IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 8 January 2024, accepted 27 January 2024, date of publication 1 February 2024, date of current version 26 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361031

RESEARCH ARTICLE

Exploring Cardiac Rhythms and Improving ECG Beat Classification Through Latent Spaces

ALBA VADILLO-VALDERRAMA[1], JACOBO CHAQUET-ULLDEMOLINS[1,2], REBECA GOYA-ESTEBAN[1], RAÚL P. CAULIER-CISTERNA[3], JUAN JOSÉ SÁNCHEZ-MUÑOZ[4], ARCADI GARCÍA-ALBEROLA[4], AND JOSÉ LUIS ROJO-ÁLVAREZ[1,5], (Senior Member, IEEE)

[1]Department of Signal Theory and Communications and Telematic Systems and Computation, Rey Juan Carlos University, Fuenlabrada, 28943 Madrid, Spain
[2]Artificial Intelligence Factory, BBVA 28050 Madrid, Spain
[3]Department of Informatics and Computing, Faculty of Engineering, Universidad Tecnológica Metropolitana, Santiago 8330378, Chile
[4]Cardiology Service, Arrhythmia Unit, Hospital General Universitario Virgen de la Arrixaca, El Palmar, 30120 Murcia, Spain
[5]D!lemmaLab Ldt, 28917 Fuenlabrada, Spain

Corresponding author: Alba Vadillo-Valderrama (alba.vadillo@urjc.es)

This work was supported by the Research Grants meHeart, LATENTIA, and PCardioTools under Grant PID2019-104356RB-C42/AEI/10.13039/501100011033, Grant PID2022-140786NB-C31/AEI/10.13039/501100011033, and Grant PID2022-140553OA-C42/AEI/10.13039/501100011033. The work of Raúl P. Caulier-Cisterna was supported by the Agencia Nacional de Investigación y Desarrollo (ANID) Fondecyt Postdoctorado 2021 under Grant 3210305, and in part by the Autonomous Community of Madrid [European Laboratory for Learning and Intelligent Systems (ELLIS) Madrid Node].

ABSTRACT In recent years, a wide variety of Machine Learning (ML) algorithms, including Deep Learning (DL) methods, have been proposed for electrocardiogram (ECG) beat classification. However, accurately discerning ECG beat types faces challenges due to noise interference and inherent imbalances among different classes. Moreover, understanding mathematical models enclosed by black-box learning systems is an issue today. Our study employed a manifold learning algorithm capable of mapping high-dimensional data into a latent space to conduct a comprehensive analysis within a neural network learning framework. This approach involved the following studies: 1) examining the intermediate high-dimensional latent space in simple architectures by studying its projection into a visualizable latent space; 2) exploring the influence of class imbalance on the configuration of the latent space; 3) evaluating and analysing the compensatory effects of employing diverse DL architectures, such as modified autoencoders and Generative Adversarial Networks (GANs), specifically in generating data augmentation through synthetic beats. The experimental results demonstrated the effectiveness of our methodology in mitigating noise and addressing inter-class imbalance, notably enhancing the diagnostic Area Under the ROC Curve (AUC) in ECG signal analysis. Implementation of GAN data augmentation techniques resulted in a 2% improvement, elevating the AUC from 0, 9332 to 0, 9520 in the biclass dataset. Similarly, the AUC values increased by 2%, from 0, 9020 to 0, 9223, for the multiclass dataset. These findings highlight the impact of appropriate data augmentation techniques on AUC improvement. Furthermore, visualizing latent spaces during beat classifier design significantly contributes to developing solid and principled multiclass beat-discriminating systems.

INDEX TERMS Electrocardiography, machine learning, unsupervised learning, neural networks, dimensionality reduction, manifold learning, generative adversarial networks, data augmentation.

I. INTRODUCTION

Cardiovascular diseases (CVD) are disorders of the heart and blood vessels. They are one of the leading causes of mortality worldwide and a serious public health problem, as noted by the World Health Organization. Among the CVDs, the arrhythmias are characterized by an altered heart rhythm. The heart generates electrical impulses to pump blood throughout the body. However, during an arrhythmia, these impulses become irregular and originate from undesired locations [1]. Early identification and characterization of arrhythmias are crucial for accurate diagnosis and risk

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

assessment in clinical practice [2] and essential for effective patient treatment. One of the most common techniques for early detection and treatment of these arrhythmias is using medical devices measuring the Electrocardiogram (ECG), which is the primary test for detecting cardiac irregularities. Furthermore, studies have shown that long-term ECG signal monitoring and analysis can significantly improve CVD diagnosis, control, and prevention. The Holter System, a portable medical device with attached cutaneous electrodes on the chest [3], records ECG signals over extended periods, ranging from 24 hours to several days, making it one of the most effective devices for detecting arrhythmias.

In recent years, significant progress has been made in the automated detection of cardiac arrhythmias and abnormal beats [4], [5]. Traditionally, cardiologists would visually inspect and annotate ECG recordings, which requires deep domain knowledge and signal preprocessing. However, this manual method is time-consuming and can sometimes lack accuracy. Digital technology has led to the development of automated systems, revolutionizing heartbeat identification. Modern systems employ advanced signal processing algorithms and Artificial Intelligence (AI) methods to analyse ECG signals automatically and detect abnormal rhythms. ECG arrhythmias detection remains a challenging problem, as evidenced by numerous studies aimed at improving heartbeat classification using Machine Learning (ML) algorithms [6], [7]. Traditional detection methods have extensively utilized supervised ML algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). These algorithms depend on labelled training data to identify patterns and classify new instances, yet their black-box nature makes interpreting them challenging.

Deep Learning (DL) is a branch of ML that has emerged as a powerful approach in the field of ECG and has earned significant attention due to its ability to extract complex features and model intricate relationships automatically, allowing for more nuanced and accurate assessments of cardiac health [8]. DL systems offer the potential for continuous, real-time monitoring and increased precision in interpreting ECG signals, thereby improving the likelihood of detecting intermittent arrhythmias. Integrating ML and DL models into ECG analysis standardizes interpretations and mitigates variability inherent in human analysis, potentially leading to improved patient outcomes.

Autoencoders (AE), since the early days of DL, have been a resourceful tool for dimensionality reduction and feature extraction. According to the literature, AE techniques have consistently enhanced computational efficiency, user satisfaction, and crucial metrics like the Area Under the ROC Curve (AUC) across various domains [9], [10], [11]. In ECG beat classification, their application has contributed to more refined and dependable results [12], [13]. However, AE algorithms typically use high-dimensional latent spaces, leading to a black-box classification process. Complex and nonlinear transformations in deep neural networks contribute to this black-box nature, obscuring the interpretation of their internal representations and decisions.

Over the last years in ECG analysis, the contributions of DL have immensely increased [14], [15], [16]. Some of these contributions have addressed the ECG denoising problem, but the most successful of them have focused on classification applications, particularly detecting and classifying arrhythmias [5], [17], [18], [19]. One of the applications that are gaining relevance in DL is the generation of synthetic heartbeats [20] or Data Augmentation (DA). The basic DA techniques of ECGs initially involved random transformations of beats, such as scaling, flipping, and noise addition. However, these basic DAs often modify the inherent properties of ECG signals, generating noise rather than enriching the dataset with meaningful samples. These augmented samples may adversely impact ECG classification, as noted in [21], where the authors reported detrimental effects of horizontal and vertical flipping DA operations on their classifier [22]. To address the limitations of basic DA techniques, advanced methods can offer a viable alternative. A highly successful technique for generating new data is the Synthetic Minority Over-sampling Technique, commonly known as SMOTE [23], [24]. This technique involves sampling data from the minority class by creating data points along the line segment connecting a randomly chosen data point and one of its KNNs. Notably, Generative Adversarial Networks (GANs) represent one such advanced technique with utility in DA. This field has garnered much attention in computer vision for its ability to generate data without explicitly needing to model the probability density function [25], [26]. GANs are a distinctive class of neural network models in which two networks are trained simultaneously in such a way that one specializes in image generation while the other focuses on discrimination. With their auspicious potential, GANs have found significant applications in medical imaging.

Due to the black-box nature of deep neural networks, it poses challenges in understanding how these models arrive at their predictions. The term black box refers to sufficiently complex models, making them challenging for humans to interpret directly. In healthcare, where decisions have life-and-death implications, the lack of interpretability in predictive models can erode trust. To address this concern, there has been a surge in research on explainable machine learning. While the potential of explainable machine learning is substantial, it is crucial for cardiologists encountering these techniques in clinical decision-support tools or research papers to critically understand both their strengths and limitations [27]. For this reason, interpreting the decisions made by DL models is crucial. The lack of interpretability limits their applicability in critical areas such as healthcare. Researchers have proposed various techniques to interpret DL models to uncover insights into their decision-making processes [28], [29], [30]. These include visualization methods such as Manifold Learning, an unsupervised set of techniques that give us visually interpretable representations and operates under the assumption that any observed data exist within a lower-dimensional manifold that is embedded in a higher-dimensional space [31]. These techniques identify low-dimensional structures within high-dimensional data [32].

In contrast to conventional dimensionality reduction methods like Principal Component Analysis (PCA), which primarily assume linear relationships, manifold learning takes a different approach. It focuses on discerning the underlying manifold or the intricate surface on which data points reside. This approach is particularly effective in capturing nonlinear dependencies, a key aspect in ECG signal analysis. Despite its potential, manifold learning techniques are still relatively unexplored in ECG analysis, with limited research papers dedicated to this area. For example, [31] demonstrated the application of manifold learning in discriminating different types of ventricular fibrillation, showcasing its utility. A notable manifold learning technique is Uniform Manifold Approximation and Projection (UMAP) [33]. UMAP is unique in its foundation, built upon principles of Riemannian geometry and algebraic topology. This combination results in a practical and scalable algorithm, making UMAP well-suited for complex, real-world data applications like those found in ECG analysis.

In this paper, we introduce an innovative methodology to address the dual challenge in ECG signal analysis: mitigating overfitting due to class distribution imbalance and enhancing signal quality for better classification. Our approach integrates two key strategies: (1) transforming the input space into a more informative latent space using AE techniques for effective feature extraction; and (2) generating new, synthetic samples with AE and GANs. By refining the latent space, we improve the signal quality, as well as the classification and diagnostic accuracy. Recognizing the critical issue of class imbalance, our use of AE and GANs focuses on creating realistic, representative synthetic samples. These samples are instrumental in balancing the dataset, improving the training process, and ensuring enhanced model performance and generalization.

The subsequent sections of this paper are structured as follows. Section II explains the methods employed in this research, which are presented in detail, encompassing the data description, the utilization of AE for feature extraction, and the implementation of various classifiers. In section III, the dataset we used is described. Section IV is devoted to presenting the experimental setup and the obtained results, highlighting the performance of the proposed approach. Finally, section V concludes the paper by summarizing the obtained results, discussing their implications, and suggesting potential avenues for future research.

