Ventricular Arrhythmia Classification Using Similarity Maps and Hierarchical Multi-Stream Deep Learning

Qing Lin, *Member, IEEE*, Dino Oglić, Michael J. Curtis, Hak-Keung Lam, *Fellow, IEEE*, and Zoran Cvetković, *Senior Member, IEEE*

Abstract*—*Objective: **Ventricular arrhythmias are the primary arrhythmias that cause sudden cardiac death. We address the problem of classification between ventricular tachycardia (VT), ventricular fibrillation (VF) and non-ventricular rhythms (NVR).** ***Methods:*** **To address the challenging problem of the discrimination between VT and VF, we develop** ***similarity maps*** **– a novel set of features designed to capture regularity within an ECG trace. These similarity maps are combined with features extracted through learnable Parzen band-pass filters and derivative features to discriminate between VT, VF, and NVR. To combine the benefits of these different features, we propose a hierarchical multi-stream ResNet34 architecture.** ***Results:*** **Our empirical results demonstrate that the similarity maps significantly improve the accuracy of distinguishing between VT and VF. Overall, the proposed approach achieves an average class sensitivity of 89.68%, and individual class sensitivities of 81.46% for VT, 89.29% for VF, and 98.28% for NVR.** ***Conclusion:*** **The proposed method achieves a high accuracy of ventricular arrhythmia detection and classification.** ***Significance:*** **Correct detection and classification of ventricular fibrillation and ventricular tachycardia are essential for effective intervention and for the development of new therapies and translational medicine.**

***Index Terms*—Cardiac arrhythmias, hierarchical classification, residual convolutional neural networks, ventricular fibrillation, ventricular tachycardia.**

I. Introduction

GLOBALLY, cardiovascular disease is recognized as the leading cause of death [1]. Ventricular Fibrillation (VF) is one of the cardiovascular diseases with the highest mortality, manifesting as sudden death if the patient is not resuscitated within a few minutes [2]. Defibrillation is currently considered the only effective treatment for VF. However, due to poor discrimination between VF and VT, all tachyarrhythmias are typically treated with DC shocks [3]. Inappropriate shocks can be harmful and the response to interventions may vary according to the type of tachyarrhythmias [4]. Therefore, the correct detection and classification of VF and VT are important for effective intervention. Moreover, this is also essential for developing new therapies and translational medicine. However, the discovery of new therapies and their translation is hampered by a lack of consistency in diagnostic criteria for distinguishing between VF and VT [5]. Aside from inconsistency over specific definitions, the ECG has intrinsic differences from person to person. Some transient and persistent forms of polymorphic VT may be confounded with VF, as VT is not always monomorphic [6]. In addition, complex arrhythmias are often segueing between different types [5]. This all makes it very challenging to distinguish between VT and VF. Therefore, exploring an automatic method is crucial to objectively distinguish VT and VF from ECG signals.

For detection and classification of ventricular arrhythmias using ECG signals, various classification approaches and feature sets have been developed. Comprehensive reviews of different low-dimensional features proposed for ventricular arrhythmia detection and classification have been presented in [7], [8], along with systematic evaluations of their ability to discriminate between VT and VF. Alwan et al. [6] improved the discrimination between VT and VF by combining most effective features from [7], [8] with high-dimensional spectral features, using temporal ensembles of SVM classifiers. Still, the highest average sensitivity obtained by this study reached only 74.7% in the three-class classification between VT, VF, and NVR, i.e., all other rhythms apart from VT and VF.

Recently, deep learning has been applied in many different fields, including medicine and bioengineering. An end-to-end deep learning approach has the ability to automatically extract high-level and informative features from the raw input data [9]. Owing to this, many approaches have been elaborated for cardiac arrhythmia identification [9], [10], [11], [12], [13], including a deep convolutional structure with 16 residual blocks that achieved cardiologist-level accuracy in classifying 12 cardiac arrhythmias [10]. Hence, we reassessed here the accuracy of classification between VT, VF and NVR using the neural architecture proposed in [10], referred to as ResNet34.

Received 14 February 2024; revised 13 June 2024 and 20 September 2024; accepted 16 October 2024. Date of publication 1 November 2024; date of current version 21 February 2025. This work was supported by the King's-China Scholarship Council Scholarship programme. *(Corresponding author: Qing Lin.)*

Qing Lin is with the Department of Engineering, King's College London WC2R 2LS London, U.K (e-mail: qing.lin@kcl.ac.uk).

Dino Oglić is with the Center for AI, Data Sciences and AI, BioPharmaceuticals R&D, AstraZeneca, U.K.

Michael J. Curtis is with the School of Cardiovascular Medicine and Sciences, King's College London, U.K.

Hak-Keung Lam and Zoran Cvetković are with the Department of Engineering, King's College London, U.K.

Digital Object Identifier 10.1109/TBME.2024.3490187

While the ResNet34 architecture has provided some positive effects in automated ventricular arrhythmias diagnosis, it was still affected by the mislabeling of data and the lack of satisfactory generalization capabilities. Typically, deep learning requires a large amount of training data accounting for different sources of variability to avoid the overfitting problem. However, there is only small amount of ventricular arrhythmia data available to the public and the ventricular arrhythmia morphology widely varies between patients. Therefore, the decisions made by the trained model may be affected by patient-related features, which makes it difficult for the model to generalize to unseen samples. This explains the limitations of existing technologies in VT and VF detection and classification. VT consists of 4 or more consecutive ventricular beats, and its QRS complexes are wide (QRS duration equal or longer 120ms) and abnormal, and the rhythm is usually regular [14]. VF is recognized as chaotic, irregular deflections of varying voltage and no distinguishable QRS complexes. This indicates that the regularity of the ECG complex is essential to distinguishing between VT and VF [5]. To address these problems, it is necessary to elaborate a new strategy that extracts information about regularity of patterns in ECG to distinguish between VT and VF.

In the experiments, we used ResNet34 [10] as the baseline classifier to distinguish rhythms into three classes: VT, VF, and NVR. NVR refers to rhythms that are not ventricular rhythm, where a ventricular rhythm is defined as ‘sustained’–that is, four or more consecutive ventricular complexes, as per the Lambeth Conventions [5]. However, we observed no improvement in comparison with the SVM approach in [6], and similar to the results in [6] a very high variability in the accuracy depending on the split of the data between train and test records. Inspection of the data revealed inconsistent labeling criteria across and within databases we considered, including gross departures from standard definitions of VT and VF [5]. Towards further assessing and improving the classification accuracy, the data we used were first relabeled according to Lambeth Conventions [5].

As the ECG signal is the time-dependent signal, a learnable Parzen filter [15] was introduced to capture variation in the frequency of the signal over time, which has led to a better identification of ventricular tachyarrhythmias. Furthermore, according to the definitions outlined in the Lambeth convention, the distinction between VT and VF is based on regularity and similarity between consecutive segments of ECG signal, rather than detecting identifiable QRS complexes. Thus, we evaluated a two-stage hierarchical approach, which divides the 3-class problem into two parts: ventricular tachyarrhythmia detection and ventricular tachyarrhythmia discrimination. To discriminate ventricular tachyarrhythmias, we propose the similarity map as a feature to capture repetition and similarity within ECG segments. In the ventricular tachyarrhythmia identification problem, derivative features of the ECG signal were introduced too as they have been proven to facilitate the detection of P waves and QRS complexes from the ECG signal [16], [17]. For combining benefits of these different features we propose a multi-stream deep architecture that employs ResNet34in individual streams. As a result of consistent labeling and novel features, the average class sensitivity reached 89.6%.

The paper is organised as follows. Section II presents details of proposed methods. Section III discusses the data collection and relabeling. The experimental evaluation is presented in Section V. Section VI summarises the experimental outcomes, which are discussed in Section VII. Conclusions are drawn in Section VIII.

II. Methods

A. Parzen Convolutional Block

As ECG signal is a time-dependent signal, a convolutional block of learnable Parzen filters [15] was introduced to extract temporal variations in the frequency domain. In the context of automatic speech recognition, a network architecture with Parzen convolutional blocks has been demonstrated to outperform feedforward models based on non-adapted features without requiring a large amount of training data [15]. Filters in the Parzen convolutional block used in this work are differentiable filters given by

$$\phi_{\eta,\gamma}(t) = \cos\left(2\pi\eta t\right)\cdot k_{\gamma}(t), \eta > 0 \tag{1}$$

where $k_{\gamma}(t)$ is the squared Epanechnikov window

$$k_{\gamma}(t) = \max\left\{0, 1-\gamma t^{2}\right\}^{2}, \gamma > 0 \tag{2}$$

and γ and η are learnable parameters that control filter bandwidths and their centre frequencies, respectively.

