A novel Transformer-based Anomaly Detection approach for ECG Monitoring Healthcare System

Thi Thuy Van Nguyen[1,2,⋆], Cédric Heuchenne[1], Kim Duc Tran[3], and Kim Phuc Tran[2]

[1] HEC Liège - Management School of the University of Liège, Belgium
{ttv.nguyen,c.heuchenne}@uliege.be
[2] University of Lille, ENSAIT, GEMTEX, Lille, France
kim-phuc.tran@ensait.fr
[3] International Research Institute for Artificial Intelligence and Data Science, Dong A University, Danang, Vietnam,
ductk@donga.edu.vn

Abstract. Anomaly detection plays a crucial role across various domains, including healthcare, where identifying deviations from normal patterns can lead to early intervention and improved outcomes. In healthcare, such as in ECG analysis, detecting anomalous signals is essential for timely diagnosis and treatment, as it can help identify potentially life-threatening conditions that might otherwise go unnoticed. In this work, by focusing on ECG anomaly detection as an illustrative healthcare application, we propose to use a transformer-based variational autoencoder network together with a MEWMA-SVDD control chart to achieve anomaly detection. By employing this approach, we can effectively control the false alarm rate, aligning with the intended goal of minimizing unnecessary alerts. Our proposed framework not only excels in terms of accuracy but also reduces the false alarm rate, making it a favorable choice compared to existing methods.

Keywords: Anomaly detection, Transformer, Variational autoencoder, MEWMA, Support vector data description, Control chart, Federated learning, Embedded artificial intelligence.

1 Introduction

Detecting anomalies is critically important in all aspects of life, as it enables us to identify and address unexpected deviations from typical or expected behavior, ensuring safety, efficiency, and overall well-being. For instance, in healthcare, timely identification of anomalous patterns in medical data like ECG signals can lead to early intervention for potentially life-threatening conditions. In finance, detecting unusual transactions can help prevent fraud and financial risks. In cybersecurity, identifying abnormal behaviors can thwart attacks and intrusions. These examples highlight the substantial contribution of anomaly detection in enhancing and ensuring effectiveness across diverse domains.

The identification of anomalies or outliers has received much attention from both the statistics and machine learning (ML) research sectors. Many studies have been carried out to provide efficient solutions for dealing with this task in the literature. In statistics, many works have been carried out to develop anomaly detection methods. For example, in Suman and Prajapati [21], the authors explored the application of Statistical Process Control (SPC) and control charts in healthcare, tracing their evolution from industrial manufacturing to patient care, and identifying a need for broader deployment across different departments and countries in the healthcare sector. An innovative, lightweight, and model-free method for online detection of cardiac anomalies, such as ectopic beats in electrocardiography (ECG) or PPG signals was introduced in the work of Lang [13]. This approach utilizes Singular Spectrum Analysis (SSA) and nonparametric rank-based cumulative sum (CUSUM) control charts for efficient anomaly detection without the need for identifying fiducial points, making it computationally less intensive than previous SSA-based methods. In [10] the authors employed control charts to quickly and effectively monitor complex frequency signals, facilitating rapid response and anomaly detection in critical processes and testing the model in both simulation and ECG datasets. In the work of Zhou and Kan [30], a novel sensor-based unsupervised framework combining Gramian Angular Difference Field (GADF), multi-linear principal component analysis (MPCA), and control charts (Hotelling T2) for ECG anomaly detection was proposed. These are just a few examples of the applications of statistics and SPC in

⋆ Corresponding author, Email:ttv.nguyen@uliege.be

healthcare, and there are numerous other instances of SPC application across diverse domains, from healthcare to manufacturing and beyond.
Recently, there has been a growing trend towards developing machine learning techniques for anomaly detection, driven by the increased availability of large and diverse datasets, advancements in machine learning algorithms, and the desire to achieve higher accuracy in identifying anomalies in various domains, such as cybersecurity [18;6], manufacturing [11], industry [4] and healthcare [8]. In the work of Nassif et al. [14], the authors presented a systematic literature review (SLR) on ML models for anomaly detection from 2000-2020, identifying 290 research articles covering 43 different applications, 29 distinct ML models, and 22 datasets used in anomaly detection experiments. The study highlights a preference for unsupervised anomaly detection, with intrusion detection, network anomaly detection, general anomaly detection, and data applications being the most frequently studied areas in anomaly detection.

While numerous state-of-the-art ML models have demonstrated high accuracy and addressed data privacy concerns in anomaly detection, the critical aspect of controlling false alarm rates remains relatively unaddressed, posing potential adverse consequences; for example, in healthcare, a false alarm can disrupt patient care, lead to resource overload, and erode trust in the anomaly detection system, ultimately compromising patient safety. In this work, we propose to integrate a transformer-based autoencoder network proposed by Raza et al. [17] together with MEWMA-SVDD control chart to detect anomalies. According to this proposal, a transformer-based variational autoencoder (VAE) network will be used to extract important information from the data while the task of detecting the abnormal objects from the input is left to the MEWMA-SVDD. The idea of applying MEWMA techniques makes the model more sensitive to small shifts or changes in the time series while employing the concept of a control chart can help in effectively managing the false alarm rate. Besides, there has been a growing trend in incorporating artificial intelligence (AI) capabilities, into embedded systems and devices. Our objective is also to develop a lightweight framework that is suitable for implementation in embedded devices with the aim of the development of embedded artificial intelligence in the future. The rest of this paper is organized as follows: In Section 2, the main concepts used in this study, including transformer, VAEs, and MEWMA-SVDD algorithm are introduced; the detail of the proposed transformer-based VAE-MEWMA-SVDD framework for anomaly detection in multivariate time series data is presented in Section 3; in Section 4, the details of the experiment used to test the performance of the proposed framework, and the results obtained are presented; the perspective to improve the proposed framework in terms of data privacy concern is presented in section 5 and finally, the conclusions are given in Section 6.

2 Related works and background

In this section, we discuss the main concepts used in this study.

