ISSN 2007-9737

Anomaly Detection in Electrocardiographic Signals Based on Machine Learning Models

Helen Alondra Pillado-Hernández[1], Yeritza Gómez-Martínez[2,*], Alfonso Martínez-Cruz[2,3]

[1] Tecnológico de Estudios Superiores de Ixtapaluca,
Ingeniería Biomédica,
Mexico

[2] Instituto Nacional de Astrofísica, Óptica y Electrónica,
Departamento de Ciencias Computacionales, Puebla,
Mexico

[3] Secretaria de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI),
Mexico

helenalondra.pillado@gmail.com, [yeritzagmz, amartinezc]@inaoep.mx

Abstract. The ECG is a crucial tool for the prevention and diagnosis of cardiovascular diseases. However, manual analysis of large volumes of data is prone to errors and generates false alarms. In this work, we propose the design of a model for detecting anomalies in ECG signals, based on machine learning models (CAE, CAE + RF, CAE + SVM). Three approaches were evaluated using performance metrics such as accuracy, F1-score, recall, precision, MCC, ROC, and AUC. According to the results obtained, the model that shows the highest performance and robustness was CAE + RF. Additionally, this model underwent a validation stage with two test sets (A and B). In the first set, segments of normal heartbeats were extracted from the MIT-BIH. Subsequently, different types of artifacts were injected, and in set B: a VAE was trained with normal heartbeats, and the same artifacts as in set A were also injected. The CAE + RF model demonstrated robustness, achieving an AUC = 0.9991, precision = 0.9623, and F1-score = 0.9810. These findings demonstrate that the model allows for robust detection, offering a favorable balance in detecting anomalies and reducing false alarms.

Keywords. Electrocardiogram (ECG), anomaly detection, machine learning.

1 Introduction

Cardiovascular diseases (CVD) are the major cause of death globally and have a significant impact on life expectancy and quality of life. In 2022, as per estimates by the World Health Organization (WHO), 19.8 million people died due to CVD. It also said that over three-quarters of these deaths took place in low- and middle-income nations [23].

The electrocardiogram (ECG) is a basic clinical tool to record the electrical functioning of heart.

Continuous ECG surveillance enables detection of arrhythmias and other abnormalities that may result in serious medical events. Nevertheless, manual signal analysis is challenged by the continually increasing amount of biomedical data, as needed time is long, analysis is complex, and mistakes are prone to happen. Moreover, signals acquired in the outpatient or long-term monitoring are usually contaminated by a myriad of non-physiological noise or artifacts, such as power line interference, muscle noise (EMG), motion artifacts, or poor electrode contact etc., which fade the signals being acquired and can ultimately obscure medically relevant information. This may cause a hamper to medical interpretation and

eventually result in the missing of important cardiac events or misdiagnosis. [22, 1].

Artifact reduction and detection in ECG signals is critically important and challenging part of biomedical signal processing, since it affects not only the quality of the recordings that are visually examined by experts but also the accuracy of the cardiac rhythm classification algorithms executed on those signals. Distortion and noise are the artifacts that have been detected by using thresholds that have been predetermined.

Nevertheless, the non-stationary and time varying nature of many of such artifacts series renders these approaches quite constraining and less effective [21]. Machine learning based methods, particularly deep networks, have demonstrated to be very effective in this challenge, for they are able to model both normal signal and perturbations introduced by noise with complex representations. Among these, convolutional autoencoders (CAEs) have emerged as a machine learning paradigm for anomaly detection. However, their performance as anomaly detectors is poor according to the preset threshold, despite of capturing latent representations of the signal.

The suppression and detection of artifacts in ECG signals remains an open problem in biomedical signal processing, as it affects the quality of the recorded signals not only for visual inspection by cardiologists, but also in terms of the performance of automated cardiac anomaly detection algorithms. The recognition of artifacts and noise was based on the application of a set of predefined thresholds. However, the variability and non-stationary nature of many of these artifacts make these methods limiting and insufficient. [21]. Machine learning-based methods, especially deep networks, have shown extraordinary promise in tackling this problem, as they can model complex representations of both the clean signal patterns and noise-induced distortions. Among these, convolutional autoencoders (CAEs) are one ML tool that has been used for anomaly detection. However, although they capture latent representations of the signal their detection accuracy is poor based on the given threshold.

This paper assesses the performance of three models (Fig. 1): the first convolutional autoencoder model is introduced as an unsupervised model that trained on the normal heartbeat segments with MIT-BIH Arrhythmia Database, which can learn the morphology of the ECG signal and identify any deviation from the normal as suspected abnormal through the reconstruction error, meanwhile the hybrid methods of CAE + Random Forest (RF) and CAE + Support Vector Machine (SVM) leverage the intermediate feature representation obtained from the CAE encoder. These are fed to RF and SVM classifiers learning to differentiate between normal and anomalous beat segments. Both models follow the same procedure of: 0.5-40Hz bandpass filtering, z-score signal normalization, heartbeat segmentation, and model testing with conventional performances (ROC-AUC, F1-score, sensitivity, precision, and confusion matrix).

Moreover, we considered Variational Autoencoder (VAE) for synthetic data generation which enabled us to produce a newly controlled evaluation dataset with various types of artifacts embedded in the synthesized signals as well as another dataset derived from the normal heartbeats of the MIT-BIH Arrhythmia dataset by embedding artifacts in the signals, thus allowing us to test how well the top-performing model out of the three proposed would perform. This evaluation with the new data synthetically generated with VAE demonstrates the effectiveness of the best approach for detecting common artifacts in ECG signals.

