A Dilated Residual Vision Transformer for Atrial Fibrillation Detection from Stacked Time-Frequency ECG Representations

Sawon Pratiher, Apoorva Srivastava, Yedla Bindu Priyatha, Nirmalya Ghosh, and Amit Patra
Department of Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

***Abstract*—Atrial fibrillation (AF), the most frequent type of cardiac arrhythmia, has no apparent clinical symptoms in most cases for patients, making it more challenging to diagnose. However, the alterations in regular heart rhythm with an absence of visible P-waves in an Electrocardiogram (ECG) are often characterized as AF symptoms. ECG signals are often employed for AF prognosis to minimize the risk of stroke, coronary artery disease, and other cardiovascular diseases. This work proposes a new vision transformer (ViT) variant, namely, Dilated Residual ViT (DiResViT), by replacing the original patchify stem in ViT with dilated convolutional stem having residual connections for improved AF detection from an ensemble of ECG time-frequency representations. Dilated convolutions facilitate dense feature representation by comprehending the multi-scale contextual information and allowing the receptive field to expand exponentially without losing resolution. The introduction of residual connections alleviates the vanishing and exploding gradients problem by enabling the gradients to bypass some of the non-linear activation functions with improved training convergence. DiResViT's exhaustive experimental validation outperforms the prior art in ECG-based AF detection, while the ablation study evinces enhanced performance compared to the existing ViT.**

***Index Terms*—ECG; AF; deep learning; vision transformer; dilated convolutions; residual connection; attention; saliency.**

I. Introduction

An estimated 32% (18.5 million) of all deaths worldwide are related to cardiovascular diseases (CVDs) [1]. Clinical studies have shown arrhythmia as the prevailing cause of CVDs, and long-term CVDs lead to heart failure. Some common arrhythmias in clinical practice include atrial flutter, premature ventricular contraction and atrial fibrillation (AF), a supraventricular tachyarrhythmia [2]. AF's prevalence accounts for 1%–2% of the total population and is predicted to triple by 2050 [3]. Generally, the risk of AF increases with age, and the lifetime risk of AF increases to 25% by the age of 40. As a result, the auxiliary AF prognosis approaches may help clinicians improve treatment methods and achieve better treatment quality, lowering AF morbidity and mortality and critical illness caused by AF [4].

With the advent of wearable sensors with a single-channel or a 12-lead ECG recording device, personalized healthcare research has gained traction. However, in wearable monitoring, daily activities will produce unwanted interferences, and the analysis of atrial activity will be challenging. Often, it's also not feasible to acquire and analyze 12-lead ECG signals. Instead, a single-lead ECG is more economical and efficient. By leveraging recent deep learning (DL) techniques, highly accurate analytical algorithms can be developed for AF detection, and the domain-expertise requirement can be alleviated. Thus healthcare costs can be reduced [5].

AF is physiologically diagnosed as irregular atrial activity patterns [6]. Application of DL methods include ECG with 1-D convolutional neural networks (CNN) in [7], [8], with R-R intervals in [9] and with entropy features in [10]. Efficacy of 2-D CNN [11] and convolutional recurrent neural networks (CRNN) [12] on ECG spectrogram has been studied. Time frequency representations (TFRs) of ECG signals are explored in [13], [14], and [15]. The performance of the prevalent AF detection methods is briefly summarized and compared with the present work in Table IV of section III-E.

A. Motivation for Attention Mechanism in Ensemble TFRs

The motivation for stacking multiple TFRs to abstract rich feature representation stems from the previous research employing continuous wavelet transform (CWT) [13], short-time Fourier transform (STFT) [14], and Chirplet transform (CT) for ECG-TFRs abstraction [15]. STFT describes the changing spectra as a function of time. CWT abstracts the simultaneous time-frequency localization and can comprehend the non-stationary R-R intervals and P-wave absence during AF episodes. The recent application of CT on ECG signals inspires to employ a family of cyclically varying frequency-modulated bases for CT-based TFR [15].