II. ALGORITHMS

This section describes the fundamentals of our methodology, algorithms, and metrics. A detailed explanation of the methods used in this study is provided.

A. AUTOENCODERS

We implemented a neural network architecture capable of unsupervised learning in terms of a simple AE architecture [11]. This neural network consists of two main components, namely, an encoder and a decoder, as shown in Figure 1. The input data vector, denoted as $\mathbf{x}$, serves as the input to the encoder layer, where $\mathbf{x} \in \mathbb{R}^n$. This primary objective of the encoder is to generate a representation with a different dimensionality implementing a function f_{enc}, using the weight matrix $\mathbf{W}_{\text{enc}}$ and bias vector $\mathbf{b}_{\text{enc}}$. This function transforms the input data into a reduced-dimensional representation, known as latent space and denoted as $\mathbf{h}$, $\mathbf{h} \in \mathbb{R}^m$, where $m < n$. This information compression aims to capture the main characteristics of the above input data and retain the crucial patterns.

The encoder function is denoted by:

$$\mathbf{h} = f_{\text{enc}}(\mathbf{x}) = \phi(\mathbf{W}_{enc}x + \mathbf{b}_{enc}) \tag{1}$$

where ϕ represents an activation function, and commonly used choices include the sigmoid function σ or the Rectified Linear Unit (ReLU) function.

The decoder function f_{dec} reconstructs the input data as approximated data $\mathbf{x}'$ from this latent space representation, $\mathbf{h}$, using the weight matrix $\mathbf{W}_{\text{dec}}$ and bias vector $\mathbf{b}_{\text{dec}}$ [31]. The decoder function is denoted by:

$$\mathbf{x}' = f_{\text{dec}}(\mathbf{h}) = \varphi(\mathbf{W}_{\text{dec}}\mathbf{h} + \mathbf{b}_{\text{dec}}) \tag{2}$$

where φ is a nonlinear activation function, possibly and sometimes different from ϕ mapping.

B. GENERATIVE ADVERSARIAL NETWORKS

GANs are unsupervised learning techniques in the ML field. They are designed to learn patterns and structures from input data automatically. GANs are crucial in generating synthetic data that closely mirrors real data distributions. This generative model comprises two key components: the generator and discriminator models, illustrated in Figure 2. The generator model creates new examples, while the discriminator model evaluates these generated examples to distinguish them from real data. Through this adversarial process, GANs continually improve the quality and realism of their synthetic data.

The generator, denoted as G, is trained to produce new samples by taking a noise input, $\mathbf{z}$, from a prior noise space, $P(\mathbf{z})$, and mapping it to the target data space $(\mathbf{x})$, where fabricated data is generated. This mapping is achieved through the function $G(\mathbf{z}; \theta_G)$, where θ_G represents the parameters of the generator:

$$\mathbf{x}_{\text{fake}} = G(\mathbf{z}; \theta_G) \tag{3}$$

where, $\mathbf{z}$ is a random variable sampled from the distribution $P(\mathbf{z})$, and $\mathbf{x}_{\text{fake}}$ represents a sample generated by the generator. The parameters θ_G correspond to the generator's parameters subject to training.

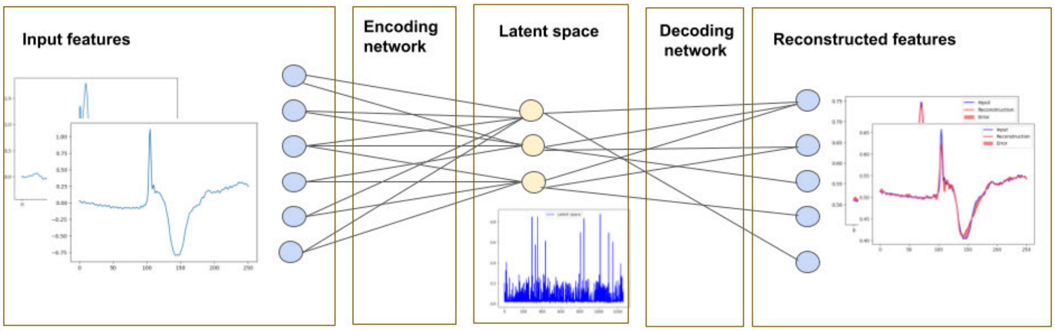


FIGURE 1. Representation of the structure with each layer of an autoencoder.

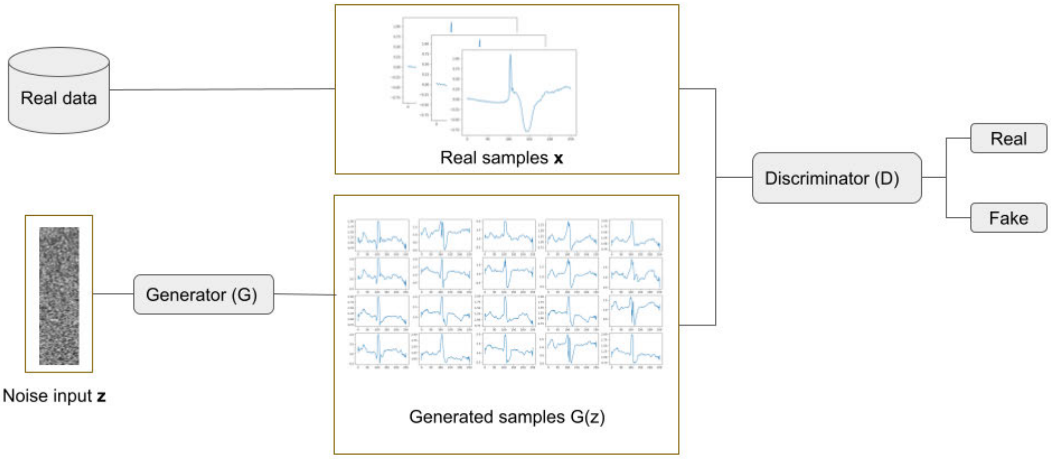


FIGURE 2. Representation of the structure with each layer of a generative adversarial network.

The discriminator, denoted as D, classifies samples as either genuine (belonging to the dataset domain) or counterfeit (generated by the generator). It takes an input $\mathbf{x}$, which can be either real or generated, and produces an output representing the probability that $\mathbf{x}$ is real. This probability estimation is achieved through the function $D(\mathbf{x}; \theta_D)$, where θ_D denotes the discriminator parameters,

$$D(\mathbf{x}) = D(\mathbf{x}; \theta_D) \tag{4}$$

where $D(\mathbf{x})$ represents the estimated probability that $\mathbf{x}$ is a real sample, while θ_D corresponds to the parameters of the discriminator that undergoes training. These two models are jointly trained in a zero-sum, adversarial game until the discriminator model is fooled about half the time, meaning the generator model generates a plausible sample [25].

C. SMOTE

The SMOTE algorithm stands out as one of the earliest and remains the most widely adopted algorithmic method for generating synthetic data in datasets [23], [34], [35]. The SMOTE algorithm is configured with two parameters: k neighbors (indicating the number of nearest neighbors to consider) and the desired count of new points to generate. Each iteration of the algorithm involves the following steps:

1) Randomly choose a minority point.
2) Randomly select any of its k nearest neighbors from the same class.
3) Randomly assign a lambda value within the range [0, 1].
4) Generate and position a new point on the vector between the two selected points, situated lambda percent of the distance from the original point.

SMOTE operates by pairing minority class observations and creating synthetic points along the connecting line. Its approach to selecting minority points is relatively liberal, potentially encompassing outliers.

D. SUPPORT VECTOR MACHINE

Support Vector Machine is an ML technique that aims to find an optimal hyperplane that separates the data points from different classes while maximizing the margin between the hyperplane and the closest data points, known as support vectors [36]. Different kernels can be implemented, and for this work, we used the radial basis function,

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{5}$$

where $\mathbf{x}, \mathbf{y}$ are two input samples, and σ is a bandwidth parameter that controls the influence of each neighboring sample in the transformed feature space.

The SVM can resolve classification problems (SVC) or regression problems (SVR). To estimate SVC, we determine the optimal combinations of hyperparameters using cross-validation. Cross-validation is a statistical technique applied to our data using machine learning to approximate the results better, avoid overfitting, and provide a more robust estimate of its generalization capabilities. This technique divides the dataset into multiple folds, iteratively trains the model on different combinations of these folds, and tests the remaining ones. We can identify the optimal hyperparameter values for the SVC model by systematically evaluating different hyperparameter configurations by cross-validation.

E. MANIFOLD APPROXIMATION AND PROJECTION

This work uses UMAP, a nonparametric graph-based dimensionality reduction algorithm, for representation and data analysis [37]. The technique constructs a graph by connecting neighboring data points based on their similarity or distance in the high-dimensional space. Then, it optimizes embedding the data points in the low-dimensional space to minimize the discrepancy between the distances in the original and projected space. The resulting UMAP embedding provides a compressed representation of the data while effectively reducing the dimensionality, preserving its local and inherent structure when projected into a lower-dimensional space.

For UMAP mathematical formulation, given a dimensional data set $\mathbf{x} = x_1 \ldots, x_N$, where $\mathbf{x} \in R^D$, in the embedding space there is a $\mathbf{z} = z_1 \ldots, z_N$,where $\mathbf{z} \in R^d$. In the high-dimensional space, when considering two data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathbf{x}_i$ selects $\mathbf{x}_j$ as one of its neighboring points based on the conditional probability $\mathbf{x}_j$, which is then defined as

$$p_{j|i} = exp\left[\frac{-d(\mathbf{x}_i, \mathbf{x}_j) - \rho_i}{\sigma_i}\right] \tag{6}$$

where σ_i is the perplexity parameter, controlling the number of effective neighbors, ρ_i can be seen as the averaged weighted distance of a given point with its neighbors, and d($\mathbf{x}_i$, $\mathbf{x}_j$) represents the distance between x_i and x_j. The value of $p_{j|i}$ is computed only for approximately n neighbors, resulting in $p_{j|i} = 0$ for all other j. UMAP uses a symmetrization of the high-dimensional probability [38], as a combination of $p_{j|i} \cdot p_{i|j}$, leading to:

$$p_{i,j} = (p_{j|i} + p_{i|j}) - p_{j|i} \cdot p_{i|j} \tag{7}$$

The subtraction term $p_{j|i}p_{i|j}$ corrects for the overlap between the two conditional probabilities, ensuring that the joint probability is not overcounting the shared information.