This work is structured as follows: Section 3 (Methodology) incorporates the methodology used in each stage: data acquisition, signal preprocessing, signal feature selection and analysis, division and training of the three approaches (CAE, CAE + RF, and CAE + SVM). It also incorporates the architecture, description of the evaluation parameters, and validation of the final model. Section 4 (Results and Analysis) breaks down the results obtained in the performance metrics, comparing the three approaches and validating the approach that obtained the best results. In addition, the scenarios of the two created datasets (normal heartbeats with artifact injection) are compared: the dataset created from MIT-BIT and the one generated from synthetic

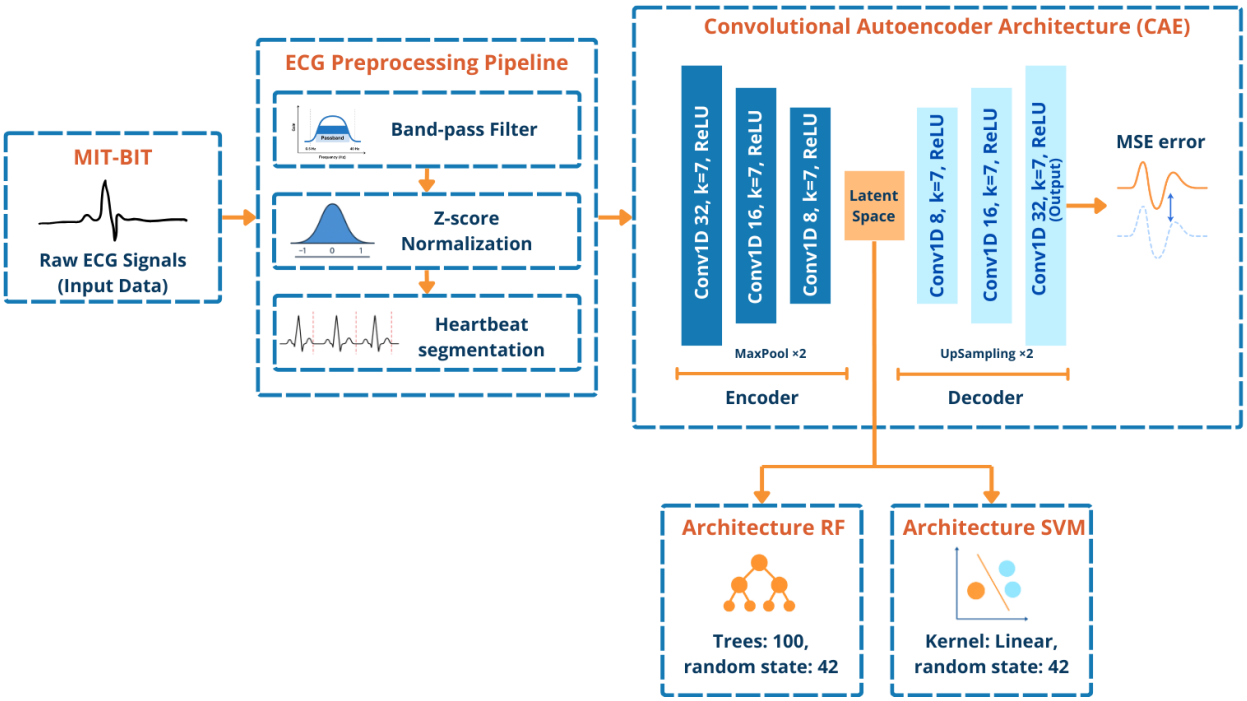

Fig. 1. General architecture of the ECG anomaly detection system

data (VAE). Finally, section 5 provides the final conclusion and future work.

2 Related Work

Some methods used to detect anomalies in ECG signals rely on traditional signal processing approaches such as the Least Squares Method (LSM) adaptive filter to reduce baseline or power line noise and obtain a cleaner signal. For example, Sharma et al. [18] propose an adaptive filter design to attenuate interference, an approach based on LSM. This perspective improves signal quality by focusing on preprocessing, without addressing automatic anomaly detection. If the step size chosed is too small, it leads to slow convergence but cleaner filtering, while if it's large, it implies fast convergence but poorer quality in the filtered signal.

Most of the literature is using the widely known MIT-BIH Arrhythmia database as a benchmark, as access to this kind of database is limited and very strict because of patient data sensitivity and protection [3, 14]. The standardization of the MIT-BIH Arrhythmia database has allowed R-wave detection algorithms, beat classifiers, and beat rhythm recognizers to be compared on a common ground [14]. Many deep learning papers utilize this database to train feedforward neural network or recurrent neural network models [10]. In 2014, Sathyapriya et al. [17] utilized the modified Pan- Tompkins algorithm to detect QRS complex in the signals obtained from the MIT-BIH Arrhythmia database to compute the number of beats per minute, an indicator for the detection of heart disease. This model has been separated into four sub-blocks: the filtering, derivation, squaring and R-peak which is used to distinguish between normal and abnormal heartbeats. It's primary contribution is to segment and preprocess heartbeats; however, no application or testing phase in a simulated clinical environment is included.

In 2001, Osowski and Linh [15] introduced a technique for ECG heartbeat recognition using a hybrid neuro-fuzzy network which integrates

a fuzzy self-organizing layer with multilayer perceptron (MPL). Their approach uses higher-order statistics which are tailored to be less sensitive to physiological differences across patients. The output of the combined neural network as a whole functions as a recognizer/classifier. The MPL serves as a so-called final classifier that decides whether the beat is normal or belongs to one out of seven arrhythmic types. However, it is not an anomaly detector and does not test its robustness against artifacts, something that is crucial for anomaly detection in real clinical scenarios. It also such as the previous one, uses MIT-BIH Arrhythmia, dataset.

A few models are concerned with protecting the privacy of the data, such as Delaney et al. [3] which in 2019 was the first to show that normal ECG signals (lead II of the dataset) can be synthesized using Generative Adversarial Networks (GANs), as well as investigating their robustness against membership inference attacks through the application of Maximum Mean Discrepancy (MMD) and Dynamic Time Warping (DTW).

The procedure is helpful as a source of synthetic signals to safeguard patient privacy and for the challenge of restricted access to public databases. One of its limitations is that it is not robust to wander, 50/60 Hz, EMG or other artifacts. On the other hand, it does not evaluate normal synthetic ECG data for training new models. Likewise, Sumalatha et al. [20] emphasize there are privacy issues associated with datasets. This work offers a comprehensive review of the use of deep learning in ECG, summarizing the progress made since the first CNNs, RNNs, LSTMs, hybrid models, attention mechanisms, and applications such as arrhythmia classification, acute myocardial infarction (AMI), heart rate variability (HRV), remote monitoring, and ECG-based biometric authentication. Moreover, it provides a mapping between reported datasets and preprocessing methods, considering robustness, real-time applicability, and privacy, and focuses on the selection of feature learning using autoencoders and further classifiers. Nevertheless, it does not assess anomaly detectors in the presence of artifacts, nor testing protocols involving controlled perturbations (noise or artifacts).

