CAN DEEP LEARNING MODELS DIFFERENTIATE ATRIAL FIBRILLATION FROM ATRIAL FLUTTER?

A PREPRINT

Estela Ribeiro
Heart Institute (InCor)
Clinics Hospital University of
Sao Paulo Medical School
Sao Paulo - SP - Brazil
estela.ribeiro@hc.fm.usp.br

Quenaz B. Soares
Heart Institute (InCor)
Clinics Hospital University of
Sao Paulo Medical School
Sao Paulo - SP - Brazil
quenaz.soares@hc.fm.usp.br

Felipe M. Dias
Heart Institute (InCor)
Clinics Hospital University of
Sao Paulo Medical School
Sao Paulo - SP - Brazil
f.dias@hc.fm.usp.br

Jose E. Krieger
Heart Institute (InCor)
Clinics Hospital University of
Sao Paulo Medical School
Sao Paulo - SP - Brazil
j.krieger@hc.fm.usp.br

Marco A. Gutierrez
Heart Institute (InCor)
Clinics Hospital University of
Sao Paulo Medical School
Sao Paulo - SP - Brazil
marco.gutierrez@incor.usp.br

ABSTRACT

This study investigates the use of standard 12-lead ECG records from six largest PhysioNet CinC Challenge 2021 databases and a private database to differentiate Atrial Fibrillation from Atrial Flutter. Image-based and one-dimensional-based Deep Learning models were considered to perform the classification using different 1D and 2D Convolutional Neural Network architectures. For 1D CNNs, LiteVGG-11 was the architecture that best performed, with Acc $77.91 \pm 1.73\%$, AUROC $87.17 \pm 1.29\%$, F1 $76.59 \pm 1.90\%$, Spe $71.69 \pm 4.73\%$ and Se $86.53 \pm 5.33\%$. For 2D CNNs, EfficientNet-B2 was the architecture that best performed, with Acc $75.20 \pm 3.38\%$, AUROC $85.50 \pm 1.14\%$, F1 $71.59 \pm 3.66\%$, Spe $74.76 \pm 13.85\%$ and Se $75.74 \pm 13.85\%$. We showed that discrimination of AFib and AFlut is not trivial, and that one-dimensional signals had better performance than two-dimensional ones. Moreover, we demonstrate that the performance of our models only on the PhysioNet CinC Challenge 2021 databases are considerably lower than only on our private dataset.

Keywords ECG · Atrial Fibrilattion · Atrial Flutter · Deep Learning · Convolutional Neural Network.

1 Introduction

Atrial Flutter (AFlut) and Atrial Fibrillation (AFib) are two types of irregular heart rhythms caused by abnormal electrical activity in the upper chambers of the heart, the Atria Cosío [2017], Ko Ko et al. [2022], and are particularly dangerous for the elderly population Shah et al. [2018]. AFib is caused by chaotic electrical activity in the atria, leading to irregular and weak atrial contractions at a rate of 350-500 beats per minute, decreasing the effectiveness of the atria in pumping blood into the ventricles, thus compromising the heart function and increasing the risk of stroke and other heart-related complications Thaler [2019], Brundel et al. [2022]. AFlut is a common supraventricular arrhythmia that is often misdiagnosed as atrial fibrillation, with the latter being more common Ko Ko et al. [2022]. It is caused by a single electrical circuit that drives the atria to contract at 250-350 beats per minute Thaler [2019]. Although the atria beat regularly, their accelerated rhythm causes a discrepancy with the rhythm of the ventricles, causing problems in the heart function Thaler [2019]. Thus, early diagnosis and treatment are crucial in managing AFib and AFlut and minimizing the risk of serious complications such as stroke Shah et al. [2018].

