TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fcvm.2022.983543

Electrocardiogram classification using TSST-based spectrogram and ConViT

Pingping Bing[1]*, Yang Liu[2], Wei Liu[2], Jun Zhou[1]* and Lemei Zhu[1]*

[1]Academician Workstation, Changsha Medical University, Changsha, China, [2]College of Mechanical and Electrical Engineering, Beijing University of Chemical Technology, Beijing, China

OPEN ACCESS

EDITED BY
Gen-Min Lin,
Hualien Armed Forces General Hospital, Taiwan

REVIEWED BY
Mehmet Akif Ozdemir,
Izmir Kâtip Çelebi University, Turkey
Dr. Roshan Martis,
Global Academy of Technology, India
Abdellah Adib,
University of Hassan II
Casablanca, Morocco

*CORRESPONDENCE
Pingping Bing
bpping@163.com
Jun Zhou
15304690053@163.com
Lemei Zhu
zhulemei1228@163.com

SPECIALTY SECTION
This article was submitted to Cardiovascular Epidemiology and Prevention,
a section of the journal
Frontiers in Cardiovascular Medicine

RECEIVED 01 July 2022
ACCEPTED 22 September 2022
PUBLISHED 10 October 2022

CITATION
Bing P, Liu Y, Liu W, Zhou J and Zhu L (2022) Electrocardiogram classification using TSST-based spectrogram and ConViT.
Front. Cardiovasc. Med. 9:983543.
doi: 10.3389/fcvm.2022.983543

As an important auxiliary tool of arrhythmia diagnosis, Electrocardiogram (ECG) is frequently utilized to detect a variety of cardiovascular diseases caused by arrhythmia, such as cardiac mechanical infarction. In the past few years, the classification of ECG has always been a challenging problem. This paper presents a novel deep learning model called convolutional vision transformer (ConViT), which combines vision transformer (ViT) with convolutional neural network (CNN), for ECG arrhythmia classification, in which the unique soft convolutional inductive bias of gated positional self-attention (GPSA) layers integrates the superiorities of attention mechanism and convolutional architecture. Moreover, the time-reassigned synchrosqueezing transform (TSST), a newly developed time-frequency analysis (TFA) method where the time-frequency coefficients are reassigned in the time direction, is employed to sharpen pulse traits for feature extraction. Aiming at the class imbalance phenomena in the traditional ECG database, the smote algorithm and focal loss (FL) are used for data augmentation and minority-class weighting, respectively. The experiment using MIT-BIH arrhythmia database indicates that the overall accuracy of the proposed model is as high as 99.5%. Furthermore, the specificity (Spe), F1-Score and positive Matthews Correlation Coefficient (MCC) of supra ventricular ectopic beat (S) and ventricular ectopic beat (V) are all more than 94%. These results demonstrate that the proposed method is superior to most of the existing methods.

KEYWORDS

ECG classification, vision transformer, convolutional neural network, time-reassigned synchrosqueezing transform, class imbalance

Introduction

Electrocardiogram (ECG) is a diagnosis and treatment technology to detect cardiac physiological activities by extracting human skin electrode signal. By analyzing ECG signal, doctors are able to correctly diagnose various arrhythmias, and then help to judge myocardial infarction, myocarditis, myocardial ischemia, pericardial effusion and other diseases. Therefore, exploring the internal characteristics of ECG is of great significance for the timely diagnosis and treatment of arrhythmia diseases (1, 2).

In the past decade, with the development of artificial intelligence, many machine learning methods mainly based on feature extraction and modal classification have achieved fruitful results in the application of ECG analysis. The works for ECG feature extraction include digital filtering (3), group optimization (4) and time-frequency analysis (5–8). Ozbay et al. combined the fuzzy C-means clustering algorithm (FCMA) and discrete wavelet transform to extract the key feature of ECG signal (9). Alickovic and Subasi used the multi-scale principal component analysis (PCA) to denoise ECG signal, and further extracted feature through autoregressive model (10). Azia et al. (11) applied empirical mode decomposition (EMD) and support vector machine (SVM) to region of interest extraction and signal denoising. In (12), the wavelet transform was utilized for data preprocessing, and then the PCA was added to project it to the lower dimensional feature space with particle swarm optimization. Marinho et al. (13) explored the combined advantages of different feature extraction methods and several classical machine learning models, and evaluated the actual achievements of Fourier transform, gerzel algorithm, higher order statistics and structural co-occurrence matrix on four types of perceptron: support vector machine, multi-layer perceptron, naive bayes model and optimum-path forest. Coast et al. (14) used the hidden Markov models to analyze cardiac arrhythmia. Osowski et al. (15) utilized the support vector machine to recognize heartbeat. Yeh et al. (16) developed a clustering method to identify ECG signal with arrhythmia. Park et al. (17) proposed the logistic regression to automatically classify the ECG interval characteristics. Li and Min (6) completed ECG classification by combining wavelet packet transform and random forests. In summary, the most commonly used machine learning methods include hidden Markov model (14), support vector machine (13, 15), clustering algorithm (16, 17), logistic regression (18), random forest (6, 19) and naive Bayes (13, 20, 21). However, the above-mentioned techniques have many limitations in practical application; for instance, they rely heavily on manual feature extraction and require a lot of time and expertise.

In recent years, due to the end-to-end learning convenience of deep learning technique, it has also made great progress in ECG classification. Kiranyaz et al. (22) introduced a 1-D convolution neural network (CNN) to deal with ECG arrhythmia classification task. Li et al. (23) presented the general regression neural network to extract correlation patterns from ECG signal. On the basis of CNN, Acharya et al. (24) added data augmentation and noise filtering technique to strengthen fitting ability of the model. Sellami and Hwang (25) paid more attention to the problem of class imbalance, and showed the solicitude for the classification of various samples in batch processing through batch-weight loss. Atal and Singh (26) developed the deep CNN, modified by rider optimization algorithm, to implement the automatic classification of ECG. In addition, some studies used the practice of machine learning for reference and combined TFA with deep learning model, which greatly improved the accuracy and robustness of the model. In order to make full use of spatial information of 2-D image, Huang et al. (7) transformed the time-domain ECG signal into time-frequency domain by STFT, and then fed the time-frequency map to the neural network as input feature. Wang et al. (27) employed continuous wavelet transform (CWT) to implement preprocessing and designed a CNN framework to achieve the automatic ECG classification from 2-D spectrum. To pursue a more readable TFR as input feature, Ozdemir et al. (28) proposed a new method for detecting and predicting seizure based on synchrosqueezing transform (SST) and CNN. Furthermore, the enhancement of TFA methods, such as STFT, CWT and Hilbert-Huang Transform (HHT), for hand gesture intelligent classification was discussed in (29). An important conclusion is that the time-frequency resolution of 2-D spectrum has a direct influence on the classification based on deep learning model. Nevertheless, these methods mentioned often simply transform the representation of ECG time-domain signal, and lack of deep excavation of its characteristics, so as to introduce a preprocessing technique in line with its attributes. Besides, the deep learning model such as deep CNN is subject to the problem of network degradation, in which the training sets are easy to be saturated due to the complexity of the deep model, and are limited by the hard inductive bias of pure convolution layers, resulting in insufficient data information mining. Finally, most of the existing studies on ECG classification do attach importance to the class imbalance in applied database, the number of normal heart rate sample is often hundreds of times that of abnormal, which will produce serious over fitting problem.