Once the data have been projected into the low-dimensional space, the similarity between high-dimensional data points is expected to be preserved in the low-dimensional representation. If we assume that the mapped positions of the high-dimensional data point $\mathbf{x}_i$ and $\mathbf{x}_j$ in the low-dimensional space are $\mathbf{h}_i$ and $\mathbf{h}_j$, respectively, the distribution $q_{i,j}$ in the low-dimensional space is defined as:

$$q_{i,j} = (1 + a||\mathbf{h}_i - \mathbf{h}_j||^{2b})^{-1} \tag{8}$$

F. METRICS

To evaluate the results obtained in the different experiments and to facilitate comparing these results, we considered the AUC. This metric provides a comprehensive assessment and facilitates a robust analysis of the outcomes of the different experiments. The AUC metric is appropriate for unbalanced datasets because it measures the ability of the classifier to discriminate between positive and negative classes without being overly sensitive to class imbalances. Unlike accuracy, which can be misleading in imbalanced datasets, AUC considers the entire range of possible thresholds, comprehensively assessing a model's performance across various operating points. This makes AUC a robust choice for evaluating classifiers on unbalanced data, where accurately capturing the minority class is often important. The AUC needs two parameters for calculation, the True Positive Rate (TPR) and False Positive Rate (FPR), described with the following equations,

$$TPR = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = \frac{FP}{TN + FP} \tag{10}$$

$$AUC = \int_0^1 TPR\, d(FPR) \tag{11}$$

where TP are the true positives, FN are the false negatives, TN are the true negatives, and FP are the false positives. For evaluating multiclass classification, we use the One-vs-Rest method. This evaluates multiclass models by comparing each class against all the others simultaneously. Doing this, we convert the multiclass classification task into a series of binary classification tasks.

AUC is complemented with additional metrics such as Accuracy and F1-score to provide a comprehensive evaluation. This enhances and strengthens the understanding of classification performance. Accuracy is a metric that measures the overall correctness of a classification model. It is calculated as the ratio of correctly predicted instances to the total number of cases in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

F1-score is a metric that balances precision and recall, providing a single value that combines both measures. It is beneficial when there is an uneven class distribution, described with the following equations:

$$F1 - score = \frac{2 \times Precission \times Recall}{Precission + Recall} \tag{13}$$

$$Precission = \frac{TP}{TP + FP} \tag{14}$$

TABLE 1. Description of the beats belonging to each class for multiclass AAMI regrouping, where *beats* stands for the total number of beats, *mean* stands for the mean number of beats per patient of each class, and *std* stands for the standard deviation for all the beats of each class per patient.

Classes	Beats	Mean	Std	MIT-BIH
Non-ectopic (N)	90.631	1.888,1	710,18	N, L, R, e, j.
Supra-vent. ectopic (S)	2.781	57,934	206,30	A, a, S, J.
Ventricular ectopic (V)	7.236	150,75	242,48	V, E.
Fusion (F)	803	16,72	73,99	F.
Unknown (U)	8.043	167,56	561,54	/, f, Q

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall and 0 indicating the worst possible F1-score. It is a valuable metric for evaluating the overall performance of a classification model, especially in situations where class distribution is imbalanced.

III. DATASET DESCRIPTION

PhysioNet is an online resource dedicated to advancing research in various biomedical and physiological signals, including areas such as cardiology and neurology. In this study, we utilize the widely recognized MIT-BIH Arrhythmia Database [39], [40], a cornerstone resource in cardiac signal research, accessible through the PhysioNet web repository [41]. This database is particularly valuable due to its comprehensive range of cardiac arrhythmia labels, making it valuable for ECG signal analysis, and many authors have used it to date.

Initially, the MIT-BIH Arrhythmia Database encompasses 19 different labels, which include: normal beat (N), left bundle branch block (L), right bundle branch block (R), bundle branch block beat (B), atrial premature beat (A), aberrated atrial premature beat (a), nodal premature beat (J), supraventricular premature or ectopic beat (atrial or nodal) (S), premature ventricular contraction (V), R-on T premature ventricular contraction (r), fusion of normal and ventricular beat (F), atrial escape beat (e), nodal (junctional) escape beat (j), supraventricular escape beat (atrial or nodal) (n), ventricular escape beat (E), paced beat (/), fusion of paced and normal beat (f), unclassified beat (Q), and beat not classified during learning (?).

In the literature, regrouping the diverse labels of ECG signals into fewer categories for more effective analysis is usual. In this study, we employ two different criteria for this purpose. The first is a multiclass labelling system, regrouping into five types (N, S, V, F, U) as recommended by the AAMI standard, a commonly used approach in many studies [1], [42], [43], [44], detailed in Table 1. This system, however, might lack clinical intuitiveness as it combines various criteria like QRS morphology and beat origin. To address this, we also employ a biclass regrouping, formulated by an expert cardiologist in our team. This approach, outlined in Table 2, primarily distinguishes between supraventricular and ventricular-originated beats, offering greater clinical relevance.

TABLE 2. Description of the beats belonging to each class for biclass regrouping where *beats* stands for the total number of beats, *mean* stands for the mean number of beats per patient of each class, and *std* stands for the standard deviation for all the beats of each class per patient number.

Classes	Beats	Mean	Std	MIT-BIH
Class 1	93.410	1.946	690	N, A, a, e, J, j, L, R.
Class 2	15.943	332,14	595	V, F, f, /.

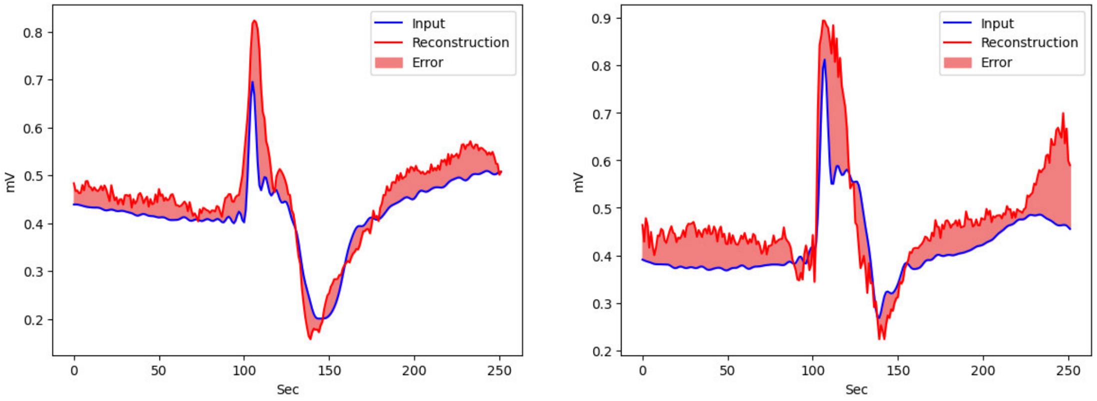
Cardiac signals often contain noise from interference, electrode contact issues, or physiological factors. This noise can significantly distort the signals, complicating their analysis and subsequent interpretation [45]. To counter these challenges and obtain meaningful results from ECG analysis, it is crucial to have an effective digital preprocessing pipeline. Our focus is on cleaning the noise while preserving the signal morphology. This preprocessing step is essential as every ECG in our database is a collection of consecutive beats, where maintaining the integrity of each beat is crucial for accurate analysis.

IV. PROPOSED METHODOLOGY

This section describes the core elements of the methodology. Firstly, it provides a comprehensive explanation of the data preparation steps, which is essential to support a wide range of experiments. Secondly, it describes a series of sequential steps to measure the impact of using latent space and data augmentation techniques to assess noise reduction and overfitting on the data set prepared in the previous point.

A. DATA PREPARATION

For our experiments with the MIT-BIH database, we employed three distinct approaches to analyse the performance and generalization capabilities of our algorithms in different scenarios. The first experiment used the entire dataset, with patient beats divided into train beats and test beats. This division ensured no overlap of patients between the training and test sets, aiming to mitigate the risk of overfitting. The second experiment utilized the original dataset but excluded data from five patients. In this subset, we did not separate the beats of patients into distinct train and test sets, allowing for the possibility of the same patient's beats appearing in both sets. However, we ensured different beats in the train and test sets. The third experiment focused exclusively on the data from the five excluded patients. This targeted selection was designed to uncover unique patterns or characteristics specific to these patients. By applying these varied methodologies, our study comprehensively evaluates how these different data handling strategies affect classification accuracy in the context of ECG beat classification.



(a)

(b)

FIGURE 3. Representation of the input signal in a blue line, the reconstructed signal in a red line, and their difference: (a) Representation implementing AEs; (b) Representation implementing GANs.

B. SEQUENTIAL METHODOLOGICAL BREAKDOWN

Our proposed methodology comprises four key steps, each designed to address specific challenges in ECG classification: (1) Data augmentation for the minority class to balance the dataset; (2) Establishing high-dimensional latent representation to capture complex patterns; (3) Applying nonlinear dimension reduction to simplify the latent space; And (4) Quantifying classification performance and providing graphical representations for in-depth analysis.

C. STEP 1: DATA AUGMENTATION FOR THE MINORITY CLASSES

The first step in our methodology tackles the twin challenges of overfitting and class imbalance, which often lead to variations in classification accuracy, especially when models are tested on independent datasets. We observe that while models perform well on samples from training patients, their accuracy diminishes with beats from different patients. This necessitates strategies to enhance model generalization across a diverse patient population. To this end, we generate new synthetic samples using advanced data augmentation techniques, explicitly targeting the minority class beats. Employing AEs and GANs, we create synthetic samples that are both representative and diverse, enriching our training datasets [10], [46].

Figure 3(a) illustrates our AE implementation, showing a visual comparison between an original beat (in blue) and its AE reconstruction (in red). Our focus here is no on noise reduction using AEs, but on generating and scrutinizing latent variable representations to guide classifier design and enhance understanding. The discrepancy between the two beats is indicated in pale red, representing the error. Figure 3(b) demonstrates using GANs to augment regular heartbeats. In both cases, the reconstructed signals maintain minimal error, indicating effective augmentation. Our approach, concentrating on the minority class, aims to alleviate class imbalance issues and enhance classification performance. We create multiple datasets incorporating augmented data to allow for comprehensive comparative analyses. These synthesized beats are integrated with the original dataset, fostering a more balanced data representation during training. Through this, we aim to develop models with improved generalization capabilities, ensuring robust performance across varied patient demographics.

D. STEP 2: HIGH-DIMENSIONAL LATENT REPRESENTATION

The second step in our methodology involves transforming the actual and augmented datasets into a latent space. This latent space, an abstract multidimensional realm, encodes significant internal representations of observed events. While not immediately interpretable, these feature values in the latent space compress or augment the data, ensuring that similar data points in this space are also closer in the input space. Our initial hypothesis posited that leveraging a latent space for beat samples would aid in noise reduction and signal refinement, isolating the core information of interest.

