


One-Class Classification and Cluster Ensembles for Anomaly Detection and Diagnosis in Multivariate Time Series Data

Adil Mukhtar ✉ 

Institute of Software Technology, Graz University of Technology, Austria

Thomas Hirsch ✉

Institute of Software Technology, Graz University of Technology, Austria

Gerald Schweiger ✉ 

Vienna University of Technology, Austria

Abstract

Real-world automated systems such as building automation, power plants, and more have benefited from data-driven learning methodologies for anomaly detection and diagnosis. Typically, these methodologies heavily rely on prior knowledge related to abnormal operations, i.e., data points labeled as anomalies. However, in practice, such labelled data points are often unavailable which poses challenges in effective anomaly detection, particularly in diagnosis. In this paper, we propose One-class Classification Cluster ENsembles (OCCEN) anomaly detection and diagnosis approach for multivariate time series data. OCCEN utilizes one-class classification learning methods for anomaly detection followed by the decomposition of anomalies into multiple clusters. Then each cluster is treated as a binary classification problem and classifiers are trained to learn cluster representations. These trained models in combination with explainable AI models are used to generate a ranked list of diagnoses, i.e., features. Finally, we re-rank those features to account for temporal dependencies through the dynamic time-warping technique. The practical evaluation of OCCEN for air handling units (AHU) demonstrates its effectiveness in identifying faults. The framework consistently outperforms the baseline in fault diagnosis, as higher scores are observed for detection and diagnostic evaluation metrics, including F1 score, intersection over union, *HitRate@k*, and *RootCause@k*.

2012 ACM Subject Classification Computing methodologies → Machine learning algorithms; Computing methodologies → Artificial intelligence; Computing methodologies → Model development and analysis; Hardware → Fault models and test metrics

Keywords and phrases Anomaly Detection and Diagnosis, Machine Learning, Explainable AI, One-class Classification

Digital Object Identifier 10.4230/OASICS.DX.2024.14

Funding The work described in this paper is funded by the ECom4Future project “FIWARE DRIVEN ENERGY COMMUNITIES FOR THE FUTURE” under contract number FFG: 903927.

1 Introduction

Automated systems such as building automation [24] and industrial plants [10] are equipped with numerous sensors and control signals. These sensors play a vital role in helping operators keep track of the system’s condition by monitoring an array of multivariate time series data produced by many integrated components [15, 3]. A key aspect of this monitoring is the identification of abnormalities within the system, enabling operators to conduct diagnostic analyses to pinpoint malfunctioning components [12]. However, this process presents significant challenges in practical scenarios due to several factors [18]:

1. The system’s complexity often limits clear insights into its workings, including how sensors, controllers, and actuators are operating.



© Adil Mukhtar, Thomas Hirsch, and Gerald Schweiger;
licensed under Creative Commons License CC-BY 4.0

35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024).

Editors: Ingo Pill, Avraham Natan, and Franz Wotawa; Article No. 14; pp. 14:1–14:19

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2. Anomalies might not only result from components that are failing or exhibiting abnormal behavior but could also be due to a variety of operation modes.

As a result, it is important to have a methodology that not only detects these anomalies but also offers a list of potential diagnoses. Such a list might include the likelihood of failure in specific sensors or controllers, and indications of malfunctions, thereby enabling a more efficient response to system irregularities.

Over the years, the exponential development of large-scale monitored equipment has motivated researchers to propose a variety of data-driven automated anomaly detection and diagnosis methodologies [7, 1]. A majority of these techniques heavily rely on the prior knowledge of faulty and non-faulty operations called *supervised* learning. One of the significant limitations of this learning approach is the scarcity of anomalous data points in historical data, consequently, rendering such learning methods infeasible for real-world applications. On the other hand, *unsupervised* learning methods require no prior anomalous data points and heuristically extract distinctive operation patterns within the historical data. Despite their effectiveness, these techniques may not be well suited for effective anomaly detection and diagnosis due to the following reasons:

- Most proposed methodologies are designed to target specific types of faults and rely on the physical characteristics of the system, resulting in limited generalizability and scalability [15].
- The prior knowledge of faulty operations in the data, i.e., target labels, is usually not available in real-world scenarios, therefore, supervised learning techniques are not considered feasible [8].
- Unsupervised clustering methods [12], including k-means clustering, hierarchical clustering, and density-based clustering, create different groups based on feature similarities. Yet, these techniques do not reveal the complex, non-linear interrelations among features, which can lead to suboptimal performance in diagnostic performance.
- In real-world scenarios, it is desirable to assist operators with a ranked list of diagnoses. However, existing methods fall short of providing a nuanced, ranked list of diagnoses that aligns with the operators' requirements.
- Finally, the adaptability of the existing methods is usually limited due to their reliance on specific data structures, predefined models, and the need for extensive labeled datasets in supervised learning, or the lack of interpretability in unsupervised learning.

In this paper, we tackle the previously mentioned challenges and propose One-class Classification Cluster Ensembles (OCCEN) technique that orchestrate already existing unsupervised and supervised learning methods for anomaly detection and diagnosis. Specifically, OCCEN utilizes a one-class classification approach to pinpoint anomalous data points, an approach that does not require priors. Following the anomaly detection task, we employ clustering techniques to group these anomalies into various groups. Subsequently, each cluster is treated as a separate binary classification learning problem to uncover the non-linear relationships among features within each cluster and further learn distinct features of clusters. The intuition is that partitioning anomalies into clusters will enable the learning model to extract stable patterns among anomalous data points. To generate a ranked list of diagnoses, i.e., the impact of specific features on the detected anomaly, cluster assignments are treated as input to the explainable AI models such as Local Interpretable Model-Agnostic (LIME) and Shapley Additive Explanations (SHAP). These models identify and rank the most influential features, or diagnoses, for each instance during a specific time step. Once the features are identified for a certain number of time steps, i.e., m , we measure the distance of each ranked feature through *Fast Dynamic Time Warping* (FastDTW) [28], which measures the

distance and accounts for temporal dependencies in time series data. In simpler terms, the main motivation behind this step is to rank the most relevant features, i.e., diagnosis, of the cluster assignments. This additional step in the ranking process allows for a more nuanced understanding of how certain features influence anomalies over time. Through this comprehensive approach, we aim to create a more effective and contextually aware system for anomaly detection and diagnosis, leveraging both unsupervised and supervised learning techniques.

We evaluated our approach on the time series dataset published by Granderson et al. [11]. The simulated dataset contains minute recordings of the Air Handling Unit (AHU) sensors, capturing both their faulty and normal operational states. For the task of fault detection, we considered fault-free recordings, i.e., normal operations, as input to the one-class classification methods and the evaluation with 10-fold cross-validation shows that Elliptic Envelope performs well and identifies faults of all types with average metric scores of 89.4% F1, 94.2% precision, and 92.6% recall. As for the fault diagnosis, we rely on k -means clustering technique to group predicted faults and evaluate various binary classification methods on each cluster. We found that the Random Forest classifier performs well, on average, across all evaluation metrics. Finally, for the baseline comparison in diagnosing faults, we evaluated our method against a standard approach that combines the one-class classification technique with explainable models, measuring how well each method identifies and ranks features (diagnoses). We found that our method significantly outperforms the standard approach in terms various diagnosis metrics, i.e., Intersection over Union (IoU), HitRate@k, and RootCause@k. In summary, we contribute the following with our work:

- OCCEN combines learning aspects of both supervised and unsupervised techniques for improved performance. Unlike existing methods, which rely primarily on labeled faulty instances for fault diagnosis, our methodology relies only on fault-free time series data.
- We introduce a novel methodology of combining one-class classification with clustering techniques incorporating temporal dependencies to boost unsupervised fault diagnosis performance. Furthermore, we provide a ranked list of diagnoses to assist the operator in effectively diagnosing faults. To the best of our knowledge, our methodology is the first that considers combining learning methods in this way for anomaly detection and diagnosis in time series data.
- We evaluated OCCEN on the simulated data for real-world application of detecting and diagnosing faults in air handling units. The evaluation suggests that the proposed methodology consistently achieves better performance compared to the baseline.