B. Similarity Maps

In the field of ventricular arrhythmia diagnosis, experts typically consider the regularity of abnormal heartbeats when analyzing patients’ ECG records [5]. In order to discover pattern repetition, or the lack thereof, in ECG signals, all possible patterns in the ECG trace need to be collected by means of a sliding window. Given an ECG segment $x = \{x_1, x_2, \ldots, x_N\}$ consisting of N samples, which we will refer to as *observation length*, we first obtain all sub-sequences through the sliding window of length $l+1$ and shift by 1. The sub-sequences W_i obtained in this manner are given by: $W_i = \{x_i, x_{i+1}, \ldots, x_{i+l}\}$. For each pair of sub-sequences, we evaluate their similarity using two measures:

- Euclidean Distance: $I_{i,j} = \sqrt{\sum_{k=1}^{l} |W_{i,k} - W_{j,k}|^2}$, where $W_{i,k}$ denotes the kth sample in W_i,
- Cosine Similarity: $I_{i,j} = \dfrac{< W_i, W_j >}{\|W_i\|\|W_j\|}$.

For each sub-sequence W_i, we define one channel of features using its similarity values with all other subsequences. The length of each feature channel is equal to L, where $L = N - l$ is the total number of subsequences. The i-th feature channel can be represented as: $I_i = \{I_{i,1}, I_{i,2}, \ldots, I_{i,L}\}$, where $I_{i,j}$ is the similarity value between subsequences W_i and W_j. Fig. 1 shows a schematic diagram of this sliding window approach to generate the i-th feature channel. The similarity map feature $\mathbf{I}$

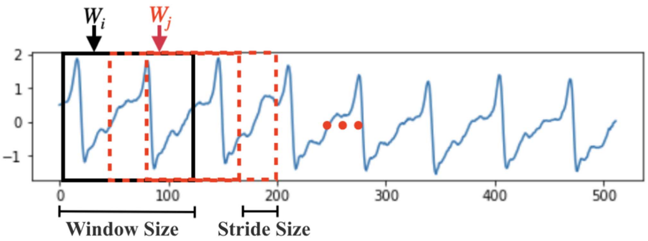

Fig. 1. An example for the generation process of similarity maps.

is thus the $L \times L$ matrix of all similarity scores $I_{i,j}$:

$$\mathbf{I} = \begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1,L-1} & I_{1,L} \\ I_{2,1} & I_{2,2} & \cdots & I_{2,L-1} & I_{2,L} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I_{L,1} & I_{L,2} & I_{L,3} & I_{L,L-1} & I_{L,L} \end{bmatrix}$$

Similarity feature maps were only used to distinguish VT and VF. This is because the similarity feature maps discard some morphological features of ECG signal, such as RR intervals, QT segments, QRS complexes, which is crucial for the identification of most non-ventricular rhythms.

C. Cyclically Shifted Similarity Maps

To obtain the average similarity map of an ECG segment, we performed circular shifts of all channels of its similarity feature map so that $I_{i,i}$ reached the first position in each of the channels. We then removed the first column of the similarity map after circular shift operation, which gave the following variant of the similarity map:

$$\mathbf{I}' = \begin{bmatrix} I_{1,2} & I_{1,3} & \cdots & I_{1,L-1} & I_{1,L} \\ I_{2,3} & I_{2,4} & \cdots & I_{2,L} & I_{2,1} \\ I_{3,4} & I_{3,5} & \cdots & I_{3,1} & I_{3,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I_{L,1} & I_{L,2} & \cdots & I_{L,L-2} & I_{L,L-1} \end{bmatrix}$$

Finally, we formed vector $\mathbf{I}'_{avg}$ which is the average of the rows of $\mathbf{I}'$.

D. Derivative Features of ECG Signal

According to the Lambeth Convention [5], ventricular arrhythmia is defined as a series of ventricular complexes, where the Q, R, S, and T waves are not necessarily detected, or even only a single oscillation exists. Therefore, the QRS peak detection is critical for the classification between ventricular arrhythmias and NVR. The derivatives are commonly used to identify the features of the signal like peaks, inflexion points, maxima and minima, which have been proven to have the potential to improve the accuracy of QRS peak detection [18]. Thus, we employed the differentiation to extract additional features for ventricular arrhythmia classification.

The first derivative term is implemented as the difference between two successive samples of the discrete signal $x = \{x_1, x_2, \ldots, x_N\}$:

$$x'_t = x_t - x_{t-1},\ for\ 1 < t \leq N \tag{3}$$

The main aim of applying the first derivative in ECG analysis is to emphasize signal changes by reducing low-frequency region waves and enhancing the high-frequency QRS complex [19], [20]. Such an enhanced focus on the QRS complex improves the accuracy of R peak detection. In the case of ventricular arrhythmias, characterized by rapid and irregular heartbeats, the first derivative aids in highlighting these quick signal changes, thus facilitating the identification of the ventricular arrhythmia. In distinguishing between VT and VF, the ability of the first derivative to recognize the QRS complex is crucial, especially in identifying the presence or absence of R peaks [21]. VT typically presents with identifiable R peaks, although at a high rate, while VF often lacks clear R peaks, making the first derivative a critical instrument in rhythm analysis.

The second derivative of signal can be calculated as follows:

$$x''_t = x_{t+1} - 2x_t + x_{t-1},\ for\ 1 < t \leq N - 1. \tag{4}$$

The second derivative is effective for identifying ventricular arrhythmias by accurately detecting R peaks in ECG signals [22]. However, it has limitations in distinguishing between VT and VF. This arises from the second derivative being more sensitive to high-frequency noise compared to the first derivative [18]. Consequently, this approach is more vulnerable to errors introduced by high-frequency components in the ECG signal, making it less suitable for differentiating between VT and VF.

III. Data and Reprocessing

A. Data Preparation

ECG signals were obtained from five publicly available arrhythmia databases: MIT-BIH Arrhythmia Database (MITDB) [23], the MIT-BIH Malignant Ventricular Arrhythmia Database (VFDB) [24], the European ST-T Database (EDB) [25], the Creighton University Ventricular Tachyarrhythmia Database (CUDB) [26] and the extended American Heart Association Database (AHADB). Since the term ventricular flutter is ambiguous and not acknowledged as an entity in the Lambeth Conventions [5] and labeled in different ways between databases, any records with clearly marked ventricular flutter rhythms were excluded from this work. As a result, 342 records were used in our experiments.

Mislabeling is commonplace in public databases because of a lack of consistency in the use of diagnostic criteria for distinguishing between VF and VT. In our exploration of databases, we have encountered unequivocal examples of the following (using the Lambeth Conventions [5] definitions of arrhythmias as the criteria): VT labeled as VF; VF labeled as VT; electrical noise labeled as VF; asystole labeled as VF. In addition, we have noted that the mislabeling of traces is not systematic or consistent. This is most evident in traces where a rhythm starts out as VT, then changes to VF, and then changes back to VT or asystole or normal sinus rhythm where the curators have in some instances labeled the different phases as different rhythms, and in other cases labeled the entire bloc of arrhythmia as a single rhythm (i.e., VT or VF). Consequently, the collected

TABLE I
THE AMOUNT OF DATA BEFORE AND AFTER RELABELING

Class	Original dataset	Relabeled dataset
VT	5,514 s	11,607 s
VF	16,267 s	3,994 s
NVR	1,085,887 s	1,092,038 s

data in the present study was rigorously relabeled according to Lambeth conventions [5] by professionals who come from the School of Cardiovascular Medicine and Sciences, King's College London. After relabeling, 78 records were identified as containing ventricular arrhythmias, while the remaining 264 records contained only NVR.

Table I shows the total duration of each category present across used records before and after the relabeling. The relabeled data set will be made publicly available.

All collected records were resampled to 250Hz and processed by a high-pass Butterworth filter with a cutoff of 0.5Hz to remove the ECG baseline wander. Each record was normalized to zero mean and a standard deviation of 1.

