Explainable Transformer-Based Approach for ECG Anomaly Detection

Thi Thuy Van Nguyen[1,2(✉)], Cédric Heuchenne[1], Kim Duc Tran[3], Guillaume Tartare[2], and Kim Phuc Tran[2]

[1] HEC Liège - Management School, University of Liège, Liège, Belgium
{ttv.nguyen,c.heuchenne}@uliege.be
[2] GEMTEX, ENSAIT, University of Lille, Lille, France
{guillaume.tartare,kim-phuc.tran}@ensait.fr
[3] IAD, Dong A University, Da Nang, Vietnam
ductk@donga.edu.vn

Abstract. Detecting anomalies is critical in various fields, especially in healthcare. The ability to identify abnormal patterns from normal ones can help clinicians make early interventions and improve patient outcomes. In Electrocardiogram (ECG) analysis, timely detection of unusual signals is crucial for diagnosing and treating serious, life-threatening health problems. However, although many AI-based anomaly detection models offer high performance, they often function as "black boxes", making us difficult to interpret the results. In this paper, we propose integrating explainable artificial intelligence (XAI) into a transformer-based network combined with a Support Vector Data Description control chart and multivariate exponential weighted moving average technique (MEWMA-SVDD chart) to obtain a robust ECG anomaly detection model. By incorporating XAI, we aim to enhance the transparency and reliability of our model, providing clear and interpretable results. We will demonstrate our approach's effectiveness by using a well-known ECG dataset and provide important insights into the detection mechanism. This approach illustrates the importance of combining advanced deep learning techniques with XAI to improve the reliability and efficiency of anomaly detection systems in monitoring healthcare.

Keywords: Explainable artificial intelligence · Anomaly detection · Transformer · Variational autoencoder · Control chart · Support vector data description · Multivariate exponentially weighted moving average

1 Introduction

Detecting anomalies is vital across various fields since it reveals unexpected deviations from normal patterns. Addressing these irregularities in time will help us improve safety, efficiency, and overall well-being. It is especially important in critical fields such as healthcare, where early detection of anomalies can help to timely intervention and increase the chances of survival for patients. One of the

S. Thomassey et al. (Eds.): RAIDS 2024, LNICST 673, pp. 85–108, 2026.
https://doi.org/10.1007/978-3-032-14055-5_7

most important applications of anomaly detection in healthcare is Electrocardiogram (ECG) monitoring, which provides vital insights for assessing cardiac health. Early detection of anomalies in ECG signals can help to prevent serious conditions like heart attacks and arrhythmias, stroke, and sudden cardiac arrest. Therefore, accurate and timely analysis of ECG data is essential for effective cardiac care (see Amini et al. [4]; Muzammil et al. [23]).

Over the years, various methodologies have been developed to detect anomalies, from traditional statistical approaches, theories from cognitive psychology to advanced machine learning (ML) techniques. Control charts, a fundamental tool in Statistical Process Control (SPC), offer us the advantage of interpretability and have been widely used for anomaly detection in healthcare and other fields (Suman and Prajapati [36]; Lang [15]; Zhou and Kan [43]). However, these methods often struggle with the complexity and high dimensionality of modern datasets. Besides SPC, cognitive psychology also plays an important role in anomaly detection by providing insides into human observational limitations. Theories such as inattentional blindness and the tunnelling effect explain why crucial details are often overlooked when the focus is too narrow or misdirected. To overcome this, training programs and optimized design principles have been implemented to improve detection rates in environments relying on human monitoring. These strategies enhance the performance by ensuring the crucial signs are more likely to be noticed (see Kreitz et al. [14]; Gao and Jia [10]; Saint-Lot et al. [29]). However, despite these improvements, such non-automated methods still face challenges in consistency and scalability, especially when handling large and complex datasets. Conversely, deep learning (DL) models have demonstrated their impressive accuracy in identifying anomalies through the use of large and complex datasets with sophisticated algorithms (Bolhasani et al. [6]; Nguyen et al. [24]). However, despite their effectiveness, these advanced models are often difficult to understand and provide little clarity regarding their decision-making processes. The lack of transparency in these AI-based models leads to significant challenges, especially in vital areas such as health care. In these fields, understanding how a model makes predictions is also as important as the predictions themselves. For example, in ECG monitoring, a "black box" model may correctly detect an anomaly, but it may not provide the necessary explanation for why the anomaly was flagged. This lack of interpretability can make it difficult for healthcare professionals to trust the model's results and make the right decisions. As a result, it may slow down medical decision-making processes and finally affect patient outcomes. Therefore, improving the transparency in AI-based models is crucial to building trust, increasing efficiency, and achieving better healthcare outcomes.

To address this issue, Explainable AI (XAI) offers a solution to improve the clarity of AI models and the transparency of their predictions. The techniques used in XAI aim to provide clear and understandable explanations of how models reach their conclusions, enhance user trust, and enable better decision-making. To date, many widely used methods have been developed to improve the transparency of AI models, with a strong focus on improving their explainability.

One commonly used method is Local Interpretable Model-Agnostic Explanations (LIME), which locally simplifies complex models with simpler, easier-to-understand models to explain individual predictions (Ribeiro et al. [28]). Another popular technique is SHapley Additive exPlanations (SHAP), which uses game theory to assign a significance value to each feature for a particular prediction, providing a consistent and globally interpretable explanation model (Lundberg and Lee [20]). Gradient-weighted Class Activation Mapping (Grad-CAM) is also a frequently employed technique for creating visual explanations, particularly with convolutional neural networks (CNNs), as it identifies the important areas in the input image that contribute to the prediction of the model (Selvaraju et al. [30]). In addition to these post-hoc methods, XAI-by-design has also received attention from researchers recently. XAI-by-design focuses on creating models that are naturally easy to interpret by designing them with transparent structures and understandable decision-making processes. A prime example in this category is the attention mechanism in neural networks, particularly transformers. This mechanism highlights the specific parts of the input data that influence the predictions of the model, providing straightforward explanations of how the model works (Vaswani et al. [40]). In healthcare, the use of XAI methods has greatly enhanced the interpretability of DL models, leading to remarkable improvements. For instance, in a recent study, Prendin et al. [26] used SHAP to assess the effect of different features on glucose level predictions in Type 1 Diabetes management, showing that when embedded in a decision support system, this helped to make better treatment decisions and blood sugar control. Alabi et al. [2] used SHAP and LIME to analyze how different factors influence survival predictions for nasopharyngeal cancer, providing personalized insights into risk factors for each patient. In [25], the authors incorporated Grad-CAM into their deep transfer learning algorithm to interpret the detected COVID-19 cases using CT-scan images and X-rays. The work by Albahri et al. [3] provided a detailed overview of XAI techniques used in healthcare, including methods like SHAP, LIME, Grad-CAM, and various data fusion techniques. The review highlights the crucial importance of transparency and reliability in AI applications. Therefore, integrating XAI into healthcare models is essential for improving clinical decision-making and building trust among healthcare professionals and patients.

