

ECG Arrhythmia Classification Using 1D Convolutional Neural Network with Data Augmentation: A Comparative Study

Aliya Naeem^{1*}, Hiyam Hatem¹

¹Department of computer science, College of Computer science & Information Technology, University of Sumer, Thi-Qar, Iraq.

* Email address of the Corresponding Author Aliyanaeem@hs.uos.edu.iq

Phone number: 07748533353

Abstract

Electrocardiogram signals (ECG) are an essential component of automated detection of heart rhythm disorders. Despite the widespread use of 1D-CNN NSO models, most current studies evaluate performance using balanced datasets or clinically rare cardiac disorders. A clear methodological gap in understanding how deep learning models are acting under realism between groups. This study addresses this gap by providing a precise systemic comparison between a 1D-CNN core-trained model without any imbalance strategy, and a developed model that combines an increase in signal-level data (Added Gaussian noise, and displacement). Circular and amplitude change) with the use of the category weight loss function, using the highly unbalanced MIT-BIH ARRHYTHMIA database, which contains 40,937 heartbeats distributed over five classes according to Aami standard. A stratification was applied at 80/20 pulse-level to keep the distribution of categories between the training and testing groups. The 1D-CNN architecture was selected due to its computational efficiency and proven suitability for temporal physiological signals processing. Quantitative results showed a significant improvement in performance, with a total accuracy increased from 45.39% to 83.71%, and the Macro F1-Score coefficient improved from 0.3413 to 0.5379. However, it is important to note that the performance of rare clinical-important groups remains a major challenge; Class A (early atrial contractility) achieved a value of F1-Score close to 0.20. A very rare training example. Grad-Cam's analysis of idiomy's ability to learn meaningful and distinctive physiological representations of each category across all five categories of CPR has also been confirmed.

Keywords: ECG; convolutional neural networks; data augmentation; heart rhythm disorders data imbalance; MIT-BIH; Grad-Cam

تصنيف اضطرابات نظم القلب باستخدام الشبكة العصبية التلافيفية أحادية البعد

مع زيادة البيانات: دراسة مقارنة

علياء نعيم^{1*}، هيام حاتم¹

قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة سومر، ذي قار، العراق

الخلاصة

تُعدّ إشارات تخطيط القلب الكهربائي (ECG) عنصراً أساسياً في الكشف الآلي عن اضطرابات نظم القلب. تقدّم هذه الدراسة نموذجاً قائماً على الشبكة العصبية التلافيفية أحادية البعد (1D-CNN) لتصنيف خمس فئات من اضطرابات نظم القلب باستخدام قاعدة بيانات MIT-BIH. تتكوّن مجموعة البيانات من 40,937 نبضة قلبية موزعة على خمس فئات وفق معيار AAMI، مع وجود اختلال شديد في التوازن بين هذه الفئات. ولمعالجة هذه المشكلة، تم اعتماد استراتيجية مزدوجة تجمع بين زيادة البيانات إضافة ضوضاء Gaussian، والإزاحة الدائرية، وتغيير السعة واستخدام أوزان للفئات داخل دالة الخسارة. أظهرت النتائج تحسناً ملحوظاً في الأداء، حيث ارتفعت الدقة من 45.39% إلى 83.71%، كما تحسّن معامل F1 الكلي من 0.24 إلى 0.54. بالإضافة إلى ذلك، أظهر تحليل Grad-CAM أن النموذج يركّز على منطقة QRS، مما يشير إلى أن النموذج يتعلم خصائص فسيولوجية ذات معنى.

1. Introduction

Cardiovascular disease (CVDs) is the leading cause of death globally, accounting for more than 30% of the total annual deaths [1]. According to the World Health Organization, it is estimated that about 17.9 million people lose their lives annually due to these diseases, and it is expected that this number will continue to rise in the coming decades if effective and timely diagnostic interventions are not available. Electrocardiography (ECG) is the gold standard for non-surgical heart monitoring, providing valuable real-time information about the electrical activity of the heart. Its low cost, wide spread, and non-interventional nature make it the most practical cardiac screening tool in both clinical and mobile environments [2]. Heart rhythm disorders are known as impaired rhythm or heart rate, and may range from simple palpitations to dangerous, life-threatening conditions such as ventricular fibrillation. Early and accurate detection of these disorders is critical to guide timely medical intervention, prevent sudden cardiac death, and improve patient outcomes. Despite its clinical importance, the manual ECG signaling is a stressful and time-consuming process, and is subjective and highly dependent on the expertise of a specialist. Inter-observer variability is a well-known problem, especially when cardiologists review long hours of continuous Holter recordings [2]. These limitations have led to increased interest in developing automated ECG signaling systems that are able to accurately identify pathological patterns.

Traditional machine learning methods, such as SVM support machines, decision trees, random forests, and the KNN algorithm, have been widely used in the classification of ECG signals. However, these methods are based mainly on the extraction of manually engineered features, which require great experience in the field, and are difficult to generalize through various recording devices, various patient groups, or different measurement conditions [3]. The performance of these systems is closely related to the quality and completeness of the extracted properties, which makes them weak when facing unexpected changes in the signal. In contrast, deep learning, especially convolutional neural networks (CNNs), has made a radical shift in biomedical signal analysis, thanks to its ability to learn hierarchical representations of properties directly from raw data [4].

Single-dimensional (1D-CNNs) bypass networks are particularly suitable for electrocardiogram processing, as sliding filters can capture local time patterns and morphological features of the signal, such as P, QRS and T waves, without the need for Manual extraction of properties [5]. This comprehensive approach to end-to-end is not only to reduce geometric voltages, but also enable the detection of precise patterns that the human expert may not be able to easily notice. However, the Class Imbalance issue remains a huge and influential challenge in classifying ECG signals. In realistic clinical databases, the number of natural impulses significantly exceeds the number of rare arrhythmia classes. This unbalanced distribution leads the models to bias towards the larger category, resulting in total precision that appears apparently high, but it hides almost complete failure to detect rare cases of clinical importance [6]. Therefore, addressing this problem by increasing data, rebalancing, or cost-sensitive learning is not just a technical improvement, but a real clinical necessity. Although there is an increase in the number of studies in this area, there is still a fundamental research gap; The majority of current ECG signaling studies either evaluate models on synthetic balanced datasets, excludes completely rare categories from experimental design, or only report overall accuracy, a misleading measure when class distribution. Highly skewed Furthermore, only a few studies measure the impact of each category of data-increasing techniques, especially for rare and clinically high-risk arrhythmia classes such as early ventricular contraction (Class V) and early atrial contraction (Class A).

This gap reduces the clinical significance of the published results, and makes it difficult to assess whether the proposed models are actually able to detect dangerous cardio events in realistic environments. This study offers two main contributions. First, a systematic comparison is made between a basic 1D-CNN-trained model without data-increasing techniques, and a developed model that combines a signal-level augmentation and the use of Classes weights within the loss function. This comparison contributes to directly measuring the impact of data increase on both overall accuracy and model performance for each category. Second, the Grad-CAM technology (gradient-dependent activation map) is used to provide a clear and clinically understandable explanation about the time zones in the ECG signal on which the model relies on classification decisions. As for the rest of the research, the data set and methodology are presented in the third section, the experimental and comparative results are presented in the fourth section, the results are discussed in the fifth section, and the research is concluded in the sixth section.