Among recent contributions from 2025, Galvis-Chacón et al. [7], propose filtering the ECG signal using wavelets (Daubechies-type filters) and automatically adjusting the threshold with search algorithms such as Particle Swarm Optimization (PSO), Differential Evolution (DE), and Genetic Algorithms (GA). They then classify the heartbeats with eXtreme Gradient Boosting (XGBoost) using MIT-BIH signals injected with simulated noise (electrical and white noise), with the aim of verifying the robustness of their noise removal proposal to improve the preprocessing stage in predictive models based on neural networks.

Its main limitations are that it does not address other common types of noise such as EMG, powerline, clipping, and baseline wander; nor does it validate the proposal under stress from specific artifacts. In the same year, Quezada-Próspero et al. [16] proposed training a deep autoencoder with spectral representations with a final sofmax layer to classify atrial fibrillation (AF). The model architecture is lightweight, reduces training times, and offers high accuracy at a lower computational cost. This study focuses on supervised classification and does not evaluate robustness, addressing classification but not anomaly detection.

In general, good advance in abnormality classification and detection in ECG signals by deep learning such as CNN, RNN, or hybrid models, unsupervised or semi-supervised approaches based on autoencoders, as well as signal synthesis (GAN / VAE) to augment the data. Hence, we introduce an anomaly detection model, which is learned from the normal data, and tested on two sets: (A) real segments with injected artifacts, and (B) synthetic datasets (VAE) with the same artifacts. The following section details the research methodology.

3 Methodology

This section presents a description of data and methods used during the development of ECG anomaly detection model and the model where synthetic data is generated for detector validation.

Table 1. Heartbeat class of AAMI and MIT-BIH

Category	Annotations
N	Normal; Left/Right bundle branch block; Atrial escape; Nodal escape
S	Atrial premature; Aberrant atrial premature; Nodal premature; Supraventricular premature
V	Premature ventricular contraction; Ventricular escape
F	Fusion of ventricular and normal
Q	Paced; Fusion of paced and normal; Unclassifiable

3.1 Structure of the MIT-BIH Arrhythmia Database

This database is considered the first set of standard test material available for evaluating arrhythmia detectors. It includes 48 half-hour ambulatory electrocardiographic recordings, obtained from 24-hour two-channel recordings in 47 subjects (records 201 and 202 belong to the same subject), studied by the BIH arrhythmia laboratory between 1975 and 1979.

Twenty-three records were randomly selected from a set of 4,000 records collected from a mixed population of hospitalized patients. The recordings were sampled at 360 samples per second on each channel, with a resolution of 11 bits in a range of 10 mV. The subjects included were 25 men and 22 women, aged between 32 and 89 years.

Among the selection criteria for recordings was that complex interactions of rhythm, morphology, and noise were present in the content [14, 8]. Table 1 shows the MIT-BIH heartbeat classes corresponding to the annotations and the AAMI EC57 (Association for the Advancement of Medical Instrumentation) heartbeat classes [14, 11, 19].

3.2 Computational Environment and System Architecture

This study was conducted using Python version 3 as the main programming language in a Google Colab environment, which provided access to high-performance computational resources including Graphics Processing Units (GPU T4) to accelerate the training of machine learning models (CAE and VAE).

To optimize performance and computational efficiency in the RF classifier, parallelization was configured with $n\_jobs = -1$ to utilize all available CPU cores for training, significantly reducing execution time. Similarly, the Deep Learning models were trained using batch processing, i.e., subsets (256 for CAE, 64 for VAE), optimizing GPU memory usage. An EarlyStopping callback was also implemented during CAE training to prevent overfitting and reduce unnecessary training time.

Table 2 shows the main libraries for development, training, and validation.

Table 2. Python libraries used

Library	Usage Description
TensorFlow / Keras	Construction, training, and evaluation of Deep Learning models (CAE, VAE)
scikit-learn	Implementation of classifiers (SVM and RF), preprocessing, data splitting, and metric calculation
NumPy	Mathematical operations and array handling
Pandas	Manipulation and analysis of tabular data, dataset metadata, and evaluation results
WFDB	Handling and reading physiological data from the MIT-BIH database (.dat, .atr, .hea)
Matplotlib / Seaborn	Libraries for data visualization: ECG signals, ROC-AUC, confusion matrices, and presentation of results
SciPy	Library used for signal processing, design, and application of bandpass filters

Fig. 2 illustrates the flow of methodology from which the proposed CAE, CAE + RF, and CAE + SVM models are constructed. The stages are described below, from data acquisition, through model design and training, to the generation of synthetic data using VAE and dataset generation.

Lastly, the procedure to test the final model performance is described.

3.2.1 ECG Signal Processing

The selected ECG signals underwent the following preprocessing steps to improve signal quality:

— Bandpass filter: With cutoff frequencies of 0.5-40 Hz. This range is widely used to maintain relevant ECG signal information (QRS complex, P wave, and T wave) while attenuating baseline noise (low frequencies) and high-frequency noise such as electromyogram (EMG) noise.

— Z-score normalization: The signals were normalized using the z-score technique [4].

This technique transforms each data point based on its distance from the mean. It is expressed as:

$$x' = \frac{x - \mu}{\sigma}. \tag{1}$$

Where x is the original value, μ is the mean of the data set, and σ is the standard deviation. Normalization ensures that signals from different recordings and leads have comparable amplitude scales, which is crucial for training machine learning models that are sensitive to data scale.

3.2.2 Feature Selection and Analysis

With the use of MIT-BIH annotations, the signals were divided into individual heartbeats, each segment of fixed length (256 samples) centered around each R-wave. To obtain a uniform length, heartbeats near the ends of the signal were padded with zeros. The beats were segregated into two broad classes, namely, normal and abnormal, for classification where normal is represented by the MIT-BIH label “N” and abnormal is with the AAMI-aligned annotation.