ECG records were split into non-overlapping segments for training and testing. The length of ECG segment is referred to as observation length. To deal with the class imbalance problem, we randomly subsampled the largest class (NVR class) to the size of the second-largest class.

IV. DATASET RELABELING

A. The Effect of Data Labeling on Classification With ResNet

To investigate the impact of data relabeling, we applied ResNet34in a three-class scenario using raw ECG segment as input from both the original and relabeled datasets; the results are shown in Table II. Experiments were conducted with an observation length of 512 samples (2 s). The relabeling procedure led to a notable increase in VT sensitivity, which more than doubled, contrasting with a substantial decrease in VF sensitivity. The greater quantity of VT samples in the relabeled dataset suggests a class imbalance, leading to a bias in the model towards VT over VF. Consequently, the overall average class sensitivity observed an improvement due to the more significant rise in VT sensitivity compared to the decline in VF sensitivity. The specificity for VT and VF increased, while the specificity for NVR decreased, contributing to an overall reduction in specificity after relabeling. This drop is attributed to relabeled NVR recordings that included rhythms resembling ventricular rhythms, which were initially misclassified as ventricular rhythms but corrected after relabeling. However, these rhythms remain prone to misclassification by the current classifier. Despite the shifts in specificity, the average sensitivity saw a considerable improvement, rising from 73.72% to 77.06%. Additionally, the average F1 score, an integrated measure of precision and recall, increased from 60.37% to 60.99% after relabeling, indicating an overall enhancement in classification accuracy. The comparative analysis presented in Table II was conducted using 10 bootstrap resamples both before and after relabeling.

B. Dataset Distribution Comparison

In order to further assess the effect of relabeling, we first study the difference in data distribution between the original and relabeled databases via t-Distributed Stochastic Neighbor Embedding (t-SNE) [27]. Then, maximum mean discrepancy (MMD) [28] is introduced as a distance metric to quantify the effect of relabeling on distribution mismatch between training and test data.

1) T-Distributed Stochastic Neighbor Embedding (t-SNE): For the original and relabeled datasets, we randomly selected 3000 1s ECG segments from VT and VF. Then, the collected 6000 segments are projected to 30 dimensions by PCA to decrease the memory requirement for calculation. After dimensionality reduction, we embed these data in 3D spaces by t-SNE.

As seen from Fig. 2, there is a set of outliers in the t-SNE plot that form a circle in the VT class of the original and relabeled datasets. These are VT signals that exhibit clear periodicity. Apart from this subset of the data, the data distributions in the 3D t-SNE plots of the relabeled data appear slightly more discriminative than in the case of the original dataset.

To gain further insight into the effect of the relabeling on the classification accuracy, we applied t-SNE on the feature vectors at the output of the trained residual blocks of ResNet34 (i.e. the input of dense layers of the trained model). We first split the dataset by records, 80% of which were used for training the model and the rest for testing. Then we collect feature vectors, which are the output of the trained residual block, by feeding the test data to the corresponding trained model. All collected feature vectors are projected to 30 dimensions by PCA to decrease the memory requirement for calculation. After dimensionality reduction, we embed these data in 3D spaces by t-SNE. This procedure is repeated 5 times on the dataset before and after relabeling. Results are shown in Fig. 3. We can observe better separation of the relabeled data.

Additional insights are illustrated in Fig. 4, depicting the difference in the embedding of distributions for the same target class relative to ResNet34 features, before and after relabeling. These embeddings are obtained by inputting the same test recordings into a ResNet model trained on the original labeled training records. Each column represents different train-test recording splits. Table IV indicates that VT shows a higher degree of separation between the original and relabeled dataset across the different data splits. This suggests that the relabeling process has significantly changed labeling of VT, primarily because episodes that had been originally labeled as VF were relabeled as VT according to the Lambeth Conventions definitions [5]. Additionally, entire blocs of arrhythmia previously labeled as a single rhythm (e.g., VT or VF) but containing a mix of VT and VF (and sometimes asystole or evident electrical noise due to typical human ECG electrode failure) were segmented and re-labeled as specific rhythms (VT, VF or NVR). The main changes in labeling, therefore, were from VT to VF, and from VF to electrical noise or asystole. Despite these changes, there remained a significant concordance between the original and relabeled datasets in rhythms classed VF, indicating that the identification of VF was similar before and after relabeling.

TABLE II
The Effect of Relabeling on Individual and Average Class Sensitivities

Method	Dataset	VT			VF			NVR			SEN_{avg} %	$SPEC_{avg}$ %	$F1_{avg}$ %
		SEN %	$SPEC$ %	$F1$ %	SEN %	$SPEC$ %	$F1$ %	SEN %	$SPEC$ %	$F1$ %			
ResNet34	original	42.23	98.78	18.43	80.98	98.60	63.76	97.95	96.79	98.92	73.72	98.06	60.37
ResNet34	relabeled	83.18	99.32	65.11	49.67	98.80	18.73	98.33	95.36	99.13	77.06	97.83	60.99

Results derived from the test set.

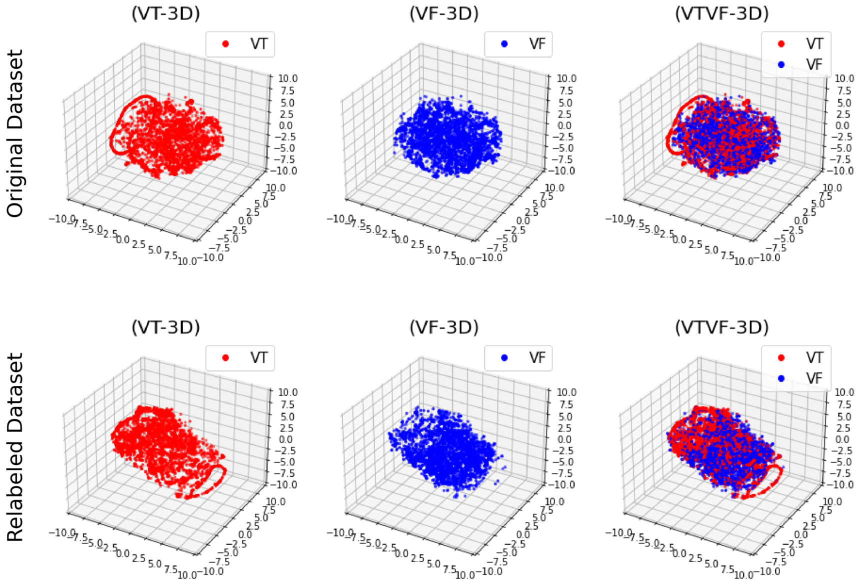

Fig. 2. Embedding of 1s ECG segments of the original and relabeled datasests to a 3D space by t-SNE.

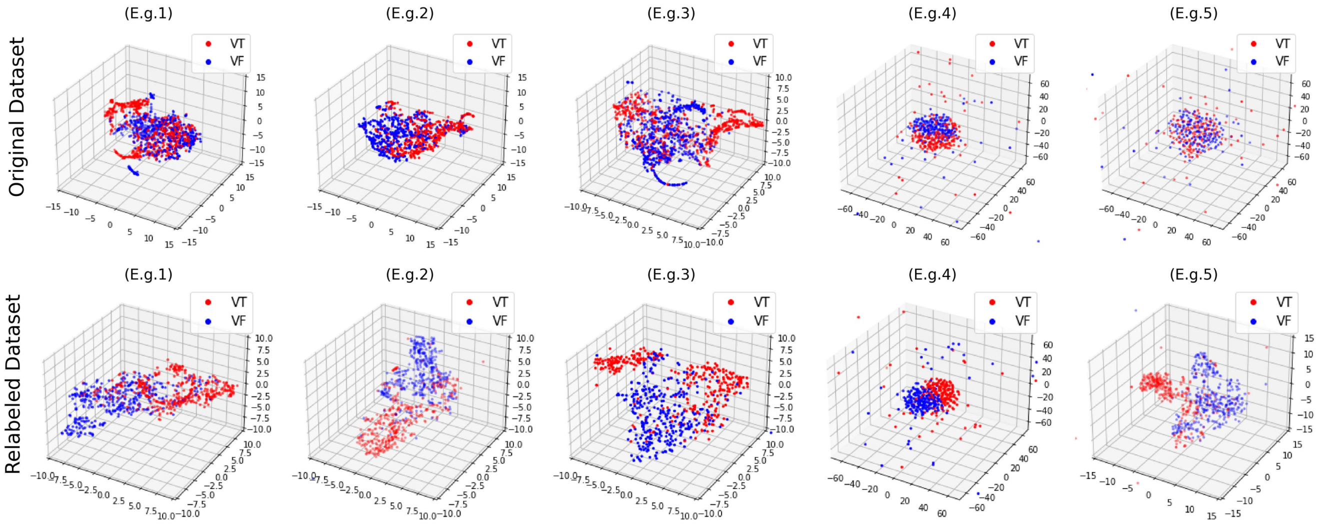

Fig. 3. Embedding of the ECG feature vectors from the original and relabeled datasets into 3D space using t-SNE. These feature vectors are generated by feeding the test data to the corresponding trained ResNet 34 model. The column represents different data splits.