2. Related work

The application of deep learning in electrocardiogram (ECG) signaling has undergone tremendous growth over the past decade, driven by advancements in computer hardware and the increase in the availability of large and explained ECG databases. A systematic review study conducted by LIU et al., analyzed 368 well-defined studies, and found that CNN-based models accounted for 58.7% of the total deep learning methods used. in the classification of heart rhythm disorders. It is noteworthy that 61% of these studies were based exclusively on the MIT-BIH Arrhythmia database in evaluation, while only 28% dealt with data-increasing strategies to address rare Classes, indicating a systematic gap. clear in this area. Feyisa [5] proposes a model of the 1D-CNN, a three-blocking frail block, which included the entire Batch Normalization, Max-Pooling, and Dropout. Competitive performance in the classification of 4 Classes of cardiac arrhythmia using the MIT-BIH base. The study showed that Batch Normalization not only contributes to the stability of the training process by reducing the internal change in distribution (Internal Covariate Shift), but also helps in accelerating the convergence of transactions, enabling the model to reach high accuracy through Epochs. On the other hand, Rasti et al. [8] developed a deliberately simple 1D-CNN structure of the ECG binary rating, and proved that simple structure could achieve good accuracy with a significant reduction in the complexity of the model and inference time, which is particularly important. For applications on wearable devices with limited resources. Kiranyaz et al. [9] provided a 1D-CNN model dedicated to each patient, with an accuracy of 98.90% by resetting a general model using individual patient data. Although this approach highlights the possibilities of deep personal learning, it is based on the availability of sufficient titled data for each patient, and it uses an intra-patient assessment that may overestimate the model's ability to generalize to new patients. Similarly, Ullah and others reported a resolution of 97.38% to 99.02% on the MIT-BIH base using an intra-patient assessment, confirming the need for stricter patient-based assessment protocols to measure the ability to measure Real clinical application. The category imbalance issue has been addressed through several different strategies.

Ding et al. [10] studied the integration of SMOTE technology with CNN models across 24 unbalanced datasets, and achieved a resolution of 99.08%, and showed that smote excels the random increase continues by generating new synthetic samples in the properties space, rather than just repeating existing samples. Shchetinin and Glushkova [11] showed that the combination of rebalancing and deep learning leads to a remarkable improvement in the detection of rare systemic disorders in unbalanced ECG data. On the structure level, the

hybrid models that integrate CNN with the BIGRU and multi-headed attention mechanisms have shown additional improvements through their ability to capture local morphological features and long-term dependencies in multiple ECG recordings. Poles [12]. Ansari and others [4] review for the period between 2017 and 2023 confirmed that convolutional networks (CNNs) are still the most widely used structure in ECG signal analysis due to their high-capacity extraction ability. However, the increased interest in Transformer structures and hybrid models such as CNN-LSTM reflects researchers' awareness that local characteristics alone may not be sufficient to represent the long-term time patterns that characterize some types of system disorders. In the field of model interpretation, Selvaraju et al. [16] introduced the Grad-Cam technology that enables visually interpreting deep network decisions based on gradients. Kovalchuk et al. [15] applied explicable artificial intelligence techniques to classify ECG disorders, and proved that providing visual explanations for model decisions is a prerequisite for doctors' acceptance and adoption of these systems in applications. Clinical.

3. Materials and Methods

3.1 Dataset

The MIT-BIH Arrhythmia is the most widely used reference standard for electrocardiogram (ECG) signaling research, containing 48 recordings from 47 subjects, and it was recorded with a sampling rate of 360 Hz and a 11-bit resolution within a ± 10 mV domain [13]. These recordings were collected at MIT Biomedical Engineering Laboratory, and each recording was about 30 minutes long. Annotation of the heartbeat was performed by two or more cardiologists independently, providing an accurate and high-quality reference (Ground Truth) to evaluate the algorithms. According to the AAMI EC57 standard, the titled pulse types are grouped into five main clinical classes: Normal beat (N), Left Bundle Branch Block(L), Right Bundle Branch Block(R), Atrial Premature contraction(A), and Premature Ventricular Contraction(V).

The processed data used in this study consisted of 40,937 heartbeat sections. Each ECG pulse was divided into a fixed-length window of 200 samples, centered around the location of the R, with the aim of capturing the complete P-QRS-T complex within each heart rate. With a sampling rate of 360Hz in the MIT-BIH database, a 200-sized sample covers approximately 0.55 seconds, a sufficient time to contain a complete heart cycle for most possible physiological heart rates. The length of this window is consistent with commonly adopted common practices in the eCG classification literature [5][9], and a balance between time coverage and computational efficiency. In addition, the division with a constant length ensures the input dimensions required for the CNN. Stratified split 80% was applied for training and 20% test, while maintaining the relative distribution of each category within the two groups, resulting in 32,757 training samples and 8,187 test samples. Table (1) displays the distribution of categories before the application of data increase techniques.

Table 1- Class distribution of the MIT-BIH dataset before augmentation

Class	Label	Description	Train Samples	Percentage
1	N	Normal Beat	25,060	76.52%
2	L	Left Bundle Branch Block	3,688	11.26%
3	R	Right Bundle Branch Block	2,941	8.98%
4	A	Atrial Premature Contraction	1,037	3.17%
5	V	Premature Ventricular Contraction	31	0.07%

3.2 Signal Preprocessing

Each raw ECG pulse underwent a two-stage standard pre-treatment chain before it was introduced into the neural network. In Phase I, a 4-degree Butterworth-degree band pass filter [0.5–40] is applied using the zero-phase Filtfilt function, with the aim of removing the baseline. Wander) resulting from breathing and patient movement (less than 0.5 Hz), as well as a reduction in the high frequency noise generated by muscle activity and the interference of electrical power lines (above 40 Hz). The use of the zero-phase filter ensures that there is no time-phase deformation that may result when using one-way filters, which is critical to maintaining the time-relationships between the various components of the signal, such as the P wave, the QRS complex, and the T wave. Using Z-Score independently on each heartbeat using the following equation: $x_{\text{norm}} = (x - \mu) / \sigma$ where μ and σ represent the arithmetic mean and the standard deviation of the signal window samples made up of 200 samples in a row. This type of normalization removes the differences in amplitude between patients and the recording channels, which may result from different electrodes' locations, body resistance, or gain settings during recording, ensuring that CNN receives inputs with homogeneous statistical properties. In order to avoid the occurrence of divisions on zero, a precautionary measure was adopted to set $\sigma = 1$ in cases where $\sigma < 1 \times 10^{-6}$, especially in flat or semi-fixed sections. Figure (1) shows representative examples of ECG signals before and after pretreatment of all five Classes, clearly showing the effect of removing baseline deflection and *amplitude* uniformity across various patterns of heart rhythm disorders.