3.2.3 Dataset Division and Model Training

The heartbeat segments were divided into training, validation, and test sets. This stratification was essential for optimizing the unsupervised training phase of the CAE and for enabling the subsequent supervised training and evaluation of the discriminative classification models (RF and SVM) of the total set of heartbeats:

— 80% of normal heartbeats reserved for the CAE training phase.

— 10% of the remaining normal heartbeats were allocated to the validation set, which was used to evaluate performance during the training process and to configure the threshold.

— 10% of normal heartbeats were assigned to the test set to evaluate the predictive capacity of the model.

At the same time, the abnormal heartbeats were distributed as follows:

— 50% of the abnormal heartbeats were incorporated into the validation set, complementing the normal heartbeats.

— 50% of abnormal heartbeats were integrated into the test set to balance both classes in the final evaluation.

The network architecture for CAE training can be found in Table 3. These values were chosen to ensure that the model could learn powerful latent representation of normal ECG segments, which leads to the best performance as a representation model for anomaly detection. The CAE model is built to operate on 256-sample ECG segments in two leads (input: 256,2). It uses Conv1D layers with filters of a specific kernel size (7) to detect patterns in the signal, applying Rectified Linear Unit (ReLu) activation functions for complexity or linear for direct reconstruction.

The first approach to CAE training was carried out using hyperparameters crucial to its efficiency and stability. The Adam optimizer was selected for its robustness in convergence [12], operating on the Mean Squared Error (MSE) loss function

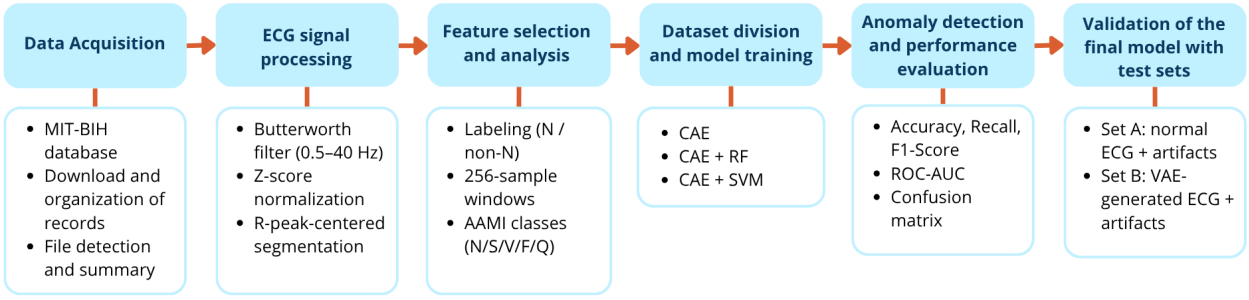

Fig. 2. Methodology of the proposed anomaly detection system

Table 3. Convolutional autoencoder (CAE) architecture

Layer	Filters	Kernel size	Activation	Padding	Pool/UpSampling size
Encoder					
Input	–	–	–	–	(input shape: 256, 2)
Conv1D	32	7	ReLU	same	–
MaxPooling1D	–	–	–	same	2
Conv1D	16	7	ReLU	same	–
MaxPooling1D	–	–	–	same	2
Conv1D	8	7	ReLU	same	–
MaxPooling1D	–	–	–	same	2
Decoder					
Conv1D	8	7	ReLU	same	–
UpSampling1D	–	–	–	–	2
Conv1D	16	7	ReLU	same	–
UpSampling1D	–	–	–	–	2
Conv1D	32	7	ReLU	same	–
UpSampling1D	–	–	–	–	2
Conv1D (output)	2	7	linear	same	–

[6, 2], used to quantify the quality of ECG signal reconstruction.

The training process was run for a maximum of 250 epochs, with a batch size of 256 for each iteration. To prevent overfitting in the model and optimize resources, essential callbacks were implemented: EarlyStopping, responsible for monitoring val_loss and stopping training after 15 epochs without improvement, and ModelCheckpoint, responsible for storing the model version with the minimum val_loss value.

For the second approach, the CAE + RF pipeline leverages CAE feature extraction, using its encoder to transform ECG segments into a lower-dimensional latent space. These extracted features become the input to the supervised RF classifier, whose objective is binary discrimination between normal and abnormal segments.

Training was carried out on a dataset combining the flattened features of normal (0) and abnormal (1) heartbeats from the validation set. The training set was randomized to ensure a uniform

distribution of classes during learning. The model was configured with n_estimators = 100 and random_state = 42 for reproducibility, using n_jobs=-1 to correctly utilize all CPU cores.

Additionally, class_weight=balanced was applied to mitigate the imbalance between classes in the training data, automatically weighting the samples of each class.

In the last approach (CAE + SVM), as in the previous pipeline, it uses the latent features extracted from the encoder, which are then flattened to adapt them to the SVM input. The SVM seeks to find the optimal hyperplane that maximizes the margin between both classes (normal and abnormal heartbeats). For this study, the model was initialized with kernel=linear, random state=42 to ensure the reproducibility of the model results, and probability=True, which allows the model to estimate the probabilities of belonging to the class for the calculation of the ROC-AUC curve.

RandomizedSearchCV was used to adjust the hyperparameters to optimize C (regularization parameter), which explores different combinations of values to find the best configuration and maximize model performance, and gamma (kernel coefficients for nonlinear kernels), which defines the influence of the training instances. The search space ranged from 0.001 to 100 (np.logspace(-3, 2, 6)), while linear and rbf were explored for the kernel. For gamma, strategies such as scale, auto, and a numerical range were included. Additionally, the fitting process was configured with n_iter=10 to sample 10 parameter combinations and cross-validation cv=3 folds, with the aim of finding a balance between exploring the hyperparameter space and computational efficiency. The best hyperparameters identified by this search were applied to the final SVM model.

3.2.4 Anomaly Detection and Perfomance Evaluation

The testing phase allows us to measure the performance of the three models: CAE, CAE + RF, and CAE + SVM. Standard metrics, presented in equations (2-6), were used for this step.

