Journal of Electrocardiology 89 (2025) 153876

Contents lists available at ScienceDirect

Journal of Electrocardiology

journal homepage: www.jecgonline.com

Residual-attention deep learning model for atrial fibrillation detection from Holter recordings

Md Moklesur Rahman [a,*], Massimo Walter Rivolta [a], Martino Vaglio [b], Pierre Maison-Blanche [c], Fabio Badilini [b,d], Roberto Sassi [a]

[a] Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
[b] AMPS-LLC, New York, NY, USA
[c] Department of Cardiology, Hôpital Bichat, Paris, France
[d] University of California at San Francisco, San Francisco, CA, USA

ARTICLE INFO

Keywords:
Atrial fibrillation detection
Deep learning
AI in ECG
Residual temporal attention

ABSTRACT

Background: Detecting subtle patterns of atrial fibrillation (AF) and irregularities in Holter recordings is intricate and unscalable if done manually. Artificial intelligence-based techniques can be beneficial. In fact, with the rapid advancement of AI, deep learning (DL) demonstrated the capability to identify AF from ECGs with significant performance. However, further development and validation on larger cohorts is still needed.
Purpose: The main purpose of this study was to develop a Residual-attention DL model by considering a large cohort of 2-lead Holter recordings.
Methods: We developed a residual DL model by collecting a large dataset of 661 Holter recordings, which was labeled manually by an expert cardiologist. The DL model leveraged attention mechanisms, allowing it to capture long-range dependencies and intricate temporal relationships crucial for identifying subtle patterns indicative of AF.
Results: Experimental results demonstrated that our model achieved a sensitivity (detection of AF) of Se = 0.928 and a specificity of Sp = 0.915, with an AUC-ROC of AUC = 0.967 on our dataset. Additionally, when evaluated with an external test dataset, specifically IRIDIA-AF, our DL model obtained Se = 0.942, Sp = 0.932, and AUC = 0.965. Finally, when compared under similar experimental conditions with other state-of-the-art models, our DL model achieved slightly better performance overall.
Conclusion: The Residual-attention DL model we proposed offers a promising solution for AF detection. The validation on external datasets contributes to its potential for deployment in clinical settings, providing clinicians with a valuable decision support system.

Introduction

Atrial fibrillation (AF) stands as the most common cardiac arrhythmia, contributing significantly to morbidity and mortality worldwide [1]. Its detection and management pose substantial challenges to healthcare systems, necessitating accurate detection methods, specifically at early stage. Holter recordings, offering prolonged cardiac monitoring, emerge as invaluable resources for detecting AF, providing continuous electrocardiogram (ECG) data over extended periods [2].

Traditionally, AF detection from Holter recordings relies heavily on manual interpretation by trained clinicians. However, this process is labor-intensive, time-consuming, and prone to inter-observer variability [3]. Moreover, the increasing prevalence of AF mandates scalable and efficient detection methods. Herein lies the imperative for computerized systems, which can automate AF detection, enhance diagnostic accuracy, and expedite patient care [2].

The advent of machine learning (ML) techniques revolutionized medical diagnostics, offering promising performance for automated AF detection [4]. Traditional ML techniques necessitate manual feature engineering, which is time-consuming and requires domain expertise. Additionally, these techniques may struggle to capture complex patterns and relationships within high-dimensional ECG signals. The dynamic

* Corresponding author.
E-mail addresses: md.rahman@unimi.it (M.M. Rahman), massimo.rivolta@unimi.it (M.W. Rivolta), vaglio@amps-llc.com (M. Vaglio), badilini@amps-llc.com (F. Badilini), roberto.sassi@unimi.it (R. Sassi).

https://doi.org/10.1016/j.jelectrocard.2025.153876
Received 20 May 2024; Received in revised form 30 November 2024; Accepted 3 January 2025
Available online 10 January 2025
0022-0736/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

nature of ECG data, influenced by factors like patient movement, environmental noise, demographics, and disease prevalence, poses challenges for traditional ML techniques.