In this study, since the signal characteristics corresponding to arrhythmia are usually reflected in the pulse of ECG, a TFA technique called time-reassigned synchrosqueezing transform (TSST) which can highlight the characteristics of pulse signal that will be used to extract ECG information, which transforms ECG in the time domain into time-frequency domain with the high frequency resolution. Then, the two-dimensional signal is transformed into picture and input into the convolutional vision transformer (ConViT) for classification. Aiming at the class imbalance problem mentioned previously, the smote algorithm is adopted to synthesize some small sample data for soft balance, and the focal loss (FL) is performed to further make up for the defect of class imbalance. The contributions of this paper are expressed as follows: (1) the TSST is employed for ECG data preprocessing to make full use of pulse information; (2) the ConViT with convolutional architecture and self-attention mechanism is used for ECG classification; (3) the smote algorithm and FL are adopted to deal with the ECG class imbalance problem.

The rest of this paper is organized as follows. Section Theory describes the fundamental principle of TSST algorithm, ConViT framework and treatments of imbalance problem. In Section

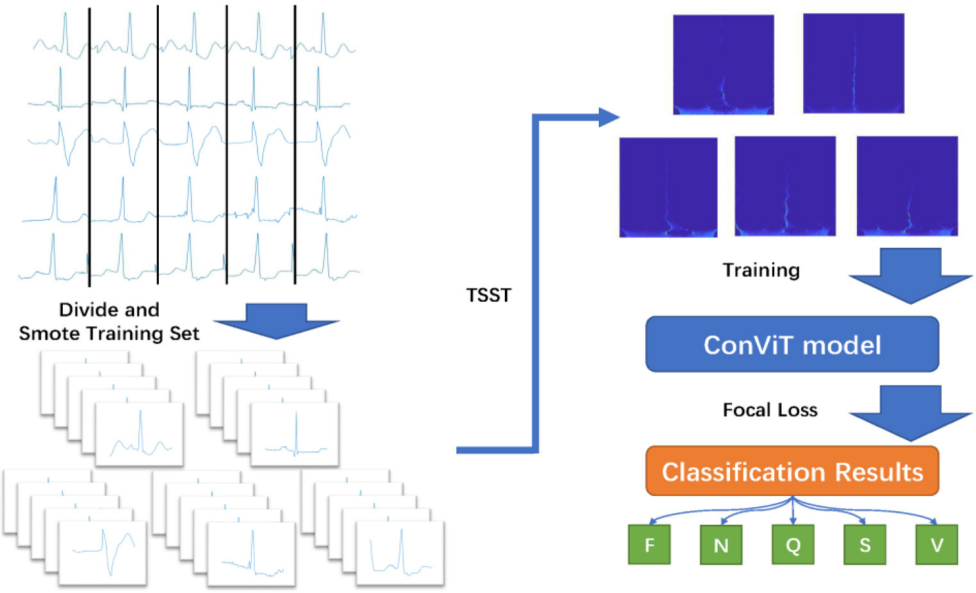


FIGURE 1
Flow chart of ECG classification based on TSST and ConViT.

Experiment, the experimental results and discussions are shown. The conclusions are drawn in Section Discussion.

Theory

Method overview

The overall framework of the proposed ECG classification method in the paper is shown in Figure 1. The test data comes from MIT-BIH arrhythmia database (30). According to the R-wave position in the annotation file, a total of 300 points within the selected interval are taken as a time domain sample, and the data are enhanced by a small number of samples in the training set. Then, the TSST is utilized to transform the one-dimensional time-domain signal into two-dimensional time-frequency map, which will be input into ConViT with FL. Under the recommendations from Association for the Advancement of Medical Instrumentation (AAMI) (31), we will divide the original samples into five categories: fusion (F), non-ectopic beat (N), unknown (Q), supra ventricular ectopic beat (S) and ventricular ectopic beat (V), showing in Table 1, for the model processing.

Time reassigned synchrosqueezingtransform

TSST is a newly developed time-frequency decomposition algorithm (32). It reassigns the time-frequency coefficients along the time direction by calculating the group-delay estimator, so as to extract the transient characteristic of pulse signal, which is highly suitable for processing ECG signal. The definition and property of TSST are stated below.

TABLE 1 Details of MIT-BIH arrhythmia database.

AAMI heartbeat class	MIT-BIH heartbeat type	MIT-BIH arrhythmia label
F	fVN	F
N	N, LBBB, RBBB, AE, NE	N, L, R, e, j
Q	P, fPN, U	/, f, U
S	AP, aAP, NP, SP	A, a, J, S
V	PVC, VE	V, E

The STFT of a signal$x$is defined as a function of time t and frequency ω computed with a Gaussian window$g$.

$$F_x^g(t,\omega)=\int_{-\infty}^{+\infty}x(\tau)\,g^*(t-\tau)e^{-j\omega\tau}d\tau \qquad (1)$$

where $g(t)=1/\sqrt{2\pi}\ e^{-t^2/2}$, and g^* denotes the complex conjugate of g. The time-frequency representation (TFR) corresponds to $\left|F_x^g(t,\omega)\right|^2$.

In order to further improve the resolution of TFR, a time reassignment step moves the energy of the signal according to the map $(t,\omega)\rightarrow\left(\hat{t}_x(t,\omega),\omega\right)$, herein, $\hat{t}_x(t,\omega)$ is the group

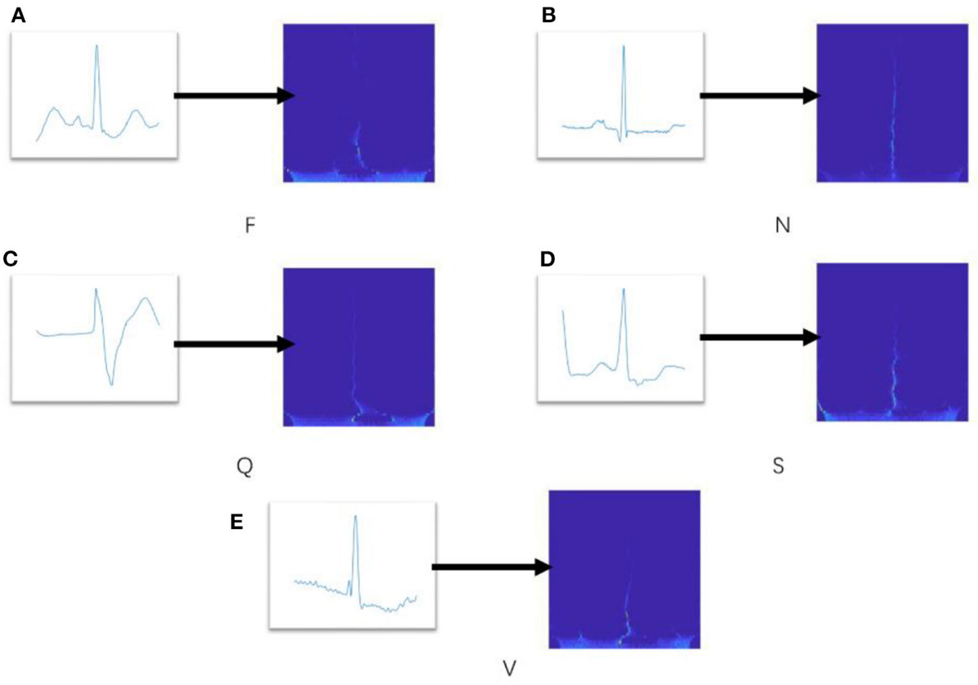


FIGURE 2
(A–E) Spectrograms of several ECG signals via TSST decomposition.