TABLE III
Statistical Analysis of T-Test Outcomes for ECG Feature Vector Embeddings From Original and Relabeled Datasets

	Original VT vs Original VF	Relabeled VT vs Relabeled VF	Original VT vs Relabeled VT	Original VF vs Relabeled VF
t-statistic	3.338	12.924	5.315	2.670
p-value	0.060	8.38E-4	0.022	0.297

These feature vectors are generated by feeding the test data to the corresponding trained ResNet 34 model. The results are calculated as an average across 10 separate data splits.

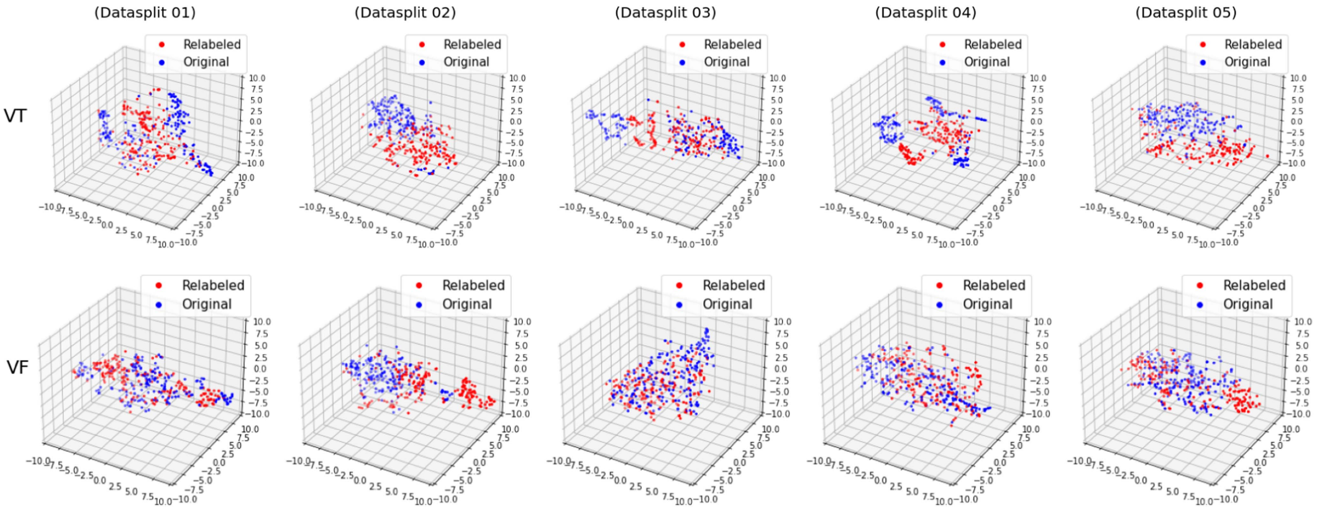

Fig. 4. Comparative t-SNE Analysis of VT and VF Embedding Features Across Original and Relabeled Datasets. These feature vectors are generated by feeding the test data to the corresponding trained ResNet 34 model. The column represents different data splits.

TABLE IV
THE INTER- AND INTRA-SET DISTANCES FOR VT AND VF, AS WELL AS THE DISTANCES ACROSS VT AND VF, IN BOTH THE ORIGINAL DATASET AND THE RELABELED DATASET

feature space	dataset	VT_MMD_{AB}	VT_MMD_{AA}	$\lvert VT\_MMD_{AB} - VT\_MMD_{AA}\rvert$	VF_MMD_{AB}	VF_MMD_{AA}	$\lvert VF\_MMD_{AB} - VF\_MMD_{AA}\rvert$	$VTVF\_MMD_{AB}$
		$d(VT_A, VT_B)$	$d(VT_A, VT_A)$		$d(VF_A, VF_B)$	$d(VF_A, VF_A)$		$d(VT_A, VF_B)$
waveform	original	0.31	0.13	0.18	0.16	0.14	0.02	0.24
	relabeled	0.19	0.14	0.04	0.21	0.14	0.07	0.34
feature	original	0.65	0.17	0.47	0.23	0.18	0.05	0.58
	relabeled	0.31	0.18	0.13	0.42	0.18	0.25	0.67

The distances are computed between two different sets of records, denoted as A and B. Both time domain vectors and feature domain vectors are used to calculate the distances.

2) T-Test: Apart from t-SNE, we also employed the t-test to quantify the observed differences in distributions associated with the target classes. Table III displays the t-test results obtained from the first principal component of the PCA on embeddings of the test recordings processed through ResNet models. We repeated this entire procedure for 10 different train-test recording splits to calculate the average t-statistic and p-value, ensuring a comprehensive and statistically robust analysis. The first two columns of Table III show the t-Tests comparing the differences between VT and VF in both the original and relabeled datasets. The embedding of test recordings with original labels is generated by inputting the test set into the ResNet model trained on original label training recordings. The same process is used for generating embeddings of the test data in the relabeled dataset, by utilizing the corresponding relabeled training data. The conclusion drawn here is consistent with what is depicted in Fig. 3, which shows the improved separation after relabeling. The last two columns of Table III show differences in the embedding of distributions for the same target class relative to ResNet34 features, before and after relabeling. These embeddings are obtained by inputting the same test recordings into a ResNet model trained on the original labeled training records. This corroborates the insights derived from Fig. 4, indicating a marked difference in the distribution of relabeled VT compared to its original, while the distribution for VF shows little variation.

3) Maximum Mean Discrepancy (MMD): According to the study in [29], the distribution mismatch between the train and test set has a negative effect on classification accuracy. To quantify the impact of relabeling on distribution mismatch between train and test data, we considered the MMD as a measure of the inter- and intra- class distances in train and test data.

Since the dataset is split by records for training and testing in our study, we randomly select 80% of the records as Dataset A and the rest as Dataset B. The distance between distributions of patterns in set A and set B can be computed using the distance between the corresponding distribution embeddings. To that end, we randomly select n samples $X = \{x_1, \ldots, x_n\}$ from set A and m samples $Y = \{y_1, \ldots, y_n\}$ from set B. In this experiment we used $n = m = 80$. The empirical estimate of the MMD between X and Y is then obtained as follows:

$$\mathrm{MMD}^2(X, Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i, y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j) \tag{5}$$

where $k : \mathcal{R} \times \mathcal{R} \rightarrow \mathbb{R}$ is a positive definite kernel function. In this study, the Gaussian kernel function

$$k(x, x') = \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right) \tag{6}$$

was used, where σ is a hyper-parameter allowing a degree of flexibility when defining this similarity. According to study [30], we set σ to be the median of pairwise distances between segments from different sets of records. Then we take the square root of the maximum mean discrepancy as the distance measure in our experiment:

$$d(X, Y) = \sqrt{\mathrm{MMD}^2(X, Y)} \tag{7}$$

For each data split, we repeat this procedure on 10 random selections of X and Y, yielding a list of distances $d(X_1, Y_1)$, $d(X_2, Y_2)$, ... , $d(X_{10}, Y_{10})$, and then compute the mean value as $d(A, B)$.

Table IV reports the distances across training and test sets of records in the original dataset and the relabeled dataset. Both time domain waveforms and feature domain vectors are considered. In the table, VT_MMD_{AA} and VT_MMD_{AB} columns show the intra- and inter-set distances, respectively, for VT segments in the two sets of records, whereas the $|\mathrm{VT\_MMD_{AB}} - \mathrm{VT\_MMD_{AA}}|$ column shows their difference, and analogously for VF records. The last column of the table shows the difference between the distribution of VT in the set A and VF in the set B.