In the following, we first discuss relevant works in Section 2. We formally define the problem statement and describe the building blocks of our proposed framework (OCCEN) in Section 3. The experimental setup details are provided in Section 4 followed by the discussion on obtained results in Section 5. We then discuss the potential implications and possible limitations of our work in Section 6, and finally conclude this work in Section 7.

2 Related Work

A variety of data-driven methodologies have been proposed over the years for the task of anomaly detection. Typical methods from clustering and classification types are considered the most relevant techniques for the task. For example, the k -means clustering technique [23] is usually adapted to group anomalous and normal data points into various clusters through similarity and/or distance measures [16, 19]. In this context, Li et al. [19] proposed a framework that analyzes long multivariate time series by dividing them into shorter subsequences

using a sliding window. Enhanced Fuzzy C-Means (FCM) clustering and particle swarm optimization are applied to identify structures and assign anomaly scores. Additionally, it includes a shape anomaly detection step, utilizing autocorrelation to capture shape information and address time shifts. In another work, Liu et al. [21] proposed an anomaly detection method for condition monitoring data, utilizing auxiliary feature vectors and DBSCAN clustering to differentiate between valid and invalid anomalies. This method segments the data based on interruptions, constructs auxiliary feature vectors for each segment, and applies Density-Based Spatial Clustering (DBSCAN) for accurate pattern recognition. A heuristic based on the 'number of clusters-Eps' curve is proposed to optimize DBSCAN parameters, effectively identifying normal patterns and anomalies in unlabelled, imbalanced datasets with rare anomalies. Clustering methods, while effective and not needing labeled data, are incapable of capturing temporal dependencies in multivariate time series across different time steps [31], a crucial aspect in anomaly detection and diagnosis. In our work, we address this issue by integrating temporal dependencies into the framework by combining explainable AI function output with the dynamic time-warping technique for the diagnostic process.

In addition to clustering techniques, classification methods are also applied [13, 6, 5]. Fouzi et al. [13] presented an effective approach for fault detection in Photovoltaics (PV) arrays, merging model-based strategies with One-class SVM (One-SVM) clustering. It utilizes a simulation model to replicate normal PV array behavior, applying One-SVM to the resulting discrepancies for fault identification. This approach, particularly adept at handling nonlinear features, showed superior performance in fault detection compared to other clustering methods in tests on a 9.54 kWp grid-connected PV plant. In contrast, this work proposes to combine unsupervised and OCC data-driven methods for anomaly detection and diagnosis. In another work, Beghi et al [5] proposed a novel semi-supervised, data-driven method. This technique utilizes Principal Component Analysis (PCA) to differentiate anomalies from typical operational variations, coupled with a reconstruction-based contribution method to identify fault-related variables. Fault diagnosis is facilitated through a decision table linking the impact of faults to specific features. The effectiveness of the Fault Detection and Diagnosis (FDD) algorithm is evaluated using experimental datasets from two different water chiller systems.

Finally, advanced methods such as deep learning are also used in the literature [31, 9]. Zhang et al. [31] proposed the Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) approach for detecting and diagnosing anomalies in multivariate time series data. It generates multi-resolution signature matrices to encapsulate different system status levels over time. This method effectively integrates a convolutional encoder and an attention-based ConvLSTM network to capture inter-sensor correlations and temporal dynamics, using a convolutional decoder to reconstruct and analyze these matrices for anomaly identification. In another work, Garg et al. [9] reviewed plenty of deep learning methods and evaluated the performance based on the newly proposed metric *composite F-score* (F_{c1}).

To summarize, we briefly discussed the adaptation of data-driven methodologies for anomaly detection and diagnosis in time series data. While these techniques share common characteristics such as feature generation, transformation, sampling, etc., the modeling is performed through one particular type of learning schema, i.e., clustering or classification. On the other hand, clustering methods are often preferable for real-world cases, however, such methods have limitations, i.e., the inability to capture temporal dependencies and non-linear relations among features. Therefore, in this work, we propose to combine classification and clustering methods and apply explainable AI in conjunction with the dynamic time-warping technique in a framework.

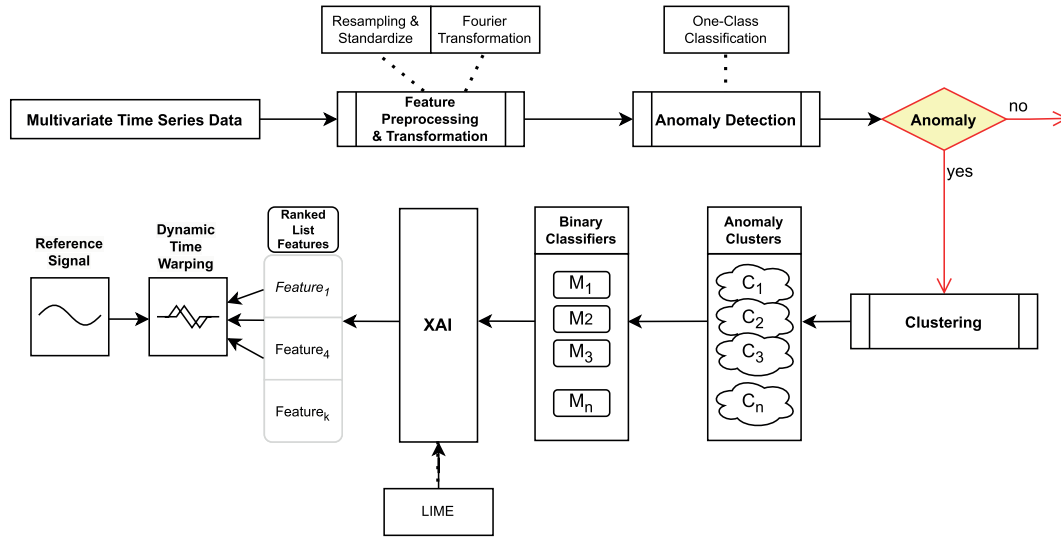
3 Framework & Methodology

In this section, we describe the problem formulation in detail along with the details regarding the building blocks of our proposed methodology.

Problem Statement: Let $T = \{t_1, t_2, t_3, \dots, t_m\}$ represents time series data observations for fault-free operations. For each data point $t_i \in T$, there is k number of recorded features, i.e., multivariate, hence, $t_i = \{x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,k}\}$.

3.1 Anomaly Detection and Diagnosis

The overall framework of OCCEN is presented in Figure 1. For the task of anomaly detection, the objective is to learn a function $f : T \rightarrow \{0, 1\}$ where $f(t_i) = 0$ represents a non-anomalous instance and $f(t_i) = 1$ indicates anomalous data point. Furthermore, it is assumed that characteristics of T represent normal operations, and deviations from normal patterns are indicative of faults.



■ **Figure 1** OCCEN Anomaly Detection and Diagnosis Framework.

The design of the diagnosis phase is provided in Algorithm 1. We now describe the algorithm in more detail. As part of the diagnosis, let $F = \{t_i | f(t_i) = 1, t_i \in T\}$ be the set of all the detected anomalies. Then clustering C can be applied to F such that $C(F) = \{c_1, c_2, c_3, \dots, c_n\}$ where n indicates the number of clusters obtained.