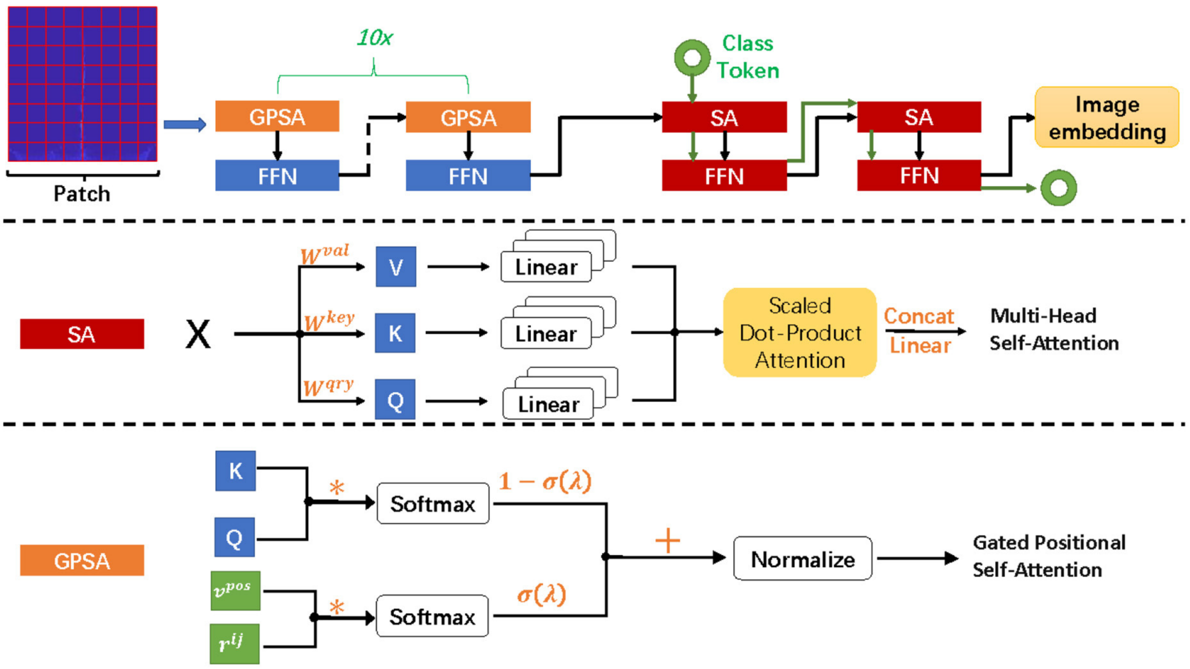


FIGURE 3
Framework of ConViT and the details of SA and GPSA.

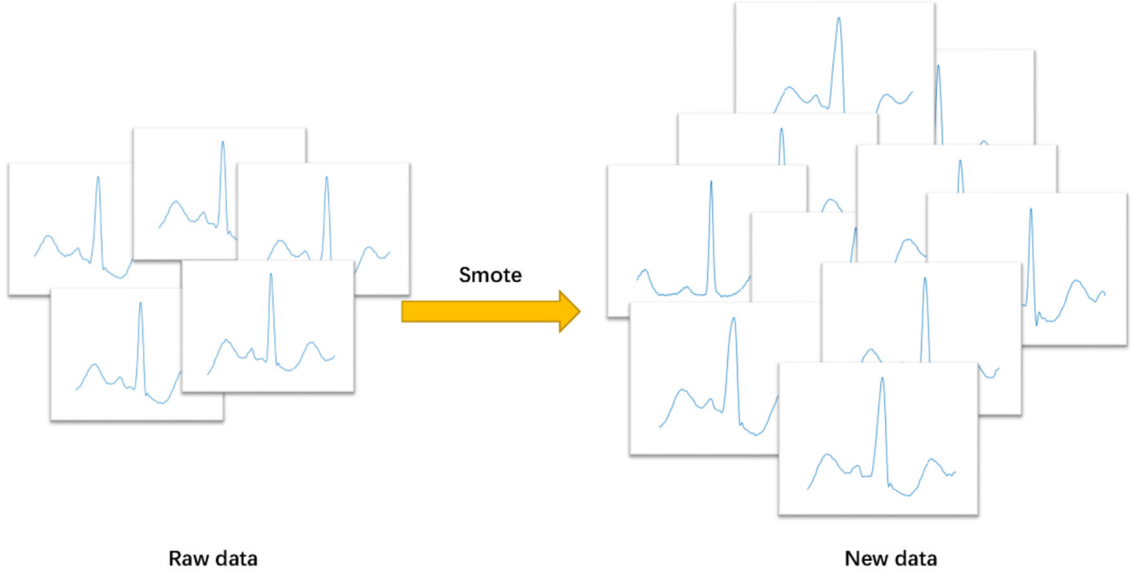


FIGURE 4
Smote result of class F samples.

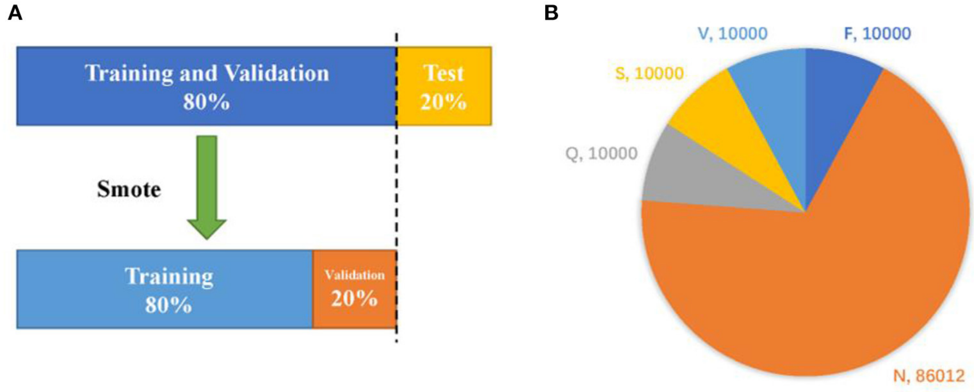


FIGURE 5
Dataset division strategy **(A)** and the quantity of samples before and after augmentation **(B)**.

delay estimation mentioned above. The time reassignment operator $\hat{t}$ can be deduced as:

$$\hat{t}_x(t,\omega) = R\left(t - \frac{F_x^{\tau g}(t,\omega)}{F_x^g(t,\omega)}\right) \quad (2)$$

where $R(Z)$ stands for the real part of Z, $\tau g(t) = tg(t)$ is a modified version of the Gaussian window function g.