In both the waveform domain and the feature domain space, the difference between the inter- and intra- set distances of VT segments, $|\mathrm{VT\_MMD_{AB}} - \mathrm{VT\_MMD_{AA}}|$, is reduced in the relabeled dataset. This means that the VTs in the relabeled dataset have less distribution mismatch across the train and test set compared with the original dataset, which may explain the higher accuracy of VT classification after the relabeling. However, the effect of the relabeling on VF data is the opposite. Compared with the increase of the difference between the inter- and intra- set distances in the case of VF after the relabeling, the decrease in the case of VT is more pronounced, and moreover the distance between VT in the training set and VF in the test is increased, suggesting that the relabeling facilitates the separation between VT and VF. Altogether, these observations support the increased VT sensitivity, decreased VF sensitivity and increased average sensitivity as a result of the relabeling, in agreement with the classification result shown in the Table II.

We conclude that the relabeling indeed resulted in more consistent labels across the available data, but also that the relabeling and ResNet34 alone are not sufficient for major gains in the classification accuracy. That motivates the work on new features developed in the previous section and classification architectures proposed in the next section.

V. Experimental Setup

A. Hierarchical Classification

As the criteria for identification of ventricular tachyarrhythmias and discrimination between VT and VF are different, we used a two-stage hierarchical approach that breaks down this 3-class problem into two sub-problems [6].

For each given ECG segment, decisions are made simultaneously by classifiers C_1 and C_2 with different sets of features. C_1 makes three types of decisions, VF, VT, or NVR, but the decision will only be accepted when the sample is classified as NVR. Otherwise, the decision is taken by the classifier C_2, which is a binary classification model that distinguishes VT and VF. This procedure is described as:

$$C(X_{c1}, X_{c2}) = \begin{cases} C_1(X_{c1}), & if\ C_1(X_{c1}) = \mathrm{NVR} \\ C_2(X_{c2}), & if\ C_1(X_{c1}) \neq \mathrm{NVR} \end{cases} \tag{8}$$

where X_{c1} and X_{c2} are the input features for classifiers C_1 and C_2, respectively. We have experimented also with C_1 making binary decisions between NVR and arrhythmia (i.e. VT or VF), but that gave inferior performance.

TABLE V
Comparative Performance of Deep Learning Architectures for Distinguishing Ventricular Tachycardia From Ventricular Fibrillation Using Similarity Maps

Model	Feature	SEN_{avg} (%)	Diagnosed as		
			Truth	VT(%)	VF(%)
1D ResNet34 [10]	Average cyclically shifted similarity map $\mathbf{I}'_{avg}$	**87.52**	VT	87.79	12.21
			VF	12.76	87.25
2D ResNet34 [31]	Similarity map $\mathbf{I}$	86.58	VT	87.82	12.18
			VF	14.66	85.34
2D ResNet50 [31]	Similarity map $\mathbf{I}$	82.75	VT	91.70	8.30
			VF	26.19	73.80
2D VGG16 [32]	Similarity map $\mathbf{I}$	79.38	VT	90.51	9.49
			VF	31.76	68.24
2D EfficientNet [33]	Similarity map $\mathbf{I}$	87.36	VT	92.26	7.74
			VF	17.55	82.45
2D MobileNetV2 [34]	Similarity map $\mathbf{I}$	87.27	VT	92.20	7.80
			VF	17.66	82.34
2D DenseNet201 [35]	Similarity map $\mathbf{I}$	84.58	VT	91.96	8.04
			VF	22.79	77.21
2D Vision Transformer [36]	Similarity map $\mathbf{I}$	85.28	VT	79.20	20.80
			VF	8.64	91.36

Results derived from the validation set.

B. Multi-Stream Deep Neural Network Architecture

To combine the benefits of different input features we propose a multi-stream deep neural network architectures for individual classifiers C_1 and C_2. Each stream consists of a preprocessing step, which extracts the corresponding feature vectors, followed by the ResNet34 [10] processing. Outputs of individual ResNet streams are then concatenated and passed to a fully connected layer to form a new set of features that are then used for classification in the softmax layer. This architecture is illustrated in Fig. 5.

The ResNet34 [10] is designed to process one-dimensional data. Therefore, we implement a 1D convolution along a single dimension of the data array. In this approach, each row of input is processed as a separate sequence, with its columns as features. This allows the 1D CNN to independently process each feature, efficiently capturing the sequential patterns within the data.

The non pre-trained 1D ResNet34 was chosen because it has been proven in [10] to achieve cardiologist-level accuracy in classifying 12 types of cardiac arrhythmias. We compared the performance of this non pre-trained 1D ResNet34 against several pre-trained 2D models, using 1D and 2D similarity maps generated by 5s ECG segment as inputs, respectively. The results, presented in Table V, indicate that the non-pretrained 1D ResNet34 outperformed the 2D pretrained models in terms of average sensitivity and balanced diagnostic accuracy for VT and VF. We note, however, that the results are not very sensitive to the particular network choice. The superior performance of the 1D model can be attributed to the cyclical rotation and averaging

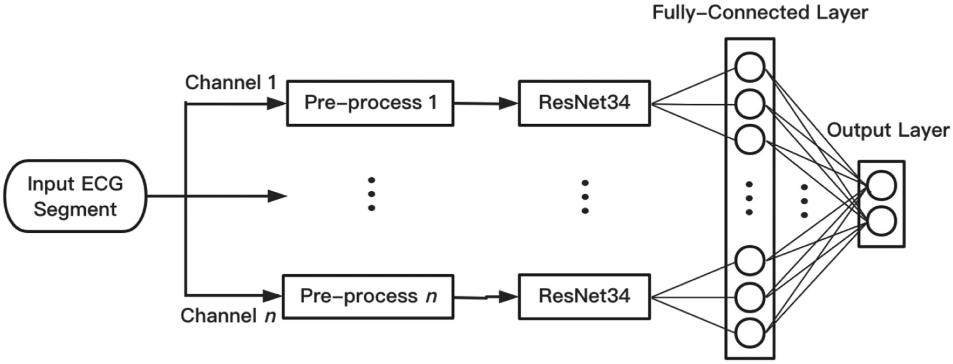

Fig. 5. The multi-channel deep convolutional neural network architecture used in classifiers C_1 and C_2.

technique used in constructing the 1D similarity maps, which provides both translation invariance and robustness to noise.

The considered input features of the classifier C_1 were

1) x: raw ECG waveforms
2) x': the first derivative of x
3) x'': the second derivative of x
4) Parzen features extracted from x using 64 trainable Parzen band-pass filters [37].

For the classifier C_2, we considered the following input features:

1) x: raw ECG waveforms
2) x': the first derivative of x
3) x'': the second derivative of x
4) Similarity map features derived from x, including the similarity map $\mathbf{I}$, cyclically shifted similarity map $\mathbf{I}'$ and average cyclically shifted similarity map $\mathbf{I}'_{avg}$.
5) Parzen features extracted from x using 64 trainable Parzen band-pass filters [37].

C. Training

After relabeling, 78 records were identified as containing ventricular arrhythmias, while the remaining 264 records contained only NVR. For the training set, 70% of the records from each group were selected. Similarly, 10% of the records from each group were allocated to the validation set, and the remaining 20% were reserved for the test set. Classification experiments were performed using 10 bootstraps resamples. In order to have a fair comparison between methods, each experiment used the same hyperparameters including learning rate and batch size. The trainable weights in ResNet34 were initialized as described in [38]. Adam optimizer with initialized learning rate $lr = 1 \times 10^{-3}$ was applied to update the weights. The categorical cross-entropy loss function was used for training the model.

The models were trained with a maximum of 50 epochs and mini-batches of size 32. We aimed to get batches that contained samples from all the classes. For each class, we first calculate the proportion of that class in the total number of training samples as $P_i = \frac{N_i}{N}$, where N is the total number of training samples and N_i is the number of samples of class i, $i \in \{VT, VF, NVR\}$. Then, the number of samples of class i in a batch was set to be $M_i = P_i \times Batch\_size$.

