A deep learning modular ECG approach for cardiologist assisted adjudication of atrial fibrillation and atrial flutter episodes

Quentin Fleury, MSc,[1,4] Rémi Dubois, PhD,[1] Sylvain Christophle-Boulard, MSc,[4] Fabrice Extramiana, MD, PhD,[2,3] Pierre Maison-Blanche, MD[2]

From the [1]IHU Liryc, Université de Bordeaux, Bordeaux, France, [2]Cardiology Department, Bichat Hospital, Paris, France, [3]Université Paris Cité, Paris, France, and [4]Microport CRM, Clamart, France.

BACKGROUND Detection of atrial tachyarrhythmias (ATA) on long-term electrocardiogram (ECG) recordings is a prerequisite to reduce ATA-related adverse events. However, the burden of editing massive ECG data is not sustainable. Deep learning (DL) algorithms provide improved performances on resting ECG databases. However, results on long-term Holter recordings are scarce.

OBJECTIVE We aimed to build and evaluate a DL modular software using ECG features well known to cardiologists with a user interface that allows cardiologists to adjudicate the results and drive a second DL analysis.

METHODS Using a large (n = 187 recordings, 249,419 one-minute samples), beat-to-beat annotated, two-lead Holter database, we built a DL algorithm with a modular structure mimicking expert physician ECG interpretation to classify atrial rhythms. The DL network includes 3 modules (cardiac rhythm regularity, electrical atrial waveform, and raw voltage by time data) followed by a decision network and a long-term weighting factor. The algorithm was validated on an external database.

RESULTS F1 scores of our classifier were 99% for ATA detection, 95% for atrial fibrillation, and 90% for atrial flutter. Using the external Massachusetts Institute of Technology database, the classifier obtains an F1-score of 97% for the normal sinus rhythm class and 96% for the ATA class. Residual errors could be corrected by manual deactivation of 1 module in 7 of 15 of the recordings, with an accuracy < 90%.

CONCLUSION A DL modular software using ECG features well known to cardiologists provided an excellent overall performance. Clinically significant residual errors were most often related to the classification of the atrial arrhythmia type (fibrillation vs flutter). The modular structure of the algorithm helped to edit and correct the artificial intelligence–based first-pass analysis and will provide a basis for explainability.

KEYWORDS Electrocardiology; Deep learning; Long-term ECG; Atrial fibrillation; Atrial flutter

(Heart Rhythm O^2 2024;5:862–872)

Introduction

Atrial Fibrillation (Afib) is the most common arrhythmia, affecting 2% to 4% of the population.[1,2] Because of population aging and increase in obesity, Afib prevalence is estimated to increase sharply in the next decade. In addition to disabling symptoms, Afib is associated with the occurrence of stroke, heart failure, cognitive decline, and increased mortality.[3] Stroke prevention with oral anticoagulant treatment and improved efficacy of rhythm control strategies have decreased the odds for thromboembolic events and mortality in case of Afib-related heart failure.[4] The implementation of early appropriate treatment requires timely accurate diagnosis of Afib and other atrial tachyarrhythmias (ATA), atrial tachycardia (AT), and atrial flutter (Afl).

Diagnosis of ATA is typically made in daily clinical practice by standard 12-lead electrocardiogram (ECG) recordings and ambulatory 24- to 48-hour Holter monitoring. These methods are appropriate in case of permanent ATA but may lead to underdiagnosis in case of intermittent ATA. Patients with symptoms have an increased likelihood of detecting an ATA. However, there is a well-known poor correlation between symptoms and ATA episodes, because a significant percentage of Afib episodes are clinically silent; up to 50% in symptomatic paroxysmal Afib patients[5,6] and up to 10% in asymptomatic patients without history of Afib included in the ASSERT study.[7] In addition, undiagnosed intermittent and asymptomatic ATA episodes are associated with an increased risk of stroke.[7] Accordingly, such subclinical Afib could represent a missed opportunity for the prevention

Address reprint requests and correspondence: Quentin Fleury, MSc, Bordeaux University Foundation, IHU Liryc, France. E-mail address: quentin.fleury@ihu-liryc.fr.

2666-5018/

https://doi.org/10.1016/j.hroo.2024.09.007

KEY FINDINGS

- We aimed to build and evaluate a deep learning (DL) modular software using ECG features well known to cardiologists to classify abnormal atrial rhythms, including both paroxysmal and sustained episodes of atrial fibrillation (AFib) and atrial flutter (Afl) with a user interface that allows cardiologists to adjudicate the results and drive a second DL analysis.
- The software provided an excellent overall performance (F1 scores of our classifier were 95% for Afib and 90% for Afl, respectively), and significant residual errors were most often related to the classification of the atrial arrhythmia type (atrial fibrillation vs atrial flutter).
- The modular structure of the algorithm helped to edit and correct the AI-based first-pass analysis and provide a basis for explainability.

of Afib-related morbidity. In addition, accumulated data from cohorts with long-term ECG recordings have built a strong case for the concept of an increased risk of Afib complication (stroke but also heart failure) with increased Afib burden (ie, the proportion of time with ATA).[8–10] Conversely, patients with a low ATA burden have low risk of thromboembolic complication and have a less favorable risk/benefit ratio with oral anticoagulant treatments.[11] The availability of long-term ECG monitoring techniques (external loop event recorders, 7–30 days beat-to-beat recorders)[12] increases the likelihood to diagnose subclinical Afib episodes and allows precise quantification of Afib burden. However, editing by technicians and physicians of such massive ECG data is a challenge.

Historically, physicians make Afib and other ATA statements on the ECG by interpreting heart rate variability (regularity, irregularity, specific pattern) together with atrial activity pattern (including rate and morphology).[13] However, integration of P-wave features is difficult on long-term ECG recording because the P wave has a low amplitude and therefore the atrial electrical signal is most often blurred in noise. Few studies have integrated the P-wave signal in Afib detection, and it does improve Afib detection but marginally.[14] Hence, most automatic software for long-term ECG recording is based on RR intervals time series, after running a cardiac beat detector and computation of RR interbeat intervals.[13,15]

Recently, many publications using deep learning approaches (DL) on resting ECG have been implemented on large, annotated ECG data sets. These algorithms demonstrated their ability to process and analyze cardiac rhythms with accuracies higher than 90%.[16,17] Although many reports are available about DL performances on resting ECG databases, results on long-term Holter recordings are scarcer, primarily because of the lack of annotated databases, because expert annotation is time consuming.

We hypothesized that a DL classifier for atrial arrhythmias based on a modular structure mimicking expert physician ECG interpretation would provide (1) an accurate atrial arrhythmias detection solution and (2) a fast and intuitive tool for physician editing.

Therefore, we aimed to (1) build a DL modular software using ECG features well known to cardiologists; (2) build a user interface that allows cardiologists to edit Holter DL outputs in a way that is familiar to them; and (3) evaluate the first-pass performance of the DL classifier as well as a second-pass performance after the physician's adjudication.

Method

Material

Holter database

Since January 2007, Holter recordings from patients referred to the Cardiac Arrhythmias Unit in Centre Medico Chirurgical Ambroise Pare (Neuilly, France) have been included in a prospective Holter Data Base (H-DB).[18] Included patients underwent Holter ECG monitoring on physician prescription, usually to document the ECG at the time of symptoms. More recently, documentation of silent atrial arrhythmias was also a recruitment criterion.