In this study, we will construct an XAI framework for ECG anomaly detection. Specifically, we will integrate an explainability module into our previously developed transformer-based network combined with a MEWMA-SVDD chart (Nguyen et al. [24]). This hybrid model was designed to detect any type of anomalies that differ from the normal ECG patterns, making it a flexible tool for detecting a wide range of heart irregularities. The model leverages the strengths of DL for feature extraction and the precise monitoring capabilities of control charts to enhance sensitivity to small changes in the time series data, reduce false alarm rates, and make it highly effective for detecting anomalies in ECG signals. We choose SHAP for its ability to provide detailed, transparent explanations at both global and individual levels, which are crucial for addressing complex data such as ECG. Specifically, we will apply the SHAP method to

analyze reconstruction error vectors obtained from the variational autoencoder (VAE) model to uncover which specific window time steps in the ECG segments significantly impact the reconstruction error. At the global level, we will identify which windows contribute the most to the overall model, while at the individual level, we will determine which windows have the most impact on specific ECG instances. This method not only enhances our understanding of the model's performance by revealing the most influential windows in ECG signals but also improves the decision-making process in clinical settings. The remaining work is structured as follows: Sect. 2 provides an overview of the related work and foundational concepts of transformers, VAE, MEWMA-SVDD algorithms, and SHAP principles. Section 3 reviews the detailed architecture of the explainable transformer-based network combined with the MEWMA-SVDD control chart, and then presents the integration of SHAP for enhanced transparency. Section 4 outlines the experiment on a real ECG dataset for evaluating the framework's performance, presents obtained results, and explains the model's interpretability through the integration of SHAP. Section 5 provides the conclusion of the work.

2 Background

In this section, we review some recent developments in ECG monitoring using DL techniques and the main background of our proposed framework, including the transformer architecture, VAE model, MEWMA-SVDD control chart, and SHAP principles.

2.1 Related Work

ECG monitoring is vital for the diagnosis and management of cardiac conditions, and the application of DL techniques has significantly advanced this field. Recently, many innovative approaches have been developed to enhance ECG monitoring with DL. In a study by Liu et al. [17], the authors utilized a vector quantized VAE to address the challenge of insufficient positive samples by employing data augmentation, improving ECG data classification performance. Shin et al. [32] proposed an enhanced AnoGAN model for arrhythmia detection, addressing data imbalance issues and demonstrating improved performance through fine-tuned decision boundaries and learning counts. Jang et al. [12] explored an unsupervised learning approach using a convolutional VAE to extract features from ECG data, showing its utility in anomaly detection and transfer learning for arrhythmia classification. Gaudilliere et al. [11] applied transformer models, typically used in natural language processing, to ECG data, achieving notable success in multi-label classification of heartbeat abnormalities. This demonstrates the versatility and strength of transformer architectures in handling sequential data like ECG signals. The advantages of using transformers include their capacity to effectively capture long-range dependencies and their scalability, which is particularly beneficial for analyzing complex ECG data. In 2023, Raza et al. [27] proposed AnoFed, a hybrid model combining

transformer-based VAE and SVDD for monitoring ECG. This model enhances privacy protection and demonstrates high performance in detecting anomalies in ECG signals. After that, Nguyen et al. [24] expanded this model by combining this transformer-based network with a MEWMA-SVDD chart for ECG monitoring. This model utilizes the capabilities of feature extraction of transformer-based VAE network, with the inclusion of the MEWMA-SVDD module further enhances the sensitivity of the model to small shifts in the data, reduces the false alarm rates. This combination makes the model very effective for detecting anomalies in ECG signals. Furthermore, Shah et al. [31] proposed a hybrid model, ECG-TransCovNet, which combines transformers and CNNs for superior temporal and spatial feature extraction, achieving very high performance in arrhythmia detection.

These studies demonstrate significant advancements in applying DL models to ECG monitoring. The strengths of transformer models and VAEs, such as their ability to handle complex data and generate high-quality representations, make them ideal candidates for further developments. Our proposed model builds on these advancements with the integration of Explainable AI techniques to enhance transparency and trustworthiness in ECG anomaly detection.

2.2 Transformer

The transformer network, initially presented by Vaswani et al. [40], has brought a revolutionary change in natural language processing and other fields due to its powerful capabilities in handling sequential data. Unlike recurrent neural networks (RNNs), which rely on sequential information processing through recurrent connections, transformers utilize self-attention techniques that enable them to recognize the dependencies throughout the whole sequence, regardless of the distance between elements. This allows transformers to process data more efficiently and effectively. To retain information about the position, positional encoding is incorporated into input embeddings, enabling the model to recognize the sequence order, which the architecture does not inherently encode. The attention scores (Att) of each component of the input are calculated using the following formula:

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

In this formula, matrices $\mathbf{K}$ (key), $\mathbf{Q}$ (query), and $\mathbf{V}$ (value) are obtained from the input, d_k is the key vector's dimension. Instead of using a single attention mechanism, transformers employ multiple attention heads. Each attention head operates on distinct projections of the matrices $\mathbf{K}$, $\mathbf{Q}$, and $\mathbf{V}$, allowing the model to recognize various patterns and correlations within the input data. These attention heads' outputs are then combined and subsequently linear transformed, resulting in the multi-head attention layer's final output.

The revolutionary multi-head attention mechanism in transformers allows for parallel processing of sequences. This parallel processing capability enables

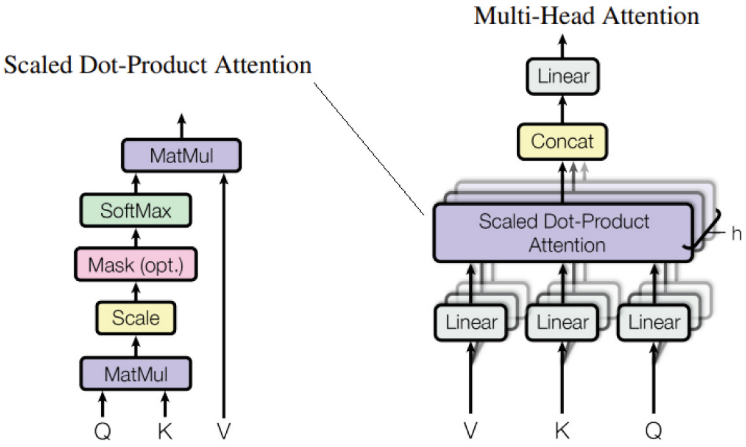

Fig. 1. Multi-Head Attention, consisting of h parallel attention layers, as introduced by Vaswani et al. [40] (Right)

transformers to understand complex data relationships, making them incredibly effective for tasks like ECG anomaly detection, where temporal and spatial dependencies are critical. Figure 1 shows how multiple attention layers run in parallel in multi-head attention introduced by Vaswani et al. [40]. Furthermore, transformers incorporate feedforward neural network layers, normalization layers, and residual connections, enhancing the model's efficiency in learning and processing sequential data. This architecture has served as the base for numerous advanced models, such as Gemini, GPT, BERT, and others, excelling in multiple NLP tasks and extending their success to other domains, as noted by Chen et al. [7]; Singh and Mahmood [34], and Acheampong et al. [1].

2.3 Variational Autoencoder (VAE)

VAEs play a crucial role in unsupervised ML, representing a significant advancement DL. Firstly introduced by Kingma et al. [13], VAEs have been widely applied in representation learning, data generation, and anomaly detection. Different from traditional autoencoders, VAEs aim to learn a probability distribution within the latent space, typically approximating a normal distribution to capture data's underlying structure.