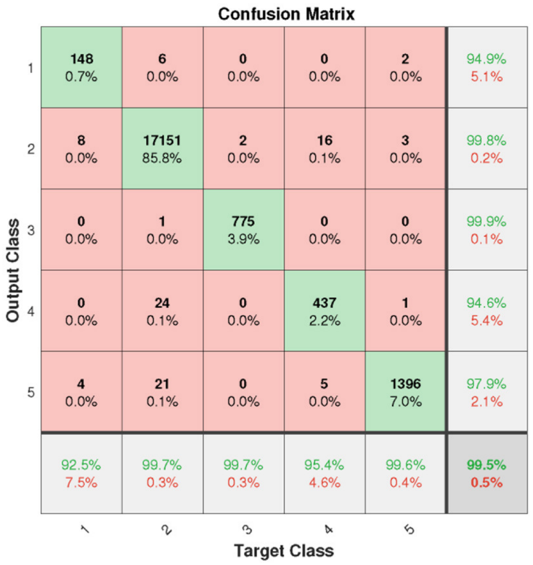
Therefore, TSST can be written as:

$$S_x^g(t,\omega) = \int_{-\infty}^{+\infty} F_x^g(t,\omega)\,\delta\left(t - \hat{t}_x(t,\omega)\right) d\tau \quad (3)$$

Next, the spectrogram $\left|S_x^g(t,\omega)\right|^2$ will be saved as picture and fed into the ConViT model as input sample. Figure 2 shows the spectrogram results, in which five representative time-domain ECG signals are transformed into two dimensional spectrograms through TSST. It can be seen that these spectrograms are characterized by high resolution in the time dimension, which is very beneficial for extracting the transient characteristics of ECG arrhythmia.

Convit structure

ConViT combines the advantages of two popular neural network frameworks, CNN and Transformer (33–36), which overcomes the shortcomings of low performance upper limit



Output Class \ Target Class	1	2	3	4	5	
1	148 0.7%	6 0.0%	0 0.0%	0 0.0%	2 0.0%	94.9% 5.1%
2	8 0.0%	17151 85.8%	2 0.0%	16 0.1%	3 0.0%	99.8% 0.2%
3	0 0.0%	1 0.0%	775 3.9%	0 0.0%	0 0.0%	99.9% 0.1%
4	0 0.0%	24 0.1%	0 0.0%	437 2.2%	1 0.0%	94.6% 5.4%
5	4 0.0%	21 0.1%	0 0.0%	5 0.0%	1396 7.0%	97.9% 2.1%
	92.5% 7.5%	99.7% 0.3%	99.7% 0.3%	95.4% 4.6%	99.6% 0.4%	99.5% 0.5%

FIGURE 6
Confusion matrix of test set (1: F, 2: N, 3: Q, 4: S, 5: V).

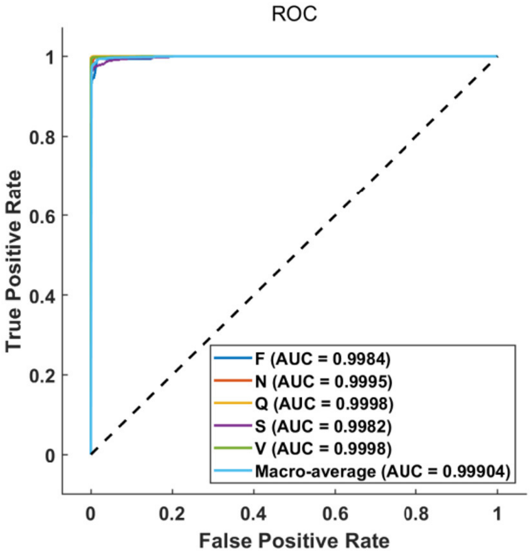


FIGURE 7
ROC of classification result and their AUCs.

caused by hard induction bias of CNN and the high dependence of Transformer on data. In the paper, the gated positional self-attention (GPSA) is employed to balance convolution and self-attention (SA) in a soft way, and its framework is shown in Figure 3. ConViT is based on vision transformer and consists of twelve propagation blocks composed of a SA layer and a two-layer feedforward network (FFN) with Gelu activation (see Figure 3). The difference is that the SA layer in the first ten blocks is replaced by GPSA layer, and the settings of SA layer are still retained in the last two blocks. In addition, the L2 regularization and dropout mechanism are applied in FNN to counter overfitting. Since the ECG spectrum is relatively simple, we set the input image with the size of 160 to 8 x 8 non-overlapping blocks of 20 x 20 pixels, and the embedding matrix dimension is 12.

For the SA layer, the essence of self-attention mechanism is to selectively manage the input through attention pooling. For single head self-attention with position, we can define it as PSA_h, and $MHSA$ performs concat and linear operations on SA_h:

$$PSA_{ij}^{h}(K,Q,V) := V^{h}\text{softmax}\left(\frac{K_i^{hT}Q_j^{h}}{\sqrt{d}} + \upsilon_{pos}^{hT} r_{ij}\right) \qquad (4)$$

$$MHSA := {}_{h\in[N_h]}^{concat}\left[SA_h(K,Q,V)\right]W^{out} + b^{out} \qquad (5)$$

where $\text{softmax}(X)_{ij} = \frac{e^{X_{ij}}}{\sum_k e^{X_{ik}}}$.

The input image is divided into multiple patches and represented as $X \in R^{D_{emb}\times N}$by embedding matrix processing. Therefore, we have $K = W^{key}X$, $Q = W^{qry}X$and $V = W^{val}X$, here W^{key}, W^{qry}, $W^{val} \in R^{D\times Demb}$, N_h is the number of attention head. Trainable embedding υ_{pos}^{h} and relative position coding r_{ij} are added to discipline position information. Then, $D_{emb} = N_h D$, $W^{out} \in R^{D_{emb}\times D_{emb}}$, $b^{out} \in R^{D_{emb}\times D}$. In (37), a PSA layer with N_h heads and a relative positional encoding of dimension $D_p \geq 3$ can express any convolutional layer with filter size of $\sqrt{N_h} \times \sqrt{N_h}$.

$$\begin{cases} \upsilon_{pos}^{h} := -\alpha^{h}\left(1, -2\Delta_1^{h}, -2\Delta_2^{h}\right) \\ r_{\delta} = \|\delta\|^{2}, \delta_1, \delta_2 \\ W^{key}, W^{qry} := 0, \quad W^{val} = I \end{cases} \qquad (6)$$

where α^h and Δ_1^h, Δ_2^h determine the width and center of each attention head, respectively. (δ_1, δ_2) is a fixed value used to define the relative offset of K and Q.

Hence, each attention head only extracts local information to achieve the effect of convolution. However, this generalized convolution operation is difficult to be carried out on ViT, so GPSA is modified to allow it to decide whether to maintain convolution.

$$GPSA^{h}(K,Q,V) := V^{h}\text{normalize}\left[A^{h}\right] \qquad (7)$$

$$\begin{aligned} A_{ij}^{h} := &\left(1-\sigma\left(\lambda_h\right)\right)\text{softmax}\left(K_i^{hT}Q_j^{h}\right) \\ &+ \sigma\left(\lambda_h\right)\text{softmax}\left(\upsilon_{pos}^{hT} r_{ij}\right) \end{aligned} \qquad (8)$$

where $\left(\text{normalize}\left[A^{h}\right]\right)_{ij} = \frac{A_{ij}}{\sum_k A_{ik}}$and $\sigma(Z) = \frac{1}{1+e^{-Z}}$.