Digital Holter devices (Spiderview, Microport, Clamart, France) were 2-lead ECG recorders at 200 samples per seconds and with an amplitude resolution of 10 μV. We selected a set of 282 Holter recordings of good quality. The quality was considered good when the automated analysis performed by the Holter Editing System (SyneScope) lasted at least 59 minutes for each hour of recording (ie, the total duration of discarded segments attributable to noise or artifacts is less than 1 min/h). Two categories of Holter recordings were selected from the H-DB, Holters with either paroxysmal or permanent (over the 24 hours) atrial arrhythmias, and Holters in sinus rhythm but with sporadic atrial of ventricular premature beats. The frequent association within a single patient of Afib with other arrhythmias was the rationale to include atrial flutter episodes in this study. No other supraventricular tachycardia (atrioventricular nodal reentrant tachycardia, atrioventricular reentrant tachycardia, AT) were included in the H-DB.

All recordings in the H-DB were de-identified, all demographic and clinical information being removed. The information used in this study was only the voltage-by-time raw digital data from the 2 ECG leads. The Holter database is compliant with European privacy regulatory requirements.

Expert annotation

Holter recordings were manually annotated on a beat-to-beat basis by a certified cardiologist, using the usual Holter editing tools (Dr P. M.-B.). Briefly, beat-to-beat editing was built on visual inspection of heart rate trends, template morphologies (normal and abnormal), long and short R-R intervals, and R-R interval ratios, supraventricular and ventricular arrhythmias (isolated extrasystoles, couplets, runs, and sustained

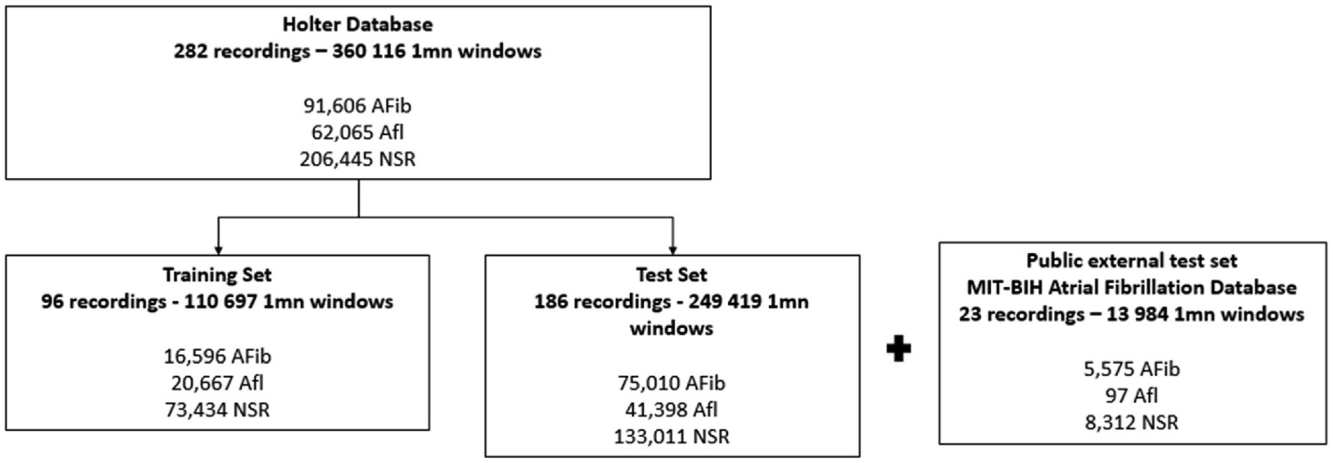

Figure 1 Description of the Holter datasets. The Holter Database contains samples of 2 channel ambulatory recordings collected at the Cardiac Arrhythmia Unit (Neuilly Sur Seine, France) and samples of the MIT-BIH database. The cohort was divided into 2 groups for training and testing, respectively. The total amount of data was 110,697 one-minute windows and 249,419 one-minute windows for training and testing, respectively. The minutes were labeled as normal sinus rhythm (NSR), or atrial fibrillation (Afib) and atrial flutter (Afl) according to the expert annotation. Numbers in each dataset box indicate the total number of NSR, Afib, and Afl minutes.

cardiac arrhythmias, by rate and by length). When a cardiac beat or an arrhythmia episode was misclassified by the system, it was manually relabeled. Over and under detections were also corrected using beat insertion or deletion tools.

ATAs were defined as any arrhythmic episode with supraventricular beats and absence of P waves of sinus origin. First, onset and offset of each atrial arrhythmia episode were visually identified using heart rate trends, full disclosure, and paging mode. Then, using electronic calipers, the specific onset and the offset were selected, and the Holter system automatically created a corresponding "time-period" of atrial arrhythmia. The duration of the episode was also calculated by the system between the onset and the offset times.

MIT-BIH atrial fibrillation database

To evaluate the performance of the proposed method on a publicly available database, and under regular noise condition, results of the MIT-Beth Israel Hospital (BIH) Atrial Fibrillation Database are also reported in this study.[19,20] It includes 23 ECG recordings lasting 10 hours sampled at 250 Hz with 10 μV resolution with manually reviewed beat annotations. The database was primarily developed to support research in the field of cardiac arrhythmia detection and analysis and is a collaborative effort between the Massachusetts Institute of Technology (MIT) and the Beth Israel Hospital (BIH).

Data sets definition and recording segmentation

The H-DB was randomly divided into 2 subsets: the first for algorithm training and the second for further testing. The training data set consisted of 96 Holter recordings with 3 subcategories according to the presence of Afib, Afl, or normal sinus rhythm (NSR) episodes. Up to 93 recordings lasted more than 1200 minutes, and the shortest duration was 598 minutes. This dataset included 12 patients with permanent Afib and 16 with permanent Afl. The testing dataset consisted of 186 Holter recordings, with up to 168 recordings longer than 1200 minutes and a minimum recording duration of 177 minutes. This dataset included 51 patients with permanent Afib and 22 with permanent atrial flutter. The MIT-BIH Atrial Fibrillation database was also used as a testing dataset. The DB characteristics are summarized in Figure 1. First, all recordings were truncated into consecutive 1-minute segments. Then windows with less than 20 valid cardiac beats labels were removed. Figure 1 shows that the testing set of the H-DB includes a total of valid 249,419 one-minute windows, 75,016 in Afib, 41,398 in Afl, and 133,011 in NSR, respectively.

The MIT-DB was used as an external testing set.