These two abnormal heart rhythms are frequently unrecognized and untreated since they typically are asymptomatic or can present symptoms that may include chest pain, lightheadedness, shortness of breath, fainting, and palpitations Shah et al. [2018], Brundel et al. [2022], related to the rapid ventricular rate that results in inadequate diastolic ventricular filling Ko Ko et al. [2022]. The development of automatic systems can aid in detecting these conditions quickly and accurately, ultimately improving the performance of medical staff and decreasing patient waiting times. This is especially beneficial for low-income hospitals with limited access to experienced cardiologists and can help alleviate the burden on their medical infrastructure.

The electrocardiogram (ECG), an essential tool for diagnosing cardiac problems with millions of exams performed every year worldwide, is widely used for AFib and AFlut detection. The ECG measures the heart's electrical activity through electrodes attached to the patient skin and is the gold standard for noninvasive diagnosis of several types of heart disorders Thaler [2019]. Clinical diagnosis for AFib and AFlut generally relies on non-invasive 12-lead ECGs, and the patterns of electrical activity seen on the ECG signal allow the distinction between these two conditions Cosío [2017], Brundel et al. [2022].

On the ECG, the electrical activity of AFib is characterized with the absence of P waves, irregular RR intervals and fibrillatory waves. On the other hand, a common pattern of AFlut is the presence of sawtooth flutter waves. However, despite their distinctive sources and patterns, AFlut is commonly misdiagnosed as AFib Ko Ko et al. [2022] due to their similar symptoms Shah et al. [2018], Brundel et al. [2022] and the prevalence of AFib. Some works suggest that AFlut can be misinterpreted as AFib, affirming that, in high rates, the ventricular activity is not regular and AFlut can appear with characteristics of AFib on surface ECGs Ghafoori et al. [2018], making specialists to misdiagnose AFlut as AFib. This misdiagnosis can lead to inappropriate treatment, as the two conditions have distinct therapeutic approaches.

In the past decades, automated AFib detection has gained interest from the research community and Deep Learning has proven to be an effective technique for automatic ECG analysis Hicks et al. [2021], Wegner et al. [2022]. In this sense, performance related to the detection of AFib against non-AFib class normally present high performance Ivanovic et al. [2019], Dias et al. [2021], Tutuko et al. [2021]. Some works even propose to join AFib and AFlut in one class Jekova et al. [2022] for this classification. However, the distinction between AFib and AFlut has received relatively little attention, and the few works that proposes to differentiate AFib from AFlut have results far from satisfactory Ivanovic et al. [2019].

Most studies on the differentiation of AFib and AFlut normally use the MIT-BIH Atrial Fibrillation Moody and Mark [1983], Goldberger et al. [2000] and MIT-BIH Arrhythmia Moody and Mark [2000], Goldberger et al. [2000] datasets with long records of only two-lead one-dimensional ECGs for analysis. Their approach is to classify ECG signals into three labels, i.e., AFib, AFlut and Normal Sinus Rhythm, from a limited number of subjects, thus needing to partition these data into small segments. However, the use of the same subject in the training and test sets can produce better results due to data leakage given the high interdependence of intra-subject heartbeats. Thus, caution is advised since it has been shown that using intra-subject paradigms may lead to overoptimistic biased classifications de Chazal et al. [2004], da S. Luz et al. [2016], Butkuvienė et al. [2021].

Moreover, most of the works aiming to classify ECG exams, regardless of the labels, are based on one-dimensional signals. However, a physician usually gives a diagnosis by visually inspecting an ECG chart and interpreting the 12-lead ECG exam. Thus, we hypothesize that bi-dimensional (images) DL models aiming to differentiate Atrial Fibrillation from Atrial Flutter would outperform one-dimensional ones. Furthermore, given that misdiagnoses of these two conditions are common, we also hypothesize that DL models designed to differentiate them would present poor performance.