Deep learning (DL) emerged as a compelling solution to address the shortcomings of traditional ML models in AF detection [5–7]. DL models are typically defined by their architectures, with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) being prominent examples. These architectures, particularly CNNs and RNNs, demonstrate remarkable prowess in learning hierarchical representations and capturing temporal dependencies from raw data [8]. Nonetheless, DL models encounter challenges in effectively leveraging long-range dependencies within sequential data, which are prevalent in ECG signals. One notable advancement in DL architectures, the attention mechanism, presents a paradigm shift in addressing these challenges [9]. The integration of residual-attention mechanisms into DL models holds profound implications for AF detection from Holter recordings. By leveraging both residual connections and attention mechanisms, these models can effectively capture long-range dependencies and salient features within ECG signals, thus likely improving performance for AF detection.

Typically, many existing methods for AF detection from Holter recordings have been validated using a small number of patients *i.e.,* MIT-BIH AF [10] and long-term AF dataset [11], limiting their applicability to real-world clinical settings. In contrast, our work aims to address this limitation by leveraging a new and clinically significant dataset comprising diverse patient populations and high-quality ECG recordings. Being very flexible in nature, DL models are prone to learning specific characteristics of the dataset used to train them, potentially resulting in a model that struggles to generalize in practice. By utilizing a larger and more representative dataset, we can enhance the generalizability and reliability of DL models. Particularly in the context of AF detection, the residual-attention DL model presents significant advantages over other DL-based methods when applied to Holter recordings. Firstly, the attention mechanism enables the model to focus on important segments of the ECG signal, enhancing interpretability and robustness [12,13]. Secondly, the inclusion of residual connections facilitates the training of deeper networks, enabling a better capture of complex temporal dependencies in ECG segments [12,13]. The main contributions of our study are as follows:

- We collected a large retrospective cohort of clinical data from the Holter device and developed a residual-attention DL for the detection of AF from the Holter recordings.
- We compared the performance of state-of-the-art DL models and the proposed model.
- We assessed the performance of the proposed DL model on different demographic groups.

Materials and methods

In our study, we used two datasets to develop and evaluate our DL model for AF detection from Holter recordings. One of the datasets is private, while the other, known as the IRIDIA-AF dataset, is publicly accessible. In the following, we provide detailed descriptions of both datasets.

The dataset

The dataset comprised 661 Holter records collected from 661 patients at Groupe Hospitalier Ambroise Paré in Paris, France. Each Holter recording had an average duration of around 23 h and was captured using a Microport Spiderview Holter recorder, a 2-lead system operating at a sampling rate of 200 Hz and an amplitude resolution of 10μV. The details about the dataset are described in [14]. The average age of the patients was approximately 60 years, with females accounting for approximately 39 % of the records. About 50 % of the records ($n = 333$) included at least one episode of AF or atrial flutter (AFL), with durations of the episodes varying from a few short events up to the entire record (chronic AF or AFL). The remaining records were entirely in sinus rhythm ($n = 195$), were characterized by a large incidence of premature ventricular contraction beats (PVC, $n = 41$), included episodes of atrial tachycardia (AT, $n = 61$) or ventricular tachycardia (VT, $n = 31$). The population distribution is illustrated in Fig. 1. In terms of atrial arrhythmia burden, our dataset presents the following distribution: AF accounts for 193,000 min, AFL for 93,000 min, and AT for 48,000 min. Notably, normal sinus rhythm (NSR) spans 180,000 min. The annotations underwent rigorous scrutiny to ensure that a minimum of 59 min per hour were edited by a single cardiologist, minimizing noise and guaranteeing a high quality of the data under analysis. The dataset was collected from two distinct batches; the first one (268 records) was used for training and validation and also included a number of AT events, whereas the second one was used exclusively for testing, and while there were no atrial tachycardia episodes there were on the contrary other challenging ventricular rhythms, including ventricular tachycardia (see Fig. 1).

IRIDIA-AF dataset

The dataset comprised 167 Holter records collected from 152 patients at an outpatient cardiology clinic located in Belgium [15]. The records were collected using a Microport Spiderview Holter recorder, which is the same device employed for our data collection (please refer to the previous paragraph). Notably, records from patients with specific conditions, such as cardiac implantable electronic devices, persistent or permanent AF, or other cardiac diseases were excluded from the dataset. These exclusion criteria were implemented to ensure the homogeneity of the dataset and to focus the analysis on paroxysmal AF. All the 167 records contained in the IRIDIA-AF dataset were considered only for testing.