2.1 Transformers

The transformer is a neural network architecture that was introduced in the paper "Attention is All You Need" by Vaswani et al. [27] in 2017 and since then, it has become a cornerstone in natural language processing (NLP) and has been used in various other fields due to its ability to handle sequential data efficiently. Similar to recurrent neural networks (RNNs), another neural network designed for handling sequential data, transformers utilize an encoder and a decoder architecture. However, transformers distinguish themselves by utilizing self-attention mechanisms, allowing them to capture dependencies across the entire sequence without considering distance, while RNNs rely on recurrent connections to sequentially pass information through time steps. Both the encoder and decoder in the transformer architecture consist of stacked layers that include multi-head self-attention, additive connections, layer normalization, and position-wise feed-forward layers. Additionally, positional encoding is applied to the input embeddings in a pre-processing step, ensuring that the model can retain the positions of elements within the sequence, as the transformer architecture does not inherently incorporate any information about sequence order. The core innovation of the transformer is the self-attention mechanism, which enables the model to assess the significance of different parts within the input sequence while processing each element. The formula for computing the attention scores for each element in the input sequence is as follows

$$\text{Attention}(\mathbf{Q},\ \mathbf{K},\ \mathbf{V}) = \text{Softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\Big)\mathbf{V} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ represent query, key, and value matrices derived from the input sequence, and d_k is the dimension of the key vectors. Instead of executing a single attention function with d_{model}-dimensional keys, the attention mechanism is employed multiple times through the use of parallel attention heads, each equipped with different projections of query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices. This design enables the model to learn diverse representations of the input sequence, thereby enhancing its ability to capture complex patterns and relationships. The outputs of the multiple attention heads are concatenated and linearly transformed by multiplying query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices with the weight matrices $\mathbf{W}$ which are computed during the training process to produce the final output of the multi-head attention layer

$$\text{Multihead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\text{Attention}_1, \text{Attention}_2, \ldots, \text{Attention}_h) \cdot \mathbf{W}^O \tag{2}$$

$$\text{Attention}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{W}_i^Q(\mathbf{K}\mathbf{W}_i^K)^T}{\sqrt{d_k}}\right)(\mathbf{V}\mathbf{W}_i^V) \tag{3}$$

where h represents the number of attention heads, and $\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{W}_i^V$ are learnable weight matrices for query, key, and value projections specific to the i^{th} head, respectively, and $\mathbf{W}^O$ is a learnable weight matrix used for the final linear transformation. Fig illustrates the multi-head attention mechanism proposed by Vaswani et al. [27] Multi-head attention is used in both the encoder and decoder to capture different aspects of the input sequence

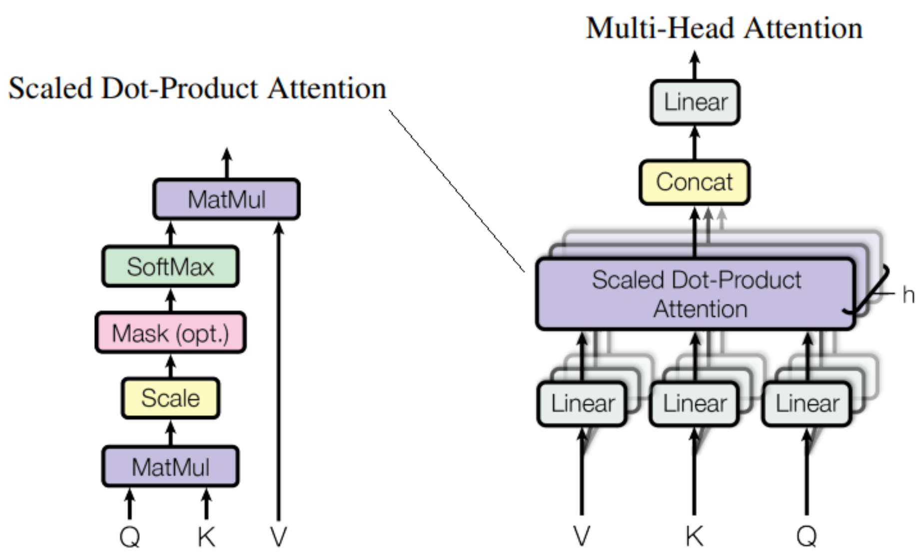

Fig. 1. (Right) Multi-Head Attention with h attention layers running in parallel proposed by Vaswani et al. [27]

simultaneously, enabling the model to understand complex relationships between words. Additionally, feedforward neural network layers and other components, such as layer normalization and residual connections, enhance the model's ability to learn and process sequential data efficiently. Transformer architecture has been the foundation for many subsequent models, including BERT, GPT, and many others, which have achieved state-of-the-art results in various NLP tasks and beyond, see Wolf et al. [28]; Chen et al. [5]; Singh and Mahmood [19], and Acheampong et al. [1], among many others.

2.2 Variational autoencoders

Variational Autoencoders (VAEs) is a type of generative model used in unsupervised machine learning and they represent a powerful paradigm in the field of deep learning. VAEs were first introduced by Kingma et al. [12] and since then, it has been widely used for various applications, including data generation, data denoising, representation learning, and anomaly detection. Unlike traditional autoencoders that learn deterministic mappings to fixed points in the latent space, VAEs shift their focus to learning a probability distribution of data in the latent space, often explicitly regularized to resemble a standard normal distribution, which is the most common choice, in order to capture the underlying data structure.

Let X denote the given input data and assume that X originates from a latent variable Z, which is not directly observable. In VAEs, the authors consider the probabilistic encoder (characterized by the conditional distribution $p_\theta(Z|X)$) and decoder (defined by $p_\theta(X|Z)$) instead of deterministic ones, where θ are the network parameters that characterize p. Assume that encoded representations Z in the latent space follow the prior distribution $p_\theta(Z)$. According to Bayes theorem, the relationship between prior $p_\theta(Z)$, the likelihood $p_\theta(X|Z)$, and the posterior $p_\theta(Z|X)$ is given by

$$p_\theta(Z|X) = \frac{p_\theta(X|Z) \cdot p_\theta(Z)}{p_\theta(X)} = \frac{p_\theta(X|Z) \cdot p_\theta(Z)}{\int p_\theta(X|Z) \cdot p_\theta(Z)\ dz} \tag{4}$$

Suppose that $p_\theta(Z)$ is a standard normal distribution and that $p_\theta(X|Z)$ is a normal distribution, we can compute $p_\theta(Z|X)$ using Bayes theorem. However, this computation is intractable in most cases due to the integral at the denominator of Eq. 4. To solve this problem, Kingma et al. [12] suggested using the variational inference technique. More precisely, the authors tried to approximate $p_\theta(Z|X)$ by another tractable distribution $q_\phi(Z|X)$, i.e, $p_\theta(Z|X) \approx q_\phi(Z|X)$, where ϕ represents the parameters of the approximate distribution, and minimize a metric describing the difference between two probability distributions named Kullback-Leibler divergence, $KL(q_\phi(Z|X)||p_\theta(Z|X))$,

$$KL(q_\phi(Z|X)||p_\theta(Z|X)) = \mathbb{E}_{q_\phi(Z|X)}\left[\log \frac{q_\phi(Z|X)}{p_\theta(Z|X)}\right] \tag{5}$$

to ensure the similarity of these two distributions.