Global structure of the DL classifier

As illustrated in Figure 2, the overall classifier was based on 3 consecutive processing steps. The first stage combined 3 separate expert neural networks, each network focusing on a given ECG pattern typical of atrial arrhythmias, (1) the raw ECG voltage by time data, (2) presence or absence of P-wave morphology, and (3) R-R interval statistics, and each network trained to recognize either NSR, Afib, or Afl episodes. Of note, the network applied to R-R intervals was lead independent, whereas the 2 others were lead dependent. The output from each network is a probability score ranging from 0 to 1, indicating the likelihood of belonging to a specific arrhythmia class, resulting in a 10-probability vector per minute of recording with 5 individual probabilities and 2 classification tasks (NSR vs Afib, and NSR vs Afl). The second stage involves a decision neural network that combines these 10 probability scores and takes a decision from it to classify the minute of recording under 1 of the 3 cardiac rhythm categories (Afib, Afl, and NSR). The final stage is essential for transitioning the truncated minute-by-minute decisions to the total duration of the recording. This is accomplished through the application of a Hidden Markov Model

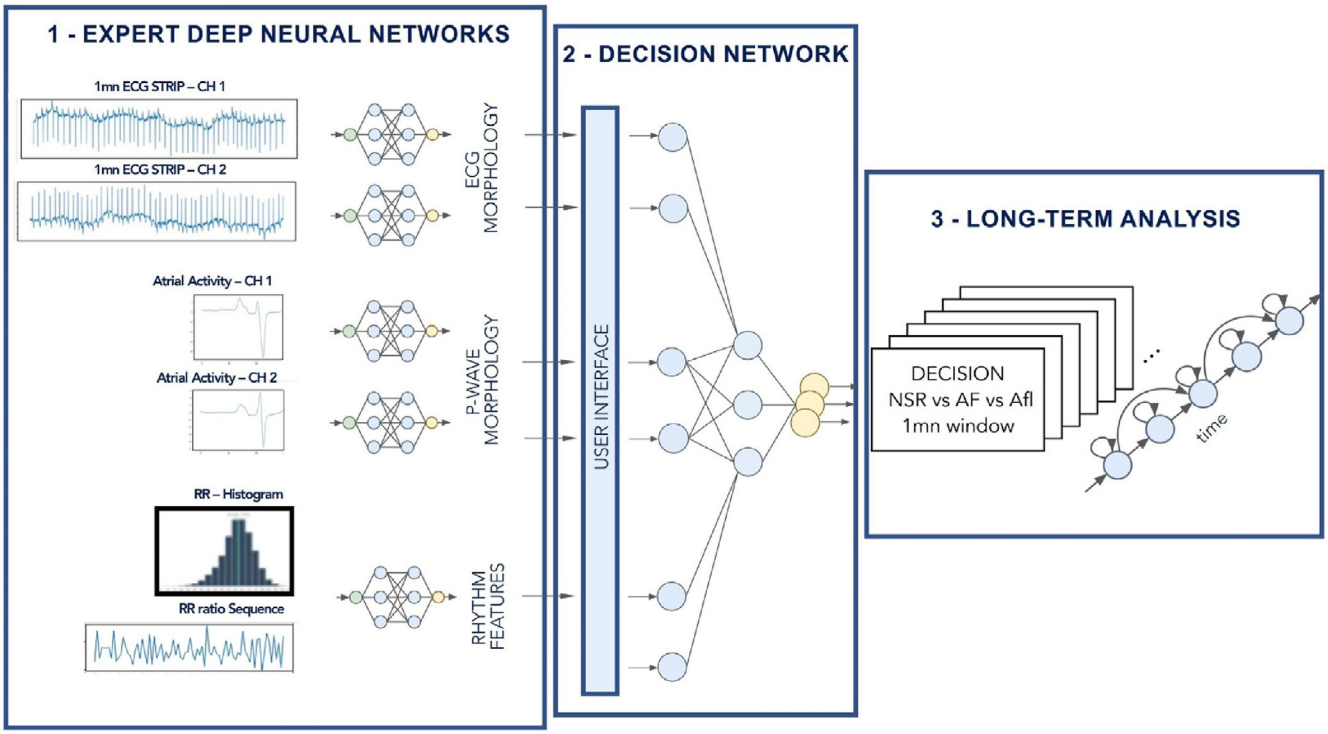

Figure 2 DL classifier global structure. The classifier that processes 1-minute window of ECG and RR embed 3 panels: • The first panel consists of expert networks that focus on global morphology, mean P-wave, and RR intervals of the window. • The second panel is the arrhythmia decision part of the classifier. In this panel, the user can interact with the network to disable expert networks. • The last panel is present to consider the temporal aspect of Afib and Afl episodes. This panel analyzes the whole Holter recording and assigns to each minute a label (Afib, Afl, NSR).

(HMM), which effectively models the temporal transitions between different states of cardiac rhythm.

See Appendix A for further details about each panel of the classifier.

Statistical analysis

The classifier's performances were evaluated using the F1 score during the learning process.[21] The F1 score is a measure of predictive performance, often used in machine learning systems, particularly for evaluating medical performances. It is calculated as a tradeoff between positive predictive value and sensitivity, thus enabling us to assess a classifier's ability to detect an event without erroneously classifying too many positive events. Strictly speaking, the F1 score is the harmonic mean of positive predictive value and sensitivity. However, its medical accuracy is reported here through 3 metrics: accuracy, sensitivity, and specificity, with 3 objectives: (1) Identifying the patients that show at least 1 episode of any ATA during the recording; (2) Determining the ATA burden (number of minutes spent in ATA) during the Holter recording; and (3) Classifying the ATA category (either Afib or Afl).

At level 1, evaluation is made at the "patient level," asserting that any patient with at least 1 minute of Afib or Afl will belong to ATA. Based on comparison between the expert annotation and the investigated DL model outputs, sensitivity and specificity were calculated (ratios including number of recordings tagged as ATA and numbers of recordings free of ATA). We considered the results according to patient classification: a patient detected as ATA was a patient with at least 1 minute of ATA labeled by the classifier overlapping with 1 minute of expert ATA, and an NSR patient was a patient for whom no minutes had been detected by the classifier as ATA.

At level 2, the focus shifts to the accuracy of minute-by-minute detection by the classifier, comparing the respective labels from both the expert and the classifier. Accuracy is a metric for evaluating classification models, and it is calculated by dividing the number of correct predictions by the total number of predictions. Here, Afib and Afl labels are pooled again under a single ATA category; therefore, level 2's performance metric distinguishes between NSR and ATA classifications. After the accuracy analysis, patients were sorted based on the accuracy level of the detected ATA burden. When the accuracy reached 100%, the DL classifiers' outputs match expert annotations perfectly, eliminating the need for any re-annotation. For those with 99% accuracy, equivalent to approximately 15 minutes of discrepancies in a 24 -hour recording (1% of data), the results are also highly reliable, necessitating minimal or even no relabeling by the expert. An accuracy of 95% indicates a relatively reliable detection process (an error for approximately 70 minutes of data), some 1-minute windows requiring relabeling. Finally, an accuracy of 90% or an error rate of 10% means a need to review the results.

Table 1 NSR and ATA results depending on the accuracy level

Accuracy	NSR recordings (Total: 80)	ATA recordings (Total: 106)	Permanent ATA recordings (Total: 74)	Paroxysmal ATA recordings (Total: 32)
100%	68	66	64	2
99%	76	96	70	26
95%	78	102	71	31
90%	79	106	74	32

ATA = atrial tachyarrhythmias; NSR = normal sinus rhythm.

Level 3 details the ATA category, differentiating and reporting the outcomes for Afib and Afl separately, providing a detailed view of the classifier's capability in distinguishing between the 2 types of ATA. The Bland-Altman analysis was used to quantify agreement between expert ATA annotations and classifier outcomes.[22] See Appendix B for further details.

Results

Automated first-pass analysis

At level 1 of the statistical analysis, ATA was detected in 105 of the 106 recordings, with ATA episodes corresponding to 99% sensitivity. The only false-negative ATA detection was related to a short 2-minute AFib episode on a 24-hour recording. When paroxysmal ATA episodes were sorted by length, this episode was the shortest one, and the next ATA episodes lasted 5 minutes and were correctly identified. Specificity for ATA detection at the patient level was 85%.