In the Receiver Operating Characteristic Curve Area (ROC AUC), the AUC measures the ability of a classification model to distinguish between classes by plotting the ROC curve. This curve shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) [5, 13, 21, 9]. The TPR, also known as sensitivity or recall, measures the proportion of true positives that were correctly identified, expressed as:

$$Recall = TPR = \frac{TP}{TP + FN}, \tag{2}$$

where TP represents true positives and FN represents false negatives:

FPR is a metric that measures how often the model incorrectly labels normal data as anomalies. Its mathematical expression is:

$$FPR = \frac{FP}{TN + FP}, \tag{3}$$

where FP represents false positives.

Precision (PPV) is the portion of positive classifications that the model labeled correctly, expressed by the equation:

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

Matthews Correlation Coefficient (MCC) is a binary classification metric that summarizes the quality of the classifier using the entire confusion matrix (TP, TN, FP, FN). Its mathematical expression is given by:

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{5}$$

Its range is (-1,1), where the perfect prediction is 1, while -1 represents predictions that are the opposite of reality, and 0 represents performance equivalent to random.

F1-score is the harmonic mean between precision and sensitivity. A high F1-score indicates that the model detects anomalies well (high sensitivity) without generating too many

false alarms (high precision). Its mathematical expression is shown below:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (6)$$

Accuracy represents the proportion of correct predictions, whether false or positive. This metric considers all possible values in the confusion matrix; it's useful when classes are balanced, but otherwise, it loses its usefulness if a class is rare. The mathematical expression is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \quad (7)$$

3.2.5 Validation of the Final Model with Test Sets

In this section, the controlled evaluation of the final pipeline was considered. After training the final model (best model), two datasets were developed, designated as sets A and B. Both sets were created by injecting artifacts into baseline (normal) ECG segments, and both sets were structured to contain 50 samples for each type of artifact and 50 samples of normal segments without artifacts.

To create dataset A, normal ECG segments obtained directly from the preprocessed training set of the MIT-BIH dataset were implemented.

Twelve different types of artifacts were injected into these real segments: baseline wander, powerline, EMG, motion step, motion tri, lead off flat, lead off rail, clipping, invert, time warp, quantize 8 bit, and midding. Each type of artifact was created with random variations in amplitude, frequency, or duration to simulate a spectrum of these interferences. Additionally, an equivalent number of normal segments without artifact injection were included.

Set B was designed to evaluate the generalization ability of the pipeline in the face of variations in normal signal morphology introduced by a generative model (VAE). Development was parallel to set A, but with a notable difference in the basis; in contrast to dataset A, artifacts were injected onto ECG segments synthetically generated by a previously trained VAE. The VAE was trained with normal heartbeat segments to learn and reproduce their probabilistic distribution of heartbeat physiology. By incorporating synthetic data generated by the VAE, variations in the normal ECG signal that the model might encounter in unseen data or data with different morphological characteristics are simulated.

For the VAE architecture (encoder):

— Input: Input layer (256,2).

— Conv1D + MaxPooling1D blocks for spatial reduction (filters: 16, 32, 64; kernel_size = 3; pool_size = 2).

— Flatten and Dense layers (64 neurons with ReLu activation).

— Separate Dense layers for z_mean and z_log_var (both with latent_dim = 16 neurons).

— Lambda layer (sampling) to sample the latent space using z_mean and z_log_var.

While the architecture of the VAE (decoder):

— Input: Shaped input layer (latent_dim,).

— Dense layer to expand to encoded size (T_enc * F_enc).

— latten and Dense layers (64 neurons with ReLu activation).

— Reshape to encoded spatial shape Encoded temporal length (T_enc), number of encoded filters (F_enc) where T_enc=32, F_enc=64).

— Conv1D + UpSampling1D blocks to reconstruct the spatial shape (filters: 64, 32, 16; kernel_size 3; upsampling size 2).

— Conv1D layer (Output): 2 filters, kernel size 3, 'linear' activation, 'same' padding.

The VAE training was performed for 100 epochs, processing the data in batches of 64. In turn, it was performed by implementing the Adam optimizer with a learning rate of 1e-3.

For reconstruction loss, it was calculated with MSE between the original input and the output reconstructed by the decoder, and KL loss, which measures the Killback-Leibler divergence between the learned latent distribution and the standard normal distribution.

4 Results and Analysis

The confusion matrices (Fig. 3) and ROC AUC curves (Fig. 4) summarize the operational performance of each approach proposed in this research for the detection of anomalies in ECG signals. The first approach (CAE), based on reconstruction error, proved to be efficient in detecting anomalies (TP) and few FPs. However, it presents a high number of omissions, i.e., it omits many anomalies as normal (FN = 6,996).

The model presented high accuracy of 94.7%, maintains a moderate recall of 0.6664, AUC of 0.8390, and MMC of 0.53, indicating that the correlation between predictions and actual value is moderate, far from being a robust classification (when it tends to 1).

In the second hybrid approach of CAE + RF, the evaluation metrics improved considerably, obtaining very few FP (11) and few FN (791). TN detection was 8,716 and TP was 20,183. This resulted in a PPV of 0.9995, accuracy of 0.973, F1-score of 0.9810, and MMC of 0.94, an index that indicates that the approach is robust. Finally, in the third hybrid approach of CAE + SVM, it is in the middle between sensitivity of 0.916, with PPV of 0.995, and accuracy of 0.9375. Similar to the previous approach, the approach proved to be robust and computationally light, although it falls short of the CAE + RF model. Compared to the ROC AUC (Fig. 4), the approach applying RF tends to have a higher AUC, indicating that the model has a better ability to distinguish between normal and anomalous classes.

For the first approach (CAE), performance was evaluated by varying the percentile (decision threshold for classifying the signal) and the number of epochs. The 92nd, 93rd, 95th and 96th percentiles were tested with epochs of 100–250.

Table 4 summarizes the values obtained in the evaluation metrics for the three approaches: CAE (varying the hyperparameters), CAE + RF, and CAE + SVM, where the hybrid CAE + RF model was the approach that stood out the most in all the metrics used for the evaluation of the approaches.