For each cluster, train a binary classifier B_j that distinguishes instances among clusters by learning distinctive cluster patterns. To train classifier B_j , let D_{c_j} be the cluster instances in c_j , then randomly select an equal number of instances from $\bigcup_{i \neq j} D_{c_j}$, i.e., the union of instances from other clusters to form D_{other} . Then the training set for B_j is $D_{c_j} \cup D_{other}$ ensuring a balanced representation of the target cluster and other clusters.

To generate a ranked list of diagnoses, let E be the explainable model, such as LIME, applied to each instance $t_i \in D_{c_j}$ such that $E(t_i)$ provides explanations, ranked list of features, for the cluster assignment. Then identify top-ranked features $F_{top}(t_i)$ that contribute to the assignment of t_i . Retain the top-ranked features $F_{top}(t_i)$ from $E(t_i)$ over a sequence of length m window. For each feature $f_r \in F_{top}(t_i)$ calculate *FastDTW* distance between f_r in the faulty sequence and its counterpart sequence in the fault-free observations T_{normal} and denote

■ **Algorithm 1** Diagnosis through Anomaly Detection, Clustering, and Feature Ranking.

Require: T : Set of all instances (both normal and anomalous)
Require: $f(t_i)$: Anomaly detection function (1 if anomalous, 0 if normal)
Require: E : Explainable model (e.g., LIME)
Require: m : Window length for feature retention
Require: T_{normal} : Set of fault-free (normal) observations
Ensure: Ranked list of critical features for root cause diagnosis

- 1: **Step 1: Anomaly Detection**
- 2: $F \leftarrow \{\}$ ▷ Set of all detected anomalies
- 3: **for** each $t_i \in T$ **do**
- 4: **if** $f(t_i) = 1$ **then**
- 5: $F \leftarrow F \cup \{t_i\}$
- 6: **end if**
- 7: **end for**
- 8:
- 9: **Step 2: Clustering**
- 10: Apply clustering C on F : $C(F) = \{c_1, c_2, \dots, c_n\}$
- 11:
- 12: **Step 3: Train Binary Classifiers for Each Cluster**
- 13: **for** each cluster $c_j \in C(F)$ **do**
- 14: $D_{c_j} \leftarrow$ instances in c_j
- 15: $D_{\text{other}} \leftarrow$ randomly select $|D_{c_j}|$ instances from all other clusters $i \neq j$
- 16: $D_{\text{train}} \leftarrow D_{c_j} \cup D_{\text{other}}$
- 17: Train binary classifier B_j on D_{train}
- 18: **end for**
- 19:
- 20: **Step 4: Feature Explanation and Ranking**
- 21: **for** each cluster $c_j \in C(F)$ **do**
- 22: **for** each instance $t_i \in D_{c_j}$ **do**
- 23: $F_{\text{top}}(t_i) \leftarrow$ top-ranked features from $E(t_i)$
- 24: Retain $F_{\text{top}}(t_i)$ over a window of length m
- 25: **end for**
- 26: **end for**
- 27:
- 28: **Step 5: Calculate FastDTW Distance for Top Features**
- 29: **for** each feature $f_r \in F_{\text{top}}(t_i)$ **do**
- 30: $DTW_{\text{distance}}(f_r, T_{\text{normal}}) \leftarrow \text{FastDTW}(f_r \text{ in faulty sequence}, f_r \text{ in } T_{\text{normal}})$
- 31: **end for**
- 32:
- 33: **Step 6: Rank Features by FastDTW Distance**
- 34: Rank features in $F_{\text{top}}(t_i)$ based on $\text{FastDTW}_{\text{distance}}(f_r, T_{\text{normal}})$ in descending order
- 35:
- 36: **Step 7: Apply Truncation**
- 37: Truncate the ranked feature list to retain only the most critical features with the highest FastDTW distances
- 38: **return** Truncated list of critical features for root cause diagnosis

this as distance $FastDTW(f_r, T_{normal})$. Now rank the features based on the $FastDTW$ distance in descending order to apply truncation. In the context of diagnoses, these features are considered critical for identifying the root cause based on the severity [31], i.e., distance, as these features show the most deviation from normal behavior. In the following sections, we now describe the learning methods employed and how these are adapted in this work.

3.2 One-class Classification (OCC)

Labeling large datasets, particularly high-resolution time series data, is very challenging in real-world situations because it requires a lot of time and effort. The sparsity of class labels (minority class), e.g., anomalies in our case, in the datasets negatively affects the learning capability and performance of the model due to the learning bias towards the dominating class labels (majority class). To address this, one-class classification (OCC) methods are often used [17]. One-class classification methods are a special case of binary- or multi-classification learning methods. These methods are extensively used for detecting anomalies and novelties in time series data [2]. The main concept involves learning the distribution of a single class, typically normal observations, and establishing decision boundaries that differentiate between inliers and outliers. Formally, boundaries for object z , i.e., each object z in the dataset is a multivariate characterization of non-anomalous characteristics, can be defined based on the following concepts [4]:

- the measure of how far away, i.e., $d(z)$, or how likely it is, i.e., $p(z)$, (similarity) that an object z resembles the target class, which is represented by a training set X_{train} .
- a threshold, either θ_d or θ_p , applied to this distance or probability (similarity).

The similarity and distance measurements are used as the decision-making function for categorizing new, unknown objects as either inliers or outliers. For example, an object is identified as an outlier if $d(z) > \theta_d$ or $p(z) < \theta_p$, and the opposite criteria apply for classifying it as an inlier. A variety of OCC methods proposed in the literature [17] utilize either one or both of the key concepts of similarity and distance analysis. For instance, OneClassSVM [29] defines boundaries by mapping objects into higher dimensional feature space to find hyperplanes characterizing normal data points. Whereas, the Local Outlier Factor (LOF) assesses the local density deviation of a given data point concerning its neighbors, identifying anomalies based on significant differences in local densities. Nonetheless, other methods like Elliptic Envelope, Isolation Forest [20], Gaussian Mixture Models (GMM) [26], and advanced methods such as auto encoders [25] primarily learn the normal representation of the given objects. Each of these methodologies applies the concepts of distance and similarity in different ways, tailored to identify anomalies or outliers in various types of data effectively. Their effectiveness can depend on the specific characteristics of the data and the context of the problem being addressed.

In our study, we utilize One-Class Classification (OCC) methods to identify anomalies within multivariate time series data, previously denoted as $T = \{t_1, t_2, t_3, \dots, t_m\}$. Each object in T is presumed to be normal and is used as input for the training of the model. The primary objective of the model is to accurately learn the characteristics of normal, anomaly-free data points. This learning is then applied to classify new data points as either normal or anomalous. The effectiveness of the model is measured by its ability to accurately classify both anomalous and non-anomalous new data points, with a particular focus on achieving high precision and recall in this classification process. Essentially, the model's success relies on its proficiency in correctly identifying anomalies while minimizing false positives and negatives.

3.3 Learning Clusters Representations

Diagnosis in non-linear dynamic systems proves difficult without prior information, particularly when anomaly detection is trivial, i.e., anomalies appear as clear outliers or sudden signal disruptions. Therefore, in this step, we design a diagnostic analysis framework that first groups anomalous objects based on their characteristics and subsequently learns each group's pattern representation. The intuition behind this methodological design is that similar faulty working conditions exhibit common anomalous characteristics, by extracting the representation of these akin faults, we can improve the association of features within each faulty condition.