When considering each 1-minute recording segment separately, the performance of our model to discriminate between NSR and ATA was high (F1 score 99% for NSR and ATA) on the test set.

Considering level 2, we calculated the accuracy for each recording. Table 1 shows categorical classification at different level of accuracy. All ATA recordings and 98.75% of NSR recordings had an accuracy $\geq$ 90%.

For recordings with paroxysmal ATA episodes, the mean difference between expert annotation and detected outcomes was 8.2 minutes (range, 0–50 min). ATA windows were correctly detected in 22 of 32 recordings.

For recordings with permanent ATA, the mean difference between expert annotation and detected outcomes was 5 minutes (range, 0–134 min). In 64 of the 74 permanent ATA recordings, all expert ATA windows were correctly detected.

The average overall error was 6 minutes (range, 0–134 min). In 86 recordings, all ATA expert-adjudicated windows were correctly detected.

Focusing on the 40 ATA recordings whose accuracy was less than 100%, the average overall difference was 15.5 minutes (range, 1–134 min), 8.5 minutes and 36 minutes in paroxysmal and permanent datasets, respectively.

Regarding NSR recordings, 12 cases had an overdetection of ATA episodes. The difference was negligible in 8 recordings (range, 2–6 minutes), and 1 patient had an error rate of 97% (1150 minutes).

Figure 3 shows the Bland-Altman plots for the 2 ATA detection dataset expert annotation and the DL-based algorithm. The systematic bias was –6.14 min with a 2SD = 172.13 min.

At level 3, we scored the classifier on its ability to distinguish between Afib and Afl patterns among detected ATA. Considering each 1-minute recording segment separately, our model could discriminate Afib and Afl episodes with F1 score of 95% and 90%, respectively. When considering only ATA minutes that had an overlap between expert annotation and detection by the classifier

(true positive minutes), 74 recordings of 106 had their arrhythmias correctly classified between Afib and Afl. The proportions of Afib and Afl minutes correctly detected by the classifier were 92% and 98%, respectively, for an overall classification success rate at 94%.

Manual second-pass analysis

Fifteen Holter recordings showed an accuracy below 90%. Detailed characteristics for each recording are given in Table 2. Of note, 11 of the 15 recordings showed a permanent atrial arrhythmia.

For 7 of 15 recordings, an intuitive single-module deactivation was enough to correct the automated analysis. For instance, case 1 in Table 2 is a permanent Afib case according to the expert, but the algorithm reports 994 Afib minutes, 194 Afl minutes, and 6 NSR minutes. An additional expert statement was permanent atrioventricular dissociation with regular rhythm and bradycardia (44 beats/min), and in this setting RR statistics are not informative for detection of ATA.

Disabling the RR module (Figure 4A) allowed relabeling Afl and NSR as Afib, so with a single click the accuracy was improved from 83.2% to 100%.

Another representative case was case 78, a permanent Afl lasting 1354 minutes, according to the expert. The classifier detected Afl for 1117 minutes and 237 minutes of Afib. Visual inspection of the ECG traces at the time of detected Afib minutes (Figure 4B) showed that the atrial electrical activity (sawtooth F-waves) was better identified in lead A. As expected, disabling lead B module and recomputing the data improved accuracy from 82.2% to 100%.

For the 7 cases in which a single manual modification was enough to improve the accuracy, the most frequent intervention was disabling a single-lead ECG. Recomputing the DL analysis took a few seconds for 24 hours of data on our computer.

External cohort validation: MIT-BIH Atrial Fibrillation database

Considering the low number of 1-minute windows of atrial flutter, we considered only 2 cardiac rhythm classes: normal sinus rhythm and ATA. The classifier obtains an F1-score of 97% for the NSR class and 96% for the ATA class.

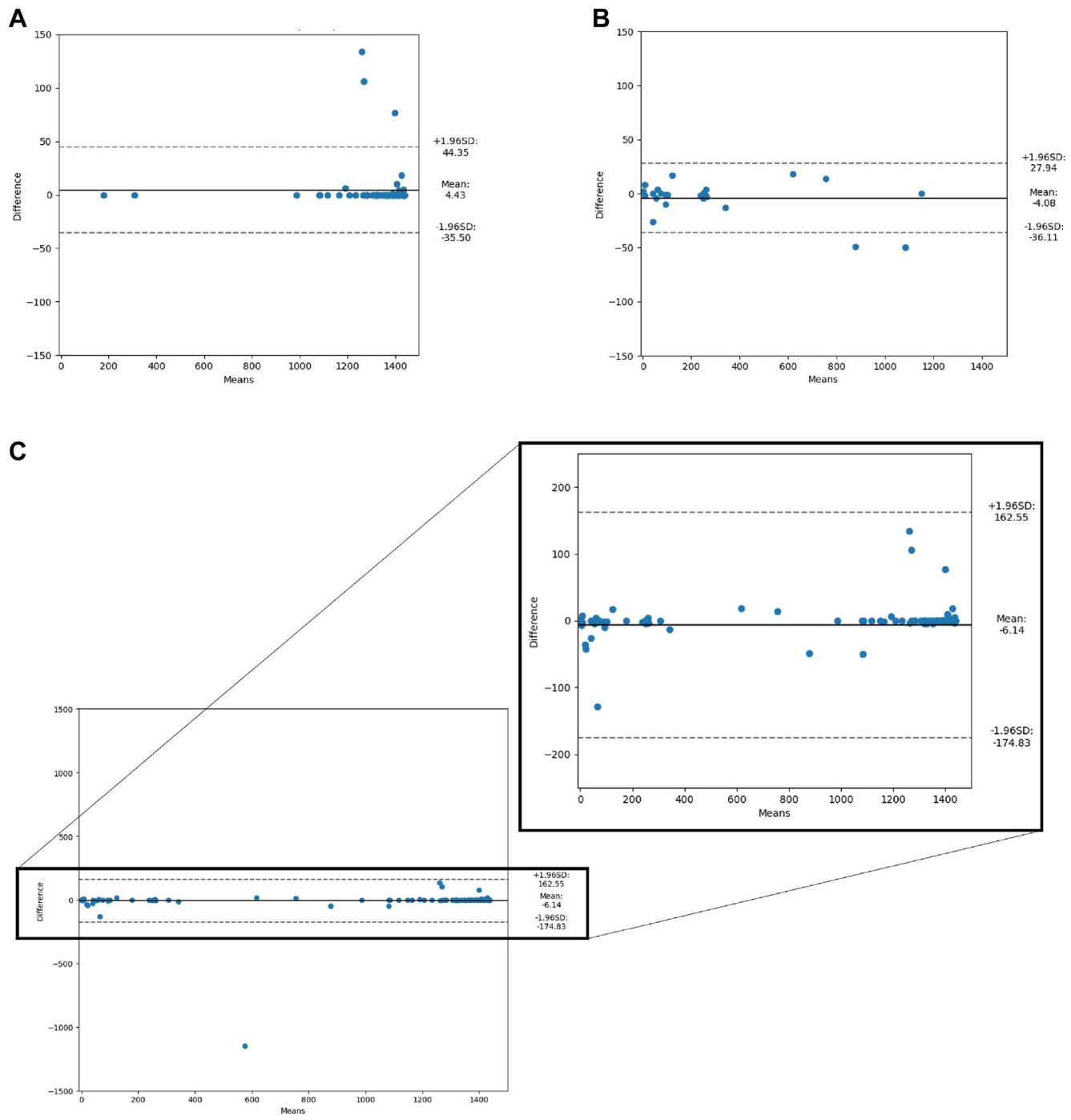

Figure 3 Bland-Altman plot between the duration of ATA of the 2 methods (expert vs DL) in case of **A:** permanent ATA (n = 74), **B:** paroxysmal ATA (n = 32), and **C:** on the whole dataset (n = 186). The mean number of minutes by the classifier is plotted along the x-axis, and the difference between the expert annotation and that of the classifier along the y-axis. The dotted gray lines represent 95% confidence intervals for the mean.