Finally, in this study, we aim to investigate the usage of 12-lead ECG records to differentiate AFib from AFlut using data from the six largest PhysioNet Cinc Challenge 2021 databases and a private database acquired from ambulatory patients of a tertiary referral hospital. We considered two distinct input data, i.e., images (2D) and one-dimensional (1D) signals, to assess which one have a better performance. This approach is different from that of other state-of-the art deep learning-based ECG classification studies. To perform our experiments, we compared different Convolutional Neural Network (CNN) architectures to evaluate their performances, being them, MobileNet Howard et al. [2017], ResNet-50 He et al. [2016], VGG16 Simonyan and Zisserman [2014], DenseNet-121 Huang et al. [2017] and EfficientNet-B2 Tan and Le [2019], for image-based input data, and LiteVGG-11 Soares et al. [2022], LiteResNet-18 Soares et al. [2022], MobileNet Howard et al. [2017], ResNet-50 He et al. [2016], VGG-16 Simonyan and Zisserman [2014], DenseNet-121 Huang et al. [2017] and EfficientNet-B2 Tan and Le [2019], for one-dimensional-based input data. To the best of our knowledge, we present here the first results on the assessment of Atrial Fibrillation and Atrial Flutter discrimination based on end-to-end Deep Convolutional Neural Networks.

2 Methods

In this section, we describe the dataset, the preprocessing steps and the deep neural networks architecture used for classifying ECG signals. The general structure of the proposed method is shown in Figure 1. All of our experiments were performed using a Foxconn High-Performance Computer (HPC) M100-NHI with an 8 GPU cluster of 32 GB NVIDIA Tesla V100 cards. The methodology was implemented using the Python framework (version 3.6.8) and Keras/TensorFlow (version 2.3.0).