In our approach to diagnosis, we initially use clustering techniques to group objects classified as anomalies by the OCC technique. These clusters capture the common characteristics of the anomalous objects, resulting in better feature associations within each cluster. However, clustering primarily groups objects based on similarities or density measures [12], and does not inherently extract complex, non-linear relationships within each cluster. To address this, we improve the cluster representations by assigning a unique label to each cluster and treating each cluster as a separate binary classification problem. This step is key in extracting inter-cluster characteristics. By doing so, the binary classifier aids in extracting more nuanced and distinct patterns from each cluster, thus providing a more precise understanding of each cluster. To achieve this, we train a separate binary classifier on each cluster. The learning data for the classifier is generated by selecting the objects of the relevant cluster and assigning the label 0 and randomly selecting an equal number of objects from other clusters, i.e., label 1, to account for the balanced representation of both classes. This process is repeated for all the clusters generated by the clustering technique.

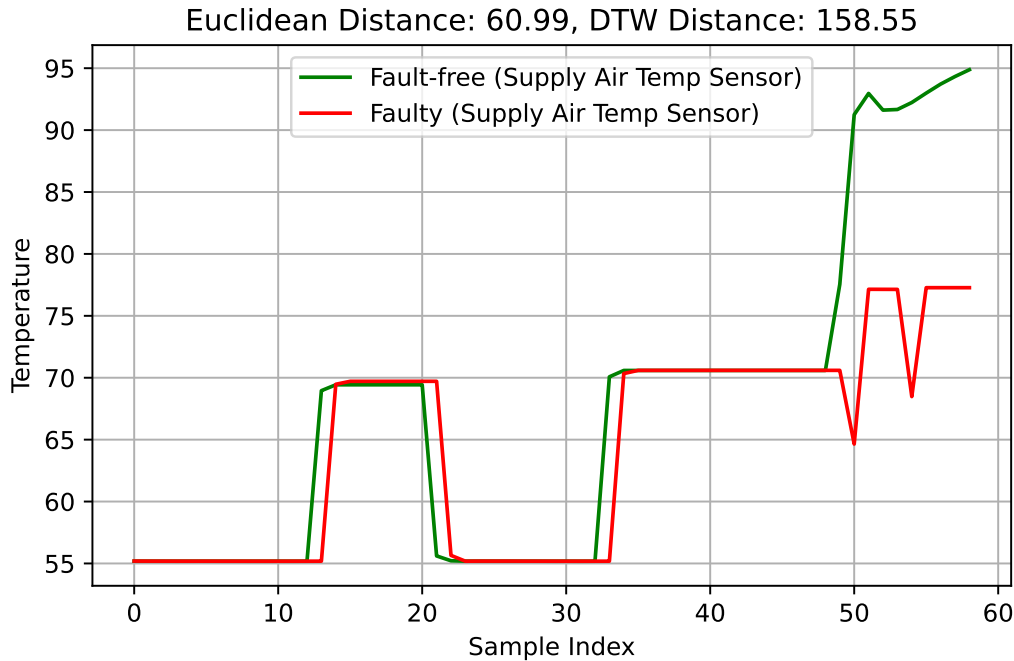
3.4 Explainable AI

Explainable AI methods such as LIME [27] and SHAP [22] are often considered useful to explain the decision made by the classifier [14]. This is accomplished by generating a ranked list of features that are indicative of the decision. These features are considered the most influential in terms of decision-making of the underlying base model. In this work, we use explainable methods as the first step toward the task of feature ranking for the anomalies. We identify the most significant features for each anomalous object in the time series. This is done by assigning the anomalous object to a particular cluster and then using the relevant binary classifier associated with that cluster, as the base model for the explainable AI method. So far, feature generation occurs at a specific time step t , and time series data often includes temporal dependencies that can impact the effectiveness of fault diagnosis. To address this, we take into account a set of features generated by explainable methods over a time sequence spanning a length of m . However, a longer time sequence will result in a large number of diagnoses, i.e., features, and it may affect the fault diagnosis performance. Therefore, to further pinpoint the diagnoses, we rely on the FastDTW to rank the features based on point-wise distance.

3.5 Fast Dynamic Time Warping (FastDTW)

FastDTW is a well-known technique to compare the similarity/distance between two given temporal time series [28]. It relies on calculating the similarity across various phases by minimizing the shifts and time distortion by “elastic” transformation. FastDTW performs better in precisely calculating and comparing two signals, particularly in time series data. Its capability to adjust for variations in time and phase differences allows for a more accurate

alignment and comparison of sequences than traditional methods [30]. This makes FastDTW an optimal choice for accurately determining the similarity or dissimilarity between two signals, especially when dealing with temporal discrepancies. For example, as shown in Figure 2, the analysis of metrics calculated via FastDTW versus Euclidean distance for both normal and faulty temperature sensor signal operations indicates better FastDTW's efficacy. In particular, FastDTW distance represents the total cost (distance) of aligning the two sequences (faulty and fault-free) through FastDTW. We note that FastDTW provides a more precise and reliable measure for detecting discrepancies in the sensor's readings, especially in distinguishing between normal and abnormal operational conditions. Therefore, we utilize FastDTW to determine the distance between features identified by the explainable method over m time steps and the corresponding normal observations of these features during the same period in a fault-free state. To achieve this, we retain the features identified by the explainable method across m time steps and then re-rank them by computing the distance between each feature and its corresponding normal state observation when the system is free of faults. Finally, a list of diagnoses is produced, ordered by their distances, with the diagnosis having the greatest distance placed at the top of the list.



■ **Figure 2** FastDTW vs Euclidean Distance Analysis.

4 Experimental Setup

4.1 Data

In this study, we employ a simulated dataset of a single-duct air handling unit (AHU), published by Granderson et al. [11]. The dataset comprises one year of operational data, incorporating both faulty and nominal system behaviors, with a total of 525,600 time samples representing each operational condition. Faults are imposed on sensors and control sequences at various biases (see Table 1), but our diagnostic approach identifies fault types without

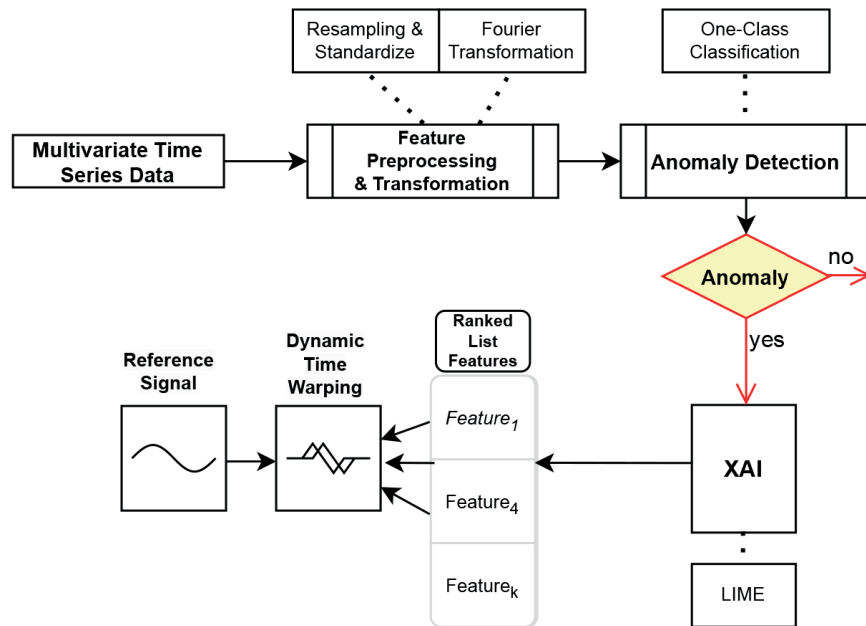
focusing on bias levels. We train OCC models for anomaly detection using only normal operations data. Pre-processing includes resampling to hourly averages, applying Fast Fourier Transformation (FFT), and feature scaling with the min-max method. These steps, along with trained models, are applied to preprocess test data, ensuring no data leakage.