Discussion

The major findings of our study are the following: Using a large, beat-to-beat annotated, 2-lead Holter database (282 Holter recordings generally lasting more than 1200 minutes) with paroxysmal and sustained atrial fibrillation and atrial flutter episodes, we built a modular DL software to classify atrial rhythm. Our model combined both preprocessed ECG data and raw ECG data to use ECG features based on expert cardiologist reasoning for ECG diagnosis. The overall performances of our classifier were excellent, with F1 scores above 0.95. Nonetheless, there were clinically significant residual errors, most often related to the classification of the atrial arrhythmia type (Afib vs Afl). Using a dedicated interface, we could easily and quickly correct approximately half of erroneous diagnostics. The most frequent successful editing process was disabling 1 of the ECG leads.

Table 2 Description of expert annotation minutes and detected minutes on the 15 patients below the 90% accuracy threshold

DB Nb	Analyzed time	Expert			Detected			Accuracy	Second pass
		Afib	Afl	ATA	Afib	Afl	ATA		
40	1182	0	0	0	0	1150	1150	0.027	
70	1307	1250	57	1307	0	1307	1307	0.043	
49	1386	1386	0	1386	176	1210	1386	0.126	
142	1387	1387	0	1387	371	1016	1387	0.267	
64	1437	1437	0	1437	423	996	1419	0.294	
124	1400	774	283	1057	596	511	1107	0.786	
78	1354	0	1354	1354	237	1117	1354	0.824	♥
1	1194	1194	0	1194	994	194	1188	0.832	♥
120	1322	0	1322	1322	87	1129	1216	0.854	
12	1413	0	1413	1413	175	1228	1403	0.869	♥
115	1437	0	1437	1437	1251	186	1437	0.87	♥
33	1328	166	1162	1328	38	1156	1194	0.881	♥
107	1433	0	1433	1433	150	1281	1431	0.893	♥
172	1164	1164	0	1164	1040	124	1164	0.893	
150	852	252	0	252	164	87	251	0.896	♥

The heart symbol in the second-pass column means that a single-module deactivation in the user interface is required to correct the automated analysis. Afib = atrial fibrillation; Afl = atrial flutter; ATA = atrial tachyarrhythmias.

Resting ECG and computerized interpretation of the ECG

Efforts to automatize ECG measurement and interpretation started in the late 1950s, leading to improvement such as the introduction of miniature digital computers and interpretative softwares. In 1990, the Common Standards for Quantitative Electrocardiography (CSE Project) under the leadership of Professor Joss Willems in Leuven in Belgium established databases containing, not only the raw digital ECG waveforms for testing measurements but also the consensus ECG interpretations performed by the CSE experts who met for over 10 years. The concept of ECG databases remain critical for the development of new ECG softwares.[23,24]

The quality of outputs by ECG interpretation programs has been consistently questioned. However, keeping in mind that millions of ECGs are collected and analyzed annually in the setting of diagnostic statements at bedside, or serial comparisons of ECG or epidemiological studies, reduction of physician reading time could be of major benefit. Computer-assisted ECG interpretation decreased analysis time by up to 24% to 28% for experienced readers.[25–27]

Regarding cardiac rhythm statements, it is generally admitted that physician overreading to correct computer-based electrocardiogram rhythm diagnoses remains mandatory.[28,29]

De Bie et al[30] evaluated the accuracy of 7 ECG interpretation programs in detecting abnormal rhythms. Digital ECGs were analyzed by the manufacturers' interpretation programs, focusing on the ability to distinguish sinus rhythm from non-sinus rhythm, and to identify atrial fibrillation/flutter and other abnormal rhythms. All programs could distinguish between sinus and non-sinus rhythms. However, false-positive rates varied from 2.1% to 5.5%. False-negative rates varied from 2.7% up to 55.9%, and the authors concluded that physicians should not rely on computer statement alone.[30]

Shah and Rubins[31] also observed frequent errors in the interpretation of nonsinus rhythms and recommends expert overreading. Published studies consistently agree that ECG algorithms are efficient to detect normal sinus rhythm. The difficulty in making a correct diagnosis of the underlying rhythm is typically linked to recognizing P waves with a small amplitude, varying P-wave morphologies, or P waves masked by underlying noise, QRS complexes, or T or U waves, paced rhythms, or tremor.[31–34]

Artificial intelligence enhanced electrocardiography

The most recent development in the field of automated ECG analysis has been the use of artificial intelligence (AI), including a variety of machine learning techniques to aid interpretation. With better machine learning algorithms, computerized interpretation of ECGs has clearly improved arrhythmia detection, achieving an accuracy close to 95%.[35,36] Hannun and coauthors[16] developed a DL approach for ECG analysis by using a deep neural network for identifying 12 rhythm abnormalities by using single-lead ECGs. When validated against independent data reported by a committee of certified cardiologists, their algorithm was shown to be superior to an average cardiologist in identifying these rhythm abnormalities (receiver operating characteristics, 0.97 vs 0.78).[16,37]

Long-term ECG

In a recent European Heart Rhythm Association position paper, experts recognized the central dilemma in evaluating optimal monitoring duration in the AFib search. Even very long monitoring (30 days) may be insufficient to detect all episodes. To define arbitrarily a standard monitoring time, the experts stated that a monitoring duration of a minimum of 2 weeks of continuous monitoring is required to maximize AF detection. Editing of such massive beat-to-beat ECG