Preprocessing

During the preprocessing step, we applied a third-order zero-phase Butterworth bandpass filter with cutoff frequencies of 0.5 Hz and 40 Hz to suppress baseline wander and reduce power line interference. Then, each recording was segmented using a 10-s window without any overlap. The number of 10-s segments for the training, validation, and testing sets are presented in Table 1. We considered NSR and AT collectively as non-atrial fibrillation (non-AF), while grouping AF and AFL as AF. AT was categorized under non-AF due to distinctions in heart-rate stability, risk level, and treatment options compared to AF and AFL.

DL model

The DL architecture encompassed various types of layers, serving the

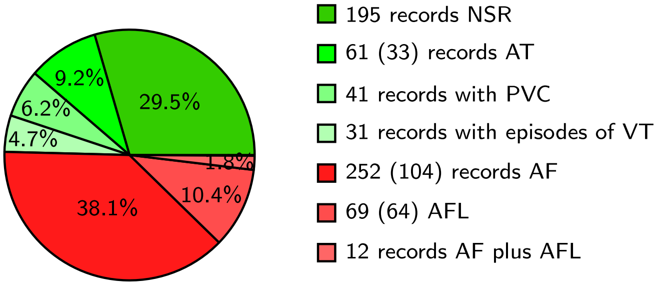

Fig. 1. Distribution of patients by records in the dataset: Records with AF/AFL are shown in red, while records without AF/AFL are shown in green. The number of "chronic" records (*i.e.*, entire records under the labeled rhythm) is indicated in parentheses. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1.
Number of 10-s ECG segments (in thousand units) for each of the two datasets considered.

Dataset	Training				Validation				Testing		
	Non-AF		AF		Non-AF		AF		Non-AF	AF	
	NSR	AT	AF	AFL	NSR	AT	AF	AFL	NSR	AF	AFL
Our dataset	748	261	848	136	24	25	66	34	2324	248	389
IRIDIA-AF	–								1872	536	–

dual purpose of feature extraction and detection. Fig. 2 elucidates the schematic representation of the residual-temporal attention (RTA) DL model proposed in this study. It highlights the integration of an RTA block and a gated recurrent unit (GRU) layer. The incorporation of the GRU layer subsequent to the RTA block facilitated proficient capture of temporal dependencies and sequential patterns inherent in 10-s ECG segments, thereby augmenting its efficacy in accurately detecting AF [16]. The RTA block was repeated six times, with the number of kernels beginning with 32, doubling every two iterations, and culminating at 128. By repeating the RTA block multiple times, the model could effectively extract hierarchical representations of the input signals, leading to improved performance in AF detection. The RTA block comprised two components, elaborated upon below.

Trunk branch

Assuming that X, a 10-s segment of ECG data, is provided as input to the trunk branch of the RTA block, the RTA block is repeated six times, with the output of each block serving as the input to the subsequent block. The input X to the trunk branch passes through a convolutional block to generate a feature map X_1, which is then fed to the attention branch.

In the trunk branch, the feature map X_1 is passed through another convolutional block to generate X_2, which is element-wise multiplied with the attention map A (obtained from the attention branch). The product $X_3 = X_2 \odot A$ results in a refined feature map, which is then combined with X_1 using a residual connection. This final refined feature map, X_3, is passed through another convolutional block to generate the feature map X_4 before proceeding to the next layer of the main DL architecture.

Attention branch

The attention branch begins with the intermediate feature map X_1 from the trunk branch. This feature map is then passed through a convolutional block to generate the feature map X_5. To capture global features, the attention branch incorporates both down-sampling and up-sampling operations. Down-sampling is performed using max-pooling, while up-sampling is achieved through nearest-neighbour interpolation. These operations facilitate the extraction of features at different scales, thereby enhancing the model's ability to capture relevant information from the input data.