■ **Table 1** Summary of Fault Imposition Methods for AHU, adapted from [11].

Fault Type	Fault Annotation	Method of Fault Imposition
Supply air temperature sensor bias	<i>Sa_temp</i>	Add a bias value to the sensor output
Outdoor air temperature sensor bias	<i>Oa_temp</i>	Add a bias value to the sensor output
OA damper stuck	<i>Dmpr_stk</i>	Automated override of outdoor air damper position to indicate it is stuck
Cooling coil valve leaking	<i>Cvlu_lkg</i>	Adjusted the coil valve position value when the control signal is zero
Cooling coil valve stuck	<i>Cvlu_stk</i>	Override of the coil valve position to indicate that the valve is stuck

4.2 Baseline

We evaluated the efficacy of different OCC methods in detecting anomalies and considered the best-performing method for the fault diagnosis analysis. To assess fault diagnosis performance, we established a simple baseline by employing explainable AI models immediately following the OCC methods, unlike OCCEN. The explainable model generates a list of ranked features, i.e., diagnoses, for a given time step. The top features ranked by the XAI model are then retained for over m time steps and further re-ranked based on distance measures calculated by FastDTW. The design of the baseline approach is shown in Figure 3. It is based on the premise that explainable AI models are generally good at identifying the key features that affect the decisions of the base model [27, 22] and FastDTW technique will further improve the ranking. The objective is now to assess how well this baseline performs in feature ranking compared to OCCEN.



■ **Figure 3** Baseline Approach Design.

4.3 Evaluation Procedure & Metrics

We experimented with multiple OCC methods for the classification of anomalies. We partitioned the dataset into two splits, using mid-August to December (30%) as a hold-out set and January to mid-August (70%) for time-series cross-validation in the anomaly detection task. Unlike traditional cross-validation, time-series cross-validation maintains the temporal order of observations, crucial for time-series data. We employed sequential 10-fold splitting, ensuring the training set includes only past data relative to the test set to prevent data leakage. For performance evaluation, we used Precision, Recall, and F1 Score, reporting the mean scores averaged over the folds for the OCC methods in Section 5. After identifying the most effective model, we used it to predict anomalies in the faulty dataset (January to mid-August) and applied various clustering methods, finding k -means most effective with an optimal cluster number of 4. This number is based on elbow method and silhouette coefficient analysis. We tested different classification models to understand each cluster's unique characteristics. The unseen hold-out test set was pre-processed and fed into our anomaly detection approach (OCCEN). Detected anomalies were categorized using k -means clustering, and corresponding trained binary classifiers with explainable models (LIME and SHAP) were applied. LIME was more efficient for explanation generation, while SHAP was more computationally intensive. We experimented with different sliding window sizes, focusing on LIME-prioritized features and using FastDTW to rank features based on their distance from normal signals.

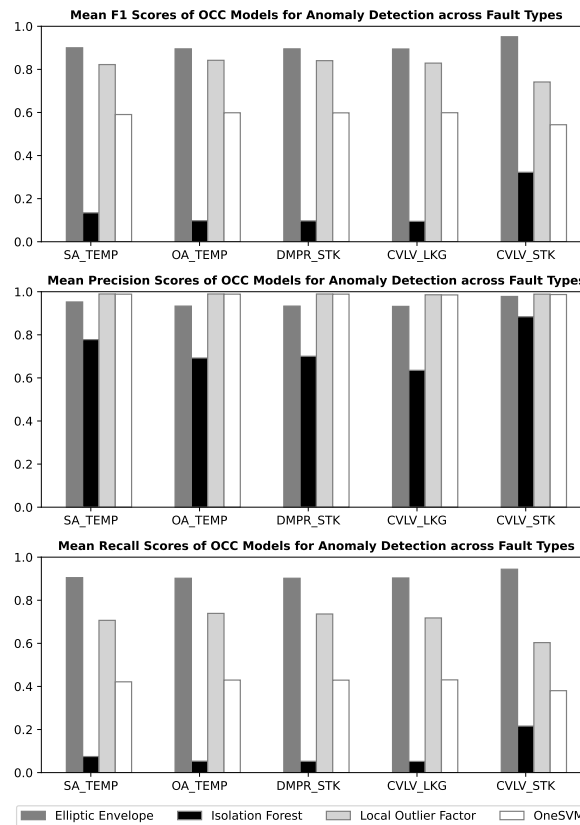
In our study, each fault type has a single root cause, but in non-linear dynamic systems, one failure can impact multiple components, making it crucial to examine these effects to identify the root cause. We consulted building automation experts to review the dataset and list possible diagnoses for each fault type. To evaluate fault diagnosis performance, we adapted two metrics: *Overlap@P* and *HitRate@k*. *Overlap@P* measures the overlap between true diagnoses and the top $P \times |GT|$ ranked diagnoses, with P as 150% or 200% and $|GT|$ as the number of true diagnoses. Most fault types have 5 true diagnoses, except *Sa_temp* with 2, capping ranked diagnoses at 10 when P is 200%. We also used *HitRate@k*, which checks if at least one true diagnosis is within the top k ranked diagnoses. Our dataset includes five fault variants: *Sa_temp*, *Oa_temp*, *Dmpr_stk*, *Cvlv_lkg*, and *Cvlv_stk*. To measure the accuracy of identifying root causes, we introduced *RootCause@k*, assessing the proportion of instances where the true cause is ranked among the top k diagnoses. We compared the effectiveness of the baseline and OCCEN using these metrics across sliding window sizes from 3 to 24 for diagnosis performance.

5 Results & Analysis

In this section, we discuss and report the results of anomaly detection and diagnosis for the experimental setup.

5.1 Anomaly Detection

The average results of the anomaly detection task are presented in Figure 4. The Elliptic Envelope model consistently shows high precision and recall across all fault types, resulting in strong F1 scores, particularly in detecting Cooling Coil Valve Stuck (*Cvlv_stk*) anomalies. Its precision above 93% and recall above 90% in most cases reflect its robustness and reliability in anomaly detection. The Isolation Forest model, however, presents a significant drop in performance, especially in recall values, which are notably low across all fault types. The



■ **Figure 4** 10-Fold Cross Validation Evaluation of OCC Methods.

Isolation Forest shows a precision of over 63% for most faults but struggles with recall, dipping as low as 5% in cases like Cooling Coil Valve Leakage (*Cvlv_lkg*) and Outdoor Air Damper Stuck (*Dmpr_stk*). The Local Outlier Factor (LOF) model exhibits high precision across all fault types, similar to the Elliptic Envelope model, but has a moderately lower recall, especially in the case of Cooling Coil Stuck (*Cvlv_stk*). Despite this, the LOF model maintains decent F1 scores, indicating a balanced trade-off between precision and recall. The One-Class SVM (OneSVM) model shows exceptional precision but considerably lower recall across all faults. This disparity results in moderate F1 scores, which are notably lower than those of the Elliptic Envelope and LOF models. The OneSVM is particularly weak in recall for Cooling Coil Stuck (*Cvlv_stk*), leading to its lowest F1 score among the fault types. Overall, the Elliptic Envelope model demonstrates the most balanced and effective performance in anomaly detection across various fault types, with the Local Outlier Factor (LOF) model following closely. In contrast, the Isolation Forest and OneSVM models, despite their high precision in some cases, lag in overall effectiveness due to significantly lower recall values.