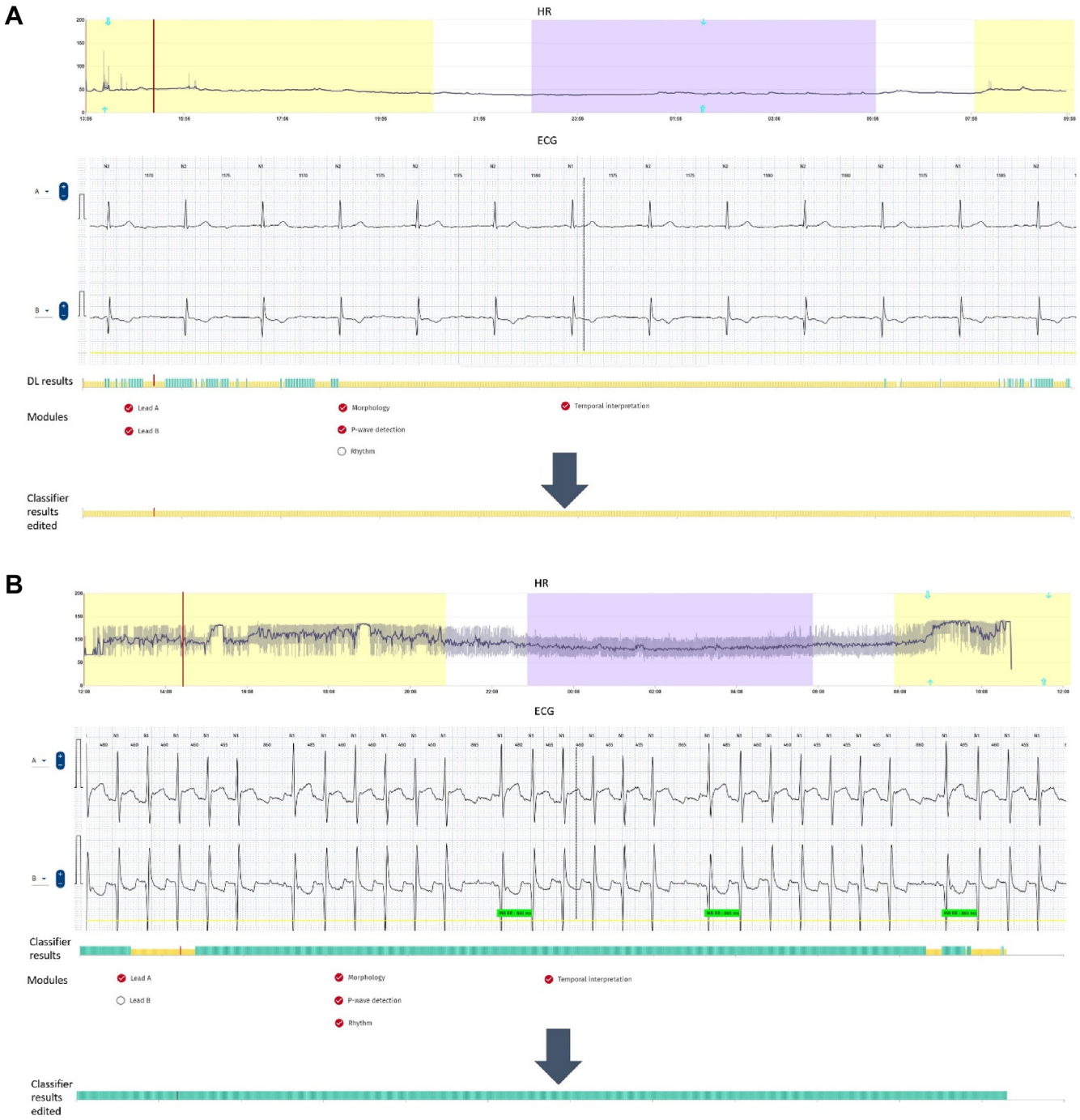

Figure 4 Configuration of the user interface screen and 2 representative use cases from the H-DB (**A** and **B**). The user interface screen layout consists of the following components, from top to bottom: Heart rate (HR) trend window to review HR at any time selected with a mowing cursor. The HR format is maximum/average/minimum. ECG strip window to review the ECG at the time of the HR cursor location. Individual beat labels and beat-to-beat RR intervals are displayed in the ECG window. DL classifier output window: The diagram represents the distribution of ATA events across the 24 hours with a color scheme, *yellow markers* for Afib minutes and *green markers* for Afl minutes. The DL outputs are locked to HR trend and ECG strips. DL control window to enable or disable the modular components of the DL classifier, from left to right: individual ECG leads A or B, individual networks morphology, P wave and RR statistics, and temporal analysis button (HMM) Figure 4a corresponds to the use case of recording 1 in Table 2. It is a permanent atrial fibrillation with atrioventricular dissociation and therefore a perfectly flat HR trend and regular RR intervals in the ECG window. Raw DL classifier outputs are both Afib and Afl segments. Disabling the meaningless RR statistics module (*vertical thick arrow*) restores a valid output with only Afib minutes. **B:** corresponds to the use case of recording 78 in Table 2. It is a permanent atrial flutter obvious in lead A. Raw DL classifier outputs are both Afib and Afl segments. Disabling the analysis on lead B (*vertical thick arrow*) restores a valid output with only Afl minutes.

data (approximately 100,000 cardiac beats per 24 hours) requires new techniques for ECG analysis, and AI could be a significant improvement.[38]

Basic Holter algorithms have not changed much, and together with handling very-long-term ECG data, noise remains a key challenge in ECG signal processing. Holter devices have been specifically developed to record the ECG during daily out-of-hospital activities, and reducing ambient noise is a major processing task. It is even more difficult because the number of ECG leads is reduced, down to 2 or only 1 lead. In this study, we included high-quality Holter recordings, and reducing noise was not a key issue. In a more standard noise environment, filtering the ECG remains needed keeping in mind that for atrial arrhythmias detection the analysis of the atrial electrical activity is crucial.[39] Finally, for P-wave signal analysis, more specialized preprocessing methods than that focusing on enhancing the QRS complex could be recommended.

DL models need large databases to train, but private or publicly available Holter databases with beat-to-beat annotations are not common.[40] Ivora et al[40] used ambulatory ECG recordings consisting of 12,111 single-lead Holter ECG recordings, but each recording was 30 seconds long, sampled at 200 Hz. In the THEW project, Holter recordings in the warehouse were provided by research academic centers and major pharmaceutical companies. Afib recordings were few and lasted only 10 minutes.[41]

Typically, DL algorithms for ambulatory ECGs are based on so-called "supervised" and "unsupervised" machine learning. Supervised learning uses labeled data for training, and unsupervised learning uses only raw data. Deep learning–based ATA detectors can therefore be categorized according to whether the R-R intervals and R-R intervals ratios or only the raw ECG serve as inputs to the network.[42–44] As far as ambulatory ECG is concerned, most studies focus performances of DL techniques to detection of Afib episodes, and the distinction between Afib and Afl is less well studied.

Ben-Moshe et al[44] have compared the performance of R-R interval inputs versus single-lead raw ECG-based ATA detectors, Afib episodes, and Afl episodes.[44] The authors collected 321 Holter recordings from 3 cardiology centers with manual beat-to-beat annotations. The authors concluded that the raw ECG network significantly outperformed the model, using R-R intervals as input. However, the performance was not similar across ECG leads, with some leads performing better for Afib episodes detection. In the Ben-Moshe et al[44] study, error analysis showed that a large percentage of false-negative windows from the raw ECG network contained AFl episodes. The authors suggested that this could be explained by the fact that DL models need a large database to train. The percentage of AFl windows in the training set was very small. In our study, we combined supervised and unsupervised networks, with 2 leads of raw ECG data. The total number of 1-minute Afl windows in the training was 20,337 from 96 recordings. Differentiating between AFib and Afl is essential because the management of patients is not similar. For Afib, patient rhythm control and rate control are recommended strategies. For most Afl patients, catheter ablation is preferred over pharmacologic therapy because of the high success rate. Atrial and ventricular arrhythmia categories are numerous, and because the prevalence of some arrhythmias is extremely low in most Holter recordings, we could not include in our study atrial tachycardia episodes, although it is obviously as critical as differentiating Afib and Afl for clinical purpose.[45] Comprehensive rhythm ECG analysis by DL-based algorithms will probably not be available in the next future. In the same way, detection of atrial extrasystoles is a conventional task for any Holter algorithm. Building a high sensitivity algorithm to discriminate arrhythmias from regular sinus rhythm is not challenging. However, some atrial arrhythmias such as frequent isolated atrial extrasystoles, or sustained atrial runs, may mimic Afib or Afl episodes, leading to false-positive detections.

