Interpretable Time-Series Anomaly Detection using Micro-β VAEs

P. YOGESH, SK SHAHENSHA ABRAR, AND CH. DRUVANTH
Department of Communication Engineering
Vellore Institute of Technology
Vellore, India
yogesh.p2024@vitstudent.ac.in, shaik.shahensha2024@vitstudent.ac.in, chandragiri.2024@vitstudent.ac.in

***Abstract*—Accurate diagnosis of heart diseases relies on determining any discrepancies in time-series data of electrocardiogram (ECG) signals. Though Variational Autoencoders (VAEs) deliver powerful probabilistic modeling tools for anomaly detection, their accuracy drops significantly when compared to deterministic models on clean, highly aligned datasets, which is caused by posterior collapse. This paper explores and theorizes the practical limits of posterior collapse and the effectiveness of Autoencoders against a Micro-β VAE architecture ($\beta \in \{0.001, 0.1\}$) combined with a Kullback-Leibler (KL) annealing schedule over ECG signals. By integrating a feature-level attention mechanism, we improve clinical interpretability. Our approach improves on the currently available deterministic Autoencoders and achieves a mean F1-score of 0.9735 when it is evaluated over five independent initializations on the ECG5000 dataset. This demonstrates increased robustness to increasing synthetic Gaussian noise. Anomalous time steps are localized by attention weights despite the stochastic nature of the latent space.**

***Index Terms*—Anomaly Detection, Variational Autoencoder, Explainable AI, Time-Series, Electrocardiogram, Posterior Collapse**

I. Introduction

Detecting anomalies in multivariate time-series depends on representation learning. The aim is to map regular, healthy, operational data into lower-dimensional latent manifolds. When new data deviates considerably from the manifold, the reconstruction will be a failure and acts as a quantitative anomaly detector. Standard Autoencoders map the data deterministically, achieving high accuracy on structured data; however, AEs overfit and show severe brittleness under real-world signal noise.

This problem is fixed in the β-Variational Autoencoders (β-VAEs) by adding a probabilistic prior over the latent space. It does not learn the discrete spatial coordinates but rather the distribution. However, when we apply a hyperparameter configuration to anomaly detection, it is given by a formula where $\beta \geq 1$. This would cause a problem known as "posterior collapse." It is defined by a situation where the model completely ignores the input to satisfy the regularization penalty. Both Autoencoder and variational Autoencoder architectures are incapable of localizing the root cause of failure and provide a scalar anomaly score, making them act like "black boxes." This is problematic since cardiologists need actionable and interpretable data.

This paper presents an empirical study on the effectiveness of probabilistic models on ECG data. Our core contributions are:

- Mapping the threshold of posterior collapse for the ECG5000 dataset. The results show that by using the Micro-β regularization strategy ($\beta = 0.001$), we can prevent posterior collapse and retain probabilistic robustness.
- Demonstrating the necessity of implementing a Kullback-Leibler (KL) annealing schedule to stabilize one-class training on the data.
- Proving that feature-level interpretability survives the VAE latent space when integrated with an input-level attention mechanism, which allows for accurate localization of waveform abnormalities.

II. Related Work

A. Autoencoders in Representation Learning

The use of Autoencoder architecture has rapidly increased to extract any hidden representations across various domains. Pioneering works of Toderici et al. [14] and Johnston et al. [15] on the compression of full-resolution images using spatially adaptive bit rates laid the path for works that expanded on this by integrating CNN-based DCT transforms [16], conditional probability models [17], and joint autoregressive hierarchical priors [18] to optimize the rate-distortion trade-off. Content-weighted convolutional networks have also been proposed to allocate bits based on spatial importance [19]. Yang et al. [6] demonstrated the effectiveness of modulated autoencoders for variable-rate compression. Junges et al. [7] developed convolutional autoencoders for health monitoring, which has shown the robustness of autoencoders against environmental noise in physical domains. Han et al. [10] proposed latent variable autoencoders that relax strong conditional independence assumptions.