5.2 Clusters Learners Performance

Analyzing the results in Table 2, we assess the performance of binary classifiers trained on each cluster using 10-fold cross-validation. High precision values indicate the model's ability to correctly identify anomalies within specific clusters, confirming the relevance and discriminative power of the features used. High recall values show the model's effectiveness in

identifying true positives, ensuring critical anomaly characteristics are captured. Consistently high F1 scores suggest a balance between precision and recall, indicating robust and reliable anomaly indicators within each cluster. Comparing model performances, Random Forest and Decision Tree models showed consistently high performance across all clusters, especially in Clusters 0 and 1, with F1 scores reaching 0.99. The Support Vector Machine also performed well, maintaining an F1 score of 0.97 across Clusters 0, 1, and 2. However, the Naive Bayes classifier showed variability, particularly in Cluster 2, where it had lower recall (0.68) and F1 Score (0.80), indicating issues with false negatives.

Since clusters are formed based on similarities in the data, each cluster likely represents different operational states or types of faults. The variation in model performance thus reinforces the idea that different clusters indeed capture distinct aspects of the system's operation, which can be utilized for more accurate fault diagnosis. The relatively low standard deviations in the metrics imply that the models' performances are consistent across different folds of the cross-validation. This consistency is important for practical applications as it indicates the models' robustness and reliability. The process of learning patterns within each cluster allows for the capture of cluster-specific patterns, which can help in better fault diagnosis. For this work, we selected the Random Forest binary classifier as our primary clusters learner as it outperforms other classifiers for most of the clusters.

■ **Table 2** 10-Fold cross validation Mean (Std) scores for the classification models learned on each cluster.

Model	Cluster Number 0			Cluster Number 1			Cluster Number 2			Cluster Number 3		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Decision Tree	0.96 (0.08)	0.99 (0.00)	0.97 (0.05)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.96 (0.11)	0.99 (0.00)	0.97 (0.06)	0.99 (0.01)	0.99 (0.00)	0.99 (0.00)
Naive Bayes	0.95 (0.08)	0.99 (0.00)	0.97 (0.05)	0.93 (0.13)	0.99 (0.00)	0.96 (0.08)	0.99 (0.02)	0.68 (0.09)	0.80 (0.05)	0.96 (0.03)	0.99 (0.00)	0.98 (0.01)
Support Vector Machine	0.96 (0.08)	0.99 (0.00)	0.97 (0.05)	0.96 (0.10)	0.99 (0.00)	0.97 (0.06)	0.96 (0.10)	0.99 (0.00)	0.97 (0.06)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
Random Forest	0.96 (0.08)	0.99 (0.00)	0.97 (0.05)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.92 (0.14)	0.99 (0.00)	0.95 (0.08)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
Logistic Regression	0.96 (0.09)	0.97 (0.01)	0.97 (0.05)	0.94 (0.10)	0.99 (0.00)	0.96 (0.06)	0.96 (0.09)	0.99 (0.00)	0.98 (0.05)	0.98 (0.02)	0.99 (0.00)	0.98 (0.01)

5.3 Diagnostic Performance

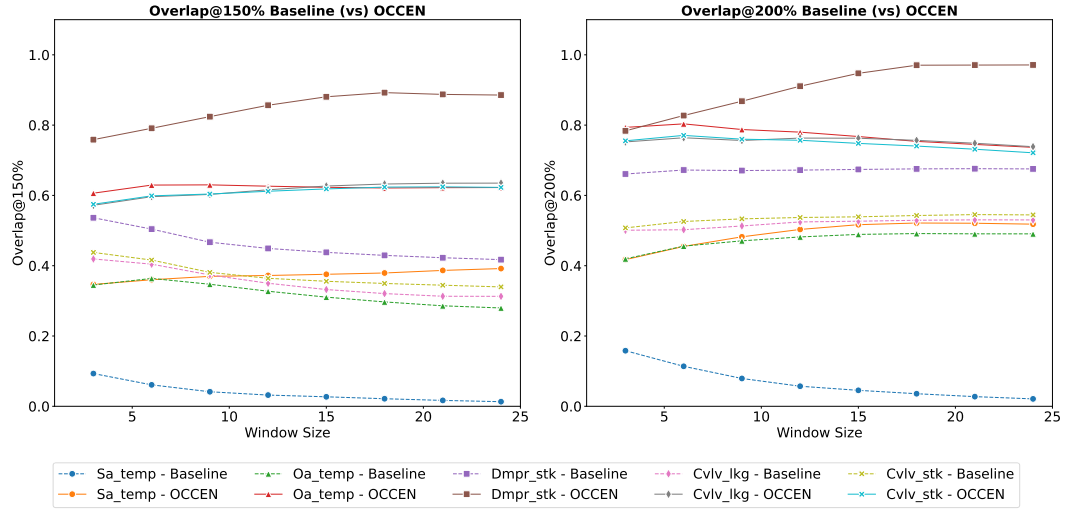
In the following, we provide a detailed analysis and discussion on the diagnostic performance OCCEN compared to the baseline across different types of faults.

5.3.1 Overlap@P

The baseline and OCCEN comparative analysis is presented in Figure 5. Overall, OCCEN consistently performs better than the baseline at ranking relevant diagnoses for all fault types and achieves higher overlap percentages. This becomes more pronounced for larger window sizes.

The diagnostic results for the fault type *Sa_temp* show that OCCEN consistently outperforms the baseline across all window sizes. The overlap metrics at 150% and 200% increase substantially with OCCEN as the window size grows, compared to the baseline performance. This indicates that larger window sizes provide more temporal information for diagnosing *Sa_temp* faults with OCCEN. Despite *Sa_temp* being a challenging fault to diagnose due to having only two true diagnoses amidst more ranked diagnoses, OCCEN

shows significant improvement over the baseline. For the *Oa_temp* fault type, OCCEN also outperforms the baseline consistently, although there is a slight decrease in overlap beyond a window size of 9, suggesting a potential saturation point. Nonetheless, OCCEN achieves 80% overlap with true diagnoses at 200% compared to only maximum of 49% for the baseline. The *Dmpr_stk* fault type shows significant performance improvement with OCCEN. The overlapping score is up to 82% and 92%, compared to the baseline's maximum of 53% and 67%. This indicates OCCEN's efficacy in diagnosing *Dmpr_stk* faults, particularly with larger window sizes. Similarly, for the *Cvlv_lkg* fault type, OCCEN consistently outperforms the baseline. The higher overlaps across all window sizes, though the increase with window size is less pronounced, suggesting smaller windows may suffice for this fault type. OCCEN achieves overlaps of up to 64% at 150% and 76% at 200%, significantly higher than the baseline. *Cvlv_stk* fault type also shows OCCEN achieving higher scores compared to the baseline across all window sizes. The overlaps increase with window size for OCCEN, reaching up to 62% at 150% and 77% at 200%, whereas the baseline reaches a maximum of 43% and 54%, respectively.



■ **Figure 5** Mean score comparison of Overlap@P metric for baseline and OCCEN.

Overall, OCCEN demonstrates significant improvements over the baseline across all fault types, consistently achieving higher overlap metrics. The varying impacts of window size across different fault types suggest that while some faults benefit from larger windows, others like *Cvlv_lkg* perform well with smaller windows. Nevertheless, OCCEN's overall performance highlights its effectiveness and robustness in diagnosing faults.

5.3.2 HitRate@k

The overall evaluation of HitRate@k metric is presented in Figure 6. Overall, the results demonstrate clear performance improvements for OCCEN in identifying at least one true diagnoses in top k ranked diagnoses and it improves as window size increases. Conversely, the baseline shows, in general, lower hit rate scores but performs relatively better for large window size.