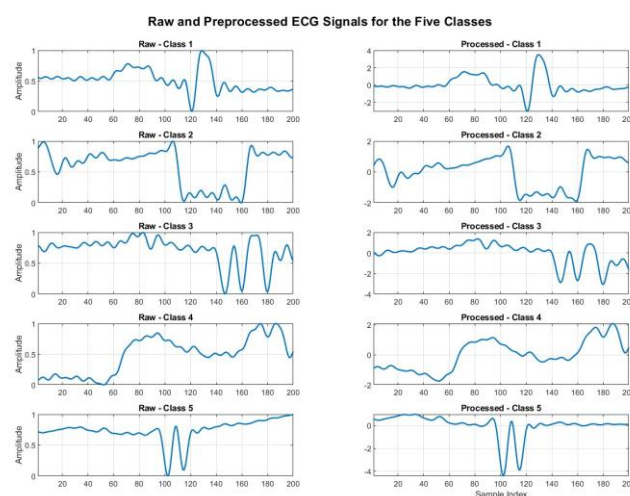


Figure -1 Raw (left column) and preprocessed (right column) ECG signals for the five AAMI classes after Butterworth bandpass filtering [0.5–40 Hz] and Z-score normalization. Preprocessing removes baseline wander and standardizes amplitude across all classes.

3.3 Data Augmentation Strategy

The MIT-BIH database shows a severe imbalance in the Classes, with the natural category (N) 76.52% of the training data, while the very rare category (V) is 0.07%, or exceeds 1,000:1. As shown in Table (1). If this problem is not addressed, any model trained in such unbalanced data will misleadingly high overall accuracy by predicting the largest class of most samples, with almost complete failure to detect rare but clinical heart rhythm disorders. To address this problem, a three-tech signal-level data increase strategy has been adopted, and has been applied exclusively to the least represented Classes (L, R, A, and V), with the aim of raising the samples in each category to 8,000 samples. In contrast, the normal category (N) was kept unchanged at 25,060 samples, in order to preserve its original properties. The three augmentation techniques were selected because they each introduce a distinct and physiologically motivated form of variability: electrode contact noise (Gaussian noise), temporal alignment uncertainty in R-peak detection (circular shift), and inter-patient amplitude variation (scaling). Together, they increase sample diversity while preserving the diagnostic morphology of each heartbeat. Importantly, these techniques operate directly in the raw signal domain without requiring distributional assumptions about the data, unlike SMOTE which interpolates in feature space and may generate unrealistic intermediate shapes for non-linear physiological signals. Similarly, GAN- or VAE-based synthesis, while potentially more diverse, requires substantially larger baseline datasets to train reliably — a condition not met for Class V (31 original samples). Future work will explore these generative approaches as additional training data becomes available. For each generating sample, one of the following data increase techniques is selected randomly and with equal probability:

(a) Adding Gaussian noises: White Gaussian noises were added to the original signal after normalization, with zero average. A standard deviation of 0.01. This process mimics the noise from the contact of the electrodes, the slight changes resulting from the patient's movement, as well as the quantization noise, which helps the model to learn the properties that are resistant to amplitude scaling instead of adapting to the properties of a specific recording noise.

(b) Circular Shift: the signal was removed in a circular random value in the field $[-5, +5]$ samples (equivalent to ± 13.9 ms at a frequency of 360). Hertz). This process simulates simple time deviations in determining the top of R or in the process of cutting the pulse, which are common in automated pulse extraction systems, and help make the model insensitive to the position of the pattern within the signal.

(c) Amplitude Scaling: The signal amplitude was multiplied by a random parameter within the domain $[0.95, 1.05]$ to simulate the physiological changes between the pulse, the gain differences between the channels, as well as the changes of the conditions. Registration. This field ($\pm 5\%$) was chosen to maintain the diagnostic form of the signal with the introduction of realistic diversity in amplitude. After applying each data increase, the Z-Score normalization was re-applied to the resulting signal to ensure that statistical properties were maintained before the training process. In addition, the weights of the Classes were calculated using the equation: $w_i = n_{\text{total}} / (n_{\text{classes}} \times n_i)$ and these weights were integrated into the loss function (Cross-Entropy), so that higher penalties were imposed on errors Classification of rare Classes during the backpropagation process, providing an additional effective layer of imbalance. The training group size after the data increased was 57,060 samples, an increase

of 74.23% compared to the original size of 32,750 samples. Table (2) shows the distribution of Classes after data increase, in addition to the calculated weights for each category.

Table 2- Class distribution after augmentation and computed class weights

Class	Before Aug.	After Aug.	Weight
1 (N)	25,060	25,060	0.4554
2 (L)	3,688	8,000	1.4265
3 (R)	2,941	8,000	1.4265
4 (A)	1,037	8,000	1.4265
5 (V)	31	8,000	1.4265
Total	32,757	57,060	—

3.4 CNN Architecture

The proposed single-dimensional neural network (1D-CNN) structure is designed to balance computational efficiency and the ability to extract properties effectively. One of the most important architectural decisions was the adoption of the Global Average Pooling (GAP) rather than the traditional style, which is based on the FLATEN, followed by the Fully Connected layers after the convolutional layers. The GAP layer reduces the spatial properties map to one value per channel by taking the average over the time dimension, resulting in a significant reduction in the number of the classification layer coefficients, and it acts as a strong Regularization mechanism to reduce the problem of hyper-adaptation. (overfitting). More importantly, GAP maintains the relationship between the properties map and time locations in the signal, and is a prerequisite for the application of interpretation techniques such as Grad-Cam. The full structure below is described and summarized in Table (3). Sequence Input Layer: It receives single-channel timelines with the lowest length of 200 samples, which represent the heartbeat windows extracted from the MIT-BIH recordings. First Convolutional Block: It consists of a CONV1D layer with a kernel size of 5 and 32 filters with SAME padding. The kernel size of 5 was deliberately chosen because at a sampling rate of 360 Hz, it covers a temporal window of approximately 13.9 ms — sufficient to distinguish the sharp transitions between adjacent P, Q, R, S, and T wave components without excessive parameter overhead.

This is followed by the Batch Normalization layer to install the properties distribution, then the RELU activation function to enter nonlinear, and finally the Maxpooling1D layer 2 and step 2 to halve the time-resolution while retaining the most important themes. Second Convolutional Block: Conv1D includes a 3-size core and 64 filters that work on a low-resolution signal to extract more accurate and abstract formal properties. It is followed by the Batch Normalization and the RELU function to ensure the stability of the training. Pooling is not used at this point to maintain sufficient time accuracy for the subsequent GAP layer. Global Average Pooling (GAP) layer compiles the 64-channel properties map over the remaining time dimension, producing a 64-interior constant vector that represents a comprehensive summary of a pulse window heart. This procedure is independent of the input length and functions as an intermediary between the syntax extractor and the classification head. Classification Head: It consists of a Fully Connected layer that contains 128 units and uses the ReLu activation function, with the aim of integrating the

extracted properties and converting them into a distinct and short representation. The Dropout layer is also included with a rate of 0.5, which randomly disables half of the units during the training process, with the aim of reducing the risk of overfitting and improving the model's generalization ability. In the end, the last Fully Connected layer consists of 5 units, followed by the SoftMax function to generate a probabilistic distribution across the five different categories.