Our results, based on a large database of long-term recordings, show that an AI-based strategy is associated with very high performances to detect atrial arrhythmias. However, although also associated with high metrics, the correct classification of atrial fibrillation and atrial flutter is associated with lower performances.

Modular architecture and future developments

Modular architectures are commonly used for AI-based algorithm. In the current study, we purposely designed our DL model with a modular setup in which each software module carries out a well-defined ECG-processing task, such as the architecture of the MEANS ECG software.[26,46] Our choice was based on the hypothesis that a DL classifier for atrial arrhythmias based on a modular structure mimicking expert physician ECG interpretation would not only provide accurate atrial arrhythmias detection but also help to edit and correct the AI-based first-pass analysis.

In our study, each network is focusing on a given ECG pattern typical of atrial arrhythmias in a totally transparent way to the cardiologists. Inputs and outputs of each network are clearly defined based on prior electrocardiography knowledge together with the goal of each network in the global analysis. The design based on prior ECG knowledge is furthermore necessary for the design of the interfaces between DL outputs and the Holter editing tasks. DL output markers at any time can be easily reviewed with associated standard editing tools, such as ECG strips or Page mode (raw ECG in the morphology module), R-R interval graphs (R-R module) and other arrhythmias markers (Figure 1 and Figure 4). Each network or each ECG lead can be enabled or disabled with instantaneous feedback after reanalysis, to identify which modules and which ECG lead contributed most to the output. Increased interaction between clinical Holter experts and AI developers is likely to improve future DL performances.[33]

The modular approach is also expected to be useful for improvement of distinct tasks in the signal analysis of

Holter recordings. Despite the overall excellent performance of our algorithm, discrimination between Afib and Afl was less perfect. The analyses of the diagnosis errors of the AI-based algorithm are highly informative. Currently, in the P wave DL network, we used coherent time averaging of normal cardiac beats. The complexes were aligned on the R peak sample (sampling rate 200 Hz, Figure Appendix A.3) and likely there was a significant variation of the trigger point over time (jittering). Alternately, in a next version of the DL model, the P-wave network will be updated by replacing time averaging with selection of reference, single individual complexes based on optimal signal-to-noise ratio of the atrial electrical activity. Another improvement will be to interface the morphology and the P-wave networks with multi-lead Holter recordings. Finally, in a revised version of the DL model, enabling and disabling each network will be automated, based on quantitative ECG features such as signal-to-noise ratio.

Finally, the modular architecture of our algorithm will allow us to assess the diagnosis value of each of its parts, providing some clues for explainability.

Limitations

The classifier's performance was assessed using the F1 score. This metric aims to evaluate a model's ability to accurately predict positive samples by calculating the harmonic mean of sensitivity and positive predictive value. However, its effectiveness can be limited by its independence from the prevalence of the predicted condition and its inability to provide information about the error distribution.

As said above, the only false-negative ATA detection was related to a short 2-minutes AFib episode on a 24-hour recording. It was the only episode of ATA during the monitoring, with otherwise regular sinus rhythm and absence of atrial extrasystole. Likely the application of HMM in such instance is acting as a "soft filter," and indeed disabling the HMM as allowed by the user interface restores in this recording Afib detection. In this study, Afib episodes of approximately 5 minutes' duration were accurately detected. A low ATA burden is associated with a lower risk of thromboembolic,[8–11] but nonetheless the DL software must be improved, and manual deactivation of the HMM cannot be considered intuitive.

Another limitation of this study lies in the spectrum of ATA episodes in the H-DB, with many permanent ATA cases or many recordings in stable sinus rhythm. In our study, there were no episodes of regular atrial tachycardia. In addition, examples with frequent atrial extrasystoles and frequent atrial runs are also missing, and the performances of our software should be reevaluated with an enriched annotated database.

Ultimately, the duration of ambulatory recordings in clinical practice is increasing up to 1 to 4 weeks, and noise levels can be high in such situations. It is therefore also imperative to collect an annotated ATA database representative of these trends.

Conclusions

Using a large, beat-to-beat annotated, 2-lead Holter databazse, we built a modular DL software based on a modular structure mimicking expert physician ECG interpretation to classify atrial rhythm. The overall performance of the classifier was excellent. Nonetheless, there were clinically significant residual errors, most often related to the classification of the atrial arrhythmia type (Afib vs Afl). The modular structure of the algorithm helps to edit and correct the AI-based first-pass analysis with a dedicated interface and will provide a first step toward explainability.

Funding Sources: This work was supported by the French National Research Agency (ANR-10-IAHU04-LIRYC).

Disclosures: Rémi Dubois, Pierre Maison-Blanche Microport CRM consultant. Quentin Fleury, Sylvain Christophle-Boulard: Microport CRM employees.

Authorship: All authors attest they meet the current ICMJE criteria for authorship.

Patient Consent: Consent was not required due to the use of de-identified data.

Ethics Statement: The Holter database is compliant with European privacy regulatory requirements.

Appendix

Supplementary data

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/10.1016/j.hroo.2024.09.007.

References

1. Di Carlo A, Bellino L, Consoli D, et al. Prevalence of atrial fibrillation in the Italian elderly population and projections from 2020 to 2060 for Italy and the European Union: the FAI Project. EP Europace 2019;21:1468–1475. https://doi.org/10.1093/europace/euz141.
2. Martin SS, Aday AW, Almarzooq ZI, et al. 2024 heart disease and stroke statistics: a report of us and global data from the American Heart Association. Circulation 2024;149:e347–e913. https://doi.org/10.1161/CIR.0000000000001209.
3. 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): the Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC). Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. European Heart Journal 2020; 42:373–498. https://doi.org/10.1093/eurheartj/ehaa612.
4. Lip GYH, Edwards SJ. Stroke prevention with aspirin, warfarin and ximelagatran in patients with nonvalvular atrial fibrillation: a systematic review and meta-analysis. Thromb Res 2006;118-3:321–333. https://doi.org/10.1016/j.thromres.2005.08.007.
5. Page RL, Wilkinson WE, Clair WK, McCarthy EA, Pritchett EL. Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. Circulation 1994;89:224–227. https://doi.org/10.1161/01.CIR.89.1.224.
6. Hindricks G, Piorkowski C, Tanner H, et al. Perception of atrial fibrillation before and after radiofrequency catheter ablation. Circulation 2005;112:307–313. https://doi.org/10.1161/CIRCULATIONAHA.104.518837.
7. Healey JS, Connolly SJ, Gold MR, et al. Subclinical atrial fibrillation and the risk of stroke. N Engl J Med 2012;366:120–129. https://doi.org/10.1056/NEJMoa1105575.