After the input data is mapped to a probabilistic distribution in the latent space in the encoder, the decoder takes samples Z from the latent space, which follow a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma}\boldsymbol{\sigma}^T$

$$Z = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\odot$ is the element-wise product, and using these samples to generate data that resembles the original input. Introducing $\boldsymbol{\epsilon}$ as an independent random variable ensures the differentiability of the sampling process, allowing gradients to flow through the sampling operation during backpropagation. This technique is commonly referred to as the re-parameterization trick, and it is a key component of variational autoencoders (VAEs) that enables the model to facilitate gradient-based optimization and training, making VAEs effective for various generative and representation learning tasks. Figure 2.2 illustrates the structure of a VAE.

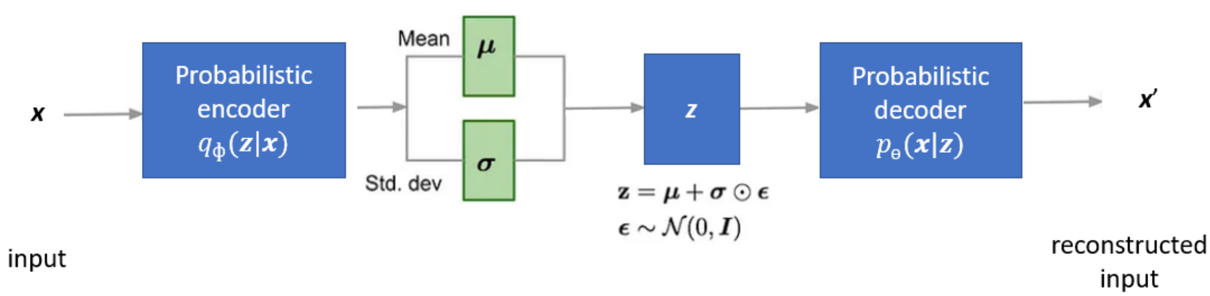

Fig. 2. The general structure of a VAE.

Let X' denote the reconstructed data. The objective function of VAEs comprises two loss terms: the reconstruction loss $L(X, X')$, which assesses the decoder's ability to reconstruct input data, and the Kullback-Leibler (KL) divergence, which quantifies the divergence between the learned distribution $q_\phi(Z|X)$ in the latent space and the true distribution $p_\phi(Z)$, which is often assumed to be a standard Gaussian distribution.

$$\text{Loss (VAEs)} = L(X, X') + \sum_i KL(q_{\phi,i}(Z|X) \| p_\phi(Z)) \tag{6}$$

The choice for the reconstruction loss depends on the type of data we are working with. VAEs training aims to minimize this objective function, often achieved through gradient descent optimization, to enhance both reconstruction accuracy and the adherence of the latent space to the desired distribution. In machine learning and artificial intelligence, VAEs have proven to be a powerful tool due to their ability to learn probabilistic representations of data. VAEs find utility across diverse domains, including generative modeling, data compression, and anomaly detection see, for example, Anstine and Isayev [2], Duan et al. [7], and Raza et al. [17].

2.3 Multivariate Exponentially Weighted Moving Average (MEWMA) and Support Vector Data Description (SVDD)

In ML, SVDD is a well-established algorithm primarily utilized for anomaly detection, effectively identifying outliers within datasets. Initially proposed by Tax and Duin [23], SVDD has garnered extensive recognition and adoption within the field of anomaly detection. One of its key advantages lies in its ability to operate without making restrictive assumptions about the underlying data distribution, as evidenced by its utilization in various studies, including those by Banerjee et al. [3], Pauwels and Ambekar [16], and Zhang and Deng [29], among numerous others. In this section, we will briefly present the main idea of the SVDD algorithm based on the MEWMA technique. Specifically, since we are focusing on detecting anomalies in a time series dataset, we will first apply the MEWMA technique by calculating the **weighted average** of the components within each individual time series, employing a memory-based approach, and then use this transformed dataset to train the SVDD model.

Recall SVDD is a one-class classification algorithm that allows us to detect abnormal observations by modeling the normal ones. This algorithm obtains a spherical-shaped boundary around the dataset, specified by its center $\mathbf{a}$ and radius R. The objective is to minimize the radius R, thereby minimizing the volume of the sphere while ensuring that it contains the majority of training objects. Suppose we have an unlabeled time series training set of size N, $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i \in \mathbb{R}^p$, p is the number of time steps, and we want to obtain a description of this dataset. The MEWMA vector $\mathbf{w_i} = (w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,p})$ can be calculated by

$$w_{i,t} = rx_{i,t} + (1-r)w_{i,t-1},\ t = 1, 2, \ldots, p \tag{7}$$

where $\mathbf{w}_0 = \mathbf{0}$, $r \in (0, 1]$ is a fixed smoothing parameter, $w_{i,t}$ is the MEWMA statistic for the time step t, and $x_{i,t}$ is the measurement at time step t of the i^{th} observation. The SVDD optimization problem to obtain a description of the dataset using MEWMA-based vectors can be described by

$$\min_{R,\mathbf{a}} \quad R^2 + C\sum_{i=1}^{N} \xi_i \tag{8}$$

$$\text{subject to} \qquad \|\mathbf{w}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, N;\ \ \xi_i \geq 0$$

where C is the trade-off parameter which is introduced to control the balance between the volume of the sphere and the errors. Introducing two Lagrange multipliers α_i, γ_i, the dual problem of problem (8) will be

$$\max_{\alpha} \quad \sum_{i=1}^{N} \alpha_i \langle \mathbf{w}_i, \mathbf{w}_i \rangle - \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{9}$$

$$\begin{aligned} \text{subject to} \qquad & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leq \alpha_i \leq C \quad \text{for} \quad i = 1, \ldots, N \end{aligned}$$

This dual problem is convex quadratic and much easier to solve compared with the original problem (8). Solving this problem and noting that the parameters γ_i are omitted during the solving process, we obtain the center $\mathbf{a}$ and the radius R of the sphere as follows

$$\mathbf{a} = \sum_{SV} \alpha_s \mathbf{w}_s \tag{10}$$

$$R^2 = \langle \mathbf{w}_k, \mathbf{w}_k \rangle - 2\sum_{i=1}^{N} \alpha_i \langle \mathbf{w}_i, \mathbf{w}_i \rangle + \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{11}$$

where SV denotes the index set of the support vectors and $\mathbf{w}_k$ in formula (11) of R^2 is any object that satisfies $\mathbf{w}_k \in SV_{<C}$, the set of support vectors having $\alpha_k < C$, see Tax and Duin [23] for more details.