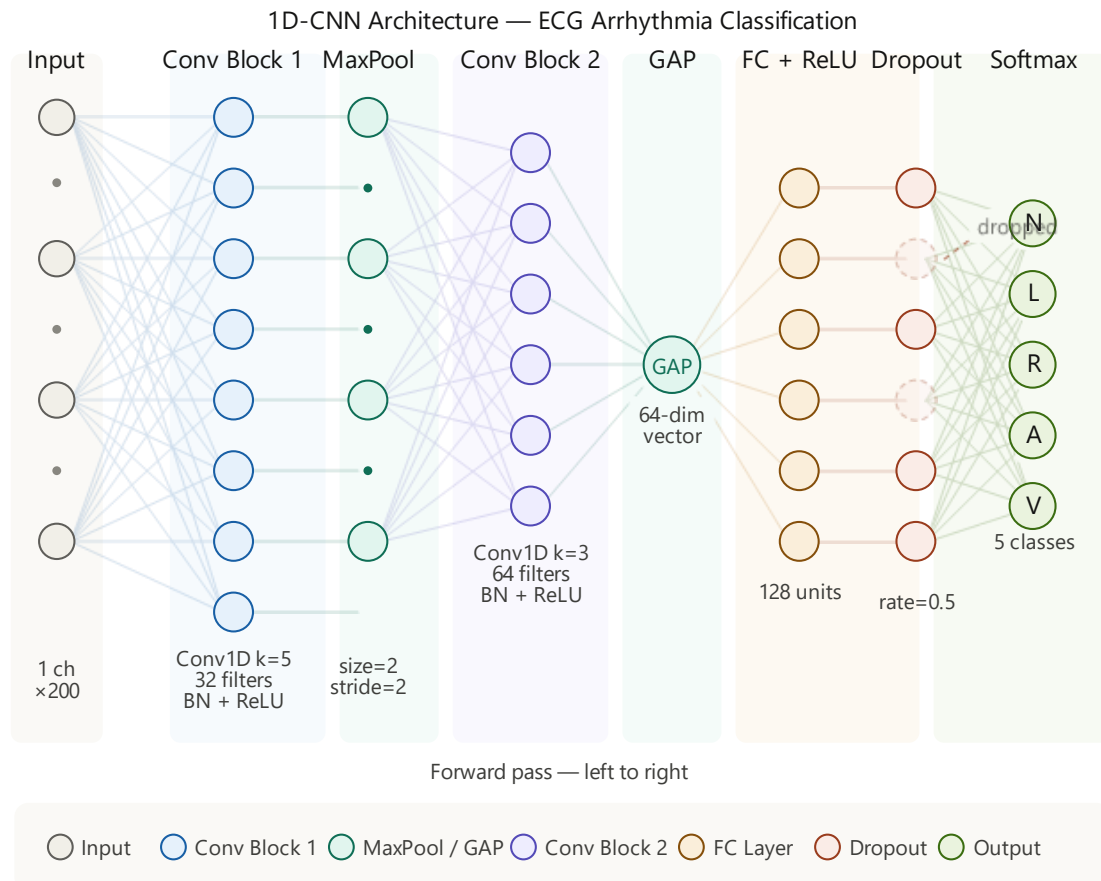


Figure -2 Proposed 1D-CNN Architecture for ECG Arrhythmia Classification

The 1D-CNN architecture proposed in this study is like a lightweight basic framework within a wider system to monitor and classify ECG signals that are still under development. The full framework is based on the integration of CNN-Bilstm hybrid deep learning models with an IoT-based system to instantly acquire ECG signals and remote monitoring. Therefore, the CNN architecture used in this work is designed to provide computational efficiency, conduct a precise comparative analysis, and ensure compatibility with future instant publishing scenarios.

3.5 Training Configuration

Both experimental cases — with and without data increased — were trained using the same hyperparameters settings to ensure a fair and accurate comparison. Adam's algorithm was chosen for its ability to adjust the learning rate in an adaptive manner, in addition to its strong practical performance in the ECG signal classification tasks. An initial learning rate of 1×10^{-3} was adopted throughout the training process without any reduction mechanism, as the initial experiments showed good stability in the convergence using this constant rate in 25

training epochs (EPOCH). The batch size (MINI-Batch) was set at 128 samples to balance the grading quality grading and memory efficiency. The training data was mixed at the beginning of each training course to reduce the bias resulting from the arrangement of the data. All experiments were carried out with one GPU. The model's performance was evaluated exclusively using the 8,187 held-out test set, which were not used during the training or transaction selection phase. Performance metrics were calculated for each category separately, including: Precision, Recall, Specificity and F1-Score, based on the Confusion Matrix. In addition, the macro-averaging averages were calculated for each scale, in order to provide an unbiased assessment of performance across all five Classes, giving equal weight to the rare and common Classes. This method of evaluation is particularly suitable for unbalanced data, as it prevents the Normal category from hiding the poor performance of the rare Classes of heart rhythm disorders.

It is important to note that no explicitly separate validation set was used in this study. This decision was made intentionally for the following reasons: (1) Hyperparameters were selected based on the standard settings widely adopted in the ECG signaling literature specifically the learning rate. 10×10^{-3} , batch size 128, and the number of training epochs 25 instead of repetitively adjusting depending on the test set. This prevents the impact of test data on forming the form. (2) The test group, comprising 8,187 samples, was strictly isolated and never used during training or during any hyperparameter tuning process

, ensuring an unbiased final evaluation. (3) The small size of the data for the rare categories, especially the V category, which contains only 31 original samples, means that additionally dividing the training group to customize an independent verification group will lead to a critical reduction in the number of samples of low-representation classes available for learning. We acknowledge that this is one of the determinants of the current study, and the future works will include the use of K-Fold-type cross verification to enhance the statistical durability and the ability to generalize the reported results. However, the absence of an independent verifiable group remains a systematic limitation in this current study.

3.6 Grad-CAM Explain ability

To gain a deeper understanding of the time zones in the ECG signal on which the CNN model is trained in classification decisions. Enhanced [16]. The Grad-CAM is a post-HOC Explainability-based gradient interpretation technology, generating a discriminatory positioning map for each category by taking advantage of the flowing gradient information from the predicted class. Returning to the properties maps in the last wrap layer inside the model. Formally, for a predicted class c , the importance weight α_k^c for the k -th feature map A^k of the final convolutional layer is computed as the global average of the gradient of the class activation score y^c with respect to each spatial location (i, j) of A^k : $\alpha_k^c = (1/Z) \sum_i \sum_j (\partial y^c / \partial A^k_{ij})$. The Grad-CAM localization map is then obtained as the ReLU of the weighted sum of all feature maps: $L^c = \text{ReLU}(\sum_k \alpha_k^c \cdot A^k)$. The ReLU operation retains only the features that positively contribute to the predicted class, suppressing negative activations that correspond to other classes. The resulting 1D importance vector is bilinearly upsampled to the original 200-sample signal length and normalized to the range [0,1] for visualization as a heatmap overlaid on the input ECG waveform.