We utilized AE techniques, renowned for compressing multidimensional data into concise latent representations, to achieve this. Our experiments focused on the latent space generated by the encoder. We experimented with various encoders, selecting the most effective based on performance. This selection process involved an extensive parameter sweep, evaluating different aspects, such as layer count, encoding dimension, epochs, and batch size, to determine the optimal configuration for our objectives. For this stage, we incorporated two architectural approaches: the first uses a regular AE. In contrast, the second adopts a modified AE approach, called the encoder plus fine-tuning method. To carry out this fine-tuning, we have implemented a transfer learning technique, which involves adding a softmax layer alongside the encoder layer for the classification task. This additional layer is then retrained using the encoder weights. This approach allows us to leverage the knowledge acquired

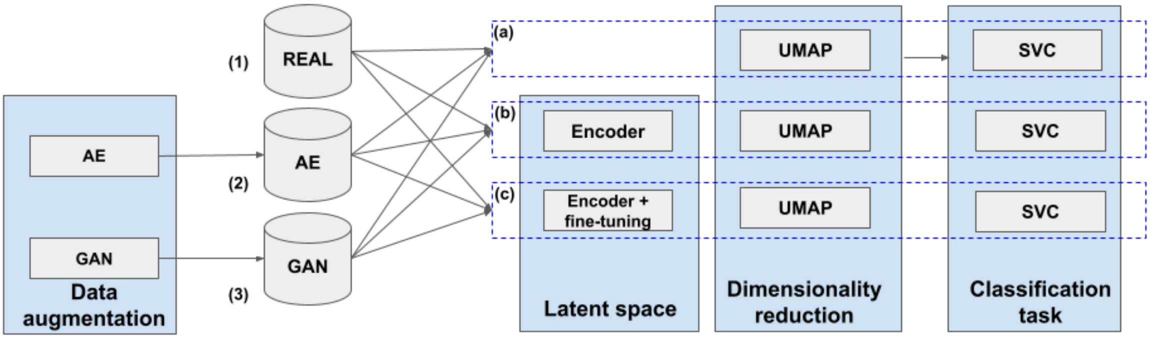


FIGURE 4. Outline of the proposed methodology.

during the initial AE phase to enhance performance in a related task.

E. STEP 3: NONLINEAR DIMENSION REDUCTION OF LATENT SPACE

The third step in our approach involves using UMAP for dimensional reduction, as instrumental in revealing complex patterns and trajectories hidden within the ECG beat samples. By implementing this technique, our goal is to discern intricate structures and relationships within the data and provide valuable insights for further analysis and interpretation in ECG signal analysis.

F. STEP 4: CLASSIFICATION PERFORMANCE AND REPRESENTATION

In the final step, we focus on the classification task, following the dimensional reduction where the distinct classes have been effectively separated. Utilizing the reduced-dimensional representation of the data, which enhances class separability, we aim to achieve accurate and reliable classification of the ECG signals. This step facilitates us to assess our methodology in addressing noise reduction and inter-class imbalance challenges, ultimately contributing to improved diagnostic AUC in ECG signal analysis.

G. EXPERIMENTAL FRAMEWORK

The experimental framework comprises distinct blocks implemented using different datasets, as seen in Figure 4. First, these datasets used the original data shown in frame (1). Second, it uses the data obtained by the autoencoder-based augmentation (AE), as shown in frame (2). Finally, it uses the data obtained using data augmentation with GAN in frame (3). Each block encompasses three sub-experiments. The initial sub-experiment involves only the dimensionality reduction coupled with a classifier task, as seen in frame (a). In the subsequent sub-experiment, we incorporate latent space, the dimensionality reduction, and the classification task, as shown in frame (b). The final sub-experiment introduces latent space with fine-tuning, followed by the implementation of the dimensionality reduction. Finally, the classification task is applied, as we can see in frame (c).

In totality, we conducted nine sub-experiments for each dataset to meticulously compare each component of the proposed methodology. The combination of these represents the nine described sub-experiments to reference each experiment in the experimentation section straightforwardly. For instance, sub-experiment 3C would indicate utilizing latent space with fine-tuning, dimensionality reduction, and classification tasks for the dataset augmented with GAN-generated data.

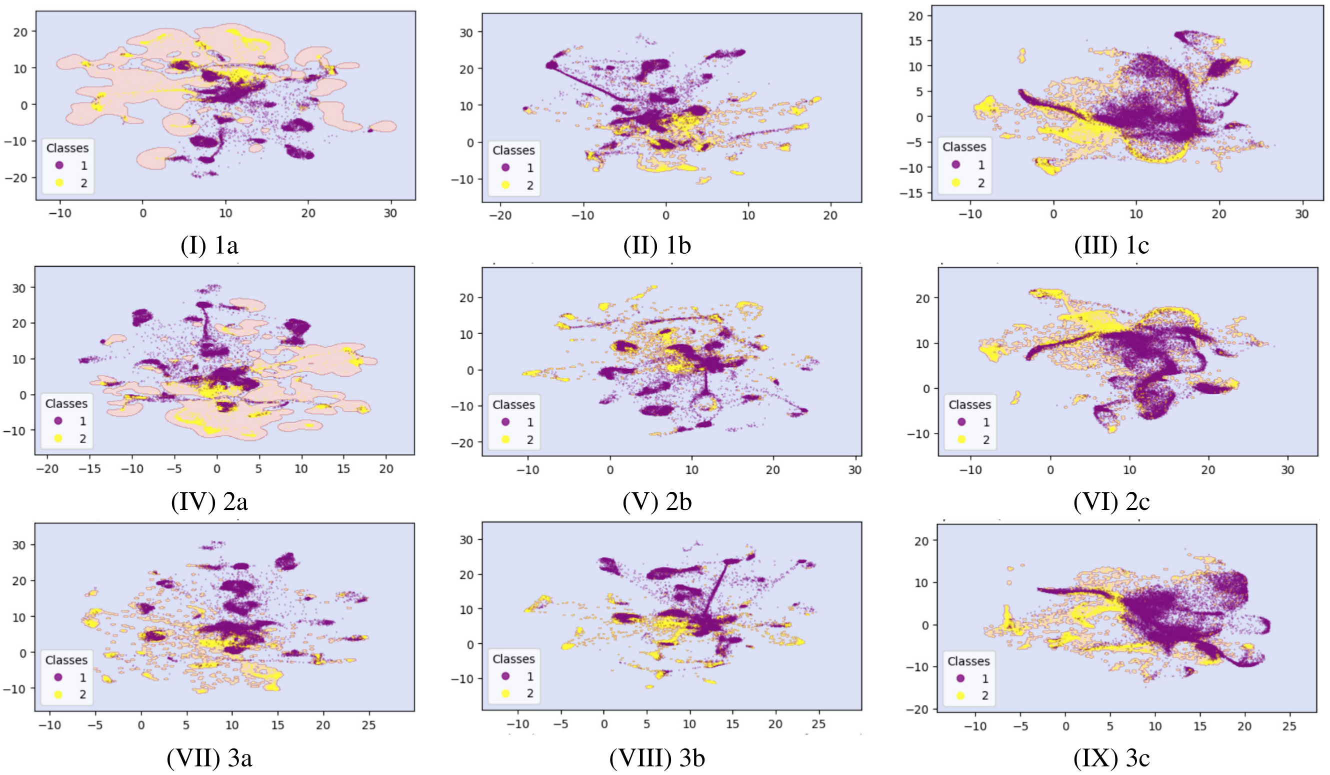
V. EXPERIMENTS AND RESULTS

In the upcoming subsections, we present a detailed account of the experimental setup and evaluation methodologies employed to assess the effectiveness of our proposed methodology. The criteria for label regrouping are outlined in [1], where the adoption of biclass regrouping is justified from a clinical standpoint. Through a series of experiments, we investigate the impact of transforming the input space into the latent space, evaluate the noise reduction achieved using AE techniques, and explore the compensatory effect of generating synthetic samples using AEs or GANs. The experimental results shed light on the potential of our methodology to simultaneously address the challenges of beating noise pollution and inter-class imbalance in ECG signal analysis.

A. BICLASS DATASET

1) EXPERIMENT 1: COMPLETE DATA

In the first experiment, we utilized the entire biclass dataset from the MIT-BIH database, carefully partitioning the patients into distinct train and test sets [1]. This meticulous segregation is pivotal to ensure that no patient data overlaps between the sets, thereby preventing potential information leakage and confounding effects during model evaluation. The primary objective was rigorously assessing the algorithm's performance and robustness across a diverse range of patient samples and cardiac signal patterns. Furthermore, this explicit separation into train and test subsets is crucial for evaluating the algorithm's ability to generalize effectively to unseen data, thereby enhancing the



(I) 1a

(II) 1b

(III) 1c

(IV) 2a

(V) 2b

(VI) 2c

(VII) 3a

(VIII) 3b

(IX) 3c

FIGURE 5. Representation of experiment one for biclass data set. The decision region corresponding to Class 2 is depicted in pink. The labels beneath each image convey the following meanings: the first number is associated with the three different approaches mentioned previously and represented in Figure 4, (1) Refers to the entire dataset; (2) Refers to the entire dataset excluding five patients; (3) Refers for the five patients. (I) Original data; (II) latent space applied to real data; (III) Fine-tuning applied to real space; (IV) AE data; (V) latent space applied to AE data; (VI) Fine-tuning applied to AE data; (VII) GAN data; (VIII) latent space applied to GAN data; (IX) Fine-tuning applied to GAN data.

reliability and validity of our findings and contributing to the advancement of cardiac signal analysis methodologies.