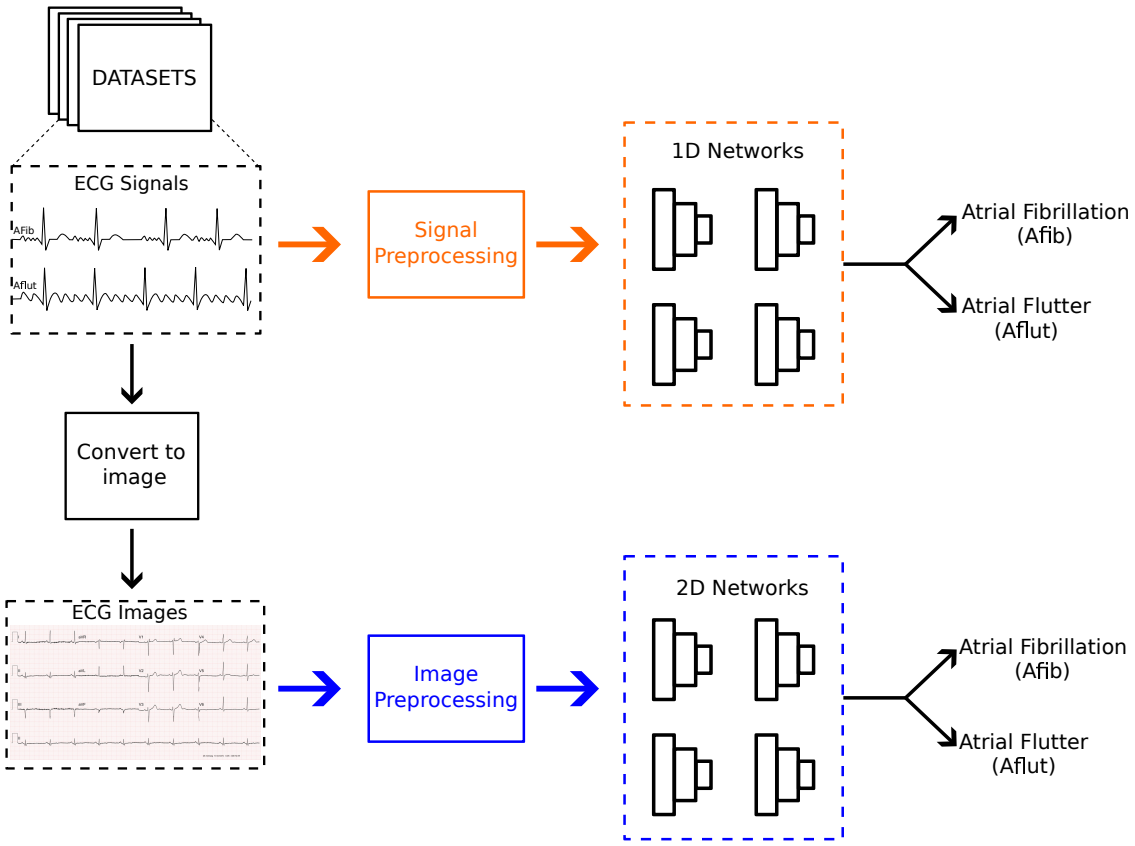

Figure 1: General structure of the proposed method for binary AFib/AFlut classification with different 1D and 2D CNN architectures.

2.1 Datasets

We used the PhysioNet CinC Challenge 2021 databases Reyna et al. [2021, 2022]. This dataset is currently the largest freely available repository of standard 12-lead ECG records, containing 30 clinical diagnoses of cardiac abnormalities. It is composed of the following datasets: public (CPSC) and unused (CPSC extra) China physiological signal challenge, St. Petersburg Institute of Cardiological Technics (INCART), Physikalisch-Technische Bundesanstalt (PTB), PTB-XL, Georgia 12-lead ECG Challenge, Chapman Shaoxing and Ningbo. 12-lead ECG exams usually present only 10 seconds of recorded signals. Moreover, since cropping records might remove the arrhythmic morphologies stated in the labels of long-term ECG exams, we chose datasets with recordings that were approximately 10 seconds in length. Therefore, the PTB and INCART datasets were not included in this study since their recordings were longer than 10 s.

We also used a private 12-lead ECGs dataset (named InCor-DB), with data collected from 2017 to 2020. The data were available from the Picture Archiving and Communication System (PACS) of a tertiary referral hospital in Brazil specialized in cardiology (Heart Institute Hospital), acquired from MORTARA TM ELI 250c machines, with 52 different clinical diagnoses of cardiac abnormalities. This private dataset complies with all relevant ethical regulations and was approved by the Research Ethics Committee.

The scope of this study was to evaluate patient data diagnosed with AFib and AFlut arrhythmia. All other records with different diagnostic annotations were not considered in this work. Furthermore, we estimate class weights to deal with

the unbalanced datasets. Table 1 summarizes the number of ECG records available in the six largest PhysioNet Cinc Challenge 2021 databases and the InCor-DB dataset.

Table 1: Number of selected 12-lead ECG records from six CinC Challenge 2021 datasets and InCor-DB private database.

Datasets	ECG records (total)	Atrial Fibrillation	Atrial Flutter
Chapman-Shaoxing 12-lead ECG	2,225	1,780	445
CPSC 2018 Training Set (CPSC 2018)	1,221	1,221	0
China 12-Lead ECG Challenge (CPSC2018-Extra)	207	153	54
Georgia 12-Lead ECG Challenge	756	570	186
Ningbo First Hospital 12-lead ECG	7,615	0	7,615
PTB-XL Electrocardiography	1,587	1,514	73
Heart Institute Hospital (InCor-DB) 12-Lead ECG	9,528	8,219	1,309
Total	*23,139*	*13,457*	*9,682*

2.2 Data preprocessing

The standard 12-lead ECG raw signals were resampled to 500 Hz when necessary, and signals length were fixed as 10 seconds by retaining only the 10 first seconds of longer signals or zero-padding shorter signals until they reach 10 seconds.

The preprocessing of the data is divided in two parts, one related to one-dimensional signals, and the other related to bi-dimensional signals, i.e. images.

The one-dimensional signals were filtered with a 2-40 Hz bandpass butterworth filter. We used the original sampling rate of 500 Hz.

To create the two-dimensional signals dataset, we mapped the 1D raw signals from the original datasets to their image counterparts by employing the MORTARA ECG image template without any signal as background and drawing the signals onto this image (original dimension 1671x3122x3). We set the MORTARA template as background to present an ECG exam to our proposed models in a manner similar to what physicians would typically encounter, since one of our hypotheses were that image-based DL models would outperform one-dimensional ones. These signals were filtered with a 60 Hz notch filter and a 0.5-100 Hz bandpass butterworth filter before converting to image. We transformed the images into grayscale and resized them to 30% of their original size to decrease the computational complexity of the models, resulting in input size of 501x936 grayscale image.