An ECG with normal sinus rhythm (NSR) exhibits quasi-periodic statistical regularity. In contrast, pathological AF symptoms include irregular R-R intervals (intermittent conduction of impulses to the ventricles) and absence of visible P waves (depolarization of the atria) [6]. The substantial variation in average RR-interval, P-wave, and QRS-complex morphology in quasi-stationary ECG signals having AF episodes are encoded as non-linear, time-varying waveforms. Highly efficient automated AF detection requires expressive power to boost ECG segments with missing P-waves and irregular R-R intervals in an input ECG stream and ignore the regular rhythms. Automated region-of-interest localization in an ECG-TFRs without explicit manual intervention requires DL models to mimic cognitive attention [16]. Recently developed ViT, which is the state-of-the-art for image classification problems, inspires us to explore a ViT variant specific to the AF identification task using an ensemble of TFRs.

B. Research contributions

This work is the first to implement a ViT variant for CVD analysis to the best of our knowledge. The design and evaluation of the framework provide the following contributions:

- A new ViT architecture, namely **DiResViT** has been designed, by integrating dilated convolutional stem with residual connections to provide the best of both worlds.
- **DiResViT** leverages the multi-head attention mechanisms of the ViT model with the multi-scale feature extraction capabilities of computationally efficient dilated convolution and the advantages of skip connections.
- **DiResViT**'s AF recognition accuracy outperforms the state-of-the-art with regards to classification scores and

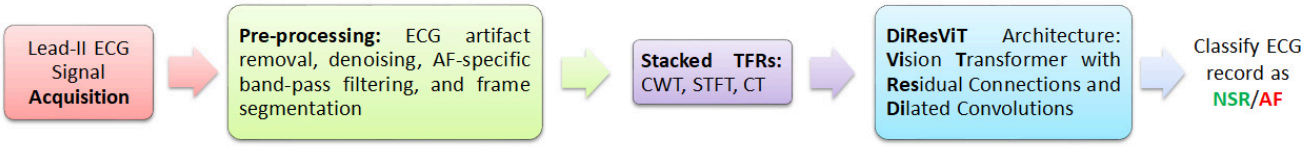

Figure 1: AF detection pipeline showing ECG pre-processing stages, abstraction of ensemble TFRs, and **DiResViT** framework.

demonstrates better convergence than pure ViT models with 2× fewer trainable parameters for same network depth, which can reduce the training cost significantly.
- An ensemble stack of multiple ECG TFRs provides robust AF saliency for subsequent representation learning.

As an overview, section II explains the different technical blocks in the workflow (refer Fig. 1) and the proposed **DiResViT** architecture. Experimental results are discussed in section III, and section IV concludes the paper.

II. Methodology

A. AF Classification Pipeline

Most of the prior DL-based ECG research followed a similar workflow [12]: Initial signal pre-processing, then often segmented into equal-length segments. It's also usual to convert the 1-D ECG signals into 2-D TFRs. After that, various data augmentation techniques are generally employed to eschew the class imbalance and then fed into DL models [17].

B. Pre-processing: AF-specific Filtering and Segmentation

Here lead-II ECG records are used for analysis. Initially, the acquired ECG signals are subjected to a Savitzky-Golay-based finite impulse response filter with an optimal 4^{th} order polynomial and a frame length of 17 samples for smoothing and differentiation [18]. For removing baseline wandering, the filtered ECG signals are detrended with a 6^{th} order polynomial and passed through a Butterworth high-pass filter with a cut-off frequency of 0.5 Hz for motion artifact removal. Subsequently, these high-pass filtered ECGs are subjected to a notch filter for power-line interference removal [19]. The prior art exemplifies the spectral content of the atrial activity, a precursor for AF anomaly, within the 5-30 Hz frequency band [6]. As such, a band-pass filter (pass-band: 5-30 Hz) has been used to extract the spectral content of the P-waves for AF identification [20]. After that, the filtered records are divided into equal-length frames using non-overlapping windows to capture the inter-sample and intra-beat correlations successfully.

C. TFR computation from the Filtered ECG Signals

A near compact region-of-support and the shape similarity to an ECG beat supports the selection of the Morlet wavelet for CWT-based TFR analysis. We have used a window length of 128 samples and 512 frequency bands for the CWT of filtered ECG signals. The STFT-based TFR plots are obtained with 1024 frequency bands and a Gaussian window of width 512 Bins. With the same parameters as STFT, a Chirplet's line frequency modulation rate of 5 Hz/sec is used for the CT.