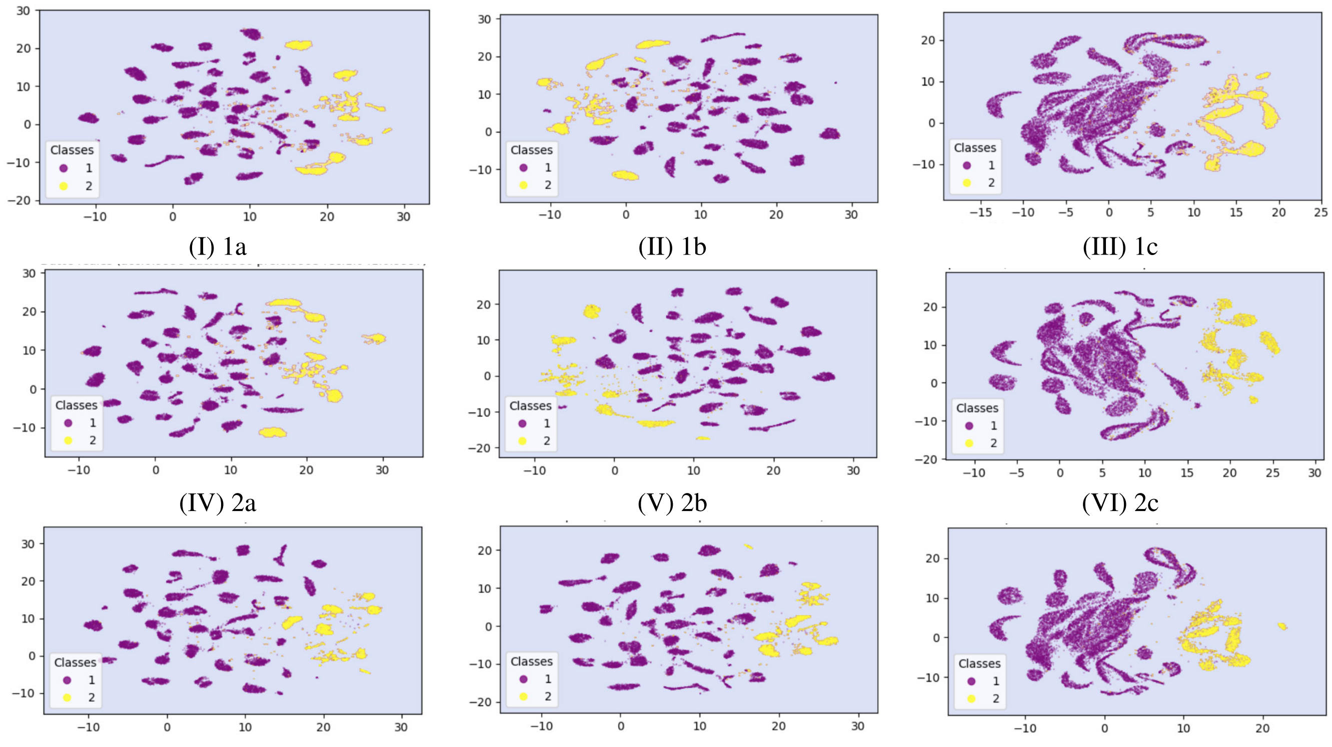
To ensure a comprehensive representation of beats from all classes in both the train and test datasets, we specifically selected patients with beats associated with less frequent classes, achieving an even distribution across both datasets. This strategic selection allows for the inclusion of samples from both classes in each dataset, reinforcing the robustness and reliability of our experimental results. As a consequence of this approach, the test set comprises 78% of beats from class 1 and 22% from class 2, the minority class. This distribution offers a detailed view of the test set composition, underscoring the imbalanced nature of the class distribution and highlighting how our methodology addresses these imbalances.

The experiment is structured into the three blocks mentioned previously in the methodology and represented in Figure 4. In the first block, the input data consists of the original database, continuing with the second block, where the input data involves an augmented database generated using AE. Finally, in the third block, the input data also involves an augmented database created using GANs.

For this experiment, we carried out nine detailed sub-experiments as previously outlined. These are visually represented in Figure 5, which provides a comprehensive illustration of the distribution patterns and boundaries of both classes. This visual representation offers insights into the spatial arrangement and segregation of data points from each class.

The first row in Figure 4 (1) corresponds to the original database, the second row (2) to the augmented database using AE, and the third row (3) to the augmented database using GAN. In the first row, three sub-experiments (paths 1a, 1b, and 1c) involve various applications of UMAP and classifiers, depicted in Figures 5 (I to III). Similarly, the second and third rows follow the same structure with paths 2a, 2b, 2c (Figures IV to VI) and 3a, 3b, 3c (VII to IX), respectively.

A significant improvement in outcomes is noted when applying augmentation techniques to the underrepresented minority class. Particularly, integrating AE and GANs shows pronounced enhancements in AUC metrics. According to Table 4, the AUC increased from 0.9332 without data augmentation to 0.9520 with GAN implementation, marking a 2% increase. These improvements are especially notable in augmentation margins, validating the efficacy of our approach. Furthermore, implementing latent spaces using fine-tuning also shows improvement; we achieve an AUC of 0.9545 with fine-tuning compared to 0.9332 without data augmentation. Regarding accuracy, a rise from 0.9609 without data augmentation to 0.9755 is observed when employing



(I) 1a (II) 1b (III) 1c

(IV) 2a (V) 2b (VI) 2c

(VII) 3a (VIII) 3b (IX) 3c

FIGURE 6. Representation of experiment two for biclass dataset, the first number is associated with the three different approaches mentioned previously and represented in Figure 4, (1) Refers to the entire dataset; (2) Refers to the entire dataset excluding five patients; (3) Refers for the five patients. (I) Original data; (II) latent space applied to real data; (III) Fine-tuning applied to real space; (IV) AE data; (V) latent space applied to AE data; (VI) Fine-tuning applied to AE data; (VII) GAN data; (VIII) latent space applied to GAN data; (IX) Fine-tuning applied to GAN data.

ENC + CLF with GAN for data augmentation, reflecting an increase of 1.5%. These results underscore the impact of data augmentation techniques in enhancing the analytical outcomes of our study.

The enhancements achieved through our methodology are evident in the confusion matrix presented in Figure 7. This matrix, highlighting the 2-class experiment (supraventricular (SV) vs. ventricular and other (V) origin), demonstrates clear improvements in both classes with increased true positives and reduced misclassifications. Particularly, we observed a notable reduction in false SV detections classified as ventricular (from 509 to 309) and false V detections classified as SV (from 1296 to 872). However, an area of concern is the decrease in true positives for the V class. Figure 5 illustrates this effect; in the first scenario (Panel (I)), the sample distribution appears more fragmented, whereas data augmentation results in a more compact and defined distribution for both classes (Panel (IX)).

Furthermore, implementing latent space via two techniques, encoder, and encoder with additional fine-tuning, shows discernible improvement. We also noticed similar results in AUC when using GAN for latent space and fine-tuning. To provide a broader context, we compared these findings with results obtained using the SMOTE data augmentation technique. As shown in Table 3, the outcomes from experiments employing AE and GAN surpass those achieved with SMOTE, indicating a higher efficacy of the methods proposed here regarding AUC, accuracy, and F1 metrics.

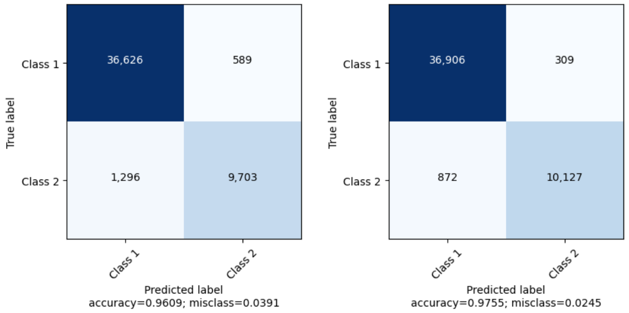


FIGURE 7. Comparison of confusion matrices: one without data augmentation and another employing ENC + CLF with GAN for data augmentation.

2) EXPERIMENT 2: ENTIRE DATASET EXCLUDING FIVE RANDOM PATIENTS

In this experiment, we employed the complete biclass dataset from the MIT-BIH database but excluded five patients. The beats of these five patients were used in the next experiment. In this selected dataset, the beats of the patients were not subject to any division into distinct train and test sets. As a

TABLE 3. Baseline results for AUC, accuracy, and F1 scores derived from biclass and multiclass experiments utilizing SMOTE for data augmentation.

	BICLASS			MULTICLASS		
	AUC	Acc	F1	AUC	Acc	F1
Exp1	0,8107	0,9056	0,9416	0,8057	0,9374	0,9281
Exp2	0,8936	0,9748	0,9862	0,9501	0,9874	0.9866
Exp3	0,8558	0,9526	0,9725	0,8235	0,8892	0,8750
Exp3'	0,9155	0,9760	0,9866	0,8017	0,9376	0,9294

result, there is a possibility of having beats from the same patients in both the test and train sets, although there are no identical beats in the test and training sets. This intermixing of patient data in the two sets can give rise to overfitting issues, where the model becomes excessively tuned to the training data features and could fail to generalize better to unseen instances.

We conducted the nine sub-experiments again using this modification of the database. The first row pertains to the first data block, represented in Figure 4 (1) using only the actual dataset. The subsequent row pertains to the second data block. It is represented in Figure 4 (2) using data augmentation with AE, and the third row pertains to the data third block, which is represented in Figure 4 (3) using data augmentation with GAN. Figure 6 represented the three sub-experiment corresponding to each row. The first row displays the results in Panels (I-II-III), the second row displays the results in Panels (IV-V-VI), and the last row displays the experiments in Panels (VII-VIII-IX). A salient observation is made during this experiment regarding certain patients whose cardiac beats can be observed in both training and test datasets, indicating the presence of overfitting. This becomes evident when examining Figure 6, where a distinct contrast between its characteristics and notably exhibits a more pronounced and refined separation of classes due to overfitting, in comparison to the distribution shown in the previous experiment, depicted in Figure 5, where the separation between the two classes was present, but not so extremely evident. As seen in Figure 4, the results obtained in this second experiment seem to exhibit significant improvements across all three blocks, with 100 % achieved for both AE and GAN, as seen in Table 4. When mixing beats from the same patient in train and test, the performance results become overly optimistic due to patient overfitting.

3) EXPERIMENTS 3 AND 3′: DATA FROM THE FIVE SELECTED PATIENTS

In these experiments, we aimed to evaluate the performance of the model on new patient data, categorized as *healthy* and *abnormal* beats. Experiment 3 involved five randomly selected patients, while Experiment 3' included five specifically chosen patients based on the prevalence of *abnormal* heartbeats. In Experiment 3, 9% abnormal beats are attributed to class 2, while in Experiment 3', 37% of beats are attributed to this class. This distinction stems from the deliberate selection of patients in Experiment 3' to include a higher proportion of *abnormal* beats.

Table 4 reveals contrasting results between these scenarios. Without data augmentation, Experiment 3 shows an increase in AUC (0.9629) compared to Experiment 1 (0.9332). However, Experiment 3' demonstrates a slightly lower AUC (0.9262) than Experiment 1. Notably, the use of GAN-generated data results in improved AUC for both Experiments 3 (0.9906) and 3' (0.9541), surpassing Experiment 1 (0.9562). This improvement, particularly in minority class detection, highlights the efficacy of GAN-generated data in enhancing classification performance. Furthermore, a distinction between Experiments 3 and 3' is evident, with the former achieving a higher AUC, mainly due to class imbalance favoring the majority class in the case of *healthy* patients. Overall, as seen in Table 4, all biclass dataset experiments exhibit improved AUC with data augmentation. Nevertheless, it is noteworthy that the difference in AUC between using AE and GAN is not significant. Additionally, our general assessment shows a substantial enhancement in AUC when implementing fine-tuning.