Let X represent the input data, which is assumed to originate from unobservable latent variable Z in latent space. In VAEs, the focus is on probabilistic encoders and decoders rather than deterministic ones. Based on Bayes' theorem, the connection between $p_\theta(Z)$, $p_\theta(Z|X)$ and $p_\theta(X|Z)$, is expressed as:

$$p_\theta(Z|X) = \frac{p_\theta(X|Z) \cdot p_\theta(Z)}{\int p_\theta(X|Z) \cdot p_\theta(Z)\ dz} \tag{2}$$

In a real-world application, calculating $p_\theta(Z|X)$ directly is often intractable due to the denominator's integral. To address this challenge, Kingma et al. [13] proposed to approximate $p_\theta(Z|X)$ with a more tractable distribution $q_\phi(Z|X)$ by using variational inference, where ϕ denotes the approximate distribution's parameters. Let $KL(q_\phi(Z|X)||p_\theta(Z|X))$ denote the Kullback-Leibler (KL) divergence, which measures the difference between two distributions:

$$KL(q_\phi(Z|X)||p_\theta(Z|X)) = \mathbb{E}_{q_\phi(Z|X)}\left[\log \frac{q_\phi(Z|X)}{p_\theta(Z|X)}\right]. \tag{3}$$

By minimizing this divergence, the similarity of $q_\phi(Z|X)$ and $p_\theta(Z|X)$ is ensured.

Once input data is mapped to a probabilistic distribution by the encoder in the latent space, the decoder samples Z to produce data that resembles input and this sample is generally assumed to follow $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^T)$, i.e., $Z = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$. Here, $\boldsymbol{\epsilon}$ follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise multiplication. By incorporating $\boldsymbol{\epsilon}$, the sampling process remains differentiable. This re-parameterization trick technique is crucial for enabling the training and optimization of VAEs, making them effective for various tasks relating to representation learning and generative. The common structure of a VAE is illustrated in Fig. 2.

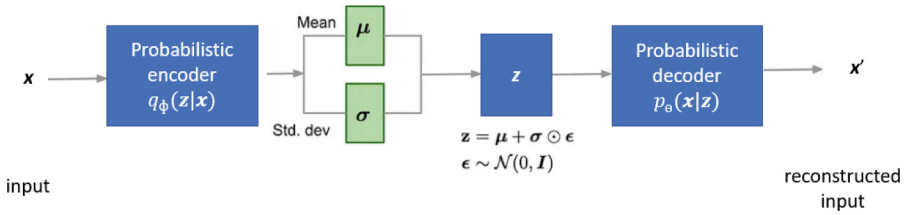

Fig. 2. VAE general structure.

The loss function of VAEs includes 2 main components: the loss from reconstruction $L(X, X')$, where X' denotes reconstructed data, and the KL divergence $KL(q_{\phi,i}(Z|X)\|p_\phi(Z))$. The reconstruction loss evaluates the ability of the decoder to reconstruct input data while the KL divergence accesses the difference between $q_\phi(Z|X)$ and $p_\phi(Z)$. The loss function of VAEs is a key component in training these models, and the goal is to minimize this function during training:

$$\text{TotalLoss (VAE)} = L(X, X') + \sum_i KL(q_{\phi,i}(Z|X)\|p_\phi(Z)). \tag{4}$$

The type of data being processed will decide the choice of reconstruction loss $L(X, X')$. Training VAEs requires minimizing this total loss function, typically via a gradient-based method, to enhance the accuracy of the reconstruction and the alignment between targeted distribution and latent space. VAEs have demonstrated their utility across various domains, such as data compression, generative modeling, and anomaly detection, as evidenced by works such as Anstine and Isayev [5], Zhang et al. [41], and Raza et al. [27].

2.4 SVDD Model Using the MEWMA Technique for Anomaly Detection

SVDD is a highly recognized algorithm in ML, especially in anomaly detection and outlier identification in datasets. First introduced by Tax and Duin [38], SVDD has become well-known for its effectiveness in various applications. One of its key strengths is its capacity to function effectively without strict assumptions regarding the data's underlying distribution, thus enhancing its versatility and robustness, as shown in many studies (Zhang and Deng [42]; Liu et al. [16]; Frusque et al. [9]). This section briefly overviews the principles of SVDD combined with the MEWMA technique, a simple yet effective method for detecting anomalies in time series data. The MEWMA approach calculates the weighted average of components in the data, utilizing a memory-based advantage to preprocess the data for utilization in the SVDD model.

SVDD is an innovative algorithm for one-class classification that excels at detecting abnormal instances by effectively modeling the normal. It establishes a sphere around the normal data, with the center's sphere denoted by a and the sphere's radius denoted by R. The main goal is to minimize the sphere's volume while still encompassing a large amount of the training samples. Given a time series dataset $S = \{x_1, x_2, \ldots, x_N\}$ where $x_i \in \mathbb{R}^p$ and p representing the count of time steps, the MEWMA vector $\mathbf{w_i} = (w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,p})$ is computed as:

$$w_{i,t} = rx_{i,t} + (1 - r)w_{i,t-1}, \tag{5}$$

where $\mathbf{w}_0 = \mathbf{0}$, r is the fixed smoothing parameter, $r \in (0, 1]$ and $t = 1, \ldots, p$. The SVDD optimization problem, aimed at describing the dataset using these MEWMA-derived vectors, is formulated as follows:

$$\begin{aligned} \min_{R,\mathbf{a}} \quad & R^2 + C\sum_{i=1}^{N} \xi_i \\ \text{st.} \quad & \|\mathbf{w}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, N, \\ & \xi_i \geq 0 \end{aligned} \tag{6}$$

where, C is a trade-off that helps to balance the sphere's volume and the errors. By introducing Lagrange multipliers α_i and γ_i to problem 6, we obtain the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i \langle \mathbf{w}_i, \mathbf{w}_i \rangle - \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \\ \text{st.} \quad & \sum_{i=1}^{N} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq C \quad \text{for} \quad i = 1, \ldots, N. \end{aligned} \tag{7}$$

This convex quadratic problem is easier to solve compared to the primal one. Upon solving this, we obtain:

$$\mathbf{a} = \sum_{s \in SV} \alpha_s \mathbf{w}_s \tag{8}$$

$$R^2 = \langle \mathbf{w}_k, \mathbf{w}_k \rangle - 2 \sum_{i=1}^{N} \alpha_i \langle \mathbf{w}_i, \mathbf{w}_k \rangle + \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{9}$$

here, SV is the support vectors index set and for any support vector $\mathbf{w}_k$ with $0 < \alpha_k < C$, the equality in Eq. 9 holds, meaning that $\mathbf{w}_k$ lies exactly on the boundary of the hypersphere.