Typical ECG signals belonging to NSR and AF, respectively, their filtered versions, and corresponding TFRs (CWT, STFT, and CT) are shown in Fig. 2. The TFRs in Fig. 2(a) belongs to NSR, portray consistent patterns. In contrast, the visual saliency highlighting the absence of P-waves and irregular R-R intervals with AF episodes is apparent from Fig. 2(b).

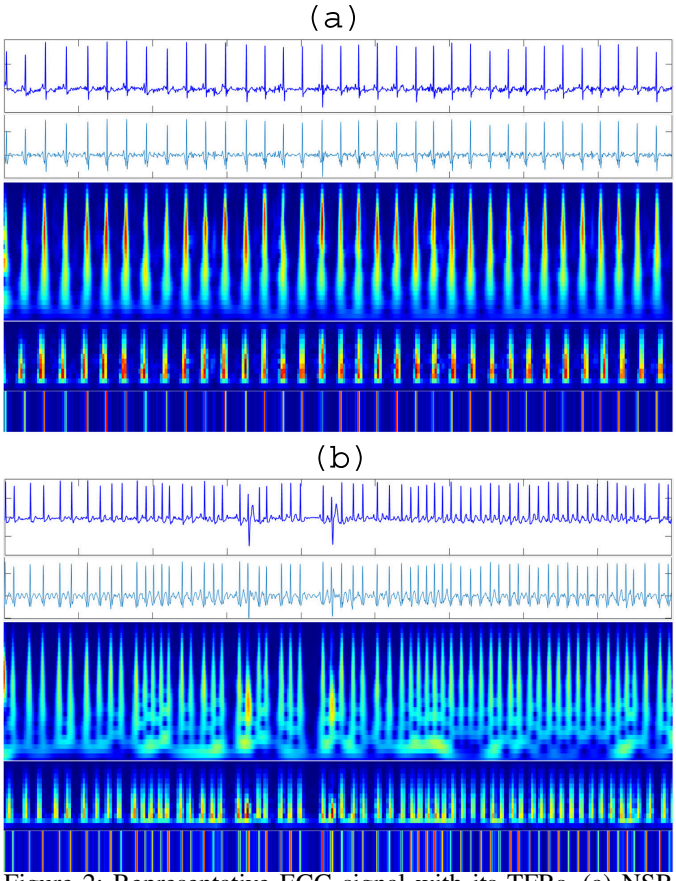

Figure 2: Representative ECG signal with its TFRs. (a) NSR and (b) AF. Here the 1^{st} row exemplifies the raw ECG signal, 2^{nd} row exemplifies the filtered signal used for TFR analysis, 3^{rd} to 5^{th} rows illustrate the TFRs of the filtered ECG signal corresponding to CWT, STFT, and CT, respectively in order.

D. Proposed *DiResViT* Architecture

ViT in [21] has been used as the building block for multi-header attention. To abstract multi-scale representation learning and stable training convergence [22], computationally efficient dilated convolutions and residual blocks have been incorporated to drastically alter its optimization behavior with significantly improved performance with a reduced number of trainable parameters. A modular level schematic of the **DiResViT** architecture is shown in Fig. 3. For the same network depth, the introduction of residual skip connections in the **DiResViT** architecture halved the number of trainable parameters, and improved convergence and performance compared to the original ViT formulation [21]. For example, in the AF identification task, having same network depth (refer Fig.

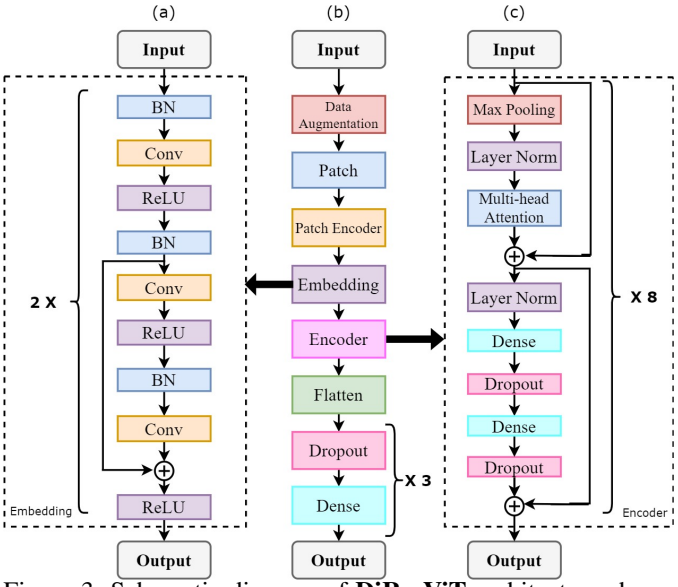

Figure 3: Schematic diagram of **DiResViT** architecture showing early dilated convolutional layers and residual connections.