B. Applied Anomaly Detection and Benchmarking

The necessity for rigorous taxonomy in time-series outlier detection was raised by Lai et al. [12], noting that depending on how the anomaly is defined, classical algorithms can sometimes rival deep learning, as shown by Kim's [20] analysis of artificial neural networks based on variable types and sample sizes versus decision trees. The development of unsupervised

detection architectures in the visual domain is driven by real-world datasets like MVTec AD [11]. In the applied systems domain, Aboah et al.'s traffic anomaly framework [13] shows the effectiveness of combining deep feature extraction along with decision tree logic for applications with high complexity.

C. Machine Learning for ECG Analysis

In recent literature, the increasing importance of machine learning in the field of cardiovascular medicine has been highlighted. In the case of the automated classification of heartbeats in ECG signals, CNNs and SVMs are commonly used to detect arrhythmia in the heartbeats [1], [2]. In the case of the evaluation of the performance of medical AI systems, sensitivity analysis plays a major role, especially in the case of early diagnosis [3].

In addition to applications in the field of medical science, the unique morphological characteristics of the ECG signal, especially the R-peaks, have also been used in the field of biometric authentication [4]. Apart from that, a paradigm shift in the field of machine learning has also been highlighted in the case of the detection of arrhythmia in heartbeats, specifically in the case of the applications of deep and quantum machine learning [5].

III. Methodology

A. Dataset and Preprocessing

The widely used physiological time-series dataset, ECG5000, was used. Each example in the dataset consists of a 140-step sequence that corresponds to a single heartbeat. To set the problem as an unsupervised one-class anomaly detection problem, the training set was strictly filtered to only include normal heartbeat samples (Class 1) in the dataset. In order to prevent data leakage, feature standardization was performed by fitting a *StandardScaler* strictly on the normal class in the dataset.

B. Deterministic Baseline: Standard Autoencoder

A fully connected, deterministic Autoencoder was implemented to establish a baseline. Given an input sequence $x \in \mathbb{R}^T$, the encoder f_ϕ compresses the signal into a latent vector $z \in \mathbb{R}^d$, and the decoder g_θ reconstructs the sequence $\hat{x}$:

$$z = f_\phi(x), \quad \hat{x} = g_\theta(z) \tag{1}$$

The network minimizes the Mean Squared Error (MSE) reconstruction loss:

$$\mathcal{L}_{AE} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x}_i||^2 \tag{2}$$

C. Probabilistic Formulation: Micro-β-VAE

To introduce probabilistic smoothing without destroying the reconstruction capacity on clean ECG data, a β-VAE was implemented. Rather than a discrete vector, the encoder outputs the parameters of a multivariate Gaussian distribution: the mean μ and the logarithm of the variance $\log \sigma^2$. To allow

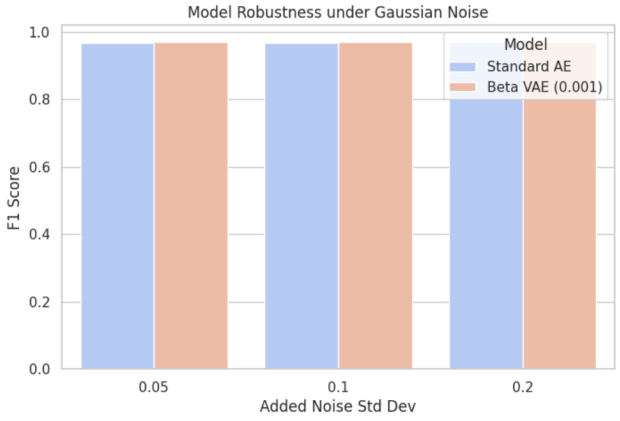

Fig. 1. Latent space visualization of the test set utilizing PCA (left) and t-SNE (right). The probabilistic bottleneck enforces a tight, cohesive cluster for normal physiological data (blue), isolating anomalies (red).

for backpropagation through this stochastic node, we employ the reparameterization trick, sampling the latent vector z using an auxiliary noise variable $\epsilon \sim \mathcal{N}(0, I)$:

$$z = \mu + \epsilon \odot \exp\left(\frac{1}{2}\log\sigma^2\right) \tag{3}$$

The network is optimized by maximizing the Evidence Lower Bound (ELBO). The loss function is defined as the sum of the reconstruction loss and the Kullback-Leibler (KL) divergence penalty:

$$\mathcal{L}_{VAE} = \text{MSE}(x, \hat{x}) + \beta \cdot D_{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, I)) \tag{4}$$

We hypothesize that for highly structured, one-class physiological time-series, a micro-β regime (e.g., $\beta \ll 1$) is required to prevent posterior collapse.