In general, to monitor a new sample $\mathbf{z}_1, \mathbf{z}_2, \ldots$, we can transform them to the MEWMA vectors using formula (7) and calculate the distance between these vectors and the center $\mathbf{a}$ of the sphere. The square of this distance can be calculated by

$$\|\mathbf{z}_i - \mathbf{a}\|^2 = \langle \mathbf{z}_i, \mathbf{z}_i \rangle - 2\sum_{i=1}^{N} \alpha_i \langle \mathbf{z}_i, \mathbf{w}_i \rangle + \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{12}$$

In practice, instead of using the inner product in the above formulas, it is common to employ kernel functions to obtain a more flexible data description. Some popular kernel functions often used in Support Vector Classifiers include linear, polynomial, Gaussian, and Laplacian kernels, see Vapnik [26]; Smola et al. [20]. If the kernel function $\mathrm{K}(\mathbf{x}, \mathbf{y})$ is used, the squared distance in Equation 12 can be recomputed by replacing the inner product by the kernel function. Denote this kernel distance by $\mathrm{D}(\mathbf{z})$, then

$$\mathrm{D}(\mathbf{z}, \mathbf{a}) = \mathrm{D}(\mathbf{z}) = \mathrm{K}(\mathbf{z}, \mathbf{z}) - 2\sum_{i=1}^{N} \alpha_i \,\mathrm{K}(\mathbf{z}, \mathbf{x}_i) + \sum_{i,j=1}^{N} \alpha_i \alpha_j \,\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{13}$$

3 Proposed framework

In this section, we will present the details of our proposed transformer-based VAE-MEWMA-SVDD framework for anomaly detection in multivariate time series data. Within this framework, we will leverage the transformer-based variational autoencoder architecture introduced by Raza et al. [17] to extract features from the input data in the form of reconstruction error vectors and then employ the MEWMA-SVDD control chart idea on these error vectors for anomaly detection. By integrating the MEWMA-SVDD control chart into our framework, we can effectively manage the false alarm rate, a critical concern in real-life applications, particularly in healthcare.

3.1 Transformer-based variational autoencoder architecture

To extract feature representations from the input data in the form of reconstruction error vectors, we propose employing the transformer-based variational autoencoder (VAE) architecture, introduced by Raza et al. [17]. This architecture has been demonstrated to outperform other deep learning frameworks in detecting abnormal signals in the ECG dataset. The extracted reconstruction error vectors will then be used to train the subsequent MEWMA-SVDD model.

Figure 3 illustrates the transformer-based VAE architecture proposed by Raza et al. [17]. In this architectural design, the encoder module includes several key components. It begins with an initial input layer that takes in multivariate time-series data, particularly an ECG segment with 140 data points. The data then flows through a transformer layer containing multiple sub-layers illustrated in Figure 3. One of these sub-layers is the augmentation layer, which introduces random yet realistic transformations to diversify the training set without substantial information loss. After augmentation, the data is normalized using a normalization layer, followed by the application of a multi-head attention layer with the self-attention mechanism. This mechanism computes weighted averages across input

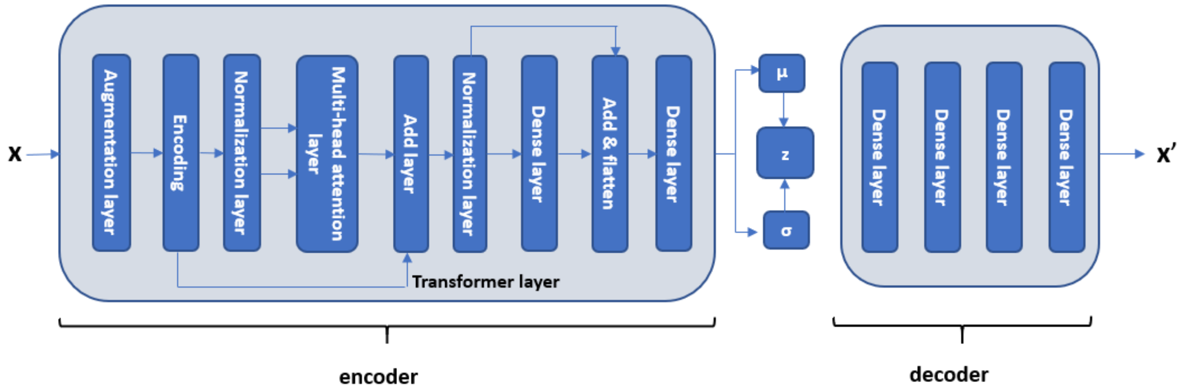

Fig. 3. Transformer-based VAE architecture proposed by Raza et al.[17].

vectors to produce corresponding output vectors. More precisely, suppose we have a set of k-dimensional input vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the self-attention mechanism calculates their corresponding output vectors $\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_n$ by

$$\mathbf{x}'_i = \sum_j w_{i,j}\mathbf{x}_j \tag{14}$$

here, the sum is taken over the whole sequence, and the sum of the weights over all indexes j is equal to one. The weights $w_{i,j}$ is computed by

$$w_{i,j} = \frac{exp(\mathbf{x}_i^T\mathbf{x}_j)}{\sum_j exp(\mathbf{x}_i^T\mathbf{x}_j)} \tag{15}$$

In this equation, the softmax function is employed to ensure the sum of weights over the whole sequence is equal to one. The resulting output from the multi-head attention layer is combined with the preceding normalization layer's output through an additional layer before passing through another normalization layer. It then enters a dense layer, applying a non-linear transformation to extract further features. The final dense layer's output of the encoder represents a latent distribution characterized by a mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. This encoding ensures that the latent distribution is close to a standard normal distribution. The latent representation then proceeds to the decoder. The decoder module of this architecture comprises four dense layers, in which, the first three layers use ReLU activation functions as the activation function while the final layer utilizes a sigmoid activation function to generate probability distributions for candidate classes.

3.2 MEWMA-SVDD module

To efficiently achieve the threshold to distinguish between "normal" and "abnormal" classes without depending on the distribution of the data, Raza et al.[17] introduced an adaptive anomaly detection technique that combines SVDD, a classifier proposed by Tax and Duin[23], with density kernel estimation to eliminate the need for fixed thresholds. The kernel function transforms input data into a higher-dimensional feature space, and the SVDD model is built in this new space. This approach employs kernel density estimation with adaptive bandwidth, determined through a grid search mechanism, to construct the SVDD module within their proposed anomaly detection framework.