To monitor new samples $\mathbf{z}_1, \mathbf{z}_2, \ldots$, we convert them into MEWMA vectors $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots$ and determine their distance to the sphere's center $\mathbf{a}$. The squared distance is calculated as follows:

$$\|\hat{\mathbf{z}}_t - \mathbf{a}\|^2 = \langle \hat{\mathbf{z}}_t, \hat{\mathbf{z}}_t \rangle - 2 \sum_{i=1}^{N} \alpha_i \langle \mathbf{w}_i, \hat{\mathbf{z}}_t \rangle + \sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle \tag{10}$$

In practical applications, kernel functions are often used instead of the inner product to achieve a more adaptable representation of data. Well-known kernel functions, such as linear, Laplacian, and Gaussian kernels, are frequently employed (Vapnik [39]; Smola et al. [35]). When using kernel function K, the squared distance is calculated as follows:

$$\mathrm{D}(\hat{\mathbf{z}}_t, \mathbf{a}) = \mathrm{D}(\hat{\mathbf{z}}_t) = \mathrm{K}(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_t) - 2 \sum_{i=1}^{N} \alpha_i \, \mathrm{K}(\mathbf{w}_i, \hat{\mathbf{z}}_t) + \sum_{i,j=1}^{N} \alpha_i \alpha_j \, \mathrm{K}(\mathbf{w}_i, \mathbf{w}_j). \tag{11}$$

2.5 SHapley Addictive ExPlanations (SHAP) - Technique

As mentioned before, among the various XAI strategies in the literature, SHAP is one of the most well-known, as proposed by Lundberg and Lee [20]. SHAP is a method based on game theory that can locally or globally interpret any machine learning or deep learning model output. This technique assigns each feature an importance value, known as Shapley scores/values, which indicate the impact of a feature over every possible combination of features. These values play a crucial role in assessing the degree of impact each feature has on the predicted outcomes of the model.

According to Lundberg and Lee [20], a method to make a complex model f more comprehensible is to approximate it with a more understandable function g. This function g is often used in explanation models, where it operates on simplified input variables z' that map to the original instances x via function h_x that ensures $x = h_x(z')$. The construction of function g is as follows:

$$g(z') = \phi_0 + \sum_{k=1}^{M} \phi_k z'_k \tag{12}$$

where $z' \in \{0,1\}^M$ a coalition of simplified input feature values z_k, where a feature's inclusion in the coalition is indicated by $z_k = 1$ (or 0, respectively), M represents the total count of simplified features, ϕ_k represents the Shapley scores of the k-th feature. These Shapley values are calculated as follows:

$$\phi_k = \sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{k\}) - v(S)] \tag{13}$$

here, N represents the entire features set, $S \subset N$ excluding k, and $v(S)$ consists of prediction value of features in S. The value ϕ_k indicates the impact of feature k on the difference in the prediction of the model when feature k is included versus when it is excluded. The factorial terms $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ ensure that all possible permutations of features are considered, providing a weighted average that reflects the marginal impact of each feature across every potential combination.

Many implementations of SHAP have been customized for different model types and use cases. Kernel SHAP employs an approach involving a special weighted local linear regression to compute Shapley values and is suitable for any type of ML model, although it may require significant computation (Lundberg and Lee [20]). On the other hand, tree SHAP is tailored for tree-based models like gradient-boosting machines, decision trees, random forests, etc., leveraging their structure to efficiently calculate Shapley values (Lundberg et al. [19]). Linear SHAP specifically addresses linear models, computing Shapley values by leveraging the model's linear structure for efficiency (Lundberg and Lee [20]). Deep SHAP is designed for DL models, combining SHAP values with the DeepLIFT algorithm to provide fast approximations of Shapley values for neural network models (Lundberg and Lee [20]; Shrikumar et al. [33]). Gradient SHAP, another variant, integrates SHAP values with gradients, approximating expected values by randomly sampling from a distribution defined by the input and a baseline, making it particularly effective for deep learning models (Lundberg and Lee [20]). These implementations, especially Deep SHAP and Gradient SHAP, are crucial for interpreting complex neural networks, where traditional methods may be inadequate due to the complexity of the model and non-linear properties.

SHAP is widely used in many research applications due to its versatility and effectiveness in explaining model predictions. Its applications span across various domains, including healthcare, where it has been used to interpret complex models for diagnosing diseases (Lundberg et al. [18,21]); finance, where it is broadly applied to various financial applications to improve model interpretability and transparency (Martins et al. [22]); and industrial control systems, where it improves AI-based anomaly detection methods (Do et al. [8]). The robustness and interpretability of SHAP make it a valuable tool for ensuring transparency and trust in AI systems, which is crucial for critical decision-making processes across different fields.

3 Proposed Explainable Transformer-Based Anomaly Detection Framework

This section introduces the proposed explainable framework for detecting anomalies in time series data. We will integrate the SHAP method into the existing framework, which combines a transformer-based VAE with the MEWMA-SVDD control chart (Nguyen et al. [24]). By including SHAP, we aim to make the anomaly detection process more transparent and the decisions of the model easier to understand. The combination of the transformer-based VAE and MEWMA-SVDD control chart has proven effective in managing false alarm rates, which is essential for practical applications, especially in healthcare. This integration aims to build a reliable and transparent framework that can be efficiently applied in healthcare settings, ensuring both accuracy and interpretability.

3.1 A Brief Review on Transformer-Based VAE Architecture

We utilized the transformer-based VAE framework in the study of Raza et al. [27] to derive reconstruction error vectors, which serve as the input for the MEWMA-SVDD chart. This network has shown an enhanced ability to identify abnormal signals in ECG datasets relative to other DL models.

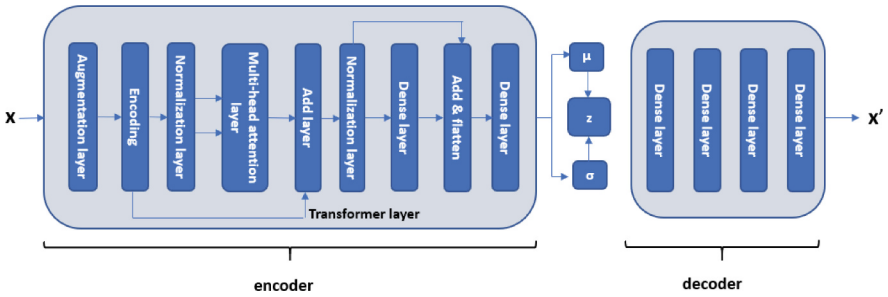

Fig. 3. Transformer-based VAE introduced in study of Raza et al. [27].

The transformer-based VAE architecture is illustrated in Fig. 3. The encoding process starts by an input layer that receives time series data. This data passes into the transformer layer, which includes several sub-layers. Subsequently, the data undergoes normalization in a normalization layer and moves to a multi-head attention layer. The weighted averages of input vectors will be calculated by this mechanism to generate the corresponding output vectors. Specifically, for a set of k-dimensional input vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, their associated output vectors $\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_n$ produced by the self-attention approach are given by:

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

where the sum is calculated across the entire sequence, and $w_{i,j}$ are determined as follows:

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

By applying the softmax function, we ensure that the weights across the sequence add up to one. The decoder consists of 4 dense layers. The last layer uses a sigmoid activation to create probability distributions for potential classes. The detail of this model can be found in the study of Raza et al. [27]. This transformer-based VAE architecture, with its advanced feature extraction capabilities and robust handling of complex time-series data, significantly enhances anomaly detection accuracy, making it particularly effective for critical healthcare applications.

3.2 A Brief Review on MEWMA-SVDD Module

To effectively identify between "normal" and "abnormal" instances without relying on data distribution and to minimize false alarm rate and heighten sensitivity to subtle changes in data, Nguyen et al. [24] proposed incorporating the MEWMA-SVDD chart to the transformer-based VAE architecture mentioned in Subsect. 3.1. Reducing false alarms is vital in healthcare to ensure the safety of patients and precise diagnostics. For instance, reducing false alarms in ECG monitoring helps avoid undue interventions and patient anxiety. The MEWMA technique, which considers every instance in the current monitoring time point, enhances sensitivity to small shifts or changes, making it particularly useful in healthcare settings.