B. MULTICLASS DATASET

In our experiments, as outlined in Section IV, we encountered a notable observation regarding certain patients whose cardiac beats appeared in both the training and test datasets. This situation raised concerns about potential overfitting. The evidence of overfitting becomes more apparent when comparing Figure 6 with Figure 5. In Figure 6, there is a more pronounced and refined separation of classes, suggesting overfitting, compared to the more blended distribution in Figure 5. In response to this finding, we adapted our experiments within the biclass dataset framework, introducing a significant variation by incorporating the multiclass dataset. This modification allowed us to validate the efficacy of our proposed methodology thoroughly. By conducting these comprehensive experiments, we aimed to gain insights into the performance of our approach across different classification scenarios.

1) EXPERIMENT 1: COMPLETE DATA

In this experiment, we applied the same procedure as in Section V-A, but this time utilizing the multiclass dataset. Our approach mirrored that of the biclass section, providing a consistent and comparable evaluation framework. This

TABLE 4. AUC, accuracy and F1 results from biclass experiments.

	Augmentation	REAL			ENC			ENC + CLF		
		AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
Exp1	-	0,9332	0,9609	0,9749	0,9341	0,9626	0,9758	0,9545	0,9754	0,9842
	AE	0,9383	0,9639	0,9768	0,9447	0,9659	0,9781	0,9556	0,9763	0,9848
	GAN	**0,9520**	0,9727	0,9825	**0,9558**	0,9702	0,9807	**0,9562**	0,9755	0,9843
Exp2	-	0,9986	0,9996	0,9997	0,9990	0,9996	0,9998	0,9996	0,9998	0,9999
	AE	0,9994	0,9998	0,9999	**1,0000**	1,0000	1,0000	1,0000	1,0000	1,0000
	GAN	0,9998	0,9999	1,0000	1,0000	1,0000	1,0000	0,9998	0,9999	1,0000
Exp3	-	**0,9629**	0,9916	0,9954	0,9643	0,9925	0,9959	0,9786	0,9959	0,9978
	AE	0,9640	0,9919	0,9956	0,9677	0,9936	0,9965	0,9883	0,9978	0,9988
	GAN	0,9643	0,9914	0,9953	0,9715	0,9943	0,9969	**0,9906**	0,9982	0,9990
Exp3'	-	0,9262	0,9444	0,9569	0,9313	0,9479	0,9597	0,9495	0,9609	0,9464
	AE	0,9275	0,9448	0,9575	0,9445	0,9568	0,9663	0,9511	0,9620	0,9702
	GAN	0,9403	0,9544	0,9646	0,9451	0,9579	0,9672	**0,9541**	0,9637	0,9715

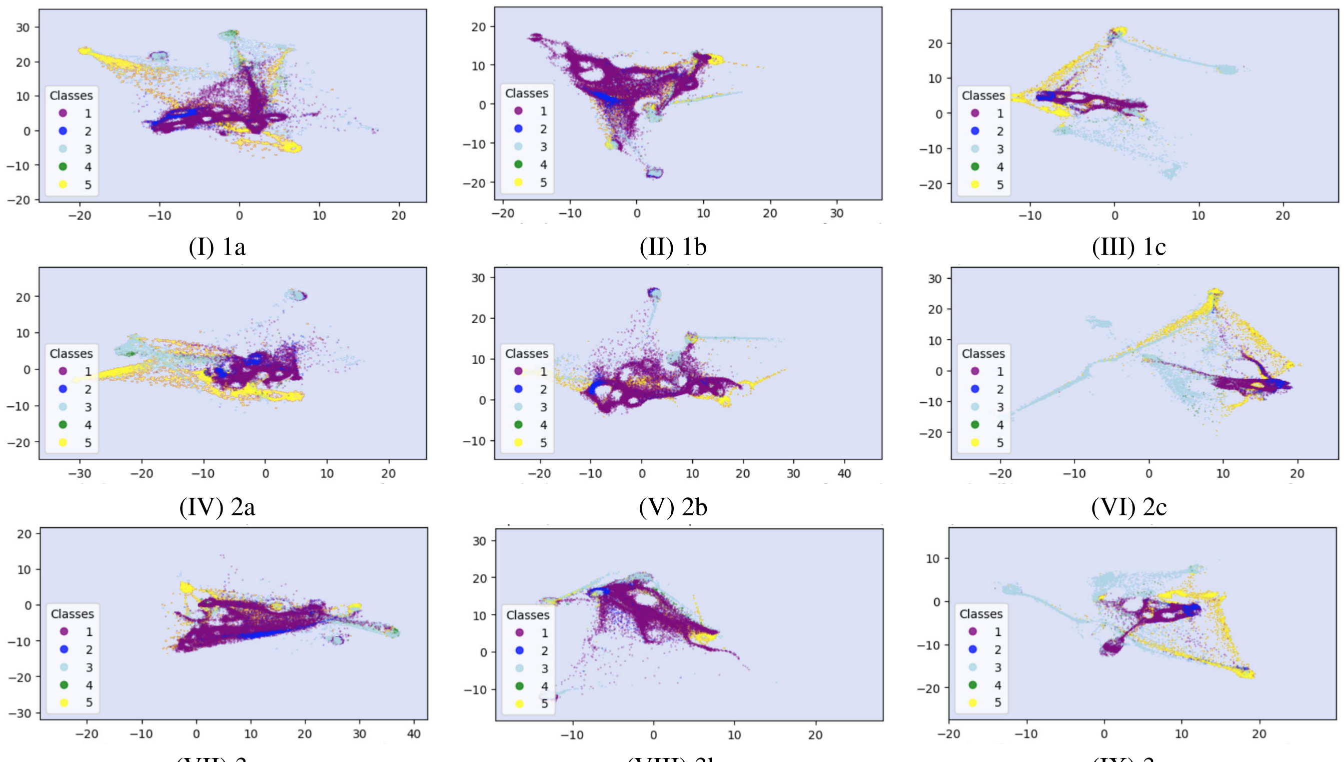
TABLE 5. AUC, accuracy and F1 results from multiclass experiments.

	Augmentation	REAL			ENC			ENC + CLF		
		AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1
Exp1	-	0,9020	0,9298	0,9275	0,9031	0,9155	0,91442	0,9179	0,9483	0,9408
	AE	**0,9223**	0,9427	0,9413	0,9394	0,9522	0,9515	0,9264	0,9539	0,9478
	GAN	0,9169	0,9325	0,9316	0,9239	0,9444	0,9416	0,9275	0,9539	0,9488
Exp2	-	0,9990	0,9940	0,9994	0,9993	0,9996	0,9996	0,9994	0,9996	0,9996
	AE	0,9992	0,9995	0,9995	**0,9998**	0,9998	0,9998	0,9996	0,9998	0,9998
	GAN	0,9994	0,9996	0,9996	0,9998	0,9999	0,9999	0,9998	0,9998	0,998
Exp3	-	0,9478	0,9860	0,9852	0,9563	0,9882	0,9876	0,9787	0,9928	0,9927
	AE	0,9574	0,9893	0,9886	0,9576	0,9896	0,9890	0,9858	0,9959	0,9958
	GAN	**0,961**	0,9901	0,9896	0,9694	0,9927	0,9923	0,9884	0,9963	0,9963
Exp3'	-	0,8692	0,8699	0,8410	0,8776	0,8829	0,8619	0,9153	0,8949	0,8681
	AE	0,8717	0,8810	0,8497	0,9228	0,9072	0,8854	0,9125	0,8945	0,8802
	GAN	0,8759	0,8786	0,8643	0,9048	0,9027	0,8817	0,9181	0,9066	0,8761

adaptation allowed us to delve into the effects of our methodology amidst a more complex and varied class distribution. Similar to Experiment 1 in the biclass dataset, we ensured that beats from all classes were adequately represented in both the train and test sets. Consequently, the distribution in the test set comprised 72% of beats belonging to class 1, with 5%, 10%, 1%, and 12% allocated to classes 2, 3, 4, and 5, respectively.

We replicated the nine sub-experiments using this modified database. These experiments followed the same structure detailed previously. The first row corresponds to the original data approach, as shown in Figure 4 (1). The second row relates to the augmented data using AE (Figure 4 (2)), and the third row involves the data augmented with GAN (Figure 4 (3)). The results of these sub-experiments are depicted in Figure 8, with the first, second, and last rows represented in panels (I-II-III), (IV-V-VI), and (VII-VIII-IX), respectively. This figure highlights the distinguishable manifolds of class one, represented in purple, across all sub-experiments. In contrast, class five, depicted in yellow and representing beats with various pathological characteristics, exhibits significant diversity, displaying up to three distinct community structures. This divergence indicates that the underlying data distribution of class five warrants further investigation and analysis.

Implementing data augmentation techniques enhances our model performance, as reflected in Table 5. For instance, we observed an increase in AUC from 0.9020 to 0.9223 when using actual AE data. This improvement highlights the effectiveness of data augmentation in boosting the model's accuracy and robustness. In terms of accuracy, the model achieved a significant increase, moving from 0.9298 without data augmentation to 0.9539 with the implementation of ENC + CLF using GAN, representing a nearly 3% improvement. The enhancements are further evident in the confusion matrix (Figure 9), which demonstrates improvements across all classes, characterized by increased true positives and a substantial reduction in misclassifications. Notably, classes 1, 3, 4, and 5 show significant improvements, although class 2 experiences a slight decline.



(I) 1a

(II) 1b

(III) 1c

(IV) 2a

(V) 2b

(VI) 2c

(VII) 3a

(VIII) 3b

(IX) 3c

FIGURE 8. Representation of experiment one for the multiclass dataset, the first number is associated with the three different approaches mentioned previously and represented in Figure 4, (1) Refers to the entire dataset; (2) Refers to the entire dataset excluding five patients; (3) Refers for the five patients. (I) Real data; (II) latent space applied to real data; (III) Fine-tuning applied to real space; (IV) AE data; (V) latent space applied to AE data; (VI) Fine-tuning applied to AE data; (VII) GAN data; (VIII) latent space applied to GAN data; (IX) Fine-tuning applied to GAN data.

Additional improvements are observed when implementing latent space through two distinct techniques: encoder and encoder with fine-tuning. Similar results in AUC were noted when using GAN for latent space and fine-tuning. Consistent with the findings in the biclass dataset, our results demonstrate the robustness of our methodology. Table 3 reveals that our AE and GAN experiments consistently outperform those using SMOTE regarding AUC, accuracy, and F1 metrics when comparing our approach to the SMOTE data augmentation technique.

2) EXPERIMENT 2: ENTIRE DATASET EXCLUDING FIVE RANDOM PATIENTS

In this experiment, we adapted the methodology of Experiment 2 from the previous section to the comprehensive multiclass dataset from the MIT-BIH database, this time excluding five random patients. This modification assessed the impact of excluding specific patient data on the performance of our model.