The gating parameter λ is learned through the model, which is utilized to balance content-based self-attention and convolution initialization position self-attention, so as to achieve the effect of soft inductive bias.

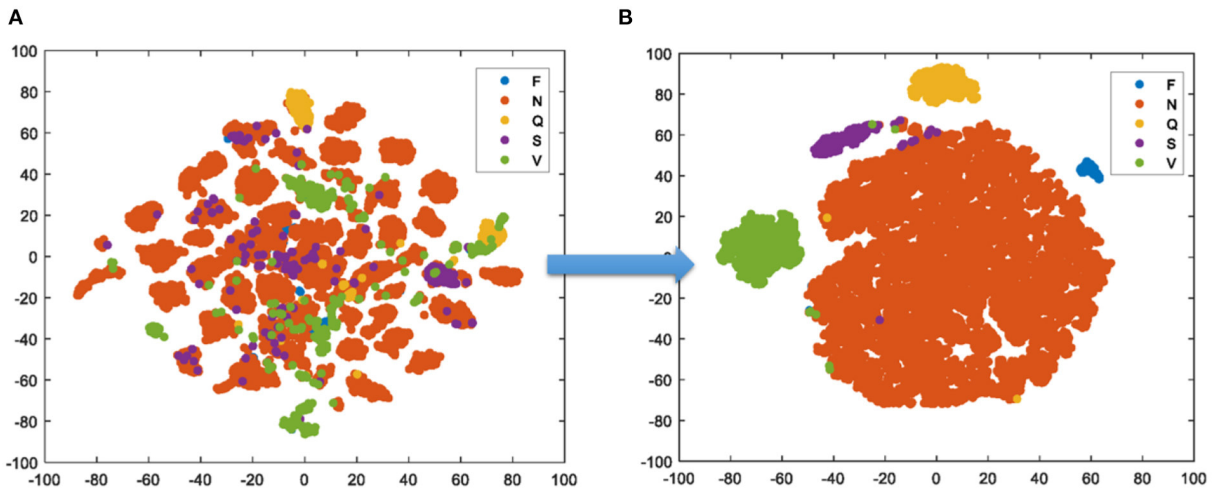


FIGURE 8
t-SNE results of input samples **(A)** and output samples **(B)**.

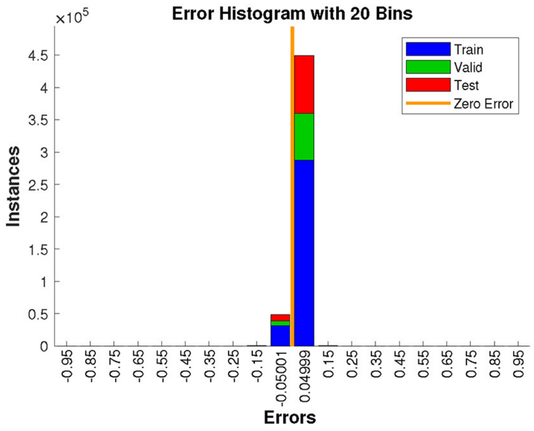


FIGURE 9
Error histogram (errors = output − target).

Treatment of class imbalance

In the actual situation, the amount of normal heart rate data is much larger than that of arrhythmia data. The problem caused by class imbalance is that the easy positive samples will make a major contribution to loss and dominate the update direction of the gradient. Hence, the model is unable to learn valid information for correct classification. In this paper, we introduce the smote algorithm and FL to combat it (38, 39). The former artificially generates a large number of scarce samples, and the latter pays attention to the samples that are difficult to be classified.

Based on the k nearest neighbor points of each sample, smote algorithm randomly selects N adjacent points to multiply the difference by a threshold in the range of [0, 1], so as to achieve the purpose of synthesizing data. The core of this algorithm is that the feature of adjacent points in feature space is similar. It does not sample in the data space, but in the feature space, so its accuracy will be higher than the traditional sampling method. Figure 4 shows the data enhancement result of smote algorithm for class F samples. The formula for constructing new sample is as follows:

$$Z_{new} = Z + \text{rand}\,(0, 1) * |Z - Z_r| \tag{9}$$

where Z indicates the original sample, and $Z_r$is the adjacent value randomly selected.

FL can be regarded as a loss function, which reduces the weight of samples easy to classify and increases the weight of samples difficult to classify. It focuses on training a sparse set of difficult samples. For multi-class classification task, FL can be defined as:

$$FL\left(p_t\right) = -\left(1 - p_t\right)\log\left(p_t\right) \tag{10}$$

$$p_t = \begin{cases} x = p & y = 1 \\ y = 1 - p & y \neq 1 \end{cases} \tag{11}$$

where p_t represents the probability predicted by the model as class t,p is the probability that the sample to be classified as positivity, and y is the output of the model. γ can adjust the rate of weight reduction of easy samples. The larger the γ, the more the loss of easy sample will be suppressed. It is worth noting that when $\gamma = 0$, FL is equal to the cross-entropy loss. In this example, $\gamma = 2$.

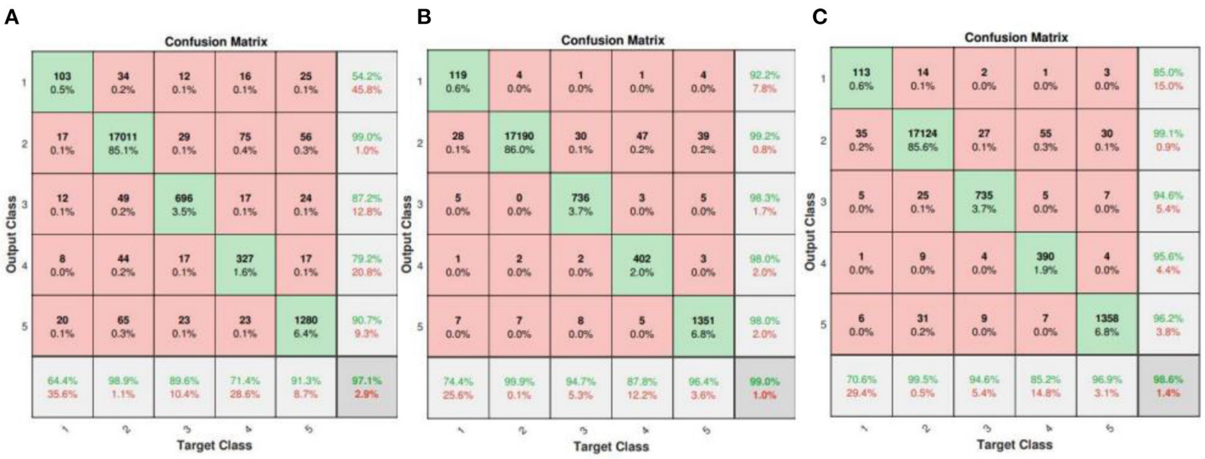


FIGURE 10
Confusion matrices of ConViT without TSST **(A)**, FL **(B)** and smote algorithm **(C)**, respectively.