Let $\mathbf{e}_i = \mathbf{x}_i - \mathbf{x}_i'$ be the reconstruction error vector from the transformer-based VAE module. The MEWMA vector $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,t}, \ldots, w_{i,p})$ is calculated using:

$$w_{i,t} = re_{i,t} + (1 - r)e_{i,t-1}, \quad t = 1, 2, \ldots, p \tag{16}$$

here, p is the number of time steps and $\mathbf{e}_i \in \mathbb{R}^p$. The MEWMA-SVDD model is fitted using $\mathbf{w}_i$ for $i = 1, \ldots, N$, where N is the training set size. The threshold h, determined by a pre-specified ARL_0, is used as the upper limit threshold of the control chart to effectively control the false alarms. An instance $\mathbf{z}$ is classified as an abnormal if its distance $\mathrm{D}(\mathbf{z}, \mathbf{a})$ from the center $\mathbf{a}$ exceeds h, where $\mathrm{D}(\mathbf{z}, \mathbf{a})$ is calculated as in Eq. 11. The threshold h of the MEWMA-SVDD chart can be obtained using bootstrap aggregating methods as described by Tang et al. [37]. Given a training set $S = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N \subset \mathbb{R}^n$, the procedure for determining h is given in Algorithm 1.

Algorithm 1: Determining the limit threshold h to manage false alarm rate.
Input: Training data set $S = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}$
Output: Threshold h for MEWMA-SVDD chart

1. Generating M bootstrap samples from S.

2. For each j^{th} bootstrap sample, where $j = 1, \ldots, M$, **do**

- Obtain the SVDD sphere by applying the SVDD algorithm in Subsection 3.2.
- Use Eq. 11 to calculate statistics $D_{j1} = \mathrm{D}(\mathbf{a}^{(j)}, \mathbf{w}_1^{(j)})$, $D_{j2} = \mathrm{D}(\mathbf{a}^{(j)}, \mathbf{w}_2^{(j)})$, $\ldots, D_{jn} = \mathrm{D}(\mathbf{a}^{(j)}, \mathbf{w}_n^{(j)})$.
- Let $\beta = \frac{1}{\mathrm{ARL}_0}$ be the chosen false alarm. Calculate $100 \times (1 - \beta)th$ percentile of n statistics $D_{jk}, k = 1, \ldots, n$, denote this value by $h^{(j)}$.

End for

3. Determine h by $100 \times (1 - \epsilon)$ percentile M values $h^{(j)}, j = 1, \ldots, M$, with ϵ is a small specified error.

This framework's performance, combining transformer-based VAE and MEWMA SVDD, has proven its effectiveness in identifying anomalies within multivariate ECG time series data (Nguyen et al. [24]). Next, we will use the XAI technique to enhance model transparency and trustworthiness. Specifically, we will incorporate the SHAP approach, as mentioned in Subsect. 2.5, into this model to provide transparent and interpretable insights into the anomaly detection process, improving decision-making and user trust.

3.3 Integration of XAI to Transformer-Based VAE Combined with MEWMA-SVDD Chart Framework

In this subsection, we detail the integration of SHAP into the framework mentioned in Subsects. 3.1, 3.2 to enhance its interpretability and transparency. SHAP is well-known for its ability to provide clear and understandable explanations for model predictions, quantifies the impact of features both globally and locally. By applying SHAP analysis globally, we assess the overall feature importance across the dataset in relation to reconstruction error, while locally, we examine individual ECG segments to understand how specific features contribute to reconstruction error in each segment. This focused application of SHAP not only deepens our understanding of how each time window influences the model's outputs but also highlights SHAP's broad applicability for providing detailed insights across all instances and for each specific instance. This approach expands the instance-specific analyses proposed by Raza et al. [27], which focuses only on specific sub-segments within individual ECG signals that contribute significantly to reconstruction loss. By providing detailed explanations at both global

and individual levels, SHAP enhances the reliability and trustworthiness of our proposed anomaly framework. This makes it an invaluable tool for advancing model in critical domains like healthcare.

Figure 4 illustrates our proposed explainable framework for ECG anomaly detection. In this proposed framework, the transformer-based VAE model extracts important feature data, and the resulting reconstruction error vectors are then utilized for anomaly detection. We apply the SHAP approach to the outputs of the VAE model both globally and locally: globally, to understand the overall feature importance across the dataset, and locally, to inspect individual ECG segments for detailed insights into the reconstruction errors. This method enhances the transparency and interpretability of the anomaly detection process facilitated by our framework.

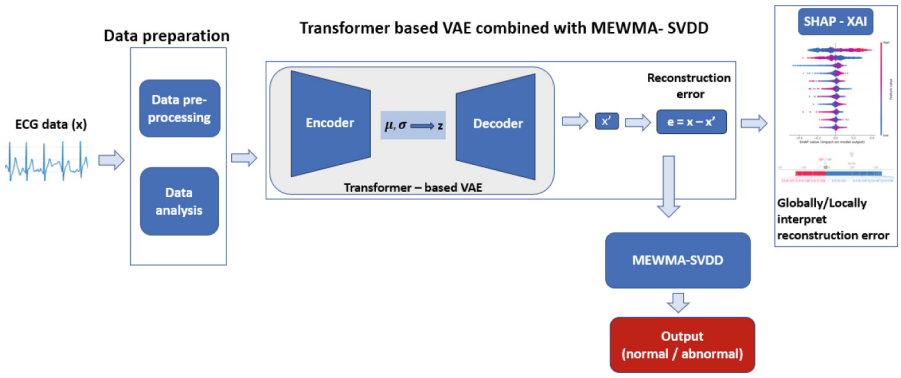

Fig. 4. The proposed explainable framework for ECG anomaly detection.

4 Experimental Procedure and Results

In this section, we first review the data utilization, experimental procedure, and the results of transformer-based VAE combined with the MEWMA-SVDD model proposed in [24]. Then, we integrate SHAP to enhance interpretability within the framework.

4.1 Brief Review of Data Utilization, Experimental Framework

In this subsection, we review the experimental results from the previously published work (Nguyen et al. [24]), which tested the effectiveness of the transformer-based VAE combined with MEWMA-SVDD chart.

Data Description. In this experiment, we combined two datasets from PhysioNet to create a hybrid dataset. The abnormal data were taken from the BIDMC Congestive Heart Failure Database, which contains ECG records from 15 patients diagnosed with severe congestive heart failure. The normal data came from the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) dataset, which includes records from 18 subjects without notable arrhythmias. Each heartbeat in these datasets was pre-labelled by the expert annotators, with the labels indicating the presence of specific arrhythmias or normal rhythms. The dataset consisting of 5,000 heartbeats, including 2,919 normal and 2,081 abnormal for training and testing, was randomly selected. Figure 5 displays visual samples from these databases, with the blue line representing a normal signal and the orange one indicating an abnormal one.