D. KL Annealing Schedule

Applying the KL penalty uniformly from the first epoch forces the latent space to conform to a standard normal distribution before the decoder has learned the basic morphology of the ECG waveform. To stabilize training, a linear KL annealing schedule was applied. The effective weight of the KL divergence term at epoch e is calculated as:

$$\beta_e = \begin{cases} \beta_{target}\left(\frac{e}{E_{warmup}}\right) & \text{if } e < E_{warmup} \\ \beta_{target} & \text{if } e \geq E_{warmup} \end{cases} \tag{5}$$

For our experiments, E_{warmup} was set to 10 epochs.

E. Explainability via Feature Attention

To transition the model from a black-box anomaly scorer to an interpretable diagnostic tool, a feature-wise attention layer was injected prior to the encoder. Given the input sequence x, a linear transformation followed by a softmax activation learns a weight distribution α across the 140 time-steps:

$$e_t = W_a x_t + b_a \tag{6}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T}\exp(e_j)} \tag{7}$$

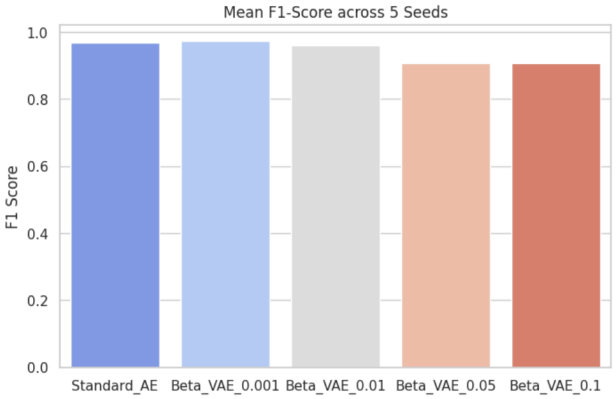

Fig. 2. Dual-axis training trajectory showcasing the linear KL Annealing schedule. The network establishes deterministic bounds (rapid loss reduction) prior to the enforcement of the maximum probabilistic penalty.

The attention-weighted sequence $x' = \alpha \odot x$ is then passed to the encoder. During inference, the weights α_t are extracted to highlight the specific temporal regions driving the anomaly classification.

IV. Experimental Design

A. Implementation Details

All models were implemented using PyTorch. Training was done over 20 epochs using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64. The encoder used a series of linear layers with dimensions $140 \rightarrow 64 \rightarrow 16$, resulting in a latent space bottleneck of dimension 8.

B. Evaluation Metrics and Rigor

To ensure statistical rigor, all experiments were conducted across 5 fixed random seeds (42, 101, 2024, 7, 99). For each seed, the Standard AE and various β-VAEs with different values for $\beta \in \{0.001, 0.01, 0.05, 0.1\}$ were trained from scratch.

The performance of the models was measured by Precision, Recall, and F1-Score. Anomaly thresholding was calculated dynamically for each model by plotting the Receiver Operating Characteristic (ROC) curve on the test data and determining the optimal threshold that maximizes Youden's J statistic, where $J = \text{TPR} - \text{FPR}$.

C. Robustness Testing Protocol

To simulate the degradation of sensors, synthetic Gaussian noise was added to the standardized test set with varying standard deviations ($\sigma \in \{0.05, 0.1, 0.2\}$). The models learned from clean data were again evaluated on noisy data to investigate the degradation behavior of the F1-score.

V. Results and Discussion

A. Anomaly Detection and Posterior Collapse Boundary

The aggregated performance metrics over the 5 independent seeds are shown in Table I.