3), **DiResViT** has 31,702,345 ($\approx 30M$) trainable parameters, compared to 60,605,961 ($\approx 60M$) in pure ViT in [21]. A brief functionality of the different modules in the **DiResViT** framework is explained below:

E. Multi-head Attention with ViT Backbone

Comprehensive ViT details for multi-head attention mechanism in image classification can be found in [21], where ViT formulation in [21] transforms the input image into patch tokens by segregating it into specific patch sizes. The stacked encoder layer of the transformer modules abstract these patches and encode them in the encoder, which is an ensemble of multi-head self-attention units in the feed-forward network (FFN) layer and produces the output $z_i = \mathcal{F}(y_i + FFN(y_i))$. Here, $y_i = \mathcal{F}(y_{i-1} + FFN(y_{i-1}))$, and $\mathcal{F}(.)$ is the output layer normalization, and y_i is the embedded series of patches encoded with its positional information.

F. Dilated Convolutions for Multi-Scale Context Aggregation

The early visual processing in ViT's patchify stem is implemented using a non-overlapping stride, and hence ViT's poor optimization compared to CNNs [22]. Early convolutional layers provide a crucial trade-off between inductive biases and transformer blocks' capacity to learn feature information and drastically alter its optimization behavior with faster convergence, stable learning rate, and improved performance. However, the limited receptive field in CNN frameworks is alleviated by incorporating dilated convolutions in **DiResViT** model. Dilated convolutions systematically combine multi-scale contextual representation without sacrificing resolution and allow the receptive field to expand exponentially on the same computation cost [23].

Mathematically, a convolution operator (C) applies a kernel (k) of size $(2r+1)^2$ to obtain the feature $(C * k)(p) = \sum_{i+j=p} C(i)k(j)$. On the contrary, dilated convolution operation (C_d) skips pixels according to the dilation factor (d), and obtains the feature $(C_d * k)(p) = \sum_{i+dj=p} C(i)k(j)$. Increasing the parameter d increases the receptive field.

G. Residual Skip Connections

The introduction of residual skip connections alleviates the vanishing gradient problem, results in fewer extra parameters for increasing network depth, and accelerates training convergence [24]. Mathematically, within the CNN layers, the input a_j is mapped to a low dimension manifold $y_k = f_\phi(a_j)$, where f_ϕ is the transformation function with parameter ϕ. The output of residual block can be depicted as $b = \mathcal{F}\ (y_k) + a_j$, with $\mathcal{F}(.)$ as the residual mapping to be learned.

III. Experimental Evaluation

A. Dataset Description

The proposed **DiResViT** model is validated on two large-scale publicly available databases, namely PhysioNet Computing in Cardiology 2017 (PCC 2017) [25], and PhysioNet Computing in Cardiology 2021 (PCC 2021) [26]. The PCC 2017 database consists of 8,528 ECG recordings, among which 5154 are normal, and 717 belong to the AF group, respectively. The ECG record length varies between 9 to 60 seconds with a sampling frequency of 300 Hz. The PCC 2021 contains 88,253 ECG recordings, among which 1506 are singly labeled as AF, and 15090 are NSR, respectively. The recording length varies from 6 seconds to 30 minutes, with a sampling frequency varying between 257 to 1000 Hz. Table I enumerate the number of ECG frames retrieved post-processing.

Table I: ECG segment statistics for different databases.

Database	# AF	# NSR	Total
*PCC 2017	732	4,976	5,708
+PCC 2021	2,435	15,960	18,395

* https://physionet.org/content/challenge-2017/1.0.0/
\+ https://physionet.org/content/challenge-2021/1.02/

B. Experimental Setup

The **DiResViT** model is implemented in Python using Tensorflow and Keras libraries in Ubuntu 20.04 LTS x64 OS. The system hardware consists of 1xTesla K80, compute 3.7, having 2496 CUDA cores GPU, 12GB GDDR5 VRAM, and 1xsingle core hyperthreaded Xeon Processors @2.3Ghz CPU. The **DiResViT** model is trained for 200 epochs with a batch size of 256 and using Adam optimizer with a learning rate of 0.001 and a weight decay factor of 0.0001. The model loss is set to sparse categorical cross-entropy, and sparse categorical accuracy is taken as the monitoring metric. The model parameters are initialized with Xaviar uniform initializer, and early stopping is employed during the training phase to avoid model overfitting. The inputs to the **DiResViT** model are resized to 128x128x3 with a patch size of 6x6 and a projection dimension of 64. The number of heads in the multi-headed attention is set to twice the projection dimension. The number of transformer layers is set as 8 and the number of units in the multilayer perceptron head is 2048x1024.