In order to leverage the advantages of SVDD, which eliminates the necessity for fixed thresholds, and enhance the model's robustness by reducing false alarms and increasing sensitivity to minor changes in time series data, we suggest incorporating the concept of the MEWMA-SVDD control chart into the framework proposed by Raza et al.[17]. Controlling the false alarm rate is important, particularly in healthcare, where patient safety and accurate diagnostics are critical. For instance, in the context of ECG monitoring, minimizing false alarms is essential to avoid unnecessary interventions and patient distress. By employing the concept of a control chart, we can effectively manage the false alarm rate within the pre-defined average run length ARL_0 for in-control scenarios. Besides, the MEWMA technique is integrated to take into account information from all observations in the time series, making the model more sensitive to small shifts or changes in the time series.

Let $\mathbf{e}_i$ denote the reconstruction error vectors obtained from the transformer-based VAE module

$$\mathbf{e}_i = \mathbf{x}_i - \mathbf{x}'_i \tag{16}$$

Suppose $\mathbf{e}_i \in \mathbf{R}^p$, where p denotes the number of time steps, the MEWMA vector $\mathbf{w_i} = (w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,p})$ can be calculated using Equation 7

$$w_{i,t} = re_{i,t} + (1-r)e_{i,t-1},\ t = 1, 2, \ldots, p$$

The EWMA-SVDD is then trained using these MEWMA vectors $\mathbf{w_i}, i = 1, \ldots, N$ where N is the size of the training set. In general, one can use the radius R of the SVDD hypersphere as the control limit. However, as stated before, we want to control the false alarm rate, so the upper limit h based on pre-specified ARL_0 is more appropriate. In this case, an observation $\mathbf{z}$ is considered abnormal if the distance $\mathrm{D}(\mathbf{z}, \mathbf{a}) > h$, where $\mathbf{a}$ is the center of the EWMA-SVDD, $\mathrm{D}(\mathbf{z}, \mathbf{a})$ as in Equation 13.

Tang et al. [22], among others, proposed to obtain the value of h given pre-specified ARL_0 by using bootstrap aggregating methods. In this work, we will apply the bootstrap percentile approach proposed in Tang et al. [22] to find the control limit when designing the EWMA-SVDD module for the proposed framework. Suppose we have a training data set of size N, $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$. The procedure to obtain upper limit threshold h using the bagging method is summarized as follows

---

Algorithm 1: Finding the upper limit threshold h

Input: A training data set of size N, $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$

Output: Upper limit threshold h of SVDD control chart

1. Generating M independent bootstrap samples of size N from the training set.
2. For each bootstrap sample $j^{th}, j = 1, \ldots, M$ **do**

- Apply the SVDD algorithm on each bootstrap sample to obtain the SVDD sphere.
- Using Eq. 13 to compute n statistics $D_{j1} = \mathrm{D}(\mathbf{a}^{(j)}, \mathbf{x}_1^{(j)}), \ldots, D_{jn} = \mathrm{D}(\mathbf{a}^{(j)}, \mathbf{x}_n^{(j)})$ where $\mathbf{a}^{(j)}$ is the center of sphere obtained by j^{th} bootstrap sample.
- Let ARL_0 be the pre-specified in-control ARL and $\beta = \dfrac{1}{\mathrm{ARL}_0}$ be the desired false alarm. Compute $100 \times (1-\beta)th$ percentile value of the set of n statistics. This value presents the estimated threshold limit obtained from the current bootstrap sample, denoted by $h^{(j)}$.

End for

3. Obtain control limit h by taking the $100 \times (1-\epsilon)$ percentile of estimated threshold limits $h^{(j)}, j = 1, \ldots, M$, where ϵ is a small pre-specified constant.

---

4 Experiment and results

In this section, we provide details of the experiment used to test the performance of the proposed framework, and the results obtained.

4.1 Data description

To illustrate and assess the performance of the proposed framework, a hybrid dataset was composed by merging two datasets from PhysioNet (see Goldberger et al. [9]). This combined dataset was also used by Raza et al. [17] to test their proposed AnoFed framework in detecting anomalies. The anomaly category was constructed using the BIDMC Congestive Heart Failure Database. This dataset includes extended ECG recordings that capture severe congestive heart failures from 15 subjects, comprising 11 males, aged 22 to 71, and 4 females, aged 54 to 63. The dataset for normal subjects consisted of recordings from the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) normal sinus rhythm dataset. This dataset contains 18 long-term ECG recordings from 18 subjects, including 5 men (aged 26 to 45) and 13 women (aged 20 to 50), all without significant arrhythmias. A total of 5,000 heartbeats (2,919 normal samples and 2,081 anomalies) were randomly selected from each dataset to facilitate the training and testing of the proposed framework. Figure 4 provides a visual representation of an exemplary sample from each of the selected databases. The orange line illustrates the abnormal signal and the blue one is the normal signal. All data were pre-processed by using interpolation to extract heartbeats of equal length.

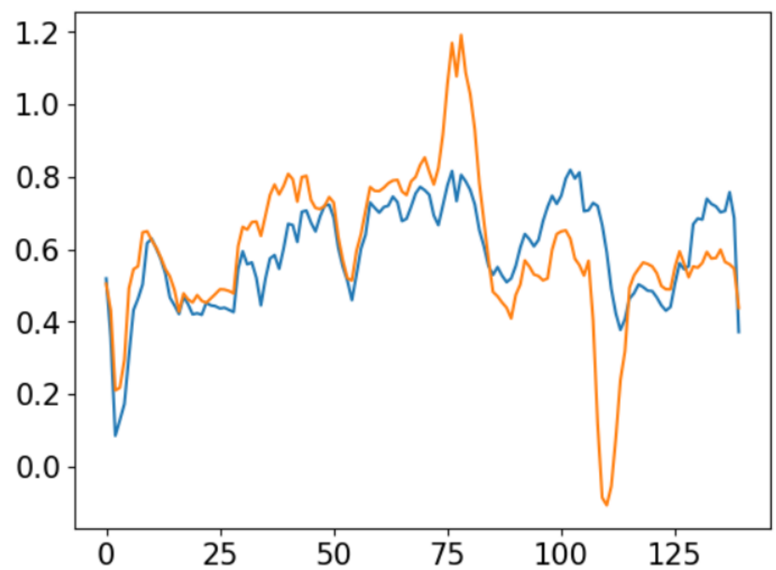

Fig. 4. An example sample from each of the two used databases.