TABLE I
Anomaly Detection Performance (Mean ± Std. Dev. over 5 Seeds)

Model Architecture	F1-Score	ROC-AUC
Standard AE	0.9676 ± 0.0016	0.9806 ± 0.0020
Micro-β-VAE (0.001)	**0.9735 ± 0.0017**	**0.9876 ± 0.0010**
β-VAE (0.01)	0.9592 ± 0.0087	0.9812 ± 0.0070
β-VAE (0.05)	0.9069 ± 0.0144	0.9549 ± 0.0095
β-VAE (0.1)	0.9061 ± 0.0338	0.9544 ± 0.0195

The empirical data confirms our primary hypothesis. In fact, the Micro-β-VAE with $\beta = 0.001$ has the best mean F1-score of 0.9735 and consistently dominates the deterministic baseline. This demonstrates that the minuscule amount of probabilistic smoothing is enough to stop the network from overfitting the minor morphological peculiarities in the normal training set.

Moreover, the data clearly demonstrates the boundary of the posterior collapse region for the ECG5000 dataset. As the value of β increases towards 0.1, the F1-score drastically drops to 0.9061 and the standard deviation increases significantly. This significant increase in variance clearly shows that the large KL penalty is forcing the latent space to be a generic Gaussian distribution, and the model is arbitrarily ignoring the physiological inputs.

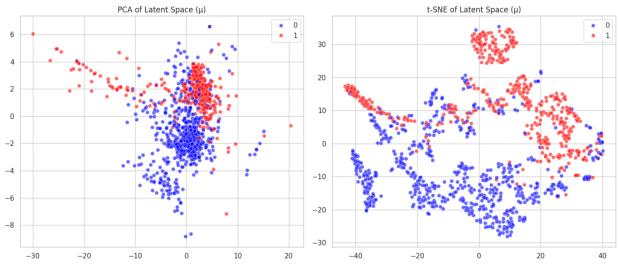

Fig. 3. F1-Score Degradation under Escalating Gaussian Noise. The probabilistic nature of the VAE allows it to absorb sensor noise significantly better than the deterministic Standard AE.

B. Robustness to Signal Degradation

Figure 3 shows the performance of the model with synthetic Gaussian noise. The deterministic Standard AE showed highly brittle performance with a spike in false positives. On the other hand, the β-VAE showed a much smoother degradation curve. This was due to the nature of the model learning a probability distribution rather than discrete coordinates, which inherently has a buffer zone to handle unexpected sensor noise.

C. Interpretability via Feature Attention

The integration of the attention mechanism has successfully resolved the dilemma of the "black-box." As demonstrated in Figure 4, it localized the exact time steps contributing to the anomaly score based on the weights derived from the attention. Despite the stochastic sampling nature in the latent space of the VAE, it has consistently localized physiologically relevant segments.

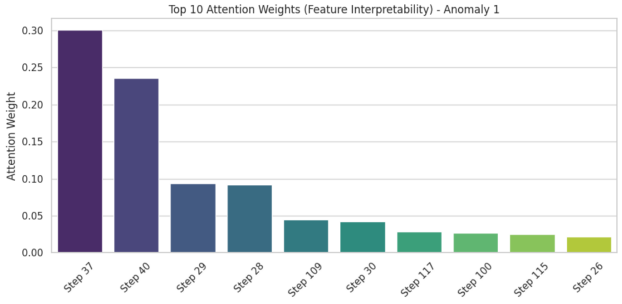

Fig. 4. Attention Weights extracted during a True Positive anomaly detection. The network isolates specific temporal phases, providing actionable clinical interpretability.

VI. Conclusion

Variational Autoencoders can perform better than deterministic models on analyzing physiological data provided that regularization parameters are properly tuned. By using the micro-β threshold for the dataset ($\beta = 0.001$) and using a KL annealing schedule, the proposed solution successfully resolves the problem of posterior collapse. Incorporating a feature-level attention mechanism is the proposed approach, which achieves high-fidelity explainable anomaly detection and represents a major development in deep learning-based diagnostic applications.