We once again conducted the nine sub-experiments using this adjusted database. The first row of results corresponds to the original data approach, as depicted in Figure 4 (1). The second row involves the augmented data using AE (Figure 4 (2)), and the third row includes the data augmented with GAN (Figure 4 (3)). The outcomes of these sub-experiments are illustrated in Figure 10, with the first row shown in panels (I-II-III), the second row in (IV-V-VI), and the last row in (VII-VIII-IX).

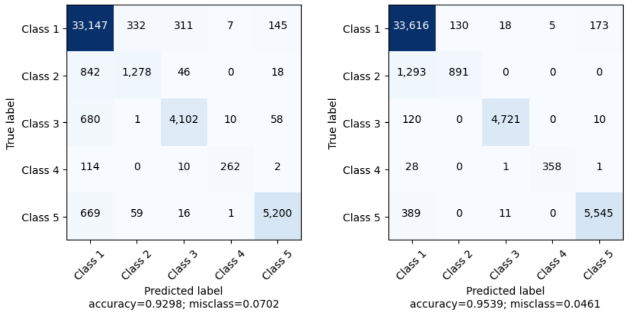
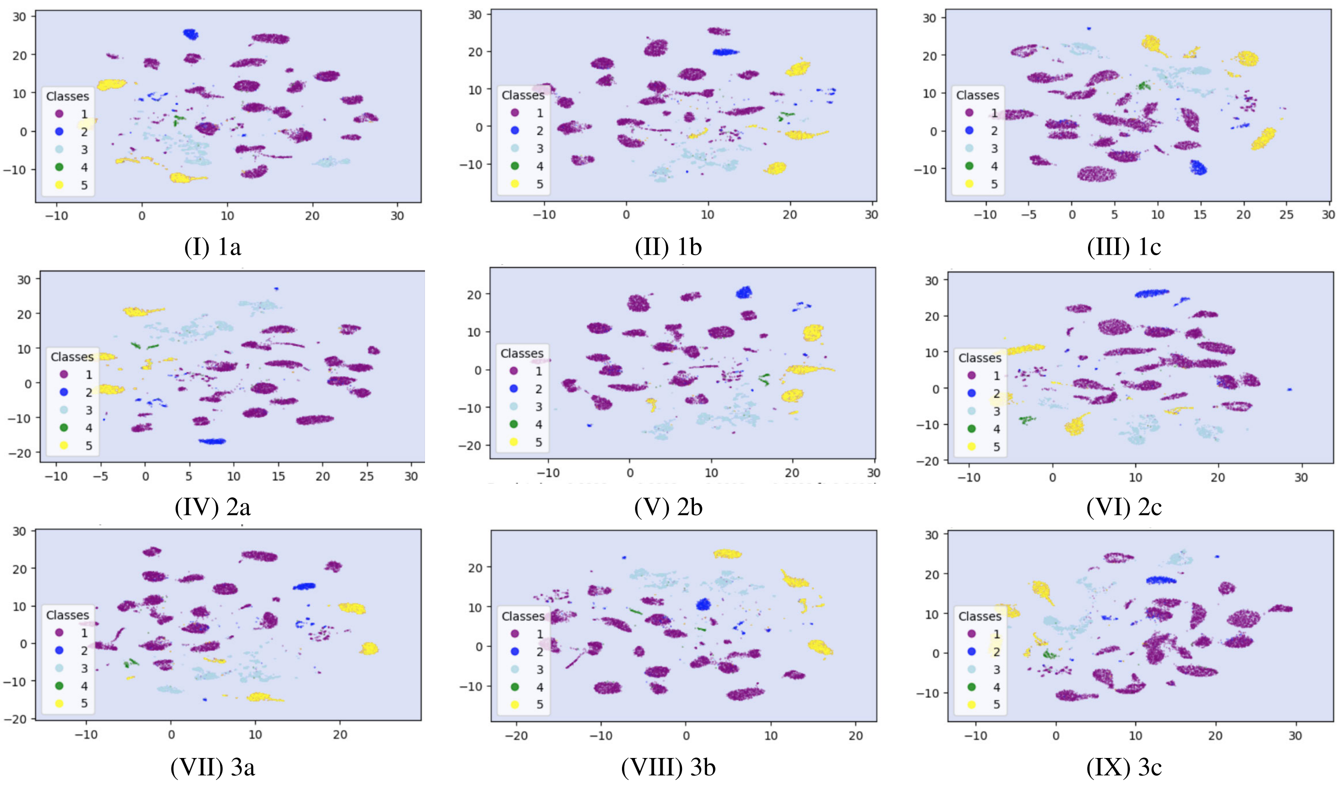


FIGURE 9. Comparison of confusion matrix in multiclass experiment: one without data augmentation and another employing ENC + CLF with GAN for data augmentation.

A key observation in this experiment pertains to certain patients whose cardiac beats appeared in both the test and training datasets, raising concerns about potential overfitting. This issue becomes evident upon examining Figure 10, which reveals a stark contrast in class separation compared to Figure 8. Notably, Figure 10 shows a more pronounced and refined class separation, indicative of overfitting, as opposed



(I) 1a (II) 1b (III) 1c

(IV) 2a (V) 2b (VI) 2c

(VII) 3a (VIII) 3b (IX) 3c

FIGURE 10. Representation of experiment two for the multiclass dataset, the first number is associated with the three different approaches mentioned previously and represented in Figure 4, where: (1) Refers to the entire dataset; (2) Refers to the entire dataset excluding five patients; (3) Refers to the five patients. (I) Real data; (II) latent space applied to real data; (III) Fine-tuning applied to real space; (IV) AE data; (V) latent space applied to AE data; (VI) Fine-tuning applied to AE data; (VII) GAN data; (VIII) latent space applied to GAN data; (IX) Fine-tuning applied to GAN data.

to the more blended distribution in Figure 8. The influence of overfitting is further reflected in the exceptionally high AUC values, reaching up to 0.9998, as shown in Table 5.

3) EXPERIMENT 3: DATA FROM FIVE SELECTED PATIENTS

Similar to Experiment 3 with the biclass dataset, the primary focus of this experiment was to assess the impact of introducing previously unseen patients to the model. Using the multiclass dataset, Experiment 3 involved *healthy vs abnormal* heartbeats predominantly from class 1, and the five patients were randomly selected. In contrast, Experiment 3’ included five patients with mainly abnormal heartbeats, with the majority of beats belonging to the other classes. We followed the same criteria as in Experiment 1 to ensure representation from all classes, resulting in a varied distribution of beats across different classes. In both Experiments 3 and 3’, we observed an increase in AUC when implementing data augmentation, including all three blocks: actual data, latent space, and fine-tuning. In Experiment 3, the AUC increased from 0.8759 to 0.961 when using GAN techniques for actual data, as indicated in Table 5.

The confusion matrices, depicted in Figure 9, present two cases: one using all patients and the other after applying fine-tuning to GAN data augmentation (shown in Figure 8, panels (I) and (IX)). Initially, we notice that N beats are often confused with S, V, and U beats. SV beats are mainly mistaken for N beats, with less confusion with other classes. However, when compared with GAN augmentation, the overall performance improves significantly. The confusion of N beats with other classes reduces, particularly for SV and V. SV beats are no longer mistaken for V, F, or U, though there is an increase in confusion with N beats. There’s a marked increase in true positives for V beats and a notable reduction in their confusion with N beats. F and U beats also improve, with increased true detections and fewer errors. The second case, with fine-tuning and GAN augmentation, substantially improves class separation. Each class exhibits more defined domains, with N beats concentrated in specific regions and clear transitions among subregions. SV beats show two distinct clusters, one in contact with N beats and another separate from them. U beats are spread across a vast domain, differentiating from N and V beats. V beats, varied in their anatomical origin, are generally distinct from N beats, except in some regions. These results illustrate the effectiveness of a well-adapted classifier and GAN data augmentation in refining class distinctions and improving classification accuracy.

VI. DISCUSSION AND CONCLUSION

The primary aim of this paper has been to underscore the pivotal role of latent space techniques in generating synthetic data, particularly in mitigating the adverse effects of inter-class imbalance and overfitting in ECG signal analysis. Our experiments across both biclass and multiclass datasets have demonstrated marked improvements in task performance through data augmentation techniques employing AE and GAN. For instance, in the biclass dataset, GAN implementation led to enhanced classification results across various scenarios, including real data, latent space utilization, and fine-tuning. It is important to note that the use of SVC in our study was mainly to provide a quantitative assessment. We focused on mapping activations to a reduced observation space and quantifying class separability and confusion using SVC, chosen for its definitive minimum that aids in metric clarity. Experiment 1 observed an AUC improvement from 0.9332 without augmentation to 0.9520 with augmentation, a 2% increase as depicted in Figure 4. Generally, both biclass and multiclass datasets showed higher AUC values with the application of AE and GAN augmentation techniques. The data augmentation techniques applied to the biclass dataset yielded positive outcomes across multiple sub-experiments.

A significant issue highlighted in our study is overfitting, which is evident in graphical representations and metrics. This was especially pronounced in experiments reaching AUC values close to 100%. Overfitting manifested in quantitative metrics and visual depictions underscored the necessity of adequately separating training and testing datasets to build models that effectively generalize to new, unseen data. Generalizing from a small dataset comprising only five patients presents notable challenges, primarily due to the limited sample size. This limitation was evident in both biclass and multiclass dataset experiments. For instance, in Experiment 2, we achieved AUC values of 100% with data augmentation in the biclass dataset and 0.9998 in the multiclass dataset. While indicative of model accuracy, these high values also point to potential overfitting issues.

When introducing previously unseen patients in Experiment 3 for both biclass and multiclass datasets, we observed an AUC increase compared to Experiment 1. For instance, the AUC for Experiment 3 was 0.9629, a 3% increase from 0.9332 in Experiment 1, as shown in Tables 4 and 5. Notably, this improvement occurred without applying data augmentation techniques to real data.

In conclusion, our research introduces a novel approach to address the dual challenge of beat noise interference and class imbalance in ECG signal analysis. We systematically improved the classification task by transforming the input space into a latent space and employing data augmentation with AE and GANs. Our findings highlight the effectiveness of innovative strategies in enhancing the accuracy and robustness of ECG signal analysis, contributing to more reliable cardiac condition diagnoses. It is crucial to separate patients into distinct training and testing sets to reduce overfitting, emphasizing the importance of robust dataset partitioning in machine learning. Our study illustrates the power of interdisciplinary approaches in advancing medical research and practice, paving the way for future innovations in this field.

REFERENCES

[1] A. Vadillo-Valderrama, R. Goya-Esteban, R. P. Caulier-Cisterna, A. García-Alberola, and J. L. Rojo-Álvarez, “Differential beat accuracy for ECG family classification using machine learning,” *IEEE Access*, vol. 10, pp. 129362–129381, 2022.

[2] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, “A survey on ECG analysis,” *Biomed. Signal Process. Control*, vol. 43, pp. 216–235, May 2018.