In this latter approach (CAE + RF), two scenarios were evaluated: normal data with artifact injection and preprocessed data with VAE. Figure 5 shows the confusion matrices for both scenarios. In both cases, the classifiers showed high capacity to avoid mislabeling normal heartbeat segments as abnormal: TN= 585, FP=15 without VAE and TN= 586 and FP=14 with VAE. The difference between the two cases is observed in the sensitivity; the model detects TP= 170 and FN=430 without VAE, and TP=165 and FN=435 with VAE. The results of examining the ROC-AUC curves in Fig. 6 show an improvement from AUC=0.821 (set A-without VAE) to AUC=0.868 (set B-with VAE), indicating better separability between classes. Table 5 summarizes the evaluation metrics for both datasets under artifact injection. The most significant difference in set B is that it ranks cases better by anomaly probability. The model, whether using VAE or not, is highly specific but not very sensitive to subtle artifacts.

Table 6 summarizes the different types of artifacts used in this article. It also measures the performance of the CAE + RF pipeline without using VAEs. The model maintained a low false alarm rate for clean signals. Furthermore, it detects electrode disconnection events with a very high rate ($lead\_off\_flat$: 100%; missing: 92%), although it shows limited sensitivity to low-amplitude or diffuse spectrum artifacts such as baseline wander, EMG, powerline, and time warp.

Table 7 presents the same method but using VAE. The results from this dataset for model validation showed a decrease in the mean probability and detection of weak artifacts such as baseline wander, EMG, and poserline, indicative of attenuation of subtle signal irregularities. The model using VAE maintains good performance in the face of electrode disconnections such as lead off, flat, and missing, and in motion artifacts (motion tri and motion step), where a slight improvement was observed compared to dataset A.

5 Conclusion

In this work, the performance of the semi-supervised hybrid model (CAE + RF) was evaluated, analyzing the impact of data preprocessing using a VAE. This approach reaches better performance than CAE and CAE + SVM,

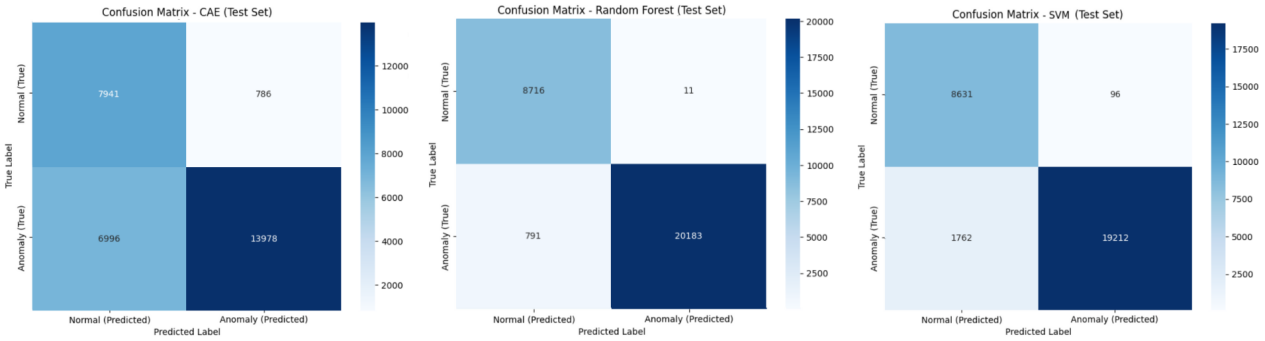

Fig. 3. Performance metrics for CAE, CAE + RF y CAE + SVM

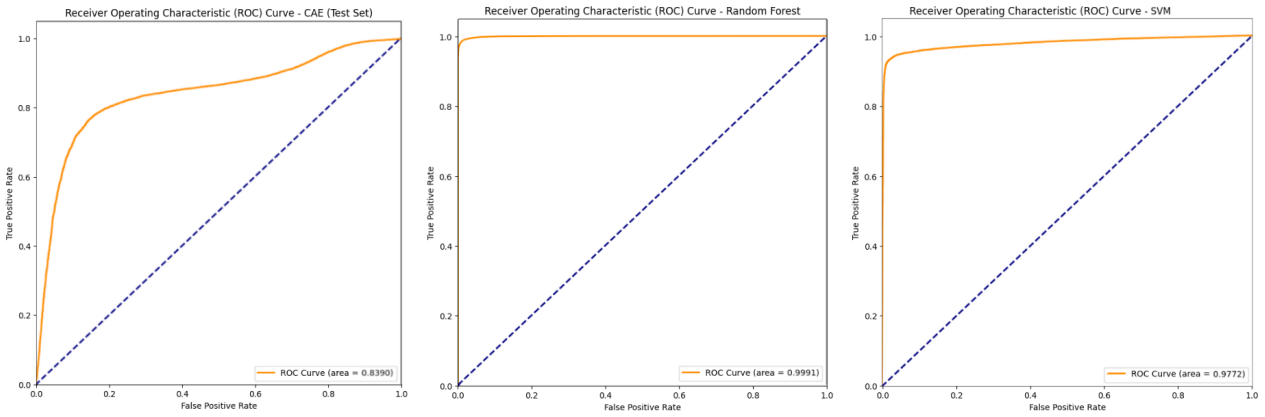

Fig. 4. ROC AUC of the models (Performance metrics: CAE, CAE + RF and CAE + SVM)

Table 4. Model comparison for different configurations

Configuration	Accuracy	Precision	Recall	F1-Score	AUC	FN	FP
CAE							
P95-100e	0.3780	0.8670	0.1340	0.2320	0.7444	18163	431
P95-250e	0.6619	0.9593	0.5443	0.6945	0.8673	9557	485
P92-250e	0.7380	0.9468	0.6664	0.7822	0.8390	6996	786
P96-200e	0.6160	0.9653	0.4709	0.6330	0.8730	9951	318
RF							
Trees: 100, random state: 42	0.9730	0.9995	0.9623	0.9810	0.9991	791	11
SVM							
Kernel: Linear, random state: 42	0.9375	0.9950	0.9160	0.9540	0.9772	1762	96

demonstrating a solid approach for detecting anomalies in ECG signals. The system maintains a low false alarm rate with an AUC of 0.821 (without VAE) and 0.868 (with VAE). Its performance is