4.2 Experiment setting

To assess the effectiveness of the proposed transformer-based VAE with MEWMA-SVDD we will use the ECG dataset mentioned before. In this dataset, 4.000 samples are randomly chosen to train the model, 1.000 remaining samples are reserved for testing the model.

In the transformer-based VAE training process, we exclusively utilized the normal dataset, aligning with our primary goal of detecting anomalies data from the normal ones through reconstruction loss. The training employed the Mean Squared Error (MSE) as the loss function, with a fixed learning rate of 0.001. To ensure training stability, a gradient threshold of 0.5 was applied. To determine the optimal bandwidth for EWMA-SVDD, the grid search module from scikit-learn was employed. The search encompassed a bandwidth range of (3, 0.2, 10), and a 30-fold cross-validation was carried out for this purpose. Additionally, the smoothing parameter for the MEWMA statistic was configured at 0.2, while the in-control average run length ARL_0 was set to 100. The establishment of model thresholds (control limits) was achieved by employing Monte Carlo simulation with 10,000 iterations in combination with the bootstrapping method.

4.3 Results

To access the classification performance of the proposed framework, we will use 4 metrics that have been widely used in the machine-learning community. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. The 4 popular metrics are defined as follows

– **Accuracy**: The number of correct predicted samples over total samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{17}$$

– **Precision**: measures the proportion of predicted positive samples that are correctly identified as positive

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

– **Recall**: measures the proportion of true positive samples correctly predicted out of all the actual positive samples

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

– **F1-score**: combine Recall and Precision

$$\text{F1-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{20}$$

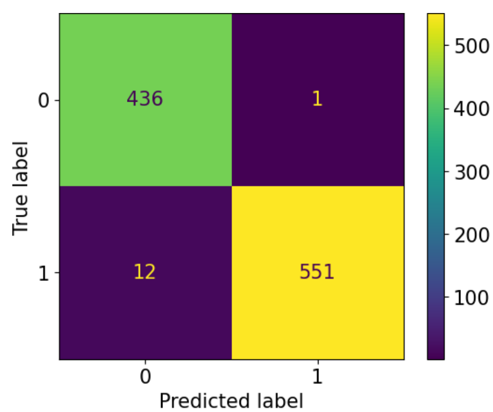

Fig. 5. Confusion matrix for the proposed framework.

Figure 5 shows the confusion matrix and Table 1 displays the anomaly detection performance of our proposed framework with the ECG dataset from Physionet. According to this confusion matrix, we can see that the model correctly classified a substantial number of normal ECG signals as normal, with 436 true negatives, while making only a single false positive prediction. Additionally, the model demonstrated a notable ability to identify anomalies, as indicated by 551 true positives, despite a relatively small number of 12 false negatives where normal signals were misclassified as anomalies. From Table 1, it is evident that our model achieved an accuracy of 98.77%, surpassing the centralized model's accuracy of 98.20% in the work of Raza et al. [17]. Notably, the number of "normal" cases misclassified as "abnormal" has been reduced to 1 in our model, as opposed to 2 cases in the study by Raza et al. [17]. Importantly, our approach does not require prior knowledge about the data distribution while simultaneously allowing us to control the in-control false alarm rate through a fixed value for ARL_0.

Figure 6 illustrates the distinct "normal" and "abnormal" classes generated by the proposed framework. In this visual representation, test samples with a distance from the center of the hypersphere exceeding the control limit (indicated by the red line) are identified as anomalies, while those with a distance less than the control limit are classified as normal. This visualization shows the efficiency of the proposed model in effectively distinguishing between the two classes. The code for this experiment can be found in Transformer-based-Anomaly-Detection-approach.

Table 1. Classification report for the proposed framework.

Class	Precision	Recall	F1-Score	Support
Normal	0.9977	0.9733	0.9853	437
Anomaly	0.9982	0.9787	0.9884	563
Accuracy		0.987		

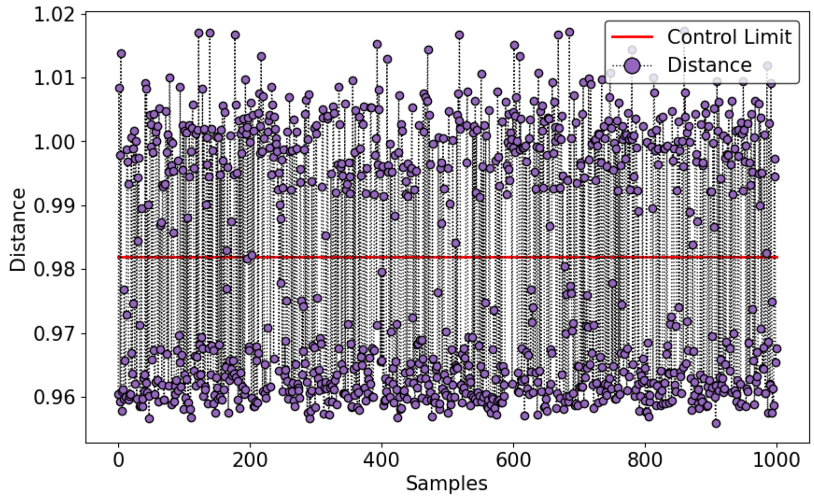

Fig. 6. Separation between two classes using the proposed framework.

5 Perspective

Together with the growing trend of ML techniques for anomaly detection, federated learning has recently gained popularity due to the increasing need for data privacy, efficiency, and collaborative machine learning. Federated learning enables model training on user devices while preserving data privacy, improving efficiency by reducing data transfer, and fostering collaboration without the need to share raw data. The integration of federated learning techniques can enhance the anomaly detection capabilities of these models and promote privacy-preserving collaborative machine learning in real-world applications. Many works have shown their state-of-the-art performance in many domains so far, for example, a robust distributed anomaly detection architecture is proposed in Truong et al.[24], combining federated learning, autoencoder, transformer, and Fourier mixing sublayer, to accurately detect irregularities in time-series data within industrial control systems (ICS) while maintaining fast learning, lightweight design, and low communication costs, making it suitable for deployment on edge devices with limited computing capacity in the Industrial Internet of Things (IIoT) context. In Truong et al.[25], authors proposed an anomaly detection architecture in a federated setting for detecting cyberattacks in Industrial Control Systems (ICS) within Industrial IoT (IIoT)-based smart manufacturing (SM), achieving improved detection performance for time series data while demonstrating feasibility and efficiency for deployment on edge computing hardware, resulting in reduced bandwidth consumption in the transmission link between edge and cloud. In Raza et al.[17], the novel framework AnoFed in a federated setting was proposed to address the challenges of anomaly detection in digital healthcare applications, providing improved privacy protection, explainability of results, and adaptive anomaly detection while demonstrating effectiveness and efficiency compared to state-of-the-art methods, making it suitable for resource-constrained edge devices. In a recent study by Nguyen et al.[15], a novel transformer-based model in a federated setting was introduced for predictive maintenance (PdM) in industrial manufacturing, demonstrating superior performance with up to a 25% improvement in accuracy compared to state-of-the-art methods. These recent advancements in federated learning-based anomaly detection highlight its potential to address challenges in diverse domains, promoting privacy, collaboration, and improved detection performance in real-world applications.