D. Evaluation Method

Due to the class imbalance, we report the sensitivity of each given category, which measures the proportion of correctly classified samples:

$$sn_s = \frac{TP_s}{TP_s + FN_s}, s \in S \tag{9}$$

where the TP_s is the number of examples in class s correctly assigned to class s, and the FN_s is the number of examples in class s incorrectly assigned to other classes.

Another metric to determine the performance of a classifier is the unweighted sensitivity, or average sensitivity, which can be calculated as:

$$SEN_{avg} = \frac{1}{|S|} \sum_{s \in S} sn_s \tag{10}$$

where $|S|$ represents the total number of classes.

VI. Experimental Results

Since the relabeling resulted in more consistently assigned labels and improved accuracy, the relabeled data set is then used in experiments reported in this section.

For Experiments 1-4, the validation set was utilized to optimize hyperparameters. After hyperparameter optimization, the final model results were evaluated and documented on the unseen test recordings in Experiments 5, 6 and 7. All experiments were conducted 10 times to obtain the average performance.

a) Experiment 1: Establishing a hierarchical model with improved performance is premised on achieving high accuracy of classifier C_1 to distinguish between ventricular arrhythmias and non-ventricular rhythms. To that end, we first explore benefits of using Parzen features in C_1 in comparison with raw ECG waveforms. These evaluations were performed using different combinations of the number of filters and maximum filter lengths with an observation length of 512 samples (2s). Table VI shows the results. Both the highest average sensitivity of 83.34% and the highest NVR sensitivity of 98.28% are achieved with 64 Parzen features extracted using long filters, of length up to 1s. Therefore, these Parzen features were used in C_1.

b) Experiment 2: We first considered the impact of different combinations of parameters for the similarity map features when

TABLE VI
He Effect of Parzen Features on Cofusion Matrices

			Diagnosed as		
Method	$SEN_{avg}(\%)$	Truth	VT(%)	VF(%)	NVR(%)
Raw ECG waveforms	79.62	VT	83.03	13.87	3.10
		VF	35.12	58.35	6.53
		NVR	0.95	1.55	97.49
Parzen 16 filters filter length < 0.5s	81.13	VT	82.80	13.61	3.59
		VF	31.36	62.56	6.08
		NVR	0.75	1.22	98.03
Parzen 16 filters filter length < 1s	82.89	VT	80.57	16.37	3.06
		VF	26.01	70.11	3.88
		NVR	0.78	1.24	97.98
Parzen 64 filters filter length < 0.5s	81.10	VT	81.31	15.27	3.42
		VF	28.85	63.84	7.31
		NVR	0.65	1.19	98.15
Parzen 64 filters filter length < 1s	**83.34**	VT	81.67	15.16	3.17
		VF	26.58	70.07	3.35
		NVR	0.39	1.33	98.28

Results derived from the validation set.

TABLE VII
Considered Similarity Map Parameters

Parameter	Options
Similarity Measure	**Euclidean distance**, cosine similarity
Feature Type	similarity map, circularly shifted similarity maps, **circularly shifted similarity maps with averaging**
Window Size	64 samples, **128 samples**, 192 samples, 256 samples

The best performing parameter combination is highlight in bold. Results derived from the validation set.

TABLE VIII
Sensitivities of the Hierarchical Architecture With Averaged Circularly Shifted Similarity Maps, Corresponding to Different Observation Lengths

Observation length	VT(%)	VF(%)	NVR(%)	$SEN_{avg}(\%)$
512 samples (2s)	83.75	77.64	98.28	86.56
768 samples (3s)	83.42	82.50	97.26	87.73
1024 samples (4s)	84.68	81.05	97.80	87.84
1280 samples (5s)	82.48	84.59	97.93	**88.33**
1536 samples (6s)	81.78	81.55	97.00	86.78
2048 samples (7s)	81.88	83.71	97.36	87.65

Results derived from the validation set.

TABLE IX
Sensitivities of the Hierarchical Approach With Averaged Circularly Shifted Similarity Maps, Corresponding to Combinations of Different Window Lengths

Window Size (samples)	VT(%)	VF(%)	NVR(%)	$SEN_{avg}(\%)$
64	81.86	84.62	97.93	88.14
128	82.48	84.59		88.33
192	79.35	86.41		87.89
256	83.64	82.59		88.05
64 & 128	82.48	86.18		88.86
64 & 128 & 256	84.53	85.63		**89.36**
64 & 128 & 192 & 256	84.98	82.91		88.61

Results derived from the validation set.

they are used alone as the input to C_2. The considered parameters are listed in Table VII with the best-performing scenario highlighted in bold. All results were obtained again with the observation length of $N = 512$ samples (2s). The highest average class sensitivity of the proposed hierarchical architecture was achieved by using averaged circularly shifted similarity maps as input features, where the window size is $l = 128$ -samples and the similarity measure is Euclidean distance. The classification results with these similarity features are shown in the first row of Table VIII.

c) Experiment 3: Next, we considered the impact of increasing the observation length from $N = 512$ samples (2s) to $N = 2048$ samples (7s). The features used in C_1 were the 64 Parzen features, whereas the input to C_2 were the similarity maps generated using windows of $l = 128$ samples, with circular shift and averaging, which achieved the best results in the previous two experiments. The results presented in Table VIII show that the accuracy is sensitive to the observation length. The highest average class sensitivity of 88.33% was achieved by increasing the observation length to $N = 1280$ samples (5s). Therefore, the observation length of 1280 samples was used for all experiments in the next step.

d) Experiment 4: As the ventricular complex varies across patients, at the input to C_2 we considered combining averaged circularly shifted similarity maps generated using different window sizes. The first four rows of Table IX show the results corresponding to individual window sizes. The highest average class sensitivity of 89.36% was obtained using the similarity maps with window sizes of 64 samples, 128 samples and 256 samples simultaneously. Hence, this combination of similarity map features was used in all subsequent experiments.

e) Experiment 5: In this experiment we considered the effect of introducing derivative features in both C_1 and C_2. Table X shows the results. To facilitate gaining insight into the effect of different features on the discrimination between VT and VF, in the first eight rows of the table we show the results obtained with Parzen features in C_1 and various streams of features in C_2. We can observe that the largest improvement, from 81.25% to 86.92%, with respect to using just the waveform features in C_2, is achieved by introducing the similarity maps. This is not surprising considering that the similarity maps capture the regularity of abnormal heartbeats, which is critical for the classification of ventricular arrhythmias. The combination of the first derivative and Parzen features with similarity maps resulted in a further increase in average sensitivity to 87.22%. However, the inclusion of waveform and second derivative features caused a degradation in the average sensitivity. With the best performing set of features in C_2, i.e. x', the similarity maps and Parzen features, fixed, we considered the benefits of including derivative features in C_1 for improved discrimination between NVR and ventricular arrhythmias. The best average sensitivity of 89.68% was achieved when x, x' and x'' features were all combined with the Parzen features. The inclusion of similarity maps in C_1 degraded the accuracy (results not shown) as it increased the confusion rate between ventricular arrhythmias and NVR.

f) Experiment 6: Table XI illustrates the improved performance of the proposed similarity map features for VT/VF discrimination by comparing them with existing state-of-the-art methods. The similarity map is an averaged circularly shifted similarity map feature, where the window size is $l = 128$ samples and the similarity measure is the Euclidean distance. This part of the study focuses exclusively on the impact of these features in classifying VT and VF. Training and testing were conducted using only ECG segments labeled as VT or VF from