2.3 Deep learning models

2.3.1 One-dimensional Classification

We used seven one-dimensional Convolutional Neural Networks to evaluate the performance of one-dimensional-based Atrial Fibrillation and Atrial Flutter classification: (i) LiteVGG-11; (ii) LiteResNet-18; (iii) MobileNet; (iv) ResNet-50; (v) VGG-16; (vi) DenseNet-121; and (vii) EfficientNet-B2. For the traditional CNNs we changed the 2D convolutions to 1D convolutions. For the Lite models (LiteVGG-11 and LiteResNet-18), we applied a lightweight CNN proposed by Quenaz et al. (2022) Soares et al. [2022] that present comparable performance with their original counterparts, with lower computing resources. Their approach uses depth-wise separable 1D convolution layers (DWConv), a reduced number of filters, a global average pooling as flattening, and a reduced number of units in the dense layers. We maintained the fully connected layers of the original models, changing only the last one for one output with a sigmoid activation. Each model was trained over up to 120 epochs using a 64 batch size. To avoid overfitting, an early stopping callback with patience of seven epochs was also added, i.e., if the model does not improve in the validation dataset for seven consecutive epochs, training is halted.

2.3.2 Image Classification

To evaluate the performance of image-based Atrial Fibrillation and Atrial Flutter classification, we used five traditional and widely used 2-dimensional Convolutional Neural Networks: (i) MobileNet; (ii) ResNet-50; (iii) VGG16; (iv) Densenet-121; and (v) EfficientNetB2. The fully connected layers were a customized 3-layer perceptron with dropout regularization of 30%, ReLU activation function in intermediate layers, and a sigmoid function in the last layer. Each model was trained over 30 epochs using an 8 batch size. As in the one-dimensional classification, an early stopping callback with patience of seven epochs was added to avoid overfitting.

2.4 Performance Evaluation

We performed a 10-fold cross-validation for all experiments, and all the results are presented in the following format: mean (std). To prevent data leakage, we ensured that exams from the same patient did not exist in different partitions of the cross-validation protocol. In order to evaluate the employed models, we assessed five different metrics, including Sensitivity (Se), Specificity (Spe), F1-score (F1), Area Under Receiver Operating Characteristic curve (AUROC) and Accuracy (Acc). Given the considerable class imbalance in the dataset, we defined our best model in terms of F1-score, which provides a more conservative evaluation of the model's performance.

2.5 Experimental Setup

We conducted experiments to evaluate the performance and generalizability of our models using two different approaches: one-dimensional ECG signals and image-based ECG exams. To assess the effectiveness of our models, we employed various evaluation setups. Our aim was to evaluate the overall performance of the models and their ability to generalize to external datasets.

The experiments were performed for both the one-dimensional and image-based ECGs using the following setups:

1. Train/Validate/Test (10-fold): PhysioNet CinC 2021 and InCor-DB;
   External validation: None;
2. Train/Validate/Test (10-fold): PhysioNet CinC 2021;
   External validation: InCor-DB;
3. Train/Validate/Test (10-fold): InCor-DB;
   External validation: PhysioNet CinC 2021;
4. Train/Validate/Test (10-fold): InCor-DB;
   External validation: PhysioNet CinC 2021 individual datasets (Chapman, CPSC, CPSC (extra), GA, Ningbo, PTB-XL).

These experiments allowed us to assess the performance of our models in different scenarios and evaluate their generalizability to diverse datasets.

3 Results

3.1 One-dimensional-based classification

Table 2 provides the performance results of the seven proposed architectures for one-dimensional-based classification of AFib and AFlut, with AFlut as the positive class. We employed PhysioNet CinC 2021 dataset and InCor-DB for training, validation, and testing using 10-fold cross-validation.