Following the down-sampling operation on X_5, a convolutional block is applied to expand the feature dimensions, yielding the feature map X_6. Subsequently, an up-sampling operation is performed, and the resulting feature map is fed into another convolutional block, which produces the feature map X_7. This feature map X_7 is then fused with the local feature map X_5 through a residual connection, resulting in the feature map X_8. The fused feature map X_8 is passed through an additional convolutional block to generate the refined feature map X_9, which further refines the

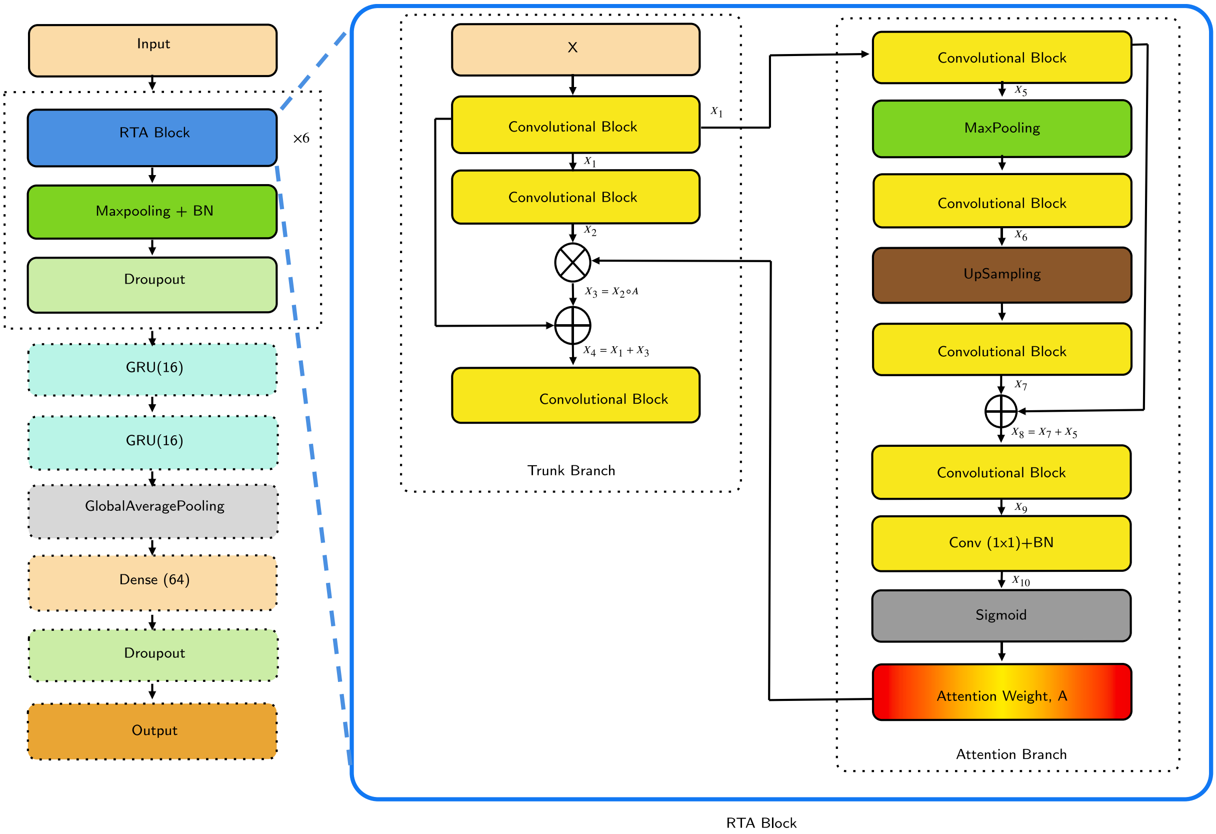

Fig. 2. Diagram of the proposed DL architecture. Here, Conv and BN refer to the convolutional layer and batch normalization, respectively.

attention map.

Finally, the feature map X_9 is processed by a convolutional layer with a 1×1 filter size and a sigmoid activation function. The 1×1 convolution performs a channel-wise transformation, aggregating information across the different feature channels at each time step. By applying this convolution, the layer effectively combines and reweights the input features at each spatial position, producing a compact, yet enriched representation. This process yields the output feature map X_{10}, which contains the temporal attention weights, denoted as A. These attention weights are then used to adjust the importance of each feature within the trunk branch, allowing the model to prioritize the most relevant features of the input data, denoted as X.