TABLE X

AVERAGE CONFUSION MATRICES OF THE PROPOSED HIERARCHICAL MODEL FOR DIFFERENT FEATURE COMBINATIONS

classifier C_1 features				classifier C_2 features					Diagnosed as				$SEN_{avg}(\%)$
x	x'	x''	Parzen	x	x'	x''	similarity maps	Parzen	Truth	VT(%)	VF(%)	NVR(%)	
			✓	✓					VT	76.92	19.87	3.22	81.25
									VF	18.86	68.55	12.59	
									NVR	0.58	1.15	98.27	
			✓	✓			✓		VT	82.69	14.10	3.22	86.92
									VF	8.29	79.12	12.59	
									NVR	0.39	1.34	98.27	
			✓	✓	✓				VT	76.50	20.29	3.22	82.15
									VF	15.72	71.69	12.59	
									NVR	0.19	1.54	98.27	
			✓	✓	✓		✓		VT	81.69	15.10	3.22	85.82
									VF	9.90	77.51	12.59	
									NVR	0.31	1.42	98.27	
			✓	✓	✓	✓			VT	76.25	20.54	3.22	81.40
									VF	17.73	69.68	12.59	
									NVR	0.50	1.23	98.27	
			✓	✓	✓	✓	✓		VT	77.49	19.30	3.22	85.12
									VF	7.82	79.59	12.59	
									NVR	0.66	1.07	98.27	
			✓		✓		✓	✓	VT	84.94	11.84	3.22	**87.22**
									VF	9.12	78.29	12.59	
									NVR	0.74	0.99	98.27	
			✓	✓	✓		✓	✓	VT	81.37	15.42	3.22	85.56
									VF	10.38	77.03	12.59	
									NVR	0.57	1.16	98.27	
✓	✓		✓		✓		✓	✓	VT	82.45	14.20	3.35	88.63
									VF	9.11	85.01	5.88	
									NVR	0.76	0.81	98.43	
✓	✓	✓	✓		✓		✓	✓	VT	81.46	15.96	2.58	**89.68**
									VF	7.27	89.29	3.45	
									NVR	0.61	1.11	98.28	
	✓	✓	✓		✓		✓	✓	VT	76.68	19.80	3.52	89.36
									VF	3.82	92.43	3.75	
									NVR	0.78	0.98	98.24	
✓	✓	✓			✓		✓	✓	VT	83.80	12.88	3.32	86.61
									VF	9.64	77.95	12.41	
									NVR	0.65	1.27	98.07	

Results derived from the test set.

TABLE XI

PERFORMANCE COMPARISON OF EXISTING FEATURES FOR IDENTIFICATION BETWEEN VT AND VF USING ECG SIGNALS

						Diagnosed as		
Model	Feature	Window Size	$SEN_{avg}(\%)$	$PRC_{avg}(\%)$	$F1_{avg}(\%)$	Truth	VT(%)	VF(%)
ResNet34 [10]	Selected features and Spectral Feature [6]	5s	76.68	74.26	74.54	VT	83.66	16.34
						VF	30.29	69.71
		8s	78.85	75.49	73.62	VT	83.68	16.32
						VF	25.97	74.03
Random Forest	Selected features [6]	5s	77.84	**81.92**	77.98	VT	92.17	7.83
						VF	36.5	63.50
		8s	68.83	75.81	67.31	VT	84.71	15.29
						VF	47.04	52.96
Linear Discriminant Analysis	Wavelets analysis [39]	5s	73.51	77.96	73.20	VT	91.47	8.53
						VF	44.45	55.55
		8s	75.11	73.41	70.70	VT	87.11	12.89
						VF	36.90	63.10
ResNet34 [10]	Pseudo Wigner Ville [41]	5s	75.78	72.70	70.85	VT	77.12	22.88
						VF	25.56	74.44
		8s	76.00	73.69	71.30	VT	73.25	26.75
						VF	21.26	78.74
ResNet34 [10]	FFREWT filter-bank [13]	5s	77.24	76.22	73.74	VT	80.50	19.50
						VF	26.03	73.97
		8s	78.66	75.02	71.53	VT	72.17	27.83
						VF	14.85	85.15
ResNet34 [10]	Proposed Similarity Map	5s	**84.73**	80.34	**80.61**	VT	84.40	15.60
						VF	14.94	85.06
		8s	81.59	76.55	76.97	VT	78.90	21.1
						VF	15.72	84.28

Results derived from the test set.

TABLE XII
AVERAGE CONFUSION MATRICES FOR THE PROPOSED MODEL'S GENERALIZATION PERFORMANCE TRAINED ON SOURCE DATASETS AND EVALUATED ON SEPARATE TARGET DATASETS

Source Datasets	Target Dataset	SEN_{avg} (%)		Diagnosed as		
			Truth	VT(%)	VF(%)	NVR(%)
MIBIH, VFDB, CUDB, EDB	AHADB	85.08	VT	90.39	0.92	8.70
			VF	20.59	75.49	3.92
			NVR	0.64	1.89	97.47
MIBIH, AHADB, EDB, VFDB	CUDB	86.96	VT	86.59	13.01	0.41
			VF	4.55	95.45	0.00
			NVR	14.29	6.91	78.80
MIBIH, AHADB, CUDB, EDB	VFDB	90.62	VT	92.22	4.12	3.66
			VF	11.03	88.97	0.00
			NVR	3.16	7.56	89.28

a relabeled dataset. The results are averages from 10 bootstrap resamples, ensuring consistency in the test sets used across the experiments.

In study [6], the authors proposed augmenting eight existing low-dimensional features which have been shown to achieve highest accuracy in ventricular arrhythmia classification [7], [8] with high-dimensional spectral features from 8-second ECG signals. We applied the 1D ResNet34 model to such augmented feature set and achieved an average sensitivity of 78.85%. Given the effectiveness and widespread use of the random forests model for engineered features, we employed the eight low-dimensional features considered in [6] with random forests. This approach yielded an averaged sensitivity of 79.21% with 5-second ECG segments. Study [39] utilized Singular Value Decomposition to extract two wavelet analysis features from ECG episodes, whilst employing an LDA model for classification; this approach attained an average sensitivity of 75.11% with 8-second ECG segments. Another study [40] proposed a pseudo-Wigner Ville time-frequency representation; this representation in combination with ResNet34 resulted in an average sensitivity of 76.00% with 8-second ECG segments. The FFREWT filter-bank features introduced in the study in [13] achieved an average sensitivity of 78.86% with a 1D ResNet34 model using 8-second ECG segments.

These comparisons highlight the improved accuracy of the ResNet34 model with similarity maps in ventricular arrhythmiaclassification, which achieved the highest averaged sensitivity of 84.96%. Even though the Random Forest model with the selected low-dimensional features achieves the highest average precision, it does not outperform our proposed method due to its poor sensitivity. Particularly critical is its VF sensitivity which is unacceptably low. Additionally, our proposed method achieves the highest averaged F1 score, indicating the best overall performance in balancing precision and sensitivity.

g) Experiment 7: Table XII illustrates the generalization performance of the proposed model trained on four source datasets and evaluated on separate target datasets, highlighting the model's ability to adapt to unseen data from a different distribution. The model achieves an average sensitivity of 85.08%, 86.96%, and 90.62% when trained on the combined datasets and tested on AHADB, CUDB, and VFDB, respectively. AHADB, CUDB, and VFDB were selected as target datasets because they included all types of episodes (VT, VF and NVR) required in this study, while MIBIH and EDB were not selected as they only contained VT and NVR episodes.

VII. Discussion

This study introduces a novel set of features, referred to as similarity maps, for discrimination between VT, VF and rhythms that are not VT or VF, and highlights the issue of data labeling, especially the discrepancies among experts concerning long recordings featuring a mix of arrhythmias. We addressed this by undertaking a systematic relabeling process. Our relabeling efforts have been proven to significantly improve our ability to train and test approaches to labeling ECG signals, allowing confidence in evaluating the benefit of incorporating similarity maps to detect recurring patterns in ECG segments. Employing similarity maps led to notable advancements in detecting VT and VF. This improvement was through the identification of repetitive patterns within electrocardiogram segments. We also performed a comparative analysis by re-implementing existing state-of-the-art features on the relabeled dataset. The results indicate that similarity maps improved sensitivities to VT and VF more than other existing features. In our research, we developed a 1D ResNet34 model to analyze similarity maps for the categorization of VT and VF. Additionally, we also discovered that utilizing pre-trained deep learning architectures such as VGG16 and EfficientNet, in conjunction with similarity maps, facilitates the differentiation of VT and VF. Our research also shows that combining derivative vectors with Parzen features effectively aids in identifying ventricular arrhythmias, laying a foundation for further enhancing the classification of VT and VF. It should be noted, however, that the hierarchical model has the potential for error propagation where mistakes at higher levels of the tree can cascade and remain uncorrected. While the hierarchical model offers a certain level of explanation in decision-making through the decision tree structure, the classification for each class depends on a black-box model that potentially hinders trust in healthcare applications. A future research direction is to explore the explainability of the proposed model.