The fault type *Sa_temp* is challenging to diagnose due to having only two true diagnoses, while the ground truth diagnoses exceed this number. Despite this complexity, OCCEN demonstrates a significant improvement over the baseline across all window sizes. The hit

rates for $k = 3, 5$, and 7 steadily increase for OCCEN. In contrast, the baseline method struggles, with hit rates peaking at 18%, 45%, and 75% for the same window size. For the fault type *Oa_temp*, OCCEN outperforms the baseline across all window sizes. The hit rates for OCCEN are significantly higher across all k -values. The baseline method, however, exhibits lower hit rates, with a maximum of 88% at a window size of 3 for $k=7$. The results for the fault type *Dmpr_stk* reveal a similar trend, with OCCEN consistently achieving higher hit rates than the baseline. OCCEN maintains near-perfect hit rates across all k -values and window sizes, with hit rates reaching up to 100% at larger window sizes. The baseline, on the other hand, peaks at 34%, 75%, and 98% for $k = 3, 5$, and 7 , respectively. These results highlight the better identification of true diagnosis by the OCCEN for *Dmpr_stk* faults.

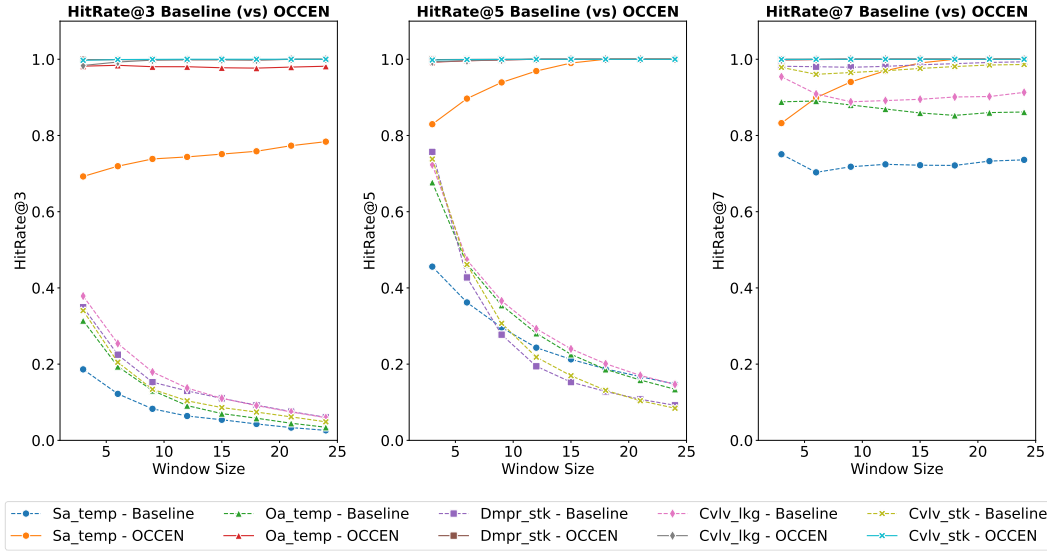


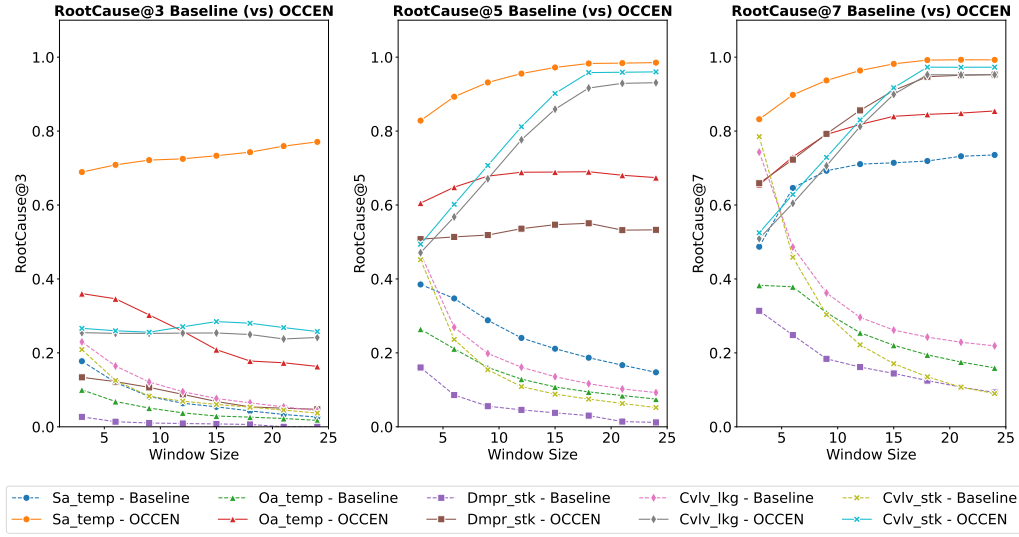
Figure 6 Mean score comparison of HitRate@k metric for baseline and OCCEN.

For the fault type *Cvlv_lkg*, OCCEN also demonstrates better performance compared to the baseline across all k -values and window sizes. The hit rates for the OCCEN are significantly higher, with perfect hit rates at $k = 7$ for window sizes of 12 and beyond. The baseline, however, achieves lower hit rates. Lastly, the fault type *Cvlv_stk* showcases a clear trend where OCCEN outperforms the baseline. The hit rates for OCCEN remain consistently high across all window sizes, reaching up to 100% for $k = 5$ and 7 across multiple window sizes. The baseline performance remains similar to the *Cvlv_lkg* performance. This is mainly because these points are part of the same component, i.e., cooling coil. Overall, the OCCEN consistently outperforms the baseline method in diagnosing the various fault types, as evidenced by higher *HitRate@k* metrics for all examined values of k and across different window sizes.

5.3.3 RootCause@k

Remember, in this study, each fault type has only one true cause. Therefore, in this evaluation procedure we are particularly interested in measuring how accurately the true fault cause is identified within the top- k ranked diagnoses. Figure 7 provides an overview of the outcomes and at first glance it can be observed that OCCEN performs better at ranking true diagnoses than the baseline.

Looking closely, for the fault type *Sa_temp*, the OCCEN method shows steady improvement and this gradually gets better as k increases. However, the baseline showing minimal improvement, if any, while OCCEN continues to perform well. Notably, OCCEN achieves a maximum RootCause@3 score of 98%, indicating significant improvement over the baseline. In contrast, the baseline method performs poorly across all window sizes and k values. Interestingly, as the window size increases, the baseline's performance declines further, with the RootCause@3 score dropping to 2% at a window size of 24. The performance disparity between the baseline and OCCEN is even more pronounced for the fault type *Oa_temp*. OCCEN begins with a score of 36% at a window size of 3 and gradually declines to 16% at a window size of 24. Despite this decline, OCCEN consistently outperforms the baseline across all window sizes and k values. OCCEN more consistently ranks the root cause for *Oa_temp* within the top 5 diagnoses compared to the baseline. In contrast, the baseline starts with a score of 1% at a window size of 3, which decreases as k increase. The baseline struggles to adapt as the window size increases, highlighting its limitations in ranking the root cause effectively.



■ **Figure 7** Mean score comparison of RootCause@ k metric for baseline and OCCEN.

For the fault type *Dmpr_stk*, OCCEN also struggles to rank the root cause within the top 3 diagnoses. However, for the RootCause@5 metric, OCCEN improves significantly, with scores between 50% and 53%, while the baseline's performance declines. The baseline method performs poorly on the RootCause@3 metric across all window sizes, scoring as low as 2% at a window size of 3 and staying below 1% even as the window size increases. For the RootCause@7 metric, the baseline's performance remains weak, starting at 31% and dropping to 9%. In contrast, OCCEN shows strong performance, starting at 65% and reaching 95% demonstrating its ability to rank the true cause among the top seven diagnoses.