Implementation details

We used the focal binary cross-entropy loss function [17] to optimize the model parameters, as depicted by the equation:

$$\text{focal loss}(\mathbf{p}) = -\sum_{j=1}^{C} \alpha_j \left(1 - p_j\right)^{\gamma} log\left(p_j\right)$$

where C is the number of classes (here, 2), p_j is the predicted probability of the j-th class, α_j denotes the balancing parameter of the j-th class to address the class imbalance, γ represents the focus parameter to down-weight easy samples. In our study, we defined the parameters as follows: $\alpha_0 = 0.8$ and $\alpha_1 = 0.2$ for the AF and non-AF classes, respectively. Additionally, we set $\gamma = 3$. The Adam optimizer with a learning rate of 0.001 was utilized to optimize the model parameters. The following hyperparameters were adopted during training: 50 epochs and a batch size of 128. To improve training efficiency, we implemented a learning rate scheduler to adjust the learning rate. Specifically, if no improvement was observed for six consecutive epochs, the learning rate decreased by 75 % of its previous value. These hyperparameter values were chosen on the basis of the performance obtained on the validation set.

Results and discussions

The performance of the DL model was evaluated using key metrics such as sensitivity (Se) and specificity (Sp). Se measures the proportion of correctly predicted positive samples (AF) out of all positive samples. Similarly, Sp measures the proportion of correctly predicted negative samples (non-AF) out of all negative samples. Furthermore, the area under the ROC curve (AUC) was calculated to provide a comprehensive measure of the overall performance of the AF detection model. As shown in Fig. 3, the model correctly classifies an AF record (a) but sometimes incorrectly classifies a non-AF record as AF (b).

Table 2 presents the performance metrics of the proposed DL model on the test set which is part of our dataset: a Se of 0.928, a Sp of 0.915, and an AUC of 0.972 for AF were obtained. We also computed the positive predictive value for our model which it was 0.750. To gauge our model's efficacy against existing state-of-the-art DL models [5–7], we re-implemented them from scratch, considering the same hyper-parameters used in our proposed model. Remarkably, our DL model demonstrated slightly improved performance compared to the other ones. Moreover, we computed the 95 % confidence interval of Se, Sp and positive predictive value using jackknife. These confidence intervals were all less than 0.005 for the three metrics.

In addition, the efficacy of our proposed DL model for AF detection was assessed using an external dataset, namely the IRIDIA-AF dataset. The results yielded a Se of 0.942 and a Sp of 0.932. In Fig. 4, the ROC curve illustrates the trade-off between Se and Sp across different threshold values. In both test sets, we achieved an AUC-ROC> 0.960 for AF detection, indicating a strong discriminatory ability of our model in distinguishing between positive and negative cases.

We further investigated the generalizability of our DL model across diverse demographic groups, particularly focusing on gender and age. The results, in terms of performance, are reported in Table 3. For gender-based analysis, the model exhibited Se rates of 0.931 for males and

Table 2
Comparing the performance of the proposed DL model with other state-of-the-art DL models on our dataset.

DL models	Se	Sp	AUC-ROC
Hannun et al. [5]	0.870	0.913	0.949
Ribeiro et al. [6]	0.851	0.921	0.944
Burke et al. [7]	0.921	0.832	0.951
Our model	0.928	0.915	0.967