4. Experimental Results and Comparison

4.1 Training Convergence

The training dynamics in both experimental cases show a clear variation in learning behavior. The Augmented model showed a smooth and stable development in the learning process, with the accuracy of the batch (Mini-Batch Accuracy) starting from 28.12% in the first iteration, and gradually increased to about 85%. At the 20th training course (Epoch 20). In parallel, the loss function (LOSS) decreased unilaterally from 1.66 at the start of the training to about 0.44 by the end of the twenty-fifth session (a total of 11,125 repetitions and a batch size of 128). The absence of significant pitfalls in the value of the loss during training indicates that the enhanced data provide balanced gradient signals, supporting a stable improvement process. In contrast, the uninterrupted model showed clearly unstable behavior, as the training was characterized by sharp fluctuations in the loss value, which sometimes reached 5.88, with unstable training dynamics and a final test accuracy of only 45.39%, significantly below the majority-class baseline of 76.52%. This is reflected in the Gradient Imbalance, which is known in unbalanced data environments, where Normal's large numerical dominance leads to this class's control over the direction of gradation during training, preventing the network. From learning the discriminatory properties of the rare Classes of heart arrhythmia. The final accuracy of the 45.39% unprocessed model on the test set was significantly worse than the majority-class baseline (76.52%), confirming almost complete failure of the model in the classification of rare Classes.

4.2 Overall Performance Comparison

Table (4) presents the overall performance measures based on macro-averaged data for both experimental conditions, using an independent test group of 8,187 samples. The baseline CNN model without data augmentation achieved an overall accuracy of 45.39%, indicating poor performance under conditions of severe class imbalance. This result was significantly lower than the baseline based on the largest class, confirming the model's failure to learn the discriminatory characteristics of underrepresented classes. After applying the proposed data augmentation strategy with class weights, the CNN model showed a significant improvement in performance, achieving an accuracy of 83.71%, representing an absolute increase of 38.32 percentage points. Furthermore, all Macro Metrics have experienced consistent improvement. Macro Precision has risen from 0.3685 to 0.5169, and Macro Recall has risen from 0.5200 to 0.6442, and Macro Specificity has risen. 0.8734 to 0.9407. Macro F1-Score also improved from 0.3413 to 0.5379. The Macro False Negative Rate ($FNR = 1 - \text{Recall}$) has also decreased from 0.4800 to 0.3558, indicating that the reinforced model is missing a far fewer number of systemic disorders. heart in general. This value is of special clinical importance; In cardiac control applications, the lack of detection of a serious systemic disorder, such as premature ventricular contraction, is much more harmful than the firing of a false alarm. For the high clinical V-Class, the FNR rate remained high at 0.50 (that is, 3 of 6 test samples were lost), while the FNR rate of Class A was about 0.69 (that is, 179 were lost 260 test samples), confirming the need for additional targeted improvements for these two low-acting categories. This consistent improvement across all assessment measures indicates that the proposed data increase strategy has effectively improved the model's ability to generalize and detect underrepresented categories. Unlike the base model, which was biased towards the dominant category, the enhanced model demonstrated a more balanced rating performance across all categories. These findings confirm that data increase is not just optional improvement, but an essential component of addressing the severe imbalance in ECG signaling heart rate disorders. A comparison was also made with the latest advanced methods (state-of-the-art methods) on the MIT-BIH database in Table (5) under Section 4.3.

Table 4- Compare the overall performance of the CNN model before and after the application of the data increase

Metric	CNN Without Augmentation	CNN With Augmentation	Improvement
Accuracy (%)	45.39	83.71	+38.32
Macro Precision	0.3685	0.5169	+0.1484
Macro Recall	0.5200	0.6442	+0.1242
Macro Specificity	0.8734	0.9407	+0.0673
Macro F1-Score	0.3413	0.5379	+0.1966
Training Epochs	25	25	—
Training Samples	32,757	57,060	+74.23%
Macro FNR (1–Recall)	0.4800	0.3558	–0.1242

4.3 Comparison with State-of-the-Art Methods

Table (5) compares the proposed augmented 1D-CNN model with representative recent studies on the MIT-BIH Arrhythmia Database. There are important methodological differences between comparative studies; Some studies are based on intra-patient split, which leads to an amplification of accuracy as a result of including the same patient data in each of the training and testing groups. Split), which provides a more conservative and realistic appreciation of the model's ability to generalize. In addition, several previous studies exclude rare categories such as A and V from the evaluation process, which leads to artificial improvement in the reported macro Metrics. These differences are noted in the table to allow fair interpretation of the results. The proposed model achieves competitive performance for a lightweight architecture, while explicitly retaining all five AAMI classes — including the clinically critical but rare Classes A and V — in the evaluation. Furthermore, the proposed 1D-CNN architecture represents a lightweight baseline component within a broader ongoing research framework that integrates a hybrid CNN–BiLSTM model with an IoT-based real-time cardiac monitoring system. In the complete framework, ECG signals are acquired through wearable sensors and processed automatically for arrhythmia classification, while critical abnormal cases trigger real-time alert notifications to a physician mobile application for remote medical monitoring and rapid clinical response.

Table 5- Comparison with representative state-of-the-art studies on MIT-BIH Arrhythmia Database. N/R = Not Reported. Note: intra-patient splits yield higher accuracy due to patient overlap between train and test sets, whereas Stratified beat-wise splits provide a more conservative and clinically realistic evaluation. The performance gap between this work and intra-patient studies is expected and reflects the stricter evaluation protocol adopted here.

Study	Method	Classes	Accuracy (%)	Macro F1	Split Type	Imbalance Strategy
Kiranyaz et al. [9] (2016)	1D-CNN (patient-specific)	5 (AAMI)	98.90	N/R	Intra-patient	Patient-specific fine-tuning
Ding et al. [10] (2023)	SMOTE + CNN	5 (AAMI)	99.08	N/R	Intra-patient	SMOTE oversampling
Feyisa [5] (2023)	1D-CNN (3 blocks)	4	97.30	N/R	Intra-patient	BatchNorm + Dropout
Bai et al. [12] (2024)	CNN-BiGRU	5	99.30	0.9820	Intra-	Class weighting +

Study	Method	Classes	Accuracy (%)	Macro F1	Split Type	Imbalance Strategy
This Work (2026)	+ Attention 1D-CNN + Augmentation + Class Weights	(AAMI) 5 (AAMI)	83.71	0.5379	patient Stratified beat-wise	hybrid model Signal augmentation + class weighting

4.4 Confusion Matrix Analysis

The CNN's augmented paradigm analysis reveals the detailed behavior of the classification process between the different Classes, which stands behind the overall performance metrics. Class N (natural pulse) showed clear control over the main diameter of the confusion matrix, with 5,224 samples correctly out of 6,264 test samples. However, some normal beats were misrapped in the L, R and A classes, indicating a formal similarity between some natural and abnormal impulses. Class L (left beam blockade) achieved a high number of correct ratings, with 841 samples correctly rated out of 922 test samples, reflecting the effectiveness of the data increase strategy in improving the representation of low-representation Classes. Likewise, the R (right beam blocking) class achieved an average rating performance, with 550 correct ratings out of 735 samples, although there was still some confusion between it and the neighboring Classes. Class A (early atrial contraction) showed relatively poor performance, with only 81 samples being correctly rated out of 260 test samples. A large number of these samples were incorrectly classified as normal impulses, reflecting the formal similarity between early atrial impulses and the normal patterns of the ECG signal. For class V (early ventricular contraction), the rarest category in the data set, it achieved a valid classification of only 3 samples out of 6 test samples. It is important to explicitly acknowledge that drawing statistically reliable conclusions from only 6 test samples is not feasible, and the reported metrics for this class should be interpreted with caution. Although the performance of this category has improved after the increase in the data, the overall results are limited due to the very small number of original training samples (only 31 samples), which significantly restricted the model's ability to generalize. Robust validation approaches such as stratified K-Fold cross-validation are recommended in future work to improve the statistical reliability of rare-class evaluation. These results confirm that the severe imbalance between groups significantly reduces the ability of CNN models to learn, and that data increases play a crucial role in improving the balanced performance of the classification of heart rhythm disorders using ECG signals.