For the fault type *Cvlv_lkg*, the baseline starts relatively well with a RootCause@3 score of 22% but its performance declines to 4%. In contrast, OCCEN consistently improves, particularly for larger k values and window sizes. The overall trend of OCCEN significantly improves for larger k values in contrast to baseline. For RootCause@5, the baseline begins at 46% but drops to 9% as the window size increases, whereas OCCEN starts at 47% and rises to 93%. Similarly, for RootCause@7, the baseline initially scores 74% but falls to 21%, while OCCEN improves from 50% to 95%. The pattern is consistent for the fault type *Cvlv_stk*,

where OCCEN significantly outperforms the baseline for larger k values and window sizes, achieving a RootCause@5 score of 96% and a RootCause@7 score of 97% at a window size of 24.

In summary, OCCEN more frequently ranks the root cause at higher positions compared to the baseline, showcasing its diagnostic capability by learning cluster representations and assigning relevant features (diagnoses) to the system's faulty operational conditions. This analysis is also consistent with the overlapping diagnoses and OCCEN's ability to rank at least one true diagnosis at the top k places.

6 Implications & Threats to Validity

In this section, we describe the implications and limitations of our work.

6.1 Implications

In real-world applications, especially dynamic systems, labeled anomalous instances are frequently unavailable due to the substantial time and effort required for their collection. Consequently, our work focuses on leveraging only the normal operations of complex dynamic systems to design an effective anomaly detection and diagnosis framework. We achieve this by combining data-driven techniques in a novel manner. Additionally, this work prioritizes and addresses the requirements of the operators, who often seek a list of diagnoses to identify and pinpoint the root cause of issues within the system. Furthermore, the framework's design enables the integration of various data-driven learning methodologies, e.g., clustering, and classification, including advanced techniques like deep learning. However, for this study, we focus solely on traditional learning methods to understand the distribution of normal system operations and to learn cluster representations for associating features with faulty states.

6.2 Threats to Validity

Like any other research work, our work has its limitations. We evaluated our approach against a simple baseline to demonstrate its effectiveness. The rationale behind this evaluation method was to showcase its practical applicability for anomaly detection and diagnosis. To our knowledge, there has been limited work in the existing literature focused on this specific use case, particularly regarding diagnostic functionality that provides a ranked list of diagnoses. While we report our findings for a multivariate time series dataset related to an air handling unit, further evaluation on other real-world datasets is needed to assess generalizability. Further, determining an optimal number of clusters for real-world applications are necessary and planned for future work. Finally, the FastDTW step relies on a reference signal of “normal” behavior, which is often unavailable without a simulation model or historical normal data. Despite these limitations, we believe the proposed framework has strong potential for similar problems.

7 Conclusion

In this study, we developed and proposed an anomaly detection and diagnosis framework called OCCEN. OCCEN relies solely on non-anomalous instances of multivariate time series. Learning cluster representation of anomalies, where each cluster can represent a fault type, the framework enhances feature significance and associations within each cluster. A ranked list of diagnoses is produced for each cluster assignment (i.e., anomaly) by employing

the XAI method (LIME) and the sequence matching technique (FastDTW) to consider temporal dependencies. The extensive evaluation of OCCEN on a synthetic dataset of the real-world use case of an air handling unit demonstrated that OCCEN outperforms the classical baseline method.

References

- 1 Anam Abid, Muhammad Tahir Khan, and Javaid Iqbal. A review on fault detection and diagnosis techniques: basics and beyond. *Artificial Intelligence Review*, 54:3639–3664, 2021. doi:10.1007/S10462-020-09934-2.
- 2 Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 8–15, 2013.
- 3 Yeonjin Bae, Saptarshi Bhattacharya, Borui Cui, Seungjae Lee, Yanfei Li, Liang Zhang, Piljae Im, Veronica Adetola, Draguna Vrabie, Matt Leach, et al. Sensor impacts on building and hvac controls: A critical review for building energy performance. *Advances in Applied Energy*, 4:100068, 2021.
- 4 Anna M Bartkowiak. Anomaly, novelty, one-class classification: a comprehensive introduction. *International Journal of Computer Information Systems and Industrial Management Applications*, 3(1):61–71, 2011.
- 5 A Beghi, R Brignoli, Luca Cecchinato, Gabriele Menegazzo, Mirco Rampazzo, and F Simmini. Data-driven fault detection and diagnosis for hvac water chillers. *Control Engineering Practice*, 53:79–91, 2016.
- 6 Efreem Heri Budiarto, Adhistya Erna Permanasari, and Silmi Fauziati. Unsupervised anomaly detection using k-means, local outlier factor and one class svm. In *2019 5th international conference on science and technology (ICST)*, volume 1, pages 1–5. IEEE, 2019.
- 7 Xuwu Dai and Zhiwei Gao. From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Trans. Ind. Informatics*, 9(4):2226–2238, 2013. doi:10.1109/TII.2013.2243743.
- 8 Laura Erhan, M Ndubuaku, Mario Di Mauro, Wei Song, Min Chen, Giancarlo Fortino, Ovidiu Bagdasar, and Antonio Liotta. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, 67:64–79, 2021. doi:10.1016/J.INFFUS.2020.10.001.
- 9 Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2021. doi:10.1109/TNNLS.2021.3105827.
- 10 Zhiqiang Ge. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, pages 16–25, 2017.
- 11 Jessica Granderson, Guanqing Lin, Yimin Chen, Armando Casillas, Jin Wen, Zhelun Chen, Piljae Im, Sen Huang, and Jiazhen Ling. A labeled dataset for building hvac systems operating in faulted and fault-free states. *Scientific Data*, 2023.
- 12 H Burak Gunay and Zixiao Shi. Cluster analysis-based anomaly detection in building automation systems. *Energy and Buildings*, 228:110445, 2020.
- 13 Fouzi Harrou, Abdelkader Dairi, Bilal Taghezouit, and Ying Sun. An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine. *Solar Energy*, 179:48–58, 2019.
- 14 Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pages 13–38. Springer, 2022.
- 15 Rolf Isermann. *Fault-diagnosis applications: model-based condition monitoring: actuators, drives, machinery, plants, sensors, and fault-tolerant systems*. Springer Science & Business Media, 2011.

- 16 Neha Kant and Manish Mahajan. Time-series outlier detection using enhanced k-means in combination with pso algorithm. In *Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India*, pages 363–373. Springer, 2019.
- 17 Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014. doi:10.1017/S026988891300043X.
- 18 Zhiling Lan, Ziming Zheng, and Yawei Li. Toward automated anomaly identification in large-scale systems. *IEEE Transactions on Parallel and Distributed Systems*, 21(2):174–187, 2009. doi:10.1109/TPDS.2009.52.
- 19 Jinbo Li, Hesam Izakian, Witold Pedrycz, and Iqbal Jamal. Clustering-based anomaly detection in multivariate time series data. *Applied Soft Computing*, 100:106919, 2021. doi:10.1016/J.ASOC.2020.106919.
- 20 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.
- 21 Hang Liu, Youyuan Wang, and WeiGen Chen. Anomaly detection for condition monitoring data using auxiliary feature vector and density-based clustering. *IET Generation, Transmission & Distribution*, 14(1):108–118, 2020.
- 22 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- 23 James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- 24 Luigi Martirano and Massimo Mitolo. Building automation and control systems (bacs): a review. In *2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pages 1–8, 2020.
- 25 Walter Hugo Lopez Pinaya, Sandra Vieira, Rafael Garcia-Dias, and Andrea Mechelli. Autoencoders. In *Machine learning*, pages 193–208. Elsevier, 2020.
- 26 Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.
- 27 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.
- 28 Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007. URL: <http://content.iospress.com/articles/intelligent-data-analysis/ida00303>.
- 29 Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. doi:10.1162/089976601750264965.
- 30 Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.
- 31 Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1409–1416, 2019. doi:10.1609/AAAI.V33I01.33011409.