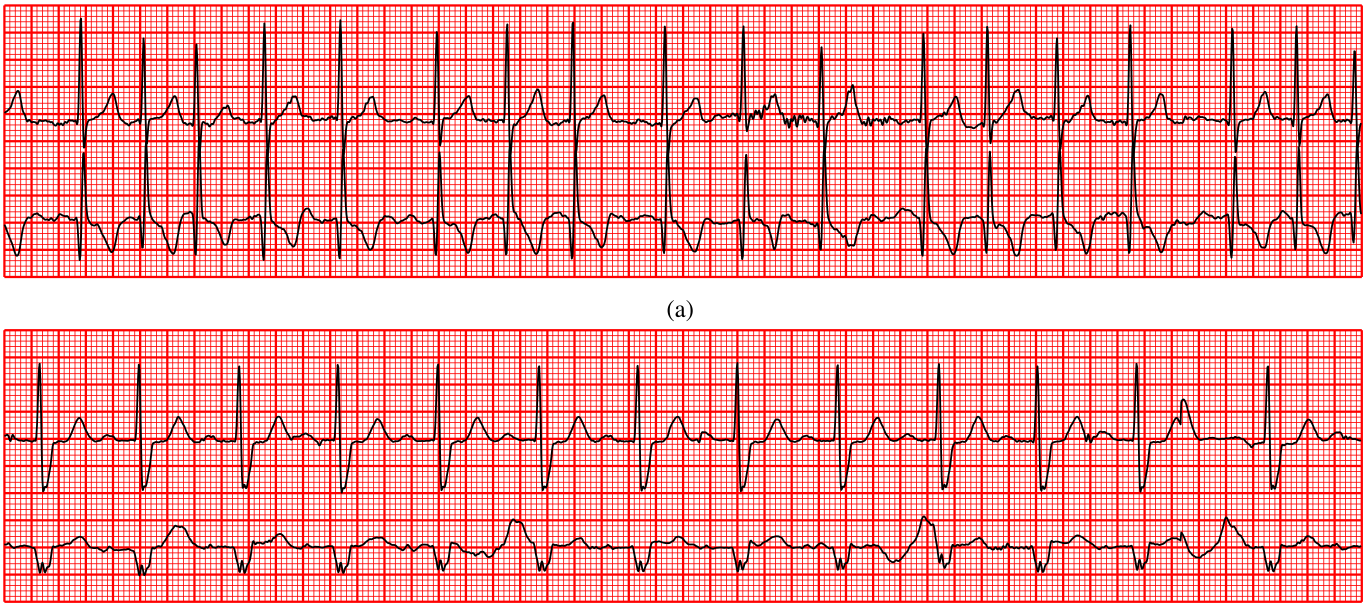

Fig. 3. Example of (a) an AF record correctly classified (true positive) and (b) a Non-AF record incorrectly classified as AF (false positive). Big horizontal squares refer to 0.2 s and big vertical squares to 0.5 mV

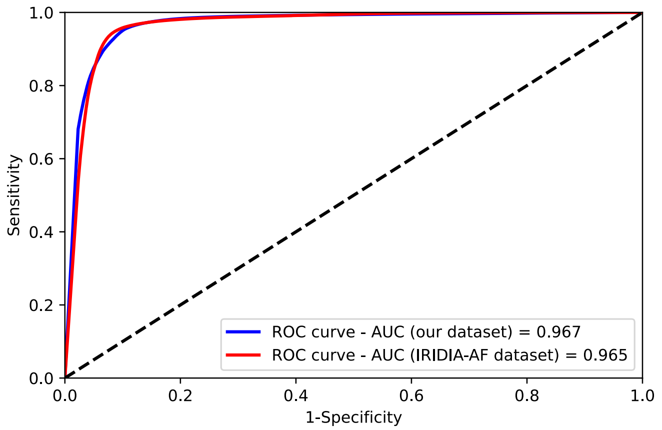

Fig. 4. ROC-curve illustrating the performance of our proposed DL model on our test set and IRIDIA-AF dataset.

Table 3.
Performance of our proposed DL model across demographic groups on the test set of our dataset.

Group	No. of records	Se	Sp
Gender: Male	258	0.938	0.897
Gender: Female	135	0.908	0.949
Age ≤ 60	164	0.904	0.934
Age > 60	229	0.945	0.901
VT	31	0.973	0.832
PVC	41	–	0.934

0.874 for females, alongside Sp of 0.912 for males and 0.951 for females. Regarding age stratification (≤ 60 and > 60), the model demonstrated a Se of 0.901 and 0.894, respectively, with a Sp of 0.922 and 0.913. Moreover, we assessed the performance of our DL model among two distinct cohorts of patients: individuals with an experience of VT or PVC. In the VT group, we achieved a Se of 0.973 and a Sp of 0.832. In the PVC group, we attained a Sp of 0.934. It is noteworthy that the number of 10-s segments for AF in VT patients was fewer than 18,000, while there were no 10-s segments of AF for PVC patients. Furthermore, we evaluated the performance of our DL model for both AF and AFL segments separately and achieved recognition rates of 0.900 and 0.975, respectively.