In future work, we suggest integrating the proposed model into a federated setting to enhance data privacy and the overall performance of the proposed framework. Specifically, the work can be done as follows: After designing

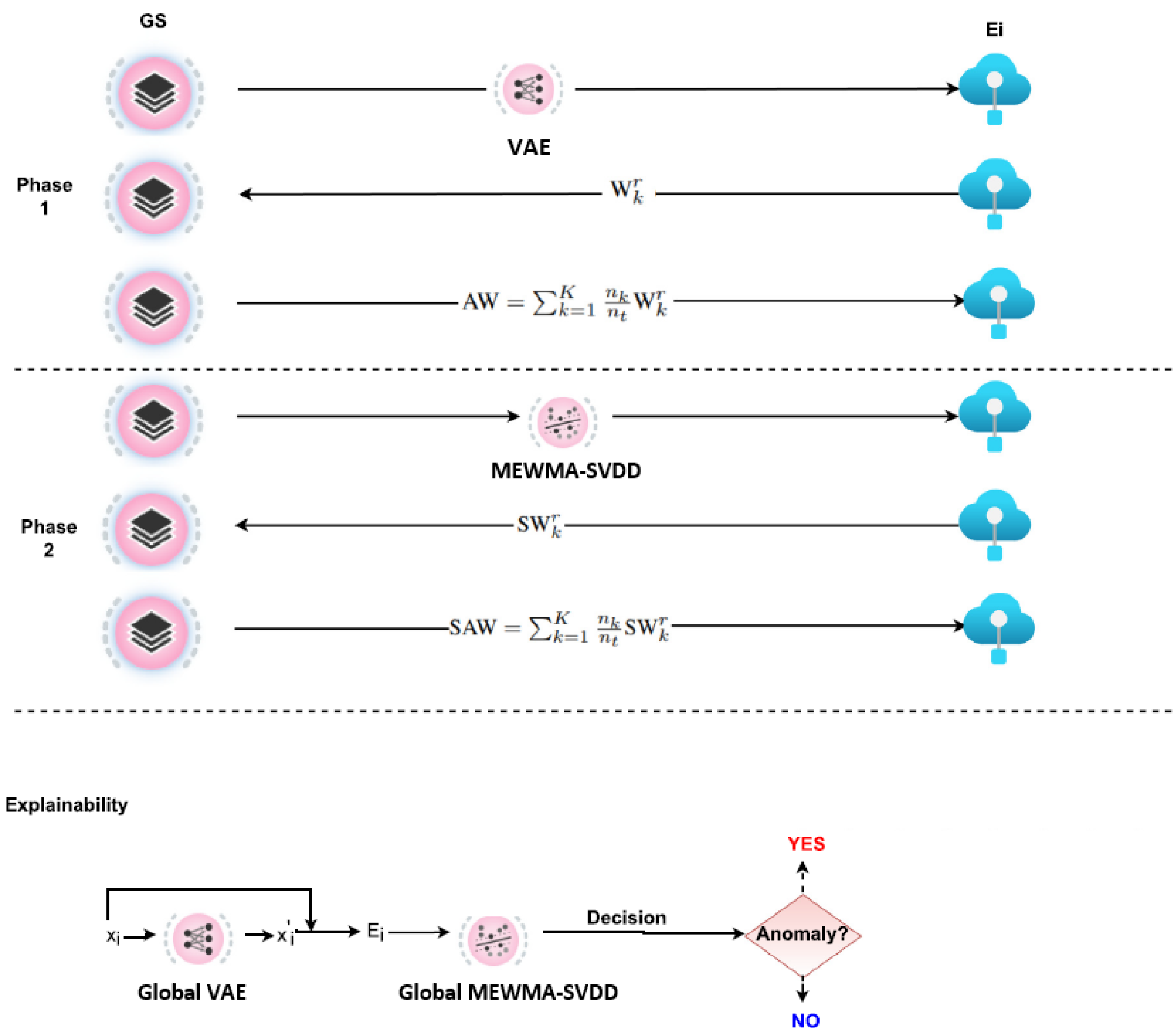

Fig. 7. Overview of the proposed framework.

the transformer-based VAE with MEWMA-SVDD architecture for anomaly detection in multivariate time series data, the next step is to further enhance its qualities through the integration of federated learning techniques. By incorporating federated learning, the model can leverage the collaborative power of distributed clients while ensuring data privacy. The general step to integrate federated learning into the proposed transformer-based VAE with MEWMA-SVDD for anomaly detection in multivariate time series data is as follows

---

Algorithm 2: Training procedure of the proposed framework in a single round

---

Input: Data from edge devices $D_1, D_2, \ldots, D_n$
Output: Trained transformer-based VAE model for anomaly detection

1. Initialize the global model (**GM**) parameters.
2. For each client device E_i:
   a. Establish a connection to the global server (**GS**).
   b. Receive the initial GM parameters from **GS**.
   c. Train the robust transformer model on $\mathbf{E}_i$ using local data $\mathbf{D}_i$
   d. Send the local model (**LM**) updated parameters back to the **GS**.
3. Aggregate the **LM** updates received from all edge devices.
4. Update the **GM** parameters based on the aggregated updates.
5. Repeat steps 3-5 for a specified number of communication rounds or until convergence is achieved.
6. Repeat steps 3-6 for emerging data.
7. Use the trained model for anomaly detection on new incoming data from edge devices.

Figure 7 demonstrates our proposed framework in future work. During the federated learning process, raw data remains on edge devices and is not totally transferred to the central server, ensuring data privacy. The aggregation of local model updates enables collaborative learning across multiple edge devices, leading to a more accurate anomaly detection model. After extracting features input in the form of reconstruction error vectors from the transformer-based VAE network, we apply the MEWMA-SVDD control chart to detect anomalies. To find the threshold, we use the Monte Carlo simulation with the bootstrapping method to find the control limit, as mentioned in section 3.2. In future work, our goal is also to deploy this proposed framework in a federated learning setting for embedded artificial intelligence applications.