TABLE 2 Classification comparison of classes S and V.

Approach	S				V				Data size
	Acc	Spe	F1-Score	MCC	Acc	Spe	F1-Score	MCC	
CNN (22)	96.6	98.1	63.2	61.5	98.4	98.7	91.4	90.6	49,557
CNN Aug (24)	97.0	98.6	92.4	90.6	97.9	98.8	94.8	93.5	452,960
2-D CNN (40)	99.3	99.7	83.4	83.1	99.3	99.6	93.6	93.3	12,548
GRNN (23)	97.4	98.9	90.2	88.8	98.4	99.4	88.8	88.0	49,661
ResNet (41)	98.8	99.9	98.4	97.3	99.4	99.9	99.7	99.7	49,564
LSTM (27)	99.3	99.6	83.4	83.1	99.3	99.5	93.6	93.3	27,789
Proposed	99.7	99.9	95.0	94.9	99.7	99.7	97.7	97.5	20,000

TABLE 3 Training parameters.

Learning rate	Batch size	Epoch	Embed dimension	Dropout rate	Decay rate
1e-4	32	120	12	0.5	0.02/10epoch

Experiment

Dataset description

In this paper, we employ MIT-BIH arrhythmia database to test the effectiveness of the proposed model, which includes 48 and a half hours of dual channel ambulatory ECG records of 47 subjects, with a sampling frequency of 360Hz and independent annotation by more than two experts.

In this example, we randomly divide the database into three parts. Firstly, the whole data is divided into training plus verification set and test set in the ratio of 8 to 2, then the former is augmented by smote algorithm and divided into training set and verification set in the same proportion. The data set division diagram and the number of samples (before and after data augmentation) (Table 1) are shown in Figure 5.

Evalution

In order to further assess the validity of the proposed model in ECG classification task, the results of the test set are evaluated in terms of accuracy (Acc), sensitivity (Sen), specificity (Spe) Positive predictive value (Ppv), $F1$-Score and Matthews Correlation Coefficient (MCC), which are expressed as follows.

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

TABLE 4 Classification results of PTB database.

Approach	Acc	Sen	Spe	Ppv	F1-Score
KNN (42)	–	92.3	88.1	–	–
HMM with GMM (43)	82.5	85.7	79.8	–	–
ANN (44)	95.6	93.3	97.9	99.3	96.2
CNN (45)	93.5	93.7	92.8	98.0	95.8
ResNet (46)	92.6	93.2	92.0	–	–
Proposed	94.6	93.6	92.1	95.9	94.0

$$Sen = \frac{TP}{TP + FN} \tag{13}$$

$$Spe = \frac{TN}{TN + FP} \tag{14}$$

$$Ppv = \frac{TP}{TP + FP} \tag{15}$$

$$F1 = \frac{Ppv \times Sen}{Ppv + Sen} \tag{16}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \tag{17}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively.

Result and discussion

In this section, the results will be discussed by means of confusion matrix, receiver operating characteristic curve (ROC), t-distributed stochastic neighbor embedding (t-SNE) and error histogram. Figure 6 shows the confusion matrix from the test set based on the proposed model. It can be clearly seen that the overall accuracy of our model is as high as 99.5%. However, due to the influence of FL on the weight of a small number of sample classes, the most class objects (class N) are probably incorrectly classified.

The ROC curve in Figure 7 further illustrates the relationship between false positive rate (FPR) and true positive rate (TPR) of various classes. As can be observed, the performance of classes F and S is slightly poor owing to the small number of samples, the ROC curves of other classes are almost perfect. Nevertheless, all the area under curves (AUCs) are larger than 0.99, which indicates that the proposed method can achieve a satisfactory classification result.

In Figure 8, the t-SNE gives the visualization result of the test set. It creates a compressed feature space, in which the similar samples are represented by the nearby points and the dissimilar samples are represented by far points with the high probability. Then, the Kullback Leibler divergence between the two distributions about the location of embedded points is minimized. Finally, the high-dimension data is simplified into a low-dimension graph with the affluent original information. One can clearly see that benefit from the feature extraction of TSST, the samples have been scattered well in space before the training, the proposed model achieves the excellent classification after the training.

In addition, Figure 9 plots the error histogram, it shows that the proposed model has less prediction error, which further demonstrates the superior performance of the presented method.

On the other hand, the confusion matrix results of ConViT without TSST (each 1D ECG signal is simply stacked into 2D image), FL and smote algorithm respectively are given in Figure 10. It can be clearly seen that the overall performance of ConViT is far inferior to the scenario with TSST, which is likely due to the fact that the information from single time series is not enough to achieve the excellent classification. In addition, the scenarios without FL and smote algorithm, shown Figures 10A,C, indicate that the ConViT without balance processing generates a bias where the data is classified into N categories. Therefore, it is concluded that the classification result of few-shot without the above mentioned tricks is poor.

Discussion

In this section, we apply our model to classification of classes S and V for comparison with other state-of-the-art methods in terms of Acc, Sen, Spe, F1-score and MCC, which is shown in Table 2. Note that the test set used in the paper consists of 20,000 beats of ECG. As illustrated in Table 2, the proposed method performs clearly better, with higher precision, which mainly benefits from the following three aspects: (1) TSST achieves the effective feature extraction on ECG signal; (2) FL and somte algorithm alleviate the conflict between the differences in various sample number; (3) Deep mining of input information by attention mechanism of ViT architecture and the CNN structure can ensure the property of small sample task. It should be mentioned that the proposed model implements 120 epochs on NVIDIA GeForce RTX 2060 about 9640s, which is suitable for a 2-D visual model with attention mechanism. Benefit from the ConViT, the model with multi-head attention mechanism can perform the fast iteration. Note that some important training parameters are listed in Table 3.

To further verify the robustness of the proposed method, we apply the trained model with binary-classification (normal and others) to PTB database (47). The dataset contains 549 records of 290 subjects with 12 leads, which records the diseases including myocardial infarction (MI), cardiomyopathy/Heart failure, bundle branch block, dysrhythmia, myocardial hypertrophy, valvular heart disease, myocarditis, miscellaneous, healthy controls (normal). Each channel is sampled at the frequency of 1 kHz with 16-bit resolution. In this experiment, we apply ECG lead II data to TSST for test, which is focused on MI

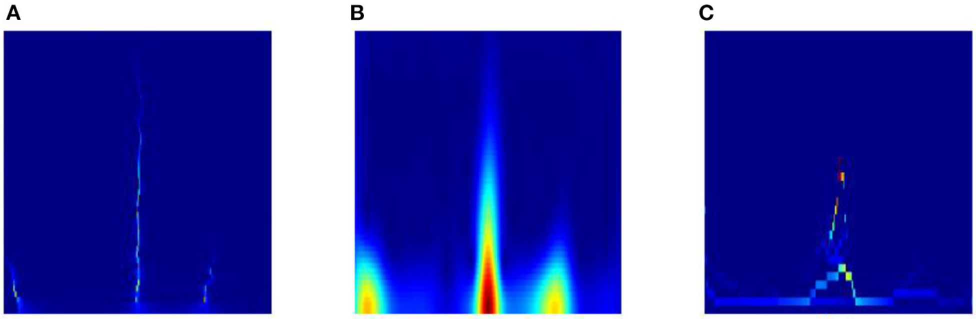


FIGURE 11
TFR of class S based on **(A)** TSST, **(B)** STFT and **(C)** EMD.