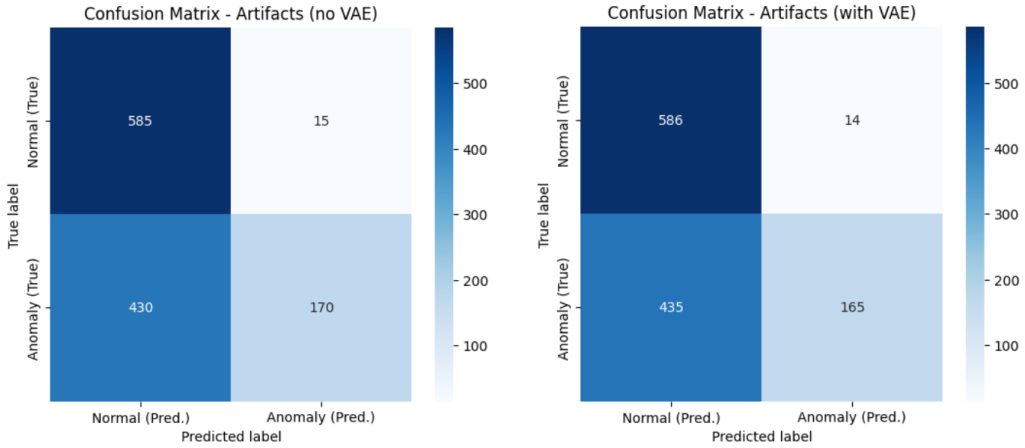

Fig. 5. Confusion matrices of CAE+RF under artifacts: without VAE and with VAE

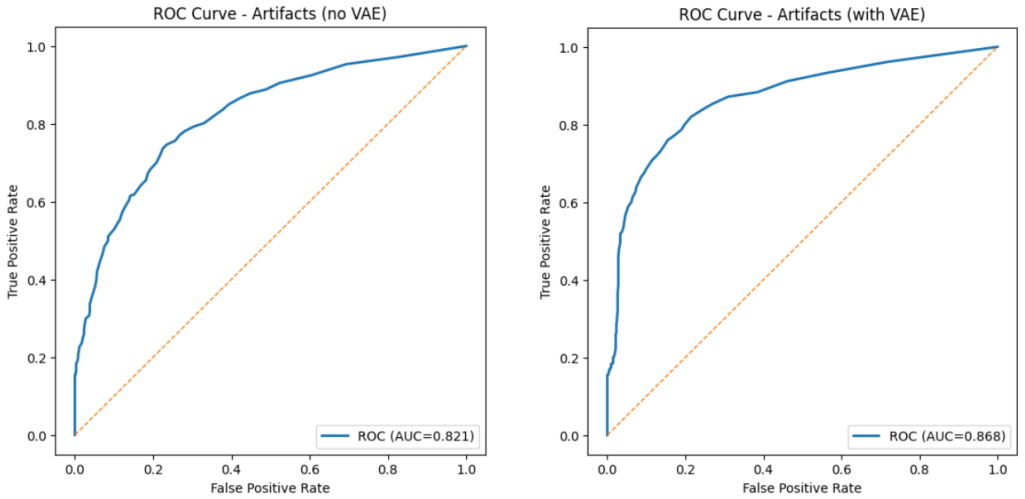

Fig. 6. CAE+RF ROC curves under artifacts: without VAE vs. with VA

Table 5. CAE + RF under artifact stress: without/with VAE preprocessing (positive class = anomaly)

Scenario	Accuracy	Precision	Recall	F1-Score	AUC	FN	FP
No VAE	0.6292	0.9189	0.2833	0.4334	0.8211	430	15
VAE	0.6258	0.9218	0.2750	0.4226	0.8681	435	14

sustainable, especially in the face of faults such as lead-off flat/rail and missing lead/signal drop, as well as in clean signals. However, the model's sensitivity decreases in the face of low-frequency disturbances such as baseline wander, power-line interference (50/60 Hz), low-level EMG, clipping level, time warping, 8-bit quantization, and motion; where the VAE tends to smooth the signal and shift the score toward a normal value. Overall, the obtained results support the semi-supervised approach as a practical basis in a clinical setting, while also highlighting the

Table 6. CAE+RF with low artifacts (without VAE). Summary by artifact type: (threshold=0.5, positive class = anomaly)

Artifact	is_anomaly_class	n_samples	mean_prob	detection_rate
baseline_wander	1	50	0.2448	0.060
clean	0	600	0.1082	0.025
clipping	1	50	0.2872	0.160
emg	1	50	0.2470	0.040
invert	1	50	0.5180	0.660
lead_off_flat	1	50	0.9856	1.000
lead_off_rail	1	50	0.4086	0.080
missing	1	50	0.8752	0.920
motion_step	1	50	0.4046	0.220
motion_tri	1	50	0.2790	0.100
powerline	1	50	0.1970	0.060
quantize_8bit	1	50	0.0522	0.000
time_warp	1	50	0.2336	0.100

Table 7. CAE+RF under artifacts with VAE. Summary by artifact type (threshold=0.5, positive class = anomaly)

Artifact	is_anomaly_class	n_samples	mean_prob	detection_rate
baseline_wander	1	50	0.1942	0.020
clean	0	600	0.0642	0.0233
clipping	1	50	0.2634	0.120
emg	1	50	0.2092	0.020
invert	1	50	0.5048	0.500
lead_off_flat	1	50	0.9900	1.000
lead_off_rail	1	50	0.4258	0.100
missing	1	50	0.8488	0.880
motion_step	1	50	0.4110	0.240
motion_tri	1	50	0.3258	0.260
powerline	1	50	0.1370	0.040
quantize_8bit	1	50	0.1124	0.060
time_warp	1	50	0.2058	0.060

need to adjust the threshold and strengthen the preprocessing in order to improve sensitivity to low-energy interference.

As future work, we propose improving the anomaly detection model by implementing an anomaly classifier and optimizing the Random Forest classifier threshold to improve the balance between FN and FP. Additionally, we propose incorporating additional features that can enrich the input information in order to improve the model's ability to distinguish subtle anomalies.