While the proposed framework exhibits remarkable accuracy in detecting anomalous ECG signals, it still has some limitations. Firstly, the aspect of data privacy remains unaccounted for, as it is only proposed for future work. Additionally, the public dataset used in the experiment is based on a relatively small population and it may limit the generalizability of the model's performance. Furthermore, an important aspect yet to be addressed is the explainability of the model (XAI), which plays an important role in understanding and interpreting the model. These limitations underscore the need for ongoing research to enhance the robustness of the proposed framework.

6 Conclusion

Anomaly detection is vital in digital healthcare and machine learning. In this paper, we proposed a transformer-based VAE with an EWMA-SVDD anomaly detection framework to detect anomaly signals in an ECG dataset. We combine a lightweight transformer-based VAE proposed by Raza et al.[17], which minimizes the reconstruction loss within three training epochs of each global round to extract feature input, together with MEWMA-SVDD to detect anomalies. The idea of integrating the MEWMA-SVDD under the concept of a control chart helps to control the false alarm rate efficiently. Lastly, we tested the proposed framework by combining two benchmark datasets from PhysioNet's repository and showed that the proposed framework achieved up to 98.77% test accuracy with 30-fold cross-validation. Besides, since deep learning models excel in detecting anomalies but their use in centralized settings raises issues like data privacy, especially in sensitive healthcare applications, we also suggest integrating federated learning into the proposed framework to achieve data privacy in future work.

Bibliography

[1] Acheampong, F.A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of bert-based approaches. Artificial Intelligence Review pp. 1–41 (2021)

[2] Anstine, D.M., Isayev, O.: Generative models as an emerging paradigm in the chemical sciences. Journal of the American Chemical Society **145**(16), 8736–8750 (2023)

[3] Banerjee, A., Burlina, P., Meth, R.: Fast hyperspectral anomaly detection via svdd. In: 2007 IEEE International Conference on Image Processing, vol. 4, pp. IV–101 – IV–104 (2007)

[4] Bertolini, M., Mezzogori, D., Neroni, M., Zammori, F.: Machine learning for industrial applications: A comprehensive literature review. Expert Systems with Applications **175**, 114820 (2021)

[5] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12299–12310 (2021)

[6] Dixit, P., Silakari, S.: Deep learning algorithms for cybersecurity applications: A technological and status review. Computer Science Review **39**, 100317 (2021)

[7] Duan, Z., Lu, M., Ma, Z., Zhu, F.: Lossy image compression with quantized hierarchical vaes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 198–207 (2023)

[8] Fahim, M., Sillitti, A.: Anomaly detection, analysis and prediction techniques in iot environment: A systematic literature review. IEEE Access **7**, 81664–81681 (2019)

[9] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)

[10] Jen, C., Wang, C.: Real-time process monitoring based on multivariate control chart for anomalies driven by frequency signal via sound and electrocardiography cases. Processes **9**(9), 1510 (2021)

[11] Kang, Z., Catal, C., Tekinerdogan, B.: Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering **149**, 106773 (2020)

[12] Kingma, D., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019)

[13] Lang, M.: A low-complexity model-free approach for real-time cardiac anomaly detection based on singular spectrum analysis and nonparametric control charts. Technologies **6**(1), 26 (2018)

[14] Nassif, A.B., Talib, M.A., Nasir, Q., Dakalbab, F.M.: Machine learning for anomaly detection: A systematic review. Ieee Access **9**, 78658–78700 (2021)

[15] Nguyen, H.D., Nguyen, H.L., Kieu, N.H., Nguyen, V.H., Truong, T.H., Tran, K.P.: Trans-lighter: A light-weight federated learning-based architecture for remaining useful lifetime prediction. Computers in Industry **148**, 103888 (2023)

[16] Pauwels, E.J., Ambekar, O.: One class classification for anomaly detection: Support vector data description revisited. In: Perner, P. (ed.) Advances in Data Mining. Applications and Theoretical Aspects, pp. 25–39, Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

[17] Raza, A., Tran, K.P., Koehl, L., Li, S.: Anofed: Adaptive anomaly detection for digital health using transformer-based federated learning and support vector data description. Engineering Applications of Artificial Intelligence **121**, 106051 (2023)

[18] Salloum, S.A., Alshurideh, M., Elnagar, A., Shaalan, K.: Machine learning and deep learning techniques for cybersecurity: a review. In: The International Conference on Artificial Intelligence and Computer Vision, pp. 50–57, Springer (2020)

[19] Singh, S., Mahmood, A.: The nlp cookbook: modern recipes for transformer based deep learning architectures. IEEE Access **9**, 68675–68702 (2021)

[20] Smola, A., Scholkopf, B., Muller, K.: The connection between regularization operators and support vector kernels. Neural Networks pp. 637–649 (1998)

[21] Suman, G., Prajapati, D.: Control chart applications in healthcare: a literature review. International Journal of Metrology and Quality Engineering **9**, 5 (2018)

[22] Tang, A., Castagliola, P., Hu, X., Xie, F.: An assessment for the conditional performance of an support vector data description (svdd)-based chart. Quality and Reliability Engineering International pp. 1–17 (2022)

[23] Tax, D.M.J., Duin, R.P.W.: Support Vector Data Description. Machine Learning **54**(1), 45–66 (2004)
[24] Truong, H.T., Ta, B.P., Le, Q.A., Nguyen, D.M., Le, C.T., Nguyen, H.X., Do, H.T., Nguyen, H.T., Tran, K.P.: Light-weight federated learning-based anomaly detection for time-series data in industrial control systems. Computers in Industry **140**, 103692 (2022)
[25] Truong, T., Ta, B., Dao, M., Tran, D., Nguyen, M., Bui, D.T., Tran, K.P., et al.: Detecting cyberattacks using anomaly detection in industrial control systems: A federated learning approach. Computers in Industry **132**, 103509 (2021)
[26] Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)
[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc. (2017)
[28] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp. 38–45 (2020)
[29] Zhang, Z., Deng, X.: Anomaly detection using improved deep svdd model with data structure preservation. Pattern Recognition Letters **148**, 1–6 (2021)
[30] Zhou, H., Kan, C.: Tensor-based ecg anomaly detection toward cardiac monitoring in the internet of health things. Sensors **21**(12), 4173 (2021)