Confusion Matrix - CNN Only

True Class \ Predicted Class	1	2	3	4	5
1	5224	132	355	490	63
2	58	841	8	14	1
3	119	12	550	29	25
4	134	14	24	81	7
5	2	1			3

4.5 Grad-CAM Explain ability Analysis

To assess whether a trained neural network (CNN) has learned physiological meaningful representations rather than reliance on superficial statistical patterns, the Grad-Cam technology was applied to the reinforced model. Figure 3 shows the Grad-Cam results of a correctly classified sample (N), with a confidence score of 0.5475. The upper part of the figure displays an ECG signal overlay with a 2D heatmap (Heat Map) resulting from the Grad-Cam using a continuous color gradient starting from blue (less important) through yellow to red (highest importance). The lower part displays the one-dimensional post-normalization curve, drawn versus the sample index, providing a quantitative representation of the time importance distribution. The Grad-Cam activation map shows an almost complete focus on the QRS region, which is approximately 100 to 130 samples, the region that represents the depolarization of the ventricular during the heart cycle. This region includes the Q wave (negative deviation before the R peak), the R (the highest positive peak), and the S (negative deviation after the R peak). The properties of the QRS complex are among the most important diagnostic criteria that cardiologists rely on to distinguish between normal rhythm and ventricular system disorders such as the V-Class. In contrast, P-wave regions (approximately 40–80 samples) and T wave (approx. 140–180 samples) receive very low significance values in the Grad-Cam map. These Findings Provides That the One-Dimensional Neural Network Model (1D-CNN) May Teach Significant Clinical Representations Consistent with Medical Practices Adopted in the Interpretation of Electrocardiogram Signals (ECG). The Spatial Cohesion of the Activation Map - The Focus on One Area of The QRS Complex the Global Average Assembly (Gav) Layer Contributes to Directing the Model Towards Extracting Properties. This Level of Explanation, Integrated and Time-Limited, is of Great Importance the Decisions of the Model by Making sure that it focuses on the same important areas in the signal to the accepting of the AI_based diagnostic systems in medical diagnosis.

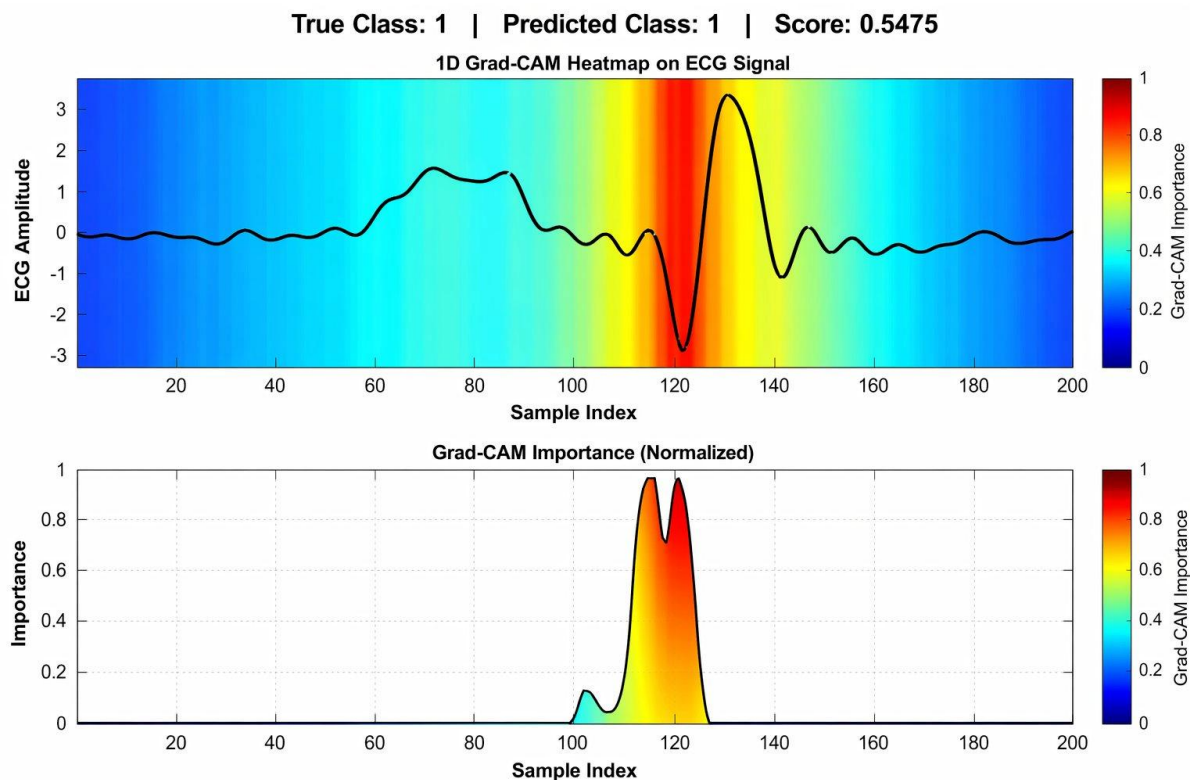


Figure -3 1D Grad-CAM visualization for a correctly classified Class 1 (Normal) sample

The Grad-Cam analysis has been expanded to exceed the N-Class and includes all five AAMI classes, as shown in Figure (4), which presents the one-dimensional Grad-Cam maps of correctly rated representative samples from each category. This multi-class analysis is necessary to verify whether the model captures the clinical morphological features of each class, rather than relying solely on the general characteristics of the QRS complex. For class N (Normal beat, score=0.3682), the activation zones around the QRS complex (samples 100–130), are consistent with the recognition of normal sinus rhythm. Class L (Left Bundle Branch Block, score=0.1625), the model showed stronger and more widespread activation patterns that extended to the ST-T region (Samples 140–190), reflecting the broad and complex shape of the QRS complex. The depolarization of ventricular depolarization associated with LBBB status is delayed. For Class R (Right Bundle Branch block, score=0.1033), the activation focuses on the final region of the QRS deflection (Samples 150–170), which corresponds to the right ventricular activation lag. Which characterizes the status of RBBB? Class A (Atrial Premature Contraction, score=0.6813), the model showed broader and less concentrated activation zones that included both the QRS region and the previous P-wave region (somes 0–50), which reflects importance Diagnostic of the abnormal shape of the P wave in the recognition of early atrial impulses. In contrast, the V-Class (Premature Ventricular Contraction, Score=0.0002) showed the Grad-Cam response weak and clearly unstable, mainly due to the very limited number of original training samples available for this category (only 31 samples). This restricted the model's ability to learn stable and clinically meaningful representations. The value of the very low confidence (0.0002) also underscores the model's difficulty in the reliable detection of this important clinical category. In general, the multi-class Grad-Cam analysis confirms that the proposed CNN model has

learned the morphological features of the ECG signals for well-represented classes, while showing the significant impact of the category imbalance problem on the interpretability and reliable feature learning of learning special characteristics. In rare Classes.