C. Training and Performance Evaluation

Each dataset (PCC 2017 and PCC 2021) is split into three partitions segment-wise. Two of them are used for training, and the third one is used for testing. This strategy is repeated across all the folds. Further, the training set for each case is split randomly in the ratio of 80:20 for training and validation. Data augmentation is used to eschew the class imbalance [17].

The different performance evaluation metrics (PEM) used for quantifying our method are accuracy (Acc.), precision (P), recall (R), F1 score (F1), and their micro-average weighted values. The **DiResViT**'s predictions can be divided into TP (true-positive), TN (true-negative), FP (false-positive), and FN (false-negative). Here, TP/FP (and TN/FN) signify the cardinality of correct/incorrect predictions for the positive (and negative) class respectively. Based on TP, TN, FP, and FN values, the PEMs are defined as: $Acc. = \frac{TP+TN}{TP+TN+FP+FN}$, $P = \frac{TP}{TP+FP}$, $R = \frac{TN}{TN+FN}$, and $F1 = \frac{2TP}{2TP+FP+FN}$.

D. AF Recognition Results and Ablation Study

The PEM scores averaged over the 5-folds are reported in Table II and III. All different PEM scores obtained by **DiResViT** outperforms the existing ViT [21] model for both the datasets. Generally, the precession, recall, and F1 scores are better for NSR class compared to the AF category (mild effect of class imbalance), with 99.92% recall obtained on the PCC 2017 database with the **DiResViT** model. A significant enhancement of 13.47% in F1 score is observed for the AF class with **DiResViT** compared to the baseline ViT [21] with ∼half trainable parameters, and justifies the clinical competency of **DiResViT** in resource-constrained environments.

Table II: PEM (%) for ViT [21] and **DiResViT**.

Architecture	Dataset	Class	P	R	F1
ViT	**PCC 2017**	**AF**	94.18	70.76	80.29
		NSR	96.44	99.03	97.17
	PCC 2021	**AF**	94.17	88.35	91.08
		NSR	98.53	98.83	98.75
DiResViT	**PCC 2017**	**AF**	**97.48**	**84.23**	**90.64**
		NSR	**98.11**	**99.92**	**99.26**
	PCC 2021	**AF**	**96.22**	**93.38**	**94.53**
		NSR	**99.27**	**99.06**	**99.13**

ViT [21] run in-house. **BOLD** signifies better performance.

Table III: Weighted PEM (%) for ViT [21] and **DiResViT**.

Architecture	Dataset	Wt. P	Wt. R	Wt. F1	Acc.
ViT	**PCC 2017**	95.75	96.14	95.21	95.57
	PCC 2021	96.54	96.72	96.47	96.15
DiResViT	**PCC 2017**	**98.05**	**98.16**	**98.69**	**97.63**
	PCC 2021	**98.19**	**98.59**	**98.54**	**98.49**

Fig. 4 demonstrate the confusion matrices corresponding to **DiResViT**'s predictions on (a) PCC 2017 and (b) PCC 2021 databases, combined across all the folds. Very high PEM values entail a nearly impeccable agreement between the predicted and the actual classes.

(a)

	Predicted AF	Predicted NSR
True AF	616 (84.16%)	116 (15.84%)
True NSR	3 (0.06%)	4990 (99.94%)

(b)

	Predicted AF	Predicted NSR
True AF	2277 (93.13%)	168 (6.87%)
True NSR	79 (0.49%)	15921 (99.51%)

Figure 4: Confusion matrices corresponding to **DiResViT**'s predictions on (a) PCC 2017 and (b) PCC 2021 databases.