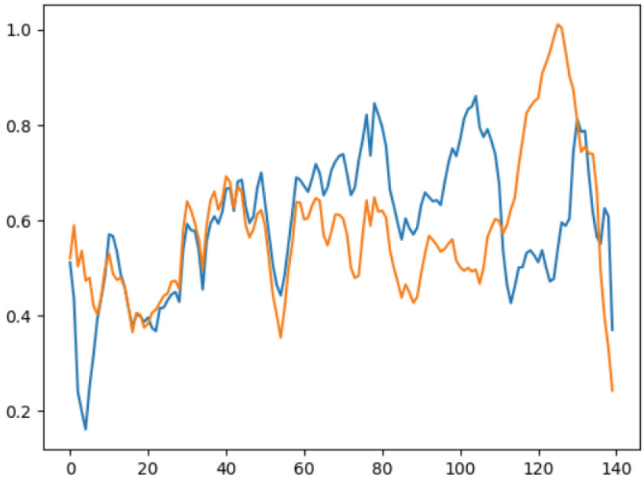

Fig. 5. An example of normal and abnormal instances from ECG dataset.

Experiment Setting. The transformer-based VAE was trained using a dataset of 4,000 randomly selected samples, reserving the remaining 1,000 samples for testing. The training process uses mean squared error to be the reconstruction loss $L(X, X')$. For the MEWMA-SVDD, the optimal bandwidth was identified through a grid search and 30-fold cross-validation. The MEWMA statistic's smoothing parameter was fixed to 0.2, and the in-control ARL ARL_0 was fixed at 100. The limit thresholds of the model were established using Monte Carlo simulations with 10,000 iterations.

Results. The effectiveness of the model was assessed using precision, recall, F1-score, and accuracy. The model obtained an impressive accuracy of 98.77%, outperforming previous approaches. The confusion matrix in Fig. 6 and the classification report in Table 1 highlight the model's effectiveness, showing a high rate of correct classifications and a significant reduction in false positives rate.

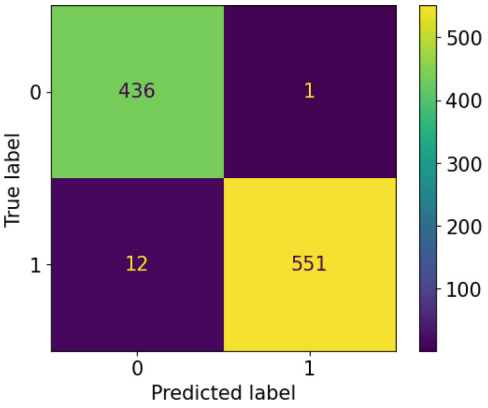

Fig. 6. Confusion matrix of the experiment.

Table 1. Classification results.

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			

Figure 7 illustrates the separation between abnormal and normal classes achieved by this framework. The test samples that lie above the threshold h are identified as anomalies, while those within the limit are identified as normal, demonstrating the model's efficiency. The code for this experiment is available at Github: Transformer-based combined with MEWMA-SVDD chart for ECG monitoring.

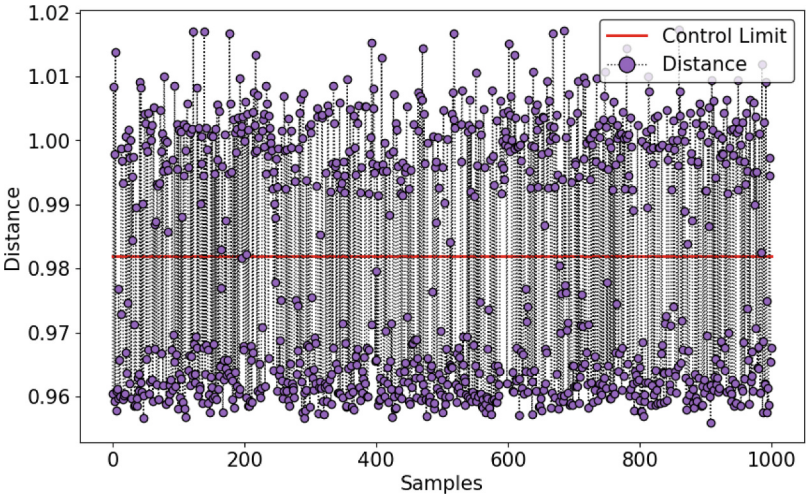

Fig. 7. Separation between normal and abnormal classes in ECG dataset.

4.2 Integration of SHAP Techniques

To enhance the proposed model's interpretability on EGG monitoring, we divided each ECG segment's 140-time steps into 14 non-overlapping windows, each containing 10-time steps. By using this windowing and averaging approach, each input feature to SHAP now represents the average reconstruction error over a window of 10 time steps rather than individual time steps. This helps in understanding which sub-segments of the ECG contribute most significantly to the model's output, simplifying the interpretation and potentially highlighting more meaningful patterns in how errors develop across the ECG sequence.

The SHAP summary plot in Fig. 8 shows the influence of each windowed segment on the reconstruction error. In the plot, each horizontal line represents a different windowed segment of the ECG, identified by labels such as "W_1-10" up to "W_131-140", with the arrangement indicating the hierarchy of their importance to the model's predictions. In this plot, the positive SHAP values indicate that the corresponding window increases the reconstruction error, contributing to the detection of an anomaly. Conversely, negative SHAP values indicate that the corresponding window decreases the reconstruction error, suggesting that the segment is less likely to be anomalous. The color coding reflects the actual values of the features, with blue representing low values and pink representing high values. This suggests that higher values within critical windows (more pink dots on the positive side of the zero line) correlate with having a greater impact on the model's prediction of anomalies. From the figure, we can see

– The plot is arranged with the most significant features at the top of the figure, indicating that "W_131-140", "W_21-30", and "W_121-130" are among the

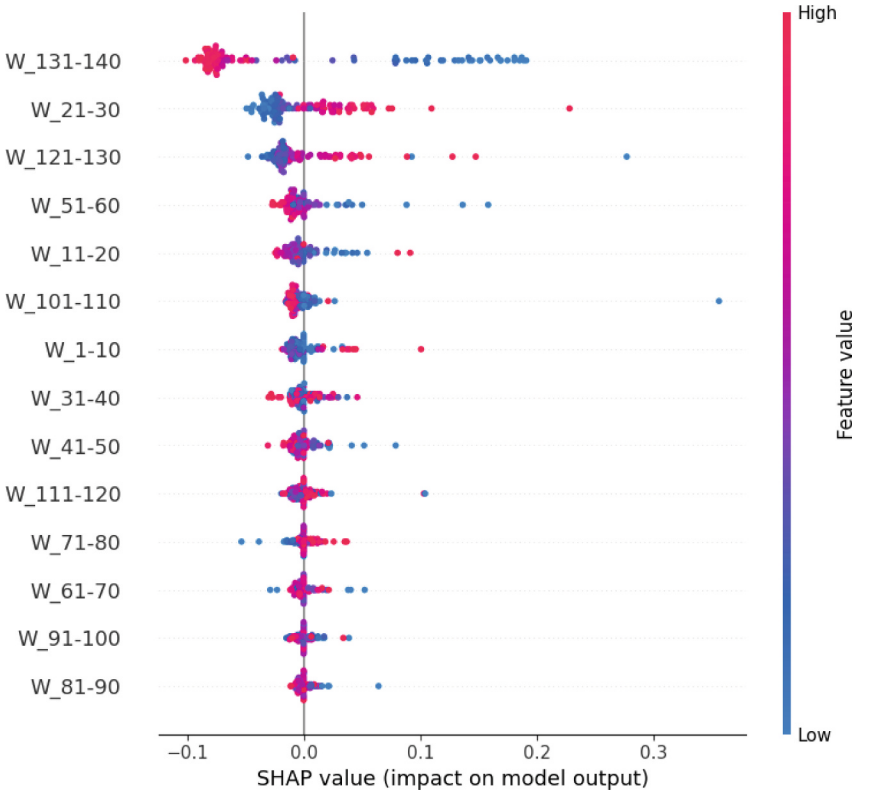

Fig. 8. SHAP plot for explaining VAE reconstruction errors.

most significant in impacting the model's output. This prioritization suggests that these windows likely contain key ECG data patterns critical for the model's predictions. For clinicians, focusing on these key segments could aid in identifying significant deviations from normal ECG patterns, which might correspond to cardiac events requiring immediate attention.