8. Link MS, Giugliano RP, Ruff CT, et al. Stroke and mortality risk in patients with various patterns of atrial fibrillation. Circulation: Arrhythmia and Electrophysiology 2017;10:e004267 https://doi.org/10.1161/CIRCEP.116.004267.
9. Van Gelder IC, Healey JS, Crijns HJ, et al. Duration of device-detected subclinical atrial fibrillation and occurrence of stroke in ASSERT. Eur Heart J 2017; 38:1339–1344. https://doi.org/10.1093/eurheartj/ehx042.
10. Mcintyre W, Benz A, Becher N, et al. Direct oral anticoagulants for stroke prevention in patients with devicedetected atrial fibrillation: a study-level meta-analysis of the noah-afnet 6 and artesia trials. Circulation 2023;149(13):981–988. https://doi.org/10.1161/CIRCULATIONAHA.123.067512.
11. Kirchhof P, Toennis T, Goette A, et al. Anticoagulation with edoxaban in patients with atrial high-rate episodes. N Engl J Med 2023;389:1167–1179. https://doi.org/10.1056/NEJMoa2303062.
12. Steinberg J, Varma N, Cygankiewicz I, et al. 2017 ISHNE-HRS expert consensus statement on ambulatory ECG and external cardiac monitoring/telemetry. Ann Noninvasive Electrocardiol 2017;22(3):e12447 https://doi.org/10.1111/anec.12447.
13. Haddi Z, Ananou B, Alfaras M, et al. Automatic atrial fibrillation arrhythmia detection using univariate and multivariate data. Algorithms 2022;15(7):231. https://doi.org/10.3390/a15070231.
14. Babaeizadeh S, Gregg RE, Helfenbein ED, Lindauer JM, Zhou SH. Improvements in atrial fibrillation detection for real-time monitoring. J Electrocardiol 2009;42:522–526. https://doi.org/10.1016/j.jelectrocard.2009.06.006.
15. Somani S, Russak AJ, Richter F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. EP Europace 2021;23:1179–1191. https://doi.org/10.1093/europace/euaa377.
16. Hannun A, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Med 2019;25 https://doi.org/10.1038/s41591-018-0268-3.
17. Minchole A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nature Med 2019;25:22–23. https://doi.org/10.1038/s41591-018-0306-1.
18. 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. Computing in Cardiology (CinC) 2022;498:1–4. https://doi.org/10.22489/CinC.2022.336.
19. Moody G, Mark R. A new method for detecting atrial fibrillation using r-r intervals. Computers in Cardiology 1983;10:227–230. https://doi.org/10.13026/C2MW2D.
20. Goldberger AL, Amaral LAN, Glass L, et al. Physiobank, physiotoolkit, and physionet. Circulation 2000;101:e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215.
21. Van Rijsbergen CJ. Information Retrieval, 2nd ed. Butterworth-Heinemann; 1979.
22. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1986;327:307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
23. Macfarlane PW, Kennedy J. Automated ecg interpretation—a brief history from high expectations to deepest networks. Hearts 2021;2:433–448. https://doi.org/10.3390/hearts2040034.
24. Willems JL, Arnaud P, van Bemmel JH, Degani R, Macfarlane PW, Zywietz C. Common standards for quantitative electrocardiography: goals and main results. CSE working party. Methods Inf Med 1990;29(04):263–271. https://doi.org/10.1055/s-0038-1634793.
25. Macfarlane P, Devine B, Latif S, McLaughlin S, Shoat D, Watts M. Methodology of ECG interpretation in the Glasgow program. Methods Inf Med 1990; 29:354–361.
26. Van Bemmel J, Kors J, Van Herpen G. Methodology of the modular ecg analysis system means. Methods Inf Med 1990;29:346–353.
27. Hongo RH, Goldschlager N. Status of computerized electrocardiography. Cardiol Clin 2006;24:491–504. https://doi.org/10.1016/j.ccl.2006.03.005.
28. Kligfield P, Badilini F, Denjoy I, et al. Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs. Am Heart J 2018;200:1–10. https://doi.org/10.1016/j.ahj.2018.02.014.
29. Poon K, Okin PM, Kligfield P. Diagnostic performance of a computer-based ecg rhythm algorithm. J Electrocardiol 2005;38:235–238. https://doi.org/10.1016/j.jelectrocard.2005.01.008.
30. De Bie J, Martignani C, Massaro G, Diemberger I. Performance of seven ECG interpretation programs in identifying arrhythmia and cardiovascular syndrome. J Electrocardiol 2020;58:143–149. https://doi.org/10.1016/j.jelectrocard.2019.11.043.
31. Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. J Electrocardiol 2007;40:385–390. https://doi.org/10.1016/j.jelectrocard.2007.03.008.
32. Bogun F, Anh D, Kalahasty G, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004;117:636–642. https://doi.org/10.1016/j.amjmed.2004.06.024.
33. Schlapfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. J Am Coll Cardiol 2017;70:1183–1192. https://doi.org/10.1016/j.jacc.2017.07.723.
34. Taggar J, Coleman T, Lewis S, Heneghan C, Jones M. Accuracy of methods for diagnosing atrial fibrillation using 12-lead ECG: a systematic review and meta-analysis. Int J Cardiol 2015;184C:175–183. https://doi.org/10.1016/j.ijcard.2015.02.014.
35. Acharya UR, Oh SL, Hagiwara Y, et al. A deep convolutional neural network model to classify heartbeats. Comput Biol Med 2017;89:389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022.
36. Li Q, Rajagopalan C, Clifford G. Ventricular fibrillation and tachycardia classification using machine learning method. IEEE Trans Biomed Eng 2013; 61(6):1607–1613. https://doi.org/10.1109/TBME.2013.2275000.
37. Zhu H, Cheng C, Yin H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet Digit Health 2020;2:e348–e357. https://doi.org/10.1016/S2589-7500(20)30107-2.
38. Kalarus Z, Mairesse GH, Sokal A, et al. Searching for atrial fibrillation: looking harder, looking longer, and in increasingly sophisticated ways—an EHRA position paper. EP Europace 2022;25:185–198. https://doi.org/10.1093/europace/euac144.
39. Thakor NV, Zhu Y-S. Applications of adaptive filtering to ECG analysis: noise cancellation and arrhythmia detection. IEEE Trans Biomed Eng 1991; 38:785–794. https://doi.org/10.1109/10.83591.
40. Ivora A, Viscor I, Nejedly P, et al. QRS detection and classification in Holter ECG data in one inference step. Sci Rep 2022;12:12641 https://doi.org/10.1038/s41598-022-16517-4.
41. Couderc J-P. The telemetric and Holter ECG warehouse initiative (thew): a data repository for the design, implementation and validation of ECG-related technologies. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE, 2010. pp. 6252–6255
42. Xue J, Yu L. Applications of machine learning in ambulatory ECG. Hearts 2021; 2:472–494. https://doi.org/10.3390/hearts2040037.
43. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836 2017. https://doi.org/10.48550/arXiv.1707.01836
44. Ben-Moshe N, Tsutsui K, Biton S, Sornmo L, Behar JA. Rawecgnet: deep learning generalization for atrial fibrillation detection from the raw ecg. arXiv preprint arXiv:2401.05411 2023. https://doi.org/10.48550/arXiv.2401.05411
45. Heeger C-H, Kuck K-H, Tilz RR. Very high-power short-duration catheter ablation for treatment of cardiac arrhythmias: insights from the fast and furious study series. J Cardiovasc Electrophysiol 2024;35:547–556. https://doi.org/10.1111/jce.16113.
46. Kors JA, Van Herpen G. Methodology of QT-interval measurement in the modular ECG analysis system (means). Ann Noninvasive Electrocardiol 2009; 14:S48–S53. https://doi.org/10.1111/j.1542-474X.2008.00261.x.