Furthermore, we propose conducting an in-depth analysis to understand the types of artifacts that are most difficult to detect. Finally,

strategies need to be developed for each of the artifacts, with the aim of making the model more sensitive, robust, and secure for application in clinical environments.

References

1. **Agrawal, S., Gupta, A. (2013).** Projection operator based removal of baseline wander noise from ecg signals. 2013 Asilomar Conference on Signals, Systems and Computers, pp. 957–961. DOI: 10.1109/ACSSC.2013.6810431.
2. **Choi, S., Choi, K., Yun, H. K., Kim, S. H., Choi, H.-H., Park, Y.-S., Joo, S. (2024).** Diagnosis of atrial fibrillation based on ai-detected anomalies of ecg segments. Heliyon, Vol. 10, No. 1, pp. e23597. DOI: 10.1016/j.heliyon.2023.e23597.
3. **Delaney, A. M., Brophy, E., Ward, T. E. (2019).** Synthesis of realistic ecg using generative adversarial networks.
4. **Developers, G. (2025).** Numerical data normalization. https://developers.google.com/machine-learning/crash-course/numerical-data/normalization. Accessed: 13-Jul-2025.
5. **Elyamani, H. A., Salem, M. A., Melgani, F., Yhiea, N. M. (2024).** Deep residual 2d convolutional neural network for cardiovascular disease classification. Scientific Reports, Vol. 14, No. 1, pp. 22040. DOI: 10.1038/s41598-024-72382-3.
6. **Esmaeili, F., Cassie, E., Nguyen, H. P. T., Plank, N. O. V., Unsworth, C. P., Wang, A. (2023).** Anomaly detection for sensor signals utilizing deep learning autoencoder-based neural networks. Bioengineering, Vol. 10, No. 4. DOI: 10.3390/bioengineering10040405.
7. **Galvis-Chacón, J., Ramos-Soto, O., Oliva, D., Valdivia, A., Rostro-González, H., Zapotecas-Martínez, S., Pérez-Cisneros, M. (2025).** Optimizing electrocardiogram denoising for enhanced cardiovascular disease detection: A metaheuristic approach. Computación y Sistemas, Vol. 29, No. 1, pp. 77–89. DOI: 10.13053/CyS-29-1-5532.
8. **Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Stanley, H. E. (2000).** Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation, Vol. 101, No. 23, pp. e215–e220. DOI: 10.1161/01.CIR.101.23.e215. RRID:SCR_007345.
9. **Google Developers (2025).** Classification: Accuracy, recall, precision, and related metrics. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=es-419. Accessed: 2025-07-17.
10. **Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., Ng, A. Y. (2019).** Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, Vol. 25, No. 1, pp. 65–69. DOI: 10.1038/s41591-018-0268-3.
11. **Kachuee, M., Fazeli, S., Sarrafzadeh, M. (2018).** Ecg heartbeat classification: A deep transferable representation. 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 443–444. DOI: 10.1109/ICHI.2018.00092.
12. **Kingma, D. P., Ba, J. (2017).** Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. DOI: 10.48550/arXiv.1412.6980. Version 9; published as a conference paper at ICLR 2015.
13. **Matias, P., Folgado, D., Gamboa, H., Carreiro, A. V. (2021).** Robust anomaly detection in time series through variational autoencoders and a local similarity score. Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) -

BIOSIGNALS, INSTICC, SciTePress, pp. 91–102. DOI: 10.5220/0010320500002865.

14. **Moody, G., Mark, R. (2001).** The impact of the mit-bih arrhythmia database. IEEE Engineering in Medicine and Biology Magazine, Vol. 20, No. 3, pp. 45–50. DOI: 10.1109/51.932724.
15. **Osowski, S., Linh, T. H. (2001).** Ecg beat recognition using fuzzy hybrid neural network. IEEE Transactions on Biomedical Engineering, Vol. 48, No. 11, pp. 1265–1271. DOI: 10.1109/10.959322.
16. **Quezada-Próspero, E., Mujica-Vargas, D., Cruz-Próspero, L. A., García-Aquino, C., Rendón-Castro, A. A. (2025).** Atrial fibrillation classification using a deep spectral autoencoder. Computación y Sistemas, Vol. 29, No. 1, pp. 15–28. DOI: 10.13053/CyS-29-1-5527.
17. **Sathyapriya, L., Murali, L., Manigandan, T. (2014).** Analysis and detection r-peak detection using modified pan-tompkins algorithm. 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp. 483–487. DOI: 10.1109/ICACCCT.2014.7019490.
18. **Sharma, I., Mehra, R., Singh, M. (2015).** Adaptive filter design for ecg noise reduction using lms algorithm. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 1–6. DOI: 10.1109/ICRITO.2015.7359333.
19. **Shin, D.-H., Park, R. C., Chung, K. (2020).** Decision boundary-based anomaly detection model using improved anogan from ecg data. IEEE Access, Vol. 8, pp. 108664–108674. DOI: 10.1109/ACCESS.2020.3000638.
20. **Sumalatha, U., Prakasha, K. K., Prabhu, S., Nayak, V. C. (2024).** Deep learning applications in ecg analysis and disease detection: An investigation study of recent advances. IEEE Access, Vol. 12, pp. 126258–126284. DOI: 10.1109/ACCESS.2024.3447096.
21. **Tian, C., Zhang, F. (2025).** Self-supervised ecg anomaly detection based on time-frequency specific waveform mask feature fusion. IEEE Access, Vol. 13, pp. 97585–97596. DOI: 10.1109/ACCESS.2025.3572484.
22. **Watson Hernández, R. A. (2022).** Interpretación del electrocardiograma normal: Electrocardiograma. Revista Ciencia y Salud Integrando Conocimientos, Vol. 6, No. 5, pp. 85–91. DOI: 10.34192/cienciaysalud.v6i5.549.
23. **World Health Organization (2025).** Cardiovascular diseases (cvds). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2025-11-10.

Article received on 31/10/2025; accepted on 15/12/2025.
**Corresponding author is Yeritza Gómez-Martínez.*