There are ECG signal analysis platforms that are currently used to predict rhythm outcomes in heart failure patients, such as 'HeartLogic' [42]. It is too early to say whether such approaches may be of value in differential diagnosis of ventricular arrhythmias themselves. Looking ahead, it would be of potential value to incorporate the proposed VF discriminatory detection into a wearable device, as such devices are still limited in their ability to detect arrhythmias other than atrial fibrillation [43]. However, effective integration requires further research.

VIII. Conclusion

In the present study, we developed a technique to improve the differentiation between VT, VF, and NVR using ECG signals. The data used for the study were obtained from five public databases, but labeling inconsistencies were observed. Consequently, we relabeled the dataset according to the Lambeth Conventions (II) [5]. We first conducted initial experiments using ResNet34 [10], which demonstrated cardiologist-level accuracy in other arrhythmia detection tasks. However, that approach was not sufficient to achieve the highest accuracy in the task of classification between VT, VF, and NVR with the relabeled dataset. Therefore, we proposed a novel feature called similarity maps, which can identify and capture the repetition and similarities present within ECG segments, leading to a significant improvement in the classification accuracy between VT and VF. To further improve performance, we then combined the similarity maps with signal waveform, Parzen features and derivative features using a multi-stream network. Compared to ResNet34 [10], the state-of-the-art network that achieves cardiologist-level accuracy on other arrhythmia tasks, our method improved the average sensitivity from 77% to 90%.

References

[1] W. H. Organization, "The top 10 causes of death; 9 december 2020," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] T. Smith and M. Cain, "Sudden cardiac death: Epidemiologic and financial worldwide perspective," *J. Interventional Cardiac Electrophysiol.*, vol. 17, no. 3, pp. 199–203, 2006.

[3] J. van Rees et al., "Inappropriate implantable cardioverter-defibrillator shocks: Incidence, predictors, and impact on mortality," *J. Amer. College Cardiol.*, vol. 57, no. 5, pp. 556–562, 2011.

[4] F. Kette et al., "What is ventricular tachycardia for an automated external defibrillator?," *J. Clin. Exp. Cardiol.*, vol. 5, 2014, Art. no. 285.

[5] M. J. Curtis et al., "The Lambeth Conventions (ii): Guidelines for the study of animal and human ventricular and supraventricular arrhythmias," *Pharmacol. Therapeutics*, vol. 139, no. 2, pp. 213–248, 2013.

[6] Y. Alwan, Z. Cvetković, and M. J. Curtis, "Methods for improved discrimination between ventricular fibrillation and tachycardia," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 10, pp. 2143–2151, Oct. 2018.

[7] Q. Li, C. Rajagopalan, and G. D. Clifford, "Ventricular fibrillation and tachycardia classification using a machine learning approach," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1607–1613, Jun. 2014.

[8] F. Alonso-Atienza et al., "Detection of life-threatening arrhythmias using feature selection and support vector machines," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 3, pp. 832–840, Mar. 2014.

[9] J. Li et al., "Deep convolutional neural network based ECG classification system using information fusion and one-hot encoding techniques," *Math. Problems Eng.*, vol. 2018, 2018, Art. no. 7354081.

[10] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, 2019, Art. no. 65.

[11] G. T. Taye et al., "Application of a convolutional neural network for predicting the occurrence of ventricular tachyarrhythmia using heart rate variability features," *Sci. Rep.*, vol. 10, no. 1, pp. 1–7, 2020.

[12] A. Picon et al., "Mixed convolutional and long short-term memory network for the detection of lethal ventricular arrhythmia," *PLoS One*, vol. 14, no. 5, 2019, Art. no. e0216756.

[13] R. Panda et al., "Detection of shockable ventricular cardiac arrhythmias from ECG signals using FFREWT filter-bank and deep convolutional neural network," *Comput. Biol. Med.*, vol. 124, 2020, Art. no. 103939.

[14] B. Surawicz et al., "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part III: Intraventricular conduction disturbances a scientific statement from the American heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the American college of cardiology foundation; and the heart rhythm society endorsed by the international society for computerized electrocardiology," *J. Amer. College Cardiol.*, vol. 53, no. 11, pp. 976–981, 2009.

[15] D. Oglic et al., "A deep 2D convolutional network for waveform-based speech recognition," *Interspeech Int. Speech Commun. Assoc.*, pp. 1654–1658, 2020.

[16] A. Malali et al., "Supervised ecg wave segmentation using convolutional LSTM," *ICT Exp.*, vol. 6, no. 3, pp. 166–169, 2020.

[17] S. Banerjee, "A first derivative based R-peak detection and DWT based beat delineation approach of single lead electrocardiogram signal," in *Proc. 2019 IEEE Region 10 Symp.*, 2019, pp. 565–570.

[18] M. S. P. Balaji et al., "Revisiting derivative based methods on QRS detections from an ECG signal," in *Proc. 2021 Int. Conf. Advancements Elect. Electron. Commun. Comput. Automat.*, 2021, pp. 1–5.

[19] S. Banerjee, "A first derivative based R-peak detection and DWT based beat delineation approach of single lead electrocardiogram signal," in *Proc. 2019 IEEE Region 10 Symp.*, 2019, pp. 565–570.

[20] S. K. Mukhopadhyay et al., "ECG feature extraction using differentiation, hilbert transform, variable threshold and slope reversal approach," *J. Med. Eng. Technol.*, vol. 36, no. 7, pp. 372–386, 2012.

[21] N. M. Arzeno, Z. -D. Deng, and C. -S. Poon, "Analysis of first-derivative based QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 2, pp. 478–484, Feb. 2008.

[22] J. Arteaga-Falconi et al., "R-peak detection algorithm based on differentiation," in *Proc. IEEE 9th Int. Symp. Intell. Signal Process.*, 2015, pp. 1–4.

[23] 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.

[24] S. D. Greenwald, "The development and analysis of a ventricular fibrillation detector," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1986.

[25] A. Taddei et al., "The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *Eur. Heart J.*, vol. 13, no. 9, pp. 1164–1172, 1992.

[26] F. M. Nolle et al., "Crei-gard, a new concept in computerized arrhythmia monitoring systems," *Comput. Cardiol.*, vol. 13, pp. 515–518, 1986.

[27] M. C. Cieslak et al., "t-Distributed stochastic neighbor embedding (t-SNE): A tool for eco-physiological transcriptomic analysis," *Mar. Genomic.*, vol. 51, 2020, Art. no. 100723.

[28] L. O'Bray et al., "Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=tBtoZYKd9n

[29] S. C. Ramirez et al., "Dataset similarity to assess semi-supervised learning under distribution mismatch between the labelled and unlabelled datasets," *IEEE Trans. Artif. Intell.*, vol. 4, no. 2, pp. 282–291, Apr. 2023.

[30] A. Gretton et al., "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[31] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 770–778.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 2015 Int. Conf. Learn. Representations. Comput. Biol. Learn. Soc.*, 2015, pp. 1–14.

[33] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[34] M. Sandler et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[35] G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[36] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[37] D. Oglic et al., "Learning waveform-based acoustic models using deep variational convolutional neural networks," *IEEE/ACM Trans. Audio Speech, Lang. Process.*, vol. 29, pp. 2850–2863, 2021.

[38] K. He et al., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. 2015 IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[39] K. Balasundaram et al., "A classification scheme for ventricular arrhythmias using wavelets analysis," *Med. Biol. Eng. Comput.*, vol. 51, no. 1, pp. 153–164, 2013.

[40] A. Mjahad et al., "Ventricular fibrillation and tachycardia detection from surface ecg using time-frequency representation images as input dataset for machine learning," *Comput. Methods Programs Biomed.*, vol. 141, pp. 119–127, 2017.

[41] A. Rosado et al., "Fast non-invasive ventricular fibrillation detection method using pseudo Wigner-Ville distribution," in *Proc. Comput. Cardiol.*, Piscataway, New Jersey, USA, 2001, vol. 28, pp. 237–240.

[42] S. Kataoka et al., "Heartlogic multisensor algorithm response prior to ventricular arrhythmia events," *J. Arrhythmia*, vol. 39, no. 5, 2023, Art. no. 826.

[43] L. Neri et al., "Electrocardiogram monitoring wearable devices and artificial-intelligence-enabled diagnostic capabilities: A review," *Sensors*, vol. 23, no. 10, 2023, Art. no. 4805.