[3] A. Galli, F. Ambrosini, and F. Lombardi, “Holter monitoring and loop recorders: From research to clinical practice,” *Arrhythmia Electrophysiol. Rev.*, vol. 5, no. 2, p. 136, 2016.

[4] M. A. Serhani, H. T. E. Kassabi, H. Ismail, and A. N. Navaz, “ECG monitoring systems: Review, architecture, processes, and key challenges,” *Sensors*, vol. 20, no. 6, p. 1796, Mar. 2020.

[5] J. Xue and L. Yu, “Applications of machine learning in ambulatory ECG,” *Hearts*, vol. 2, no. 4, pp. 473–494, 2021.

[6] M. Hassaballah, Y. M. Wazery, I. E. Ibrahim, and A. Farag, “ECG heartbeat classification using machine learning and metaheuristic optimization for smart healthcare systems,” *Bioengineering*, vol. 10, no. 4, p. 429, Mar. 2023.

[7] J. Huang, B. Chen, B. Yao, and W. He, “ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network,” *IEEE Access*, vol. 7, pp. 92871–92880, 2019.

[8] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, “Deep learning for ECG Arrhythmia detection and classification: An overview of progress for period 2017–2023,” *FrontiersPhysiol.*, vol. 14, pp. 1–20, Sep. 2023.

[9] J. Chaquet-Ulldemolins, F. J. Gimeno-Blanes, S. Moral-Rubio, S. Muñoz-Romero, and J.-L. Rojo-Álvarez, “On the black-box challenge for fraud detection using machine learning (II): Nonlinear analysis through interpretable autoencoders,” *Appl. Sci.*, vol. 12, no. 8, p. 3856, 2022.

[10] K. Babaei, Z. Chen, and T. Maul, “Data augmentation by autoencoders for unsupervised anomaly detection,” 2019, *arXiv:1912.13384*.

[11] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” 2020, *arXiv:2003.05991*.

[12] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, “Application of deep learning for heartbeats detection using ECG signal-analysis and review,” *Comput. Biol. Med.*, vol. 129, no. 129, pp. 103–113, Jun. 2020.

[13] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, “Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review,” *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103726.

[14] H. W. Loh, C. P. Ooi, S. L. Oh, P. D. Barua, Y. R. Tan, F. Molinari, S. March, U. R. Acharya, and D. S. S. Fung, “Deep neural network technique for automated detection of ADHD and CD using ECG signal,” *Comput. Methods Programs Biomed.*, vol. 241, Nov. 2023, Art. no. 107775.

[15] V. Avula, K. C. Wu, and R. T. Carrick, “Clinical applications, methodology, and scientific reporting of electrocardiogram deep-learning models,” *JACC: Adv.*, vol. 2, no. 10, Dec. 2023, Art. no. 100686.

[16] A. Subasi, S. Dogan, and T. Tuncer, “A novel automated tower graph based ECG signal classification method with hexadecimal local adaptive binary pattern and deep learning,” *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 2, pp. 711–725, Feb. 2023.

[17] X. Liu, H. Wang, Z. Li, and L. Qin, “Deep learning in ECG diagnosis: A review,” *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107187.

[18] L. E. Bouny, M. Khalil, and A. Adib, “ECG heartbeat classification based on multi-scale wavelet convolutional neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3212–3216.

[19] C. T. Wei, M.-E. Hsieh, C.-L. Liu, and V. S. Tseng, “Contrastive heartbeats: Contrastive learning for self-supervised ECG representation and phenotyping,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1126–1130.

[20] V. V. Kuznetsov, V. A. Moskalenko, D. V. Gribanov, and N. Y. Zolotykh, “Interpretable feature generation in ECG using a variational autoencoder,” *Frontiers Genet.*, vol. 12, Apr. 2021, Art. no. 638191.

[21] M. M. Rahman, M. W. Rivolta, F. Badilini, and R. Sassi, “A systematic survey of data augmentation of ECG signals for AI applications,” *Sensors*, vol. 23, no. 11, p. 5237, May 2023.

[22] J. An, R. E. Gregg, and S. Borhani, "Effective data augmentation, filters, and automation techniques for automatic 12-lead ECG classification using deep residual neural networks," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 1283–1287.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[24] M. F. Safdar, P. Pałka, R. M. Nowak, and A. A. Faresi, "A novel data augmentation approach for enhancement of ECG signal classification," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105114.

[25] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[26] Z. Wang, S. Stavrakis, and B. Yao, "Hierarchical deep learning with generative adversarial network for automatic cardiac diagnosis from ECG signals," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106641.

[27] J. Petch, S. Di, and W. Nelson, "Opening the black box: The promise and limitations of explainable machine learning in cardiology," *Can. J. Cardiol.*, vol. 38, no. 2, pp. 204–213, Feb. 2022.

[28] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić, "What makes classification trees comprehensible?" *Expert Syst. Appl.*, vol. 62, pp. 333–346, Nov. 2016.

[29] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proc. Nat. Conf. Artif. Intell.*, 2004.

[30] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Comput. Intell.*, vol. 37, no. 4, pp. 1633–1650, Nov. 2021.

[31] C. P. B. Oñate, F.-M. M. Meseguer, E. V. Carrera, J. J. S. Mu noz, A. G. Alberola, and J. L. R. Álvarez, "Different ventricular fibrillation types in low-dimensional latent spaces," *Sensors*, vol. 23, no. 5, p. 2527, Feb. 2023.

[32] T. Lin and H. Zha, "Riemannian manifold learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 796–809, May 2008.

[33] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2020, *arXiv:1802.03426*.

[34] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci.*, vol. 505, pp. 32–64, Dec. 2019.

[35] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, pp. 1–22, Jan. 2023.

[36] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.

[37] T. Sainburg, L. McInnes, and T. Q. Gentner, "Parametric UMAP embeddings for representation and semisupervised learning," *Neural Comput.*, vol. 33, no. 11, pp. 2881–2907, 2021.

[38] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study," in *Proc. Int. Conf. Image Signal Process.* Abu Dhabi, United Arab Emirates: Springer, 2020, pp. 317–325.

[39] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May/Jun. 2001.

[40] G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: A web-based resource for the study of physiologic signals," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 70–75, May/Jun. 2001.

[41] *MIT-BIH Arrhythmia Database*. Accessed: Aug. 2023. [Online]. Available: https://www.physionet.org/physiobank/database/mitdb/

[42] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomed. Signal Process. Control*, vol. 8, no. 5, pp. 437–448, Sep. 2013.

[43] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, and T. Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals," *Comput. Methods Programs Biomed.*, vol. 127, pp. 52–63, Apr. 2016.

[44] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ECG classification," *Comput. Biol. Med.*, vol. 99, pp. 53–62, Aug. 2018.

[45] L. Sörnmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, vol. 8. New York, NY, USA: Academic, 2005.

[46] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.

ALBA VADILLO-VALDERRAMA was born in Madrid, Spain. She received the B.Sc. degree in telecommunication engineering from Rey Juan Carlos University, Madrid, Spain, in 2016, and the master's degree in visual analytics and big data from the International University of La Rioja, in 2019. Her research interests include machine learning, pattern recognition, artificial neural networks, and deep learning.

JACOBO CHAQUET-ULLDEMOLINS received the B.Sc. degree in computer engineering and the M.B.A. degree from Universidad Politécnica de Madrid, Spain, in 2008 and 2012, respectively, the M.Sc. degree in artificial intelligence from Universidad Internacional Menéndez Pelayo, in 2019, and the Ph.D. degree from Rey Juan Carlos University, in 2022. He is currently the Senior Manager with BBVA. With more than 13 years of professional experience related to IT consulting in financial services. His research interest includes machine learning algorithms applied to financial services.

REBECA GOYA-ESTEBAN received the B.Sc. degree in telecommunication engineering from the Carlos III University of Madrid, Spain, in 2006, the M.Sc. degree in biomedical engineering from the University of Porto, Portugal, in 2008, and the Ph.D. degree from Rey Juan Carlos University, Madrid, Spain, in 2014. She is currently an Associate Professor with Rey Juan Carlos University. She has coauthored 29 international articles and has contributed to more than 40 conference proceedings. Her main research interests include time series analysis, bioelectrical signal processing, and statistical learning.

RAÚL P. CAULIER-CISTERNA received the Ph.D. degree in multimedia and communications from Universidad Rey Juan Carlos and the Carlos III University of Madrid, Spain, in 2018, with a specialization in bio-engineering and biomedical signal processing. He was a Postdoctoral Researcher with the Biomedical Imaging Center, Pontificia Universidad Católica de Chile. He is currently an Academician with Universidad Tecnológica Metropolitana and a collaborator in research with the Millennium Institute for Intelligent Healthcare Engineering–iHealth, Santiago, Chile, and the Rielo Institute for Integral Development. His research interests include machine learning methods for signal and image processing, artificial intelligence algorithms, non-invasive healthcare equipment, and non-invasive near-infrared spectroscopy in the spinal cord.

JUAN JOSÉ SÁNCHEZ-MUÑOZ received the M.D. and Ph.D. degrees from Universidad de Murcia, Spain, in 1987 and 2000, respectively. Since 1995, he has been a Cardiac Electrophysiologist with the Hospital General Universitario Virgen de la Arrixaca and an Assistant Professor of medicine with Universidad de Murcia, where he is currently a Consultant Cardiologist with the Arrhythmia Unit of Cardiac Electrophysiology. He has coauthored more than 40 scientific articles and more than 50 communications in cardiac electrophysiology. His research interests include arrhythmia mechanisms and cardiac signal processing in ventricular fibrillation.

ARCADI GARCÍA-ALBEROLA received the M.D. and Ph.D. degrees from Universitat de Valencia, Burjassot, Spain, in 1982 and 1991, respectively. Since 1993, he has been a Cardiologist and a Professor of medicine with the Hospital General Universitario Virgen de la Arrixaca and Universidad de Murcia, Murcia, Spain, where he is currently the Director of the Arrhythmia Unit of Cardiac Electrophysiology. He has coauthored more than 130 scientific articles and more than 60 communications in cardiac electrophysiology. His research interests include repolarization analysis, arrhythmia mechanisms, and cardiac signal processing.

JOSÉ LUIS ROJO-ÁLVAREZ (Senior Member, IEEE) received the B.Sc. degree in telecommunication engineering from Universidade de Vigo, in 1996, and the Ph.D. degree in telecommunication engineering from Universidad Politécnica de Madrid, in 2000. He is currently a Professor with the Department of Signal Theory and Communications, Rey Juan Carlos University, Spain. He has coauthored more than 140 international articles and contributed to more than 180 conference proceedings. His research interests include statistical learning methods for signal and image processing, arrhythmia mechanisms, robust signal processing methods, and data science for arrhythmias.