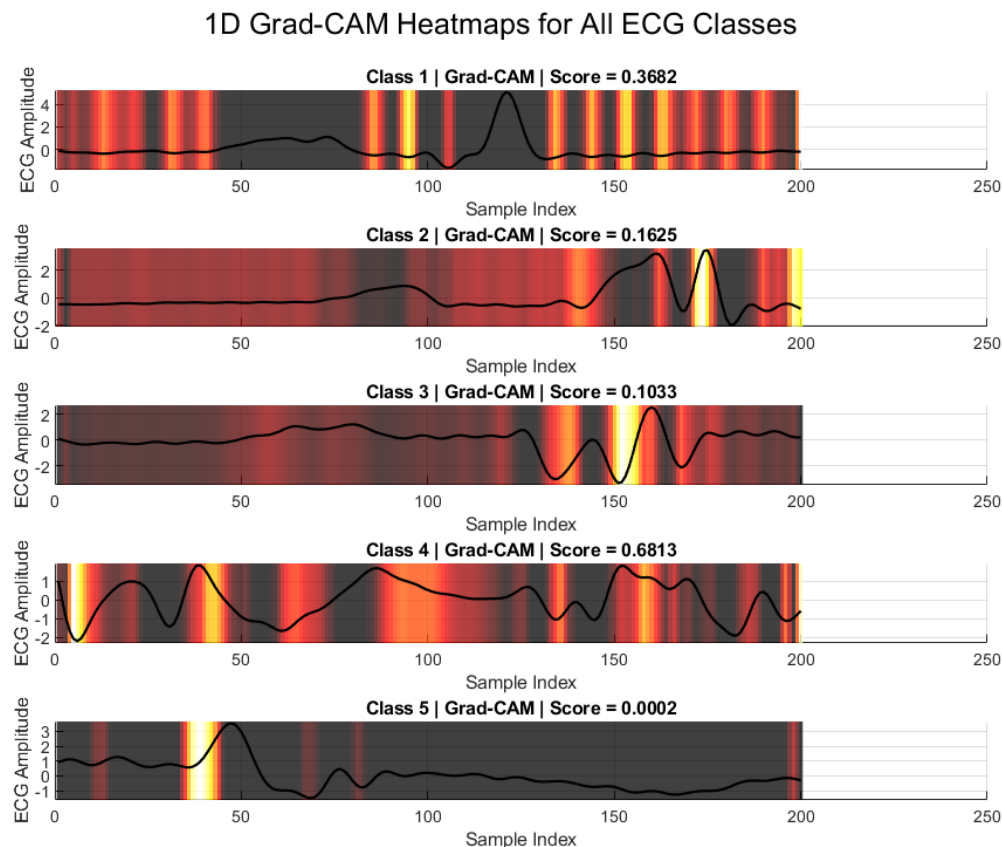


Figure -4 1D Grad-CAM Heatmaps for All Five ECG Classes (N, L, R, A, V).

5. Discussion

The experimental results of this study clearly demonstrate that data augmentation is not merely a useful addition to the training line, but a fundamental requirement for effective multi-class ECG signal classification under highly unbalanced data conditions. The 38.32 percentage point improvement in overall accuracy, along with consistent improvements across all macro-averaged metrics, definitively proves that the baseline CNN model without data augmentation learns only to predict the largest class. Although the macro F1-score of 0.5379 in the augmented model is significantly higher than that of the baseline model, it still indicates considerable room for improvement, particularly in rare classes such as A and V.

A detailed analysis of each class reveals the primary limitations when dealing with extremely rare classes. The fifth category (early ventricular contraction), which originally contained only 31 training samples and was expanded to 8,000 samples, achieved a very low F1 factor of 0.0484, despite the large number of added samples. This poor performance is better understood by a multi-cause analysis rather than just the scarcity of data: (i) the scarcity of

original training data: with only 31 real samples as basic data, the underlying morphological diversity captured is limited in nature, making Almost impossible for any model learning, strong, generalizable representation. (ii) augmentation-induced bias: the significant increase from 31 to 8000 samples means that approximately 99.6% of the 5-Series training samples are synthetically generated from a small range of original samples. Random jamming techniques, such as Gaussian noise, circular displacement, and amplitude change, produce samples that remain close to the original samples in the features space, which can lead to the model fit for a narrow, non-representative dataset. Instead of real distribution within the category. (iii) Limited Model Discrimination Capacity: The proposed bi-blocking neural network design, despite its computational proficiency, may lack the representational depth necessary to accurately distinguish between the fifth category morphology - which is characterized by With wide and exotic QRS compounds and the absence of natural P waves - and among the apparently similar patterns in other categories. Exploring deeper architectural designs (such as Resnet-1D and Inception-1D) is a promising course for future work. (4) Features interfering with class N: PVC pulses, especially when light, can show partial morphological interference with the natural pocket pulses in their fragmented representation, making class separation more difficult for the feature extractor. superficial. Class A (early atrial constriction) showed relatively poor performance, with only 81 samples out of 260 tested samples correctly classified ($F1 \approx 0.20$, $FNR \approx 0.69$). Class A's main challenge is diagnostic reliance on P-wave-wave changes - early and abnormal shapes that precede the QRS compound - precise changes that require the model to comprehend a broader time-context of features that focus on the QRS complex. and sufficient for other categories. The structure of the two-sized biblock neural network may not be able to fully comprehend these broader time dependencies. This observation is consistent with the results of previous studies, such as the study of Deng et al. [10] and the study of Nasir et al. [14], which showed that data generation technologies such as SMOTE or deep generative models such as VAE and GAN provide greater diversity and representation. Best for rare categories compared to the increase in random data. Architecturally, the simple single-dimensional symmetric neural network, comprised of two convolutional units, performs an acceptable performance with a training time of no more than 6 minutes with a single GPU. This computational efficiency is extremely important in practical applications, especially in limited-resource peripheral computing environments, such as wearable heart monitors. The use of global average pooling has proven to be useful in two main aspects: first, it significantly reduces the number of parameters (from $64 \times 100 \times 128 = 128$). 819,200 to $64 \times 128 = 8,192$), which reduces hypersensitivity; Secondly, it maintains the spatial relationship between features and time-site channels, allowing interpretation techniques such as Grad-Cam. However, it is recognized that the simplicity of the structure - two small convolutional blocks with nuclei - likely contributed to poor performance in the fifth and first categories, as these two categories of the model require the ability to monitor the subtle morphological differences (Class 5: Wide QRS compound or long-term time patterns (the first category: an abnormal P wave). The deeper structures based on residual communication (RESNET-1D) or multi-scales time-scale mechanisms have achieved an accuracy of more than 97% on the MIT-BIH dataset [12][15], and future studies should examine whether these are the structures are able to bridge the performance gap of these two rare categories while maintaining computational efficiency. A dual strategy that combines increased data and category weighting has proven effective in addressing imbalances. The process of data increase increases the sample size and enhances the morphological diversity of the rare categories, while the preferences of the groups are offset any remaining impairment in the improved training group - especially the numerical superiority of the class - by imposing greater penalties for these mistakes. Categories during