– Conversely, the features positioned at the bottom of the plot, such as W_81-90", W_91-100", and W_61-70", have less impact on the model's decision-making process. Variations within these segments, whether high or low, do not significantly influence the model's predictions of anomalies. For treatment and monitoring, anomalies detected in these lower-impact segments might require a different level of clinical urgency than those identified in the top segments. Understanding this distinction can help clinicians prioritize patient monitoring and intervention strategies based on the likelihood and potential severity of cardiac anomalies detected.

Thus, this global SHAP plot provides a comprehensive overview of the dataset, highlighting which ECG segments commonly influence anomaly detection. Clinicians can leverage these global insights to improve diagnostic precision. By

focusing on the most influential segments, they can efficiently screen for anomalies, enabling fast and precise evaluations in various medical settings. This plot's transparency about how data influences predictions boosts trust in the model. It enables clinicians to identify which ECG patterns are considered important, matching AI results with well-known medical evidence. This alignment confirms that the model's predictions are reliable and based on established practices in heart care.

Besides the global interpretation provided by the SHAP summary plot, a detailed local understanding can be gained by examining individual predictions using a SHAP force plot. This type of visualization specifically highlights how each ECG windowed segment contributes to the model's prediction for a particular case, offering precise insights into the factors contributing to anomalies in each individual instance. Figure 9 shows the force plot for an individual abnormal instance from the ECG5000 dataset and represents the impact of particular windows on the reconstruction errors. Red segments indicate that the corresponding feature increased the model's predicted reconstruction error relative to the base value, suggesting that these windows are more difficult for the VAE model to reconstruct accurately. This difficulty may be associated with anomalies or irregularities in these segments. Blue segments, on the other hand, suggest that the presence of these windows in the input data decreases the overall error, potentially indicating segments that are easier to reconstruct accurately, likely because they align more closely with the normal patterns. From the plot, we can see that window segments "$W_$131-140", "$W_$21-30" and "$W_$101-110" exhibit higher positive SHAP values, which significantly and positively impact the model's output. This suggests that these segments may contain potential sub-segments that define the abnormal pattern in an ECG. Conversely, windows such as "$W_$121-130" or "$W_$51-60" and "$W_$11-20 contribute negatively (lowering the error) to the reconstruction error, indicating areas where the model performs well, thereby suggesting these segments align more closely with the normal patterns of ECG. For a more detailed visualization, these SHAP values are plot together with the real ECG segment, as shown in Fig. 10. The top three contributing windows with the largest SHAP values, which are likely to contain abnormal patterns, are highlighted in varying intensities of red color. These highlighted segments may correspond to the region where VAE model struggles the most in reconstructing ECG signal.

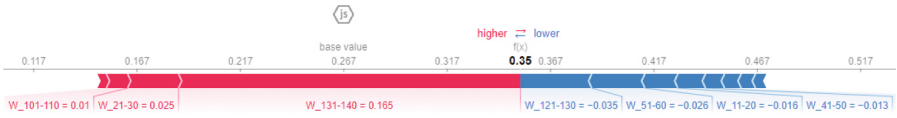

Fig. 9. SHAP force plot for explaining VAE reconstruction error in a specific instance.

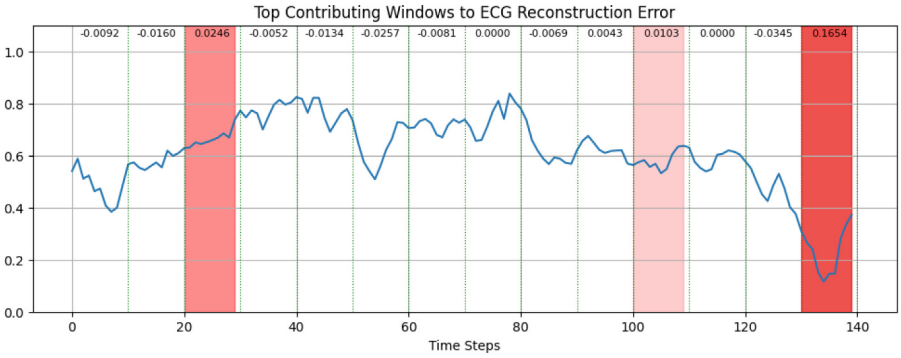

Fig. 10. SHAP Values with highlighted windows contributing to VAE reconstruction error.

To further illustration, another ECG segment is analyzed and visualized in Figs. 11 and 12. According to these figures, the window segments "$W\_131\text{-}140$", "$W\_121\text{-}130$" and "$W\_31\text{-}40$" are highlighted and may indicate the critical patterns that are potential indicator of abnormal cardiac activities. Clinicians should firstly focus on these highlighted segment during their review before further diagnostic testing.

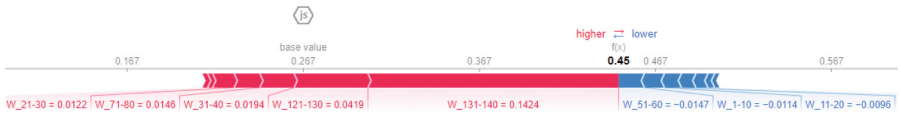

Fig. 11. SHAP force plot for explaining VAE reconstruction error in a specific instance.

Overall, the SHAP value reveals the specific windows of time steps within the ECG segments that are more influential in contributing to the reconstruction errors identified by the VAE model, both at a global and local level. By focusing on these influential windows, we can better understand which segments of the ECG signal are most contributing to anomalies, thereby improving the model's transparency and aiding clinical interpretation.

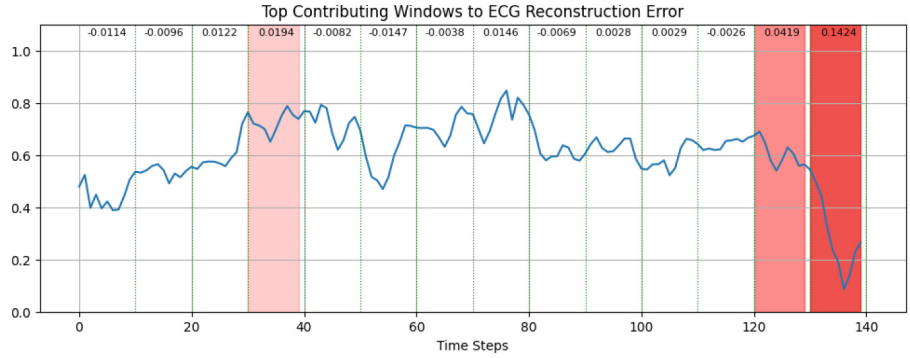

Fig. 12. SHAP Values with highlighted windows contributing to VAE reconstruction error.