E. Comparison with State-of-the-art

A comparative evaluation with the state-of-the-art AF detection methods is tabulated in Table IV, where the dominance of DL techniques is apparent. The **DiResViT** model outperforms the prevalent art on both PCC 2017 and PCC 2021 databases.

Table IV: Prior art comparison on AF detection.

Ref, YoP	Dataset	Method	Performance (%)
[12], 2017	PCC 2017	ECG signal, CRNN.	Acc. = 82.3, F1 = 79.2
[7], 2018	PCC 2017	ECG signal, 1-D CNN.	Acc. = 98.13
[11], 2018	PCC 2017	2-D CNN, ECG Spectrogram.	F1 = 83.0
[27], 2018	PCC 2017	STFT, 2-D CNN.	F1 = 80.02
[9], 2019	MIT-BIH Arrhythmia Database	R-R Intervals, 1-D CNN.	Acc. = 89.30
[8], 2020	University of Illinois Hospital & Health Science System	ECG Signal, 1-D CNN.	Acc. = 79.90
[17], 2020	PCC 2017	ECG signal, CRNN.	F1 = 87.79
[10], 2021	PCC 2017	ECG Signal, R-R Interval, Entropy features, 1-D CNN.	Acc. = 91.70
[28], 2021	PCC 2021	ECG signal, Transformer.	F1 = 93.06
[29], 2021	MIT-BIH AF Database	HRV features, Random Forest.	Acc. = 97.04, F1 = 87.03
This work	**PCC 2017**	**DiResViT, Stacked TFRs**	**P = 98.05, R = 98.16, F1 = 98.69, Acc. = 97.63**
This work	**PCC 2021**	**DiResViT, Stacked TFRs**	**P = 98.19, R = 98.59, F1 = 98.54, Acc. = 98.49**

Ref. = Reference, YoP = Year of Publication.

IV. Conclusion

This work presents **DiResViT**, a ViT variant to detect AF from a stacked ensemble of TFRs using a single-lead ECG signal. **DiResViT** merge the benefits of dilated convolution and skip connections with the multi-head attention mechanisms in ViTs. Dilated convolutions encode the multi-scale intrinsic feature representation by allowing the receptive field of the convolution kernel to be sparsely populated without increasing computational complexity. Residual connections mitigate the vanishing gradients and accuracy saturation problem with a reduced number of trainable parameters (∼ half of ViT [21] for the same network path) and highlight **DiResViT**'s computational efficiency. An exhaustive experimental validation of **DiResViT** on large-scale publicly available datasets outperforms the classical ViT and prevalent AF identification methods. In the future, we plan to explore **DiResViT** for other arrhythmia taxonomy and look into the explainable side of DL for clinically meaningful interpretations.

References

[1] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim *et al.*,

"Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[2] S. S. Chugh, J. L. Blackshear, W.-K. Shen, S. C. Hammill, and B. J. Gersh, "Epidemiology and Natural History of Atrial Fibrillation: Clinical Implications," *Journal of the American College of Cardiology*, vol. 37, no. 2, pp. 371–378, 2001.

[3] J. Kornej, C. S. Börschel, E. J. Benjamin, and R. B. Schnabel, "Epidemiology of Atrial Fibrillation in the 21st Century: Novel Methods and New Insights," *Circulation research*, vol. 127, no. 1, pp. 4–20, 2020.

[4] D. M. Lloyd-Jones, T. J. Wang, E. P. Leip, M. G. Larson, D. Levy, R. S. Vasan, R. B. D'Agostino, J. M. Massaro, A. Beiser, P. A. Wolf *et al.*, "Lifetime Risk for Development of Atrial Fibrillation: The Framingham Heart Study," *Circulation*, vol. 110, no. 9, pp. 1042–1046, 2004.

[5] J. M. Bumgarner, C. T. Lambert, A. A. Hussein, D. J. Cantillon, B. Baranowski, K. Wolski, B. D. Lindsay, O. M. Wazni, and K. G. Tarakji, "Smartwatch Algorithm for Automated Detection of Atrial Fibrillation," *Journal of the American College of Cardiology*, vol. 71, no. 21, pp. 2381–2388, 2018.

[6] C. T. January, L. S. Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa, J. C. Cleveland, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field *et al.*, "2014 AHA/ACC/HRS Guideline for the Management of Patients with Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society," *Journal of the American College of Cardiology*, vol. 64, no. 21, pp. e1–e76, 2014.