TABLE 5 Comparison results of TSST-, STFT- and EMD-based ConViT methods.

TFA	S				V			
	Acc	Spe	F1-Score	MCC	Acc	Spe	F1-Score	MCC
TSST	99.7	99.9	95.0	94.9	99.7	99.7	97.7	97.5
STFT	95.6	96.8	92.3	91.6	95.9	97.0	95.2	95.0
EMD	92.1	90.2	87.6	85.3	93.3	95.1	89.6	88.1

and healthy control data. The comparison results are listed in Table 4 Although not all indexes in the result of the proposed method are optimal, its overall performance is very competitive for an unseen dataset. The Acc of 94.6 is sufficient for MI diagnosis, which demonstrates the generalization of the proposed method again.

Third, we also list the results of class S based on TSST and traditional time-frequency analysis methods (e.g. STFT and EMD) in Figure 11. It is obvious that the TSST achieves a highly energy-concentrated TFR and highlights the pulse characteristics of ECG well compared with STFT, which helps to reduce some unnecessary convolution operations in the GPSA layer. Due to the existence of pulse points in ECG signal, EMD is easy to cause mode aliasing, as shown in the Figure 11(C), which is not conducive to feature extraction. In addition, the comparison results of TSST-, STFT- and EMD-based ConViT approaches for ECG classification using MIT-BIH dataset are shown in Table 5. The accuracy of ECG classification using TSST-based ConViT is 99.7%, which is obviously higher than STFT-based (95.6%) and EMD-based methods (92.1%). Similarly, the metrics, such as Spe, F1-Score and MCC, TSST-based ConViT also obtain the optimal values. The experiment indicates that TSST is a reliable technique for non-stationary signal, with pulse feature, processing and ECG classification in ConViT.

Actually, there are still some issues that need to be solved in the future. The first one is the adaptability of smote algorithm, traditionally used for 2-D image augmentation, for time series signals. Although the experiment (Figure 10) indicates that smote algorithm can improve ECG classification, the relevant research work is still lacking. The second one is about overfitting problem. We utilize some anti-overfitting strategies, such as L2 regularization and dropout, in the paper, but there are some differences in the classification performance for MIT-BIH and PTB datasets. Finally, more comparative experiments on the combination of TSST and deep learning models like (48) are needed, which can further illustrate the advantages of the proposed model, and this is also our future research direction.

Conclusion

In this study, we propose a novel ECG classification method, it achieves the overall accuracy of 99.5% and does a better job classifying ECG signal compared to the traditional methods. With this method, the TSST transforms one-dimension ECG signal to two-dimension time-frequency map for characterizing the pulse characteristics of arrhythmia signal. The classifier performs smote algorithm and FL to deal with the class imbalance phenomenon. The former enhances the data by feature space sampling, and the latter ensures the classification ability by increasing the weight for a few class samples. In addition, as the main architecture of the model, on the one hand, ConViT utilizes multi-head attention mechanism of Transformer for image processing to make full use of the internal related information of the input. On the other hand, the hard induction bias of CNN enables the model to achieve good result with a few samples, and greatly improves the training speed.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for this study in accordance with the local legislation and institutional requirements.

Author contributions

PB: Conceptualization and software. LZ: Validation and formal analysis. JZ: Writing—review and editing and supervision. YL: Methodology and formal analysis. WL: Writing—original draft and writing—review and editing. All authors contributed to the article and approved the submitted version.

Acknowledgments

This work was supported by young backbone teachers of Hunan province training program foundation of Changsha Medical University (Hunan Education Bureau Notice 2021 No. 29–26).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Lilly LS, Braunwald EL. "Braunwald's heart disease: a textbook of cardiovascular medicine," in *IEEE Access*, vol. 2. Amsterdam: Elsevier Health Sciences (2012).

2. Wang EK, Zhang X, Pan LY. Automatic classification of CAD ECG signals with SDAE and bidirectional long short-term network. *IEEE Access.* (2018) 6:42207—15. doi: 10.1109/ACCESS.2019.2936525

3. Mahmoud SA, Bamakhramah A, Al-Tunaiji SA. Six order cascaded power line notch filter for ECG detection systems with noise shaping. *Circ Syst Signal Process.* (2014) 33:2385–400. doi: 10.1007/s00034-014-9761-1

4. Garcia G, Moreira G, Menotti D, Luz E. Inter-patient ECG heartbeat classification with temporal VCG optimized by PSO. *Sci Rep.* (2017) 7:1–11. doi: 10.1038/s41598-017-09837-3

5. Kabir MA, Shahnaz C. Denoising of ECG signals based on noise reduction algorithms in EMD and wavelet domains. *Biomed Signal Process Control.* (2012) 7:481–9. doi: 10.1016/j.bspc.2011.11.003

6. Li T, Min Z. ECG classification using wavelet packet entropy and random forests. *Entropy.* (2016) 18:285. doi: 10.3390/e18080285

7. Huang J, Chen B, Yao B, He W. ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network. *IEEE Access.* (2019) 7:92871–80. doi: 10.1109/ACCESS.2019.2928017

8. Pokaprakarn T, Kitzmiller RR, Moorman R, Lake DE, Ashok AK, Kosorok M. Sequence to sequence ECG cardiac rhythm classification using convolutional recurrent neural networks. *IEEE J Biomed Health Inform.* (2012) 26:572–80. doi: 10.1109/JBHI.2021.3098662

9. Özbay Y, Ceylan R, Karlik B. Integration of type-2 fuzzy clustering and wavelet transform in a neural network based ECG classifier. *Expert Syst Appl.* (2011) 38:1004–10. doi: 10.1016/j.eswa.2010.07.118

10. Alickovic E, Subasi A, Effect of multiscale PCA de-noising in ECG beat classification for diagnosis of cardiovascular diseases. *Circ Syst Signal Process.* (2015) 34: 513–533, doi: 10.1007/s00034-014-9864-8

11. Aziz S, Khan MU, Choudhry ZA, Aymin A, Usman A. "ECG based biometric authentication using empirical mode decomposition and support vector machines," in *IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).* (2019), pp. 906–912. doi: 10.1109/IEMCON.2019.8936174

12. Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of ECG signals. *IEEE Trans Biomed Eng.* (2009) 56:1415–26. doi: 10.1109/TBME.2009.2013934

13. Marinho LB, Nascimento NDMM, Souza J, Gurgel MV, Reboucas Filho PP, De Albuquerque VHC. A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Future Gener Comput Syst.* (2019) 97:564–77. doi: 10.1016/j.future.2019.03.025

14. Coast DA, Stern RM, Cano GG, Briller SA. An approach to cardiac arrhythmia analysis using hidden Markov models. *IEEE Trans Biomed Eng.* (1990) 37:826–36. doi: 10.1109/10.58593