5 Conclusion

In this work, we introduced an explainable framework for ECG anomaly detection by integrating XAI into a transformer-based VAE architecture combined with the MEWMA-SVDD chart. Our approach emphasizes the critical need for transparency and interpretability in AI-based models, particularly in sensitive sectors such as healthcare. The transformer-based VAE effectively extracts features from complex ECG time-series data, while the MEWMA-SVDD control chart improves anomaly detection accuracy and reduces false alarm rates. By incorporating SHAP into this framework, we provide detailed, feature-level insights to the contributions of window time steps to the model's predictions. Our applications of SHAP at both global and local levels demonstrate its robustness in explaining the model's decision-making process. This dual application highlights the most influential window time steps in the ECG data and clarifies how these windows impact anomaly classification, offering comprehensive transparency. The detailed explanations of the experiment on the ECG dataset provided by SHAP enhance the model's usability and reliability, making it a valuable tool for healthcare professionals.

Our future works will focus on integrating our framework into federated learning environments to enhance data privacy, employing self-supervised learning techniques to improve model performance with minimal labeled data, and developing XAI-by-design models to further enhance explainability and robustness in AI systems.

Acknowledgments. The authors gratefully acknowledge HEC Liège, Management School, University of Liège, Belgium and GEMTEX, ENSAIT, University of Lille, France for providing the infrastructure and resources essential to our research. We also thank the University of Liège for financially supporting the PhD program through research grants.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Acheampong, F.A., Nunoo-Mensah, H., Chen, W.: Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif. Intell. Rev. 1–41 (2021)
2. Alabi, R.O., Elmusrati, M., Leivo, I., Almangush, A., Mäkitie, A.A.: Machine learning explainability in nasopharyngeal cancer survival using lime and shap. Sci. Rep. **13**(1), 8984 (2023)
3. Albahri, A.S., et al.: A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. Inf. Fusion (2023)
4. Amini, K., Mirzaei, A., Hosseini, M., Zandian, H., Azizpour, I., Haghi, Y.: Assessment of electrocardiogram interpretation competency among healthcare professionals and students of ardabil university of medical sciences: a multidisciplinary study. BMC Med. Educ. **22**(1), 448 (2022)
5. Anstine, D.M., Isayev, O.: Generative models as an emerging paradigm in the chemical sciences. J. Am. Chem. Soc. **145**(16), 8736–8750 (2023)
6. Bolhasani, H., Mohseni, M., Rahmani, A.M.: Deep learning applications for IoT in health care: a systematic review. Inform. Med. Unlocked **23**, 100550 (2021)
7. Chen, H., et al.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12299–12310 (2021)
8. Do, T.H., Nguyen, X.H., Nguyen, V.H., Nguyen, H.D., Truong, T.H., Kim, P.T.: Explainable anomaly detection for industrial control system cybersecurity. IFAC-PapersOnLine **55**(10), 1183–1188 (2022). 10th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2022
9. Frusque, G., Mitchell, D., Blanche, J., Flynn, D., Fink, O.: Non-contact sensing for anomaly detection in wind turbine blades: a focus-SVDD with complex-valued auto-encoder approach. Mech. Syst. Signal Process. **208**, 111022 (2024)
10. Gao, H., Jia, Z.: Detection of threats under inattentional blindness and perceptual load. Curr. Psychol. **36**, 733–739 (2017)
11. Gaudilliere, P.L., Sigurthorsdottir, H., Aguet, C., Van Zaen, J., Lemay, M., Delgado-Gonzalo, R.: Generative pre-trained transformer for cardiac abnormality detection. In: 2021 Computing in Cardiology (CinC), vol. 48, pp. 1–4. IEEE (2021)
12. Jang, J.H., Kim, T.Y., Lim, H.S., Yoon, D.: Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. PLoS One **16**(12) (2021)
13. Kingma, D., Welling, M., et al.: An introduction to variational autoencoders. Found. Trends® Mach. Learn. **12**(4), 307–392 (2019)
14. Kreitz, C., Furley, P., Memmert, D., Simons, D.: Inattentional blindness and individual differences in cognitive abilities. PLoS ONE **10**(8) (2015)
15. 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)
16. Liu, B., et al.: Adaboost-based SVDD for anomaly detection with dictionary learning. Expert Syst. Appl. **238**, 121770 (2024)
17. Liu, H., Zhao, Z., Chen, X., Yu, R., She, Q.: Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records. Comput. Methods Programs Biomed. **196**, 105639 (2020)

18. Lundberg, S.M., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020)
19. Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc. (2017)
21. Lundberg, S.M., et al.: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. **2**(10), 749–760 (2018)
22. Martins, T., De Almeida, A.M., Cardoso, E., Nunes, L.: Explainable artificial intelligence (XAI): a systematic literature review on taxonomies and applications in finance. IEEE Access (2023)
23. Muzammil, M.A., et al.: Artificial intelligence-enhanced electrocardiography for accurate diagnosis and management of cardiovascular diseases. J. Electrocardiol. (2024)
24. Nguyen, T.T.V., Heuchenne, C., Tran, K.D., Tran, K.P.: A novel transformer-based anomaly detection approach for ECG monitoring healthcare system. In: International Conference on Safety and Security in IoT, pp. 111–129. Springer, Cham (2023)
25. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos Solitons Fractals **140**, 110190 (2020)
26. Prendin, F., Pavan, J., Cappon, G., Del Favero, S., Sparacino, G., Facchinetti, A.: The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using shap. Sci. Rep. **13**(1), 16865 (2023)
27. 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. Eng. Appl. Artif. Intell. **121**, 106051 (2023)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, pp. 1135–1144 (2016)
29. Saint-Lot, J., Imbert, J., Dehais, F.: Red alert: a cognitive countermeasure to mitigate attentional tunneling. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (2020)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
31. Shah, H.A., Saeed, F., Diyan, M., Almujally, N.A., Kang, J.M.: ECG-transcovnet: a hybrid transformer model for accurate arrhythmia detection using electrocardiogram signals. CAAI Trans. Intell. Technol. (2024)
32. Shin, D.H., Park, R.C., Chung, K.: Decision boundary-based anomaly detection model using improved anogan from ECG data. IEEE Access **8**, 108664–108674 (2020)
33. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
34. Singh, S., Mahmood, A.: The NLP cookbook: modern recipes for transformer based deep learning architectures. IEEE Access **9**, 68675–68702 (2021)

35. Smola, A., Scholkopf, B., Muller, K.: The connection between regularization operators and support vector kernels. Neural Netw. 637–649 (1998)
36. Suman, G., Prajapati, D.: Control chart applications in healthcare: a literature review. Int. J. Metrol. Qual. Eng. **9**, 5 (2018)
37. Tang, A., Castagliola, P., Hu, X., Xie, F.: An assessment for the conditional performance of an support vector data description (SVDD)-based chart. Qual. Reliab. Eng. Int. 1–17 (2022)
38. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54**(1), 45–66 (2004)
39. Vapnik, V.N.: Statistical Learning Theory. Wiley, Hoboken (1998)
40. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc. (2017)
41. Zhang, C., et al.: VESC: a new variational autoencoder based model for anomaly detection. Int. J. Mach. Learn. Cybern. **14**(3), 683–696 (2023)
42. Zhang, Z., Deng, X.: Anomaly detection using improved deep SVDD model with data structure preservation. Pattern Recogn. Lett. **148**, 1–6 (2021)
43. Zhou, H., Kan, C.: Tensor-based ECG anomaly detection toward cardiac monitoring in the internet of health things. Sensors **21**(12), 4173 (2021)