[7] X. Fan, Q. Yao, Y. Cai, F. Miao, F. Sun, and Y. Li, "Multiscaled Fusion of Deep Convolutional Neural Networks for Screening Atrial Fibrillation from Single Lead Short ECG Recordings," *IEEE journal of biomedical and health informatics*, vol. 22, no. 6, pp. 1744–1753, 2018.

[8] Y.-H. Chen, A. H. Twing, D. Badawi, J. Danavi, M. McCauley, and A. E. Cetin, "Atrial Fibrillation Risk Prediction from Electrocardiogram and Related Health Data with Deep Neural Network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1269–1273.

[9] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems with Applications*, vol. 115, pp. 465–473, 2019.

[10] J. Shi, C. Chen, H. Liu, Y. Wang, M. Shu, and Q. Zhu, "Automated Atrial Fibrillation Detection Based on Feature Fusion Using Discriminant Canonical Correlation Analysis," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[11] S. Parvaneh, J. Rubin, A. Rahman, B. Conroy, and S. Babaeizadeh, "Analyzing single-lead short ecg recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation," *Physiological measurement*, vol. 39, no. 8, p. 084003, 2018.

[12] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for Electrocardiogram classification," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[13] M. Sharma and U. R. Acharya, "A New Method to Identify Coronary Artery Disease with ECG Signals and Time-Frequency Concentrated Antisymmetric Biorthogonal Wavelet Filter Bank," *Pattern Recognition Letters*, vol. 125, pp. 235–240, 2019.

[14] C. Mateo and J. A. Talavera, "Analysis of atrial and ventricular premature contractions using the short time fourier transform with the window size fixed in the frequency domain," *Biomedical Signal Processing and Control*, vol. 69, p. 102835, 2021.

[15] T. Radhakrishnan, J. Karhade, S. Ghosh, P. Muduli, R. Tripathy, and U. R. Acharya, "Afcnnet: Automated detection of af using chirplet transform and deep convolutional bidirectional long short term memory network with ECG signals," *Computers in Biology and Medicine*, vol. 137, p. 104783, 2021.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] F. N. Hatamian, N. Ravikumar, S. Vesal, F. P. Kemeth, M. Struck, and A. Maier, "The effect of data augmentation on classification of atrial fibrillation in short single-lead ECG signals using deep neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1264–1268.

[18] S. Hargittai, "Savitzky-Golay Least-Squares Polynomial Filters in ECG Signal Processing," in *Computers in Cardiology, 2005*. IEEE, 2005, pp. 763–766.

[19] L. Sharma, R. Tripathy, and S. Dandapat, "Multiscale Energy and Eigenspace Approach to Detection and Localization of Myocardial Infarction," *IEEE transactions on biomedical engineering*, vol. 62, no. 7, pp. 1827–1837, 2015.

[20] P. Langley, E. J. Bowers, and A. Murray, "Principal Component Analysis as A Tool for Analysing Beat-to-Beat Changes in Electrocardiogram Features: Application to Electrocardiogram Derived Respiration," *IEEE Trans. Biomed. Eng.*, vol. 7, p. 7, 2010.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021.

[22] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early Convolutions help Transformers See Better," *arXiv preprint arXiv:2106.14881*, 2021.

[23] F. Yu and V. Koltun, "Multi-Scale context Aggregation by Dilated Convolutions," 2016.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[25] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, "Af Classification from A Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[26] M. A. Reyna, N. Sadr, E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A. B. Rad, A. Elola, S. Seyedi, S. Ansari, H. Ghanbari, Q. Li, A. Sharma, and G. D. Clifford, "Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021," *Computing in Cardiology*, vol. 48, pp. 1–4, 2021.

[27] Z. Zhao, S. Särkkä, and A. B. Rad, "Spectro-Temporal ECG Analysis for Atrial Fibrillation Detection," in *2018 IEEE 28Th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2018, pp. 1–6.

[28] B. Wang, C. Liu, C. Hu, X. Liu, and J. Cao, "Arrhythmia classification with heartbeat-aware transformer," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1025–1029.

[29] G. Hirsch, S. H. Jensen, E. S. Poulsen, and S. Puthusserypady, "Atrial Fibrillation Detection using Heart Rate Variability and Atrial Activity: A Hybrid Approach," *Expert Systems with Applications*, vol. 169, p. 114452, 2021.