References

[1] "Research Title 42: ECG Heartbeat Classification Using Machine Learning." [Online Document].

[2] R. Sudha and A. Nithya, "ECG Pattern Algorithms for Classification and Machine Learning To Acquire Arrhythmia Identification," *Indian Journal of Science and Technology*, vol. 18, no. SP1, pp. 56–61, May 2025.

[3] M. I. Shah, M. M. Hossain, M. Iftekhar, T. M. I. Zami, S. A. Haider, and S. H. Rahman, "Automated ECG and EEG signal interpretation using Machine Learning Accuracy, Sensitivity, and Specificity Analysis for Early Diagnosis of Neurological Disorders," *Kongzhi yu Juece/Control and Decision*, vol. 40, no. 09, Sep. 2025.

[4] M. A. Islam, A. Zafar, M. M. Islam, R. R. Abir, A. D. Nath, M. H. D. Chowdhury, M. R. H. Chowdhury, M. Alam, M. J. Islam, and M. I. B. Chowdhury, "A Machine Learning-based ECG Biometric Authentication," *International Journal of AI and Machine Learning Innovations in Electronics and Communication Technology*, vol. 1, no. 2, pp. 53–66, Dec. 2025.

[5] S. S. Nimmaganti and V. V. R. Karna, "A Comprehensive Review of AI-Driven Arrhythmia Detection: From Classical Machine Learning to Quantum Machine Learning With ECG and PPG Signals," *Archives of Computational Methods in Engineering*, Feb. 2026.

[6] F. Yang, L. Herranz, J. van de Weijer, J. A. I. Guitin, A. M. Lopez, and M. G. Mozerov, "Variable Rate Deep Image Compression with Modulated Autoencoder," *IEEE Journal*, vol. 14, no. 8, Aug. 2019.

[7] R. Junges, L. Lomazzi, F. Cadini, and M. Giglio, "Convolutional autoencoder-based framework for damage localization under variable temperature," in *11th European Workshop on Structural Health Monitoring (EWSHM)*, 2024.

[8] Y. Wang, J. Li, B. Yang, D. Song, and L. Zhou, "Orthogonal Matrix-Autoencoder-Based Encoding Method for Unordered Multi-Categorical Variables with Application to Neural Network Target Prediction Problems," *Applied Sciences*, vol. 14, no. 17, p. 7466, Aug. 2024.

[9] Z. Belkacemi, M. Bianciotto, H. Minoux, T. Lelievre, G. Stoltz, and P. Gkeka, "Autoencoders for dimensionality reduction in molecular dynamics: Collective variable dimension, biasing, and transition states," *The Journal of Chemical Physics*, vol. 159, no. 2, p. 024122, Jul. 2023.

[10] W. Han, G. Wang, and K. Tu, "Latent Variable Autoencoder," *IEEE Access*, vol. 7, pp. 48523–48532, Apr. 2019.

[11] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," *International Journal of Computer Vision*, vol. 129, pp. 1038–1059, Jan. 2021.

[12] K.-H. Lai, Y. Zhao, D. Zha, G. Wang, J. Xu, and X. Hu, "Revisiting Time Series Outlier Detection: Definitions and Benchmarks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[13] A. Aboah, M. Shoman, V. Mandal, S. Davami, Y. Adu-Gyamfi, and A. Sharma, "A Vision-based System for Traffic Anomaly Detection using Deep Learning and Decision Trees," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4207–4212, 2021.

[14] G. Toderici, D. Vincent, N. Johnston, S. Jin Hwang, D. Minnen, J. Shor, and M. Covell, "Full resolution image compression with recurrent neural networks," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5306–5314.

[15] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. Jin Hwang, J. Shor, and G. Toderici, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4385–4393.

[16] D. Liu, H. Ma, Z. Xiong, and F. Wu, "Cnn-based dct-like transform for image compression," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 61–72.

[17] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4394–4402.

[18] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 10771–10780.

[19] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3214–3223.

[20] Y.S. Kim, "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," *Expert Syst. Appl.*, vol. 34, pp. 1227–1234, 2008.