi:10.1016/J.ASOC.2022.108912.

63 Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. arXiv:1707.07250.

64 Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9):1–36, 2024. doi:10.1145/3649447.

65 Zhiqian Zhao, Yeyin Xu, Jiabin Zhang, Runchao Zhao, Zhaobo Chen, and Yinghou Jiao. A semi-supervised gaussian mixture variational autoencoder method for few-shot fine-grained fault diagnosis. *Neural Networks*, 178:106482, 2024. doi:10.1016/J.NEUNET.2024.106482.

66 Jieming Zhu, Shilin He, Pinjia He, Jinyang Liu, and Michael R Lyu. Loghub: A large collection of system log datasets for ai-driven log analytics. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, pages 355–366. IEEE, 2023. doi:10.1109/ISSRE59848.2023.00071.