training. This dual approach is more efficient than using either method alone, as data augmentation addresses the sample size issue, while class weighting addresses the imbalance in the loss function. Despite the observed performance improvement, several limitations should still be considered. It is important to note that the statistical assessment of rare Classes, especially the V, is relatively limited due to the very few original samples available for testing within the MIT-BIH Arrhythmia Database database. Therefore, the results of these Classes should be interpreted with caution statistically and not considered final or fully generalized. However, these findings are still of significant scientific value, as they realistically reflect the real challenges associated with the classification of rare cardiac disorders within the actual clinical environments that are inherently deficient among groups. The results also show that the use of data-increasing techniques and category weights contributed to improving the model's ability to detect less-represented Classes compared to the basic model, although there are still clear challenges that require more advanced solutions in future studies. In future work, it is proposed that random increment techniques be replaced with advanced generation methods such as GAN or VAE, as well as the adoption of more realistic, data-driven Inter-Patient Evaluation protocols, and the expansion of Grad-CAM analysis to include rare Classes, providing a more comprehensive understanding of the model's interpretability.

6. Conclusion

This research provided a systematic and comprehensive comparative study of the classification of cardiac arrhythmias using the 1D-CNN, with and without the use of data-increased data, on the five-tier MIT-BIH database. Two experimental cases were evaluated using the same hyperparameters, with the aim of isolating the direct impact of both data and class weights in improving classification performance. Experimental results have shown conclusively that the increase in data is an essential component of the training process when dealing with unbalanced, multi-class data. The improved model achieved a total accuracy of 83.71%, and the total F1 score of 0.5379, with an absolute improvement of 38.32 percentage points and 0.1966 in the F1 scale, compared to the basic model. These findings are important to be carefully interpreted: the overall accuracy of 83.71% is a misleading measure in the context of severe class imbalance, due to its high influence on the dominant class N. Total F1 score of 0.5379 It reflects the model's performance more precisely, indicating that there is a lot of room for improvement, especially for the minority groups of clinical importance. A high F1 value for each class N (0.9003) and L (0.8944), while A ($F1 \approx 0.20$) and V ($F1 = 0.0484$) are still challenged. Great, mainly due to the intrinsic limitation of random expansion when applied to the very few original samples, as well as the limited representative capacity of the two blocks structure. The Grad-Cam analysis, which included all five categories, provided evidence of the model's ability to interpret, emphasizing that the model learns the physiological features of each category for well-represented categories, while at the same time uncovering restrictions. The ability to explain to rare categories. Combining competitive precision in common categories, high computational efficiency (under 6 minutes on a single GPU), and multi-class interpretation capacity, makes the proposed model a light and useful basis for arrhythmia research, while the limitations specified in the rare categories provide a clear roadmap for future improvements. future research directions include: (i) Replacing random expansion with advanced generative methods such as GAN or VAE to improve the morphological diversity of rare classes; (ii) Exploration of deeper structures (RESNET-1D, Inception-1D) with greater reception areas for better capture of systemic features; (iii) the adoption of evaluation protocols among patients to measure the true generalization capacity;

and (4) K-Fold stratified cross-checking application to provide statistically reliable performance estimates for rare categories.

Acknowledgements

The authors thank Physio net for providing open access to the MIT-BIH database of cardiac disorder used in this study. This website is a very valuable and reliable scientific source for biomedical researchers, as it supports repeatable research and enhances the quality of studies in biomarker analysis. The availability of such well-documented data was a key factor in developing machine learning-based heart signal ratings.

Conflict of Interest

The authors declare that they do not have any conflicts of financial or personal interests that may affect the design, conduct, analysis or reporting of this study.

References

- [1] World Health Organization (WHO). (2023). Cardiovascular diseases (CVDs) Fact Sheet. Geneva: WHO. Available at: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases>
- [2] Clifford, G.D., Azuaje, F., & McSharry, P. (2006). Advanced Methods and Tools for ECG Data Analysis. Artech House, Norwood, MA.
- [3] Kumari, A., et al. (2022). Optimized Decision Tree for Arrhythmia Classification on MIT-BIH Dataset. Computers in Biology and Medicine, 145, 105437. <https://doi.org/10.1016/j.compbiomed.2022.105437>
- [4] Ansari, Y., et al. (2023). Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017-2023. Frontiers in Physiology, 14, 1246746. <https://doi.org/10.3389/fphys.2023.1246746>
- [5] Feyisa, D.W. (2023). Classifying Cardiac Arrhythmia from ECG Signal Using 1D CNN Deep Learning Model. Mathematics, 11(3), 562. <https://doi.org/10.3390/math11030562>
- [6] Buda, M., Maki, A., & Mazurowski, M.A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259. <https://doi.org/10.1016/j.neunet.2018.07.011>
- [7] Liu, C., et al. (2023). Deep Learning-Based ECG Arrhythmia Classification: A Systematic Review. Applied Sciences, 13(8), 4964. <https://doi.org/10.3390/app13084964>
- [8] Rasti, T., Zhu, Q., et al. (2024). ECG-Based Arrhythmia Detection by a Shallow CNN Model. Series Cardiology Research, 6(1), 1-14.
- [9] Kiranyaz, S., et al. (2016). Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Transactions on Biomedical Engineering, 63(3), 664-675. <https://doi.org/10.1109/TBME.2015.2468589>
- [10] Ding, Y., et al. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. Applied Sciences, 13(6), 4006. <https://doi.org/10.3390/app13064006>
- [11] Shchetinin, E.Y., & Glushkova, A.G. (2022). Arrhythmia detection using resampling and deep learning methods on unbalanced data. Computer Optics, 46(3), 451-458.
- [12] Bai, X., Dong, X., Li, Y. et al. (2024). A hybrid deep learning network for automatic diagnosis of cardiac arrhythmia based on 12-lead ECG. Scientific Reports, 14, 24441. <https://doi.org/10.1038/s41598-024-75531-w>

- [13] Moody, G.B., & Mark, R.G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45-50. <https://doi.org/10.1109/51.932724>
- [14] Nasser, A., et al. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57, 285. <https://doi.org/10.1007/s10462-024-10884-2>
- [15] Kovalchuk, O., et al. (2025). Towards Transparent AI in Medicine: ECG-Based Arrhythmia Detection with Explainable Deep Learning. *Technologies*, 13(1), 34. <https://doi.org/10.3390/technologies13010034>
- [16] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626. <https://doi.org/10.1109/ICCV.2017.74>