Finally, the episode-related metrics for AF defined in the EC57 [18] standard were assessed on the MIT-BIH [19] and on our own dataset of 393 Holter records using the wfdb tools from Physionet [20]. For this test, we considered an episode of AF any ECG portion of at least three consecutive (AF-flagged) 10-s segments. The resulting average sensitivities and positive predictivities were 0.87/0.57 in the MIT-BIH and 0.98/0.60 in our test set. These performance metrics are somehow lower than other reported rule-based methods, and they are likely due to the crude definition of an onset/offset of the AF episode and its intrinsic limitation to be a multiple of 10-s blocks and which is one of the primary the focuses of our line of research. Overall, these findings implied the robustness and applicability of our DL model across varied demographic cohorts, indicating its potential for widespread use in clinical settings.

Limitations and future work

Our DL model is designed to analyze ECG data segmented into 10-s intervals, as demonstrated in this study using Holter recordings. Moving forward, we aim to enhance the model's performance by evaluating it at both the episode level and the patient level, addressing this limitation to improve its clinical applicability.

Conclusion

The effectiveness and reliability of our DL model are rooted in its performance on a large and clinically relevant dataset, confirming its practical value. The residual temporal attention-based DL model adeptly identifies key features in AF rhythms, resulting in robust performance across our dataset and an external test set. Notably, it surpasses slightly three existing state-of-the-art DL models, highlighting its superior capabilities. In addition, our model demonstrates consistency and resilience by analyzing diverse demographic groups. These findings underscore the importance of our model in enhancing broader applicability and generalizability in the AF detection task.

CRediT authorship contribution statement

Md Moklesur Rahman: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Massimo Walter Rivolta:** Writing – review & editing, Supervision, Investigation, Formal analysis. **Martino Vaglio:** Investigation, Formal analysis. **Pierre Maison-Blanche:** Investigation, Data curation. **Fabio Badilini:** Writing – review & editing, Supervision, Investigation. **Roberto Sassi:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors acknowledge that they have no conflict of interest.

Acknowledgments

Md Moklesur Rahman acknowledges support from a PhD fellowship funded by CardioCalm srl, Italy.

References

[1] Hindricks G, Potpara T, Dagres N, et al. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2021;42(5):373–498.
[2] Rosero SZ, Kutyifa V, Olshansky B, Zareba W. Ambulatory ECG monitoring in atrial fibrillation management. Prog Cardiovasc Dis 2013;56(2):143–52.
[3] Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet 2019;394(10201):861–7.
[4] Wegner FK, Plagwitz L, Doldi F, Ellermann C, Willy K, Wolfes J, et al. Machine learning in the detection and management of atrial fibrillation. Clin Res Cardiol 2022;111(9):1010–7.
[5] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–9.
[6] Ribeiro AH, Ribeiro MH, Paixão GM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11(1):1760.
[7] Burke D, Carey J, Doggart P, Kennedy A. Novel AI algorithm improves the automated detection of atrial arrhythmias from the apple watch. Heart Rhythm 2023;20(5):S613–4.
[8] Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 3156–64.
[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Informat Process Syst 2017;30.
[10] Moody G. A new method for detecting atrial fibrillation using rr intervals. Proceed Comput Cardiol 1983;10:227–30.
[11] Petrutiu S, Sahakian AV, Swiryn S. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. Europace 2007;9(7):466–70.
[12] Sun Y, Shen J, Jiang Y, Huang Z, Hao M, Zhang X. Mma-rnn: a multi-level multi-task attention-based recurrent neural network for discrimination and localization of atrial fibrillation. Biomed Sign Process Control 2024;89:105747.
[13] Lu X, Wang X, Zhang W, Wen A, Ren Y. An end-to-end model for ECG signals classification based on residual attention network. Biomed Sign Process Control 2023;80:104369.
[14] Vaglio M, Maison-Blanche P, Toninelli G, Isola L, Ferrari F, Badilini F. CER-S, an ECG platform for the management of continuous ECG recordings and databases. In: Computing in Cardiology (CinC). 498; 2022. p. 1–4.

[15] Gilon C, Grégoire J-M, Mathieu M, Carlier S, Bersini H. Iridia-af, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database. Scientific data 2023;10(1):714.

[16] Zhao Q, Yang L, Lyu N. A driver stress detection model via data augmentation based on deep convolutional recurrent neural network. Expert Syst Applicat 2024; 238:122056.

[17] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 2980–8.

[18] A.-A. EC57. Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms, Association for the Advancement of medical instrumentation, Arlington, VA. 1998.

[19] Moody GB, Mark RG. The impact of the mit-bih arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45–50.

[20] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–20.