15. Osowski S, Hoai LT, Markiewicz T. Support vector machine-based expert system for reliable heartbeat recognition. *IEEE Trans Biomed Eng.* (2004) 51:582–9. doi: 10.1109/TBME.2004.824138

16. Yeh Y, Chiou C, Lin H Analyzing ECG for cardiac arrhythmia using cluster analysis. *Expert Syst Appl.* (2012) 39:1000–10. doi: 10.1016/j.eswa.2011.07.101

17. Park J, Lee K, Kang K. "Arrhythmia detection from heartbeat using k-nearest neighbor classifier," in *IEEE International Conference on Bioinformatics and Biomedicine.* (2013), pp. 15–22. doi: 10.1109/BIBM.2013.6732594

18. Chazal PD, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans Biomed Eng.* (2004) 51:1196–206. doi: 10.1109/TBME.2004.827359

19. Eric M, Unai I, Javier DS, Elisabete A, Iraia I, Mikel O, et al. "ECG-based random forest classifier for cardiac arrest rhythms," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* (2019), pp. 1504–1508.

20. Sayadi O, Shamsollahi MB. A model-based Bayesian framework for ECG beat segmentation. *Physiol Meas.* (2009) 30:335. doi: 10.1088/0967-3334/30/3/008

21. Wiggins M, Saad A, Litt B, Vachtsevanos G. Evolving a Bayesian classifier for ECG-based age classification in medical applications. *Appl Soft Comput.* (2008) 8:599–608. doi: 10.1016/j.asoc.2007.03.009

22. Kiranyaz S, Ince T, Gabbouj M. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Trans Biomed Eng.* (2015) 63:664–75. doi: 10.1109/TBME.2015.2468589

23. Li P, Wang Y, He J, Wang L, Tian Y, Zhou T, et al. High-performance personalized heartbeat classification model for longterm ECG signal. *IEEE Trans Biomed Eng.* (2016) 64:78–86. doi: 10.1109/TBME.2016.2539421

24. Acharya UR, Oh SL, Hagiwara Y, Tan J, Adam M, Gertych A, et al. A deep convolutional neural network model to classify heartbeats. *Comput Biol Med.* (2017) 89:389–96. doi: 10.1016/j.compbiomed.2017.08.022

25. Sellami A, Hwang H. A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. *Expert Syst Appl.* (2019) 122:75–84. doi: 10.1016/j.eswa.2018.12.037

26. Atal DK, Singh M. Arrhythmia classification with ECG signals based on the optimization-enabled deep convolutional neural network. *Comput Methods Programs Biomed.* (2020) 196:105607. doi: 10.1016/j.cmpb.2020.105607

27. Wang T, Lu C, Sun Y. Automatic ECG classification using continuous wavelet transform and convolutional neural network. *Entropy.* (2021) 23:119. doi: 10.3390/e23010119

28. Ozdemir MA, Cura OK, Akan A. Epileptic eeg classification by using time-frequency images for deep learning. *Int J Neural Syst.* (2021) 31:2150026. doi: 10.1142/S012906572150026X

29. Ozdemir MA, Kisa DH, Guren O. Hand gesture classification using time–frequency images and transfer learning based on CNN. *Biomed Signal Process Control.* (2022) 77:103787. doi: 10.1016/j.bspc.2022.103787

30. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Magaz.* (2001) 20:45–50. doi: 10.1109/51.932724

31. Association for the Advancement of Medical Instrumentation and Others. *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms,* vol. 1998, ANSI/AAMI EC38 (1998).

32. He D, Cao H, Wang S, Chen X. Time-reassigned synchrosqueezing transform: the algorithm and its applications in mechanical signal processing. *Mech Syst Signal Process.* (2019) 117:255–79. doi: 10.1016/j.ymssp.2018.08.004

33. Dosovitskiy L, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv preprint: arXiv:2010.11929 (2020).

34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. *Adv Neural Inf Process Syst.* (2017). *arXiv preprint:arXiv:1706.03762*:5998−6008.

35. d'Ascoli S, Touvron H, Leavitt M, Morcos A, Biroli G, Sagun L. *Convit: Improving Vision Transformers with Soft Convolutional Inductive Biases.* arXiv preprint: arXiv:2103.10697 (2021).

36. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. *Stand-alone Self-attention in Vision Models.* arXiv preprint: arXiv:1906.05909 (2019).

37. Cordonnier J, Loukas A, Jaggi M. *On the Relationship Between Self-attention and Convolutional Layers.* arXiv preprint: arXiv:1911.03584 (2019).

38. Chawla NV, Bowyer KW, Lawrence LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intel Res.* (2002) 16:321–57. doi: 10.1613/jair.953

39. Lin TY, Goyal P, Girshick R, He KM. "Focal loss for dense object detection,". in *Proceedings of the IEEE International Conference on Computer Vision.* (2017), pp. 2980–2988. doi: 10.1109/ICCV.2017.324

40. Izci E, Ozdemir MA, Degirmenci M. "Cardiac arrhythmia detection from 2d ECG images by using deep learning technique," in *Medical Technologies Congress.* (2019), pp. 1–4. doi: 10.1109/TIPTEKNO.2019.8895011

41. Allam JP, Samantray S, Ari S. SpEC: A system for patient specific ECG beat classification using deep residual network. *Biocybernet Biomed Eng.* (2020) 40:1446–57. doi: 10.1016/j.bbe.2020.08.001

42. Sun L, Lu Y, Yang K, Li S, ECG. analysis using multiple instance learning for myocardial infarction detection. *IEEE Trans Biomed Eng.* (2012) 59:3348–56. doi: 10.1109/TBME.2012.2213597

43. Chang PC, Lin JJ, Hsieh JC, Wen J. Myocardial infarction classification with multi-lead ECG using hidden Markov models and Gaussian mixture models. *Appl Soft Comput.* (2012) 12:3165–75. doi: 10.1016/j.asoc.2012.06.004

44. Kojuri J, Boostani R, Dehghani P, Nowroozipour F, Saki N. Prediction of acute myocardial infarction with artificial neural networks in patients with nondiagnostic electrocardiogram. *J Cardiovasc Dis Res.* (2015) 6:51. doi: 10.5530/jcdr.2015.2.2

45. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Inf Sci.* (2017) 415:190–8. doi: 10.1016/j.ins.2017.06.027

46. Wang HM, Zhao W, Jia DY, Hu J, Li Z.Q, Yan C, et al. "Myocardial infarction detection based on multi-lead ensemble neural network," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* Berlin, Germany (2019), pp. 2614–7. doi: 10.1109/EMBC.2019.8856392

47. Bousseljot R, Kreiseler D, Schnabel A. Nutzung der ekg-signaldatenbank cardiodat der ptb uber das internet. *Biomedizinische Technik/Biomed Eng.* (1995) 40:317–8. doi: 10.1515/bmte.1995.40.s1.317

48. Ozdemir MA, Ozdemir GD, Guren O. Classification of COVID-19 electrocardiograms by using hexaxial feature mapping and deep learning. *BMC Med Inform Decis Mak.* (2021) 21:1–20. doi: 10.1186/s12911-021-01521-x