Table 2: Performance results of seven proposed architectures for one-dimensional input data.

Architectures	Acc	AUROC	F1	Spe	Se
LiteVGG-11	**77.91 (1.73)**	**87.17 (1.29)**	**76.59 (1.90)**	**71.69 (4.73)**	**86.53 (5.33)**
LiteResNet-18	76.88 (1.61)	86.64 (0.77)	73.95 (2.76)	75.30 (6.66)	79.06 (8.33)
MobileNet	77.38 (6.67)	89.52 (0.78)	76.55 (4.74)	70.10 (16.97)	87.53 (12.25)
ResNet-50	61.51 (11.42)	76.81 (9.93)	64.35 (5.78)	46.73 (28.84)	82.07 (17.11)
VGG-16	66.49 (2.04)	71.07 (2.76)	59.73 (2.77)	71.48 (3.58)	59.51 (4.39)
DenseNet-121	75.90 (0.84)	84.46 (0.92)	71.78 (1.31)	77.69 (3.64)	73.39 (4.17)
EfficientNet-B2	52.49 (6.92)	54.27 (2.78)	28.48 (28.55)	59.28 (44.05)	43.30 (45.59)

In addition to the previous results, we decided to carry out this classification problem with the following strategy: (i) Train/Validate/Test the models only with the PhysioNet CinC 2021 dataset and then perform an external validation with the InCor-DB dataset (Table 3); and the opposite by (ii) Training, validating and testing the models only with the InCor-DB dataset and then performing an external validation with the PhysioNet CinC 2021 dataset (Table 4).

Table 5 shows the results of our best model for one-dimensional-based classification, i.e., the LiteVGG-11 model, trained only with the InCor-DB dataset and perform an external validation with each individual datasets of CinC 2021 databases.

Table 3: Performance results of seven proposed architectures for one-dimensional input data. Train = CinC 2021 dataset. External Validation = InCor-DB dataset.

Test on CinC 2021 dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
LiteVGG-11	72.24 (2.13)	77.43 (1.38)	78.21 (2.95)	57.02 (7.39)	81.73 (7.25)
LiteResNet-18	70.15 (5.05)	75.85 (2.19)	75.98 (8.50)	53.49 (14.11)	80.64 (15.41)
MobileNet	**71.01 (4.54)**	**76.49 (1.69)**	**78.44 (3.92)**	**45.58 (22.44)**	**86.90 (12.13)**
ResNet-50	64.63 (9.51)	71.08 (3.71)	67.62 (22.32)	49.25 (26.80)	74.24 (29.69)
VGG-16	53.45 (9.89)	55.94 (4.73)	48.18 (31.63)	57.28 (30.64)	50.96 (34.33)
DenseNet-121	67.12 (1.38)	72.52 (1.29)	73.01 (2.20)	58.30 (5.75)	72.63 (5.36)
EfficientNet-B2	48.34 (10.84)	50.62 (2.02)	34.21 (34.65)	59.46 (45.51)	41.25 (45.96)
External Validation on InCor-DB dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
LiteVGG-11	48.93 (10.59)	83.37 (3.67)	32.98 (4.20)	42.56 (12.79)	88.95 (4.66)
LiteResNet-18	45.25 (14.84)	86.03 (3.11)	32.92 (6.19)	37.80 (18.58)	92.02 (9.55)
MobileNet	38.04 (19.33)	74.13 (4.20)	29.09 (4.39)	30.16 (24.27)	87.50 (12.30)
ResNet-50	42.42 (25.60)	66.73 (8.37)	26.20 (9.99)	36.90 (33.70)	77.08 (28.61)
VGG-16	55.35 (22.09)	54.69 (3.11)	18.07 (11.87)	56.04 (31.05)	51.05 (34.61)
DenseNet-121	58.30 (6.89)	77.67 (2.59)	35.45 (3.41)	54.52 (8.28)	82.02 (3.01)
EfficientNet-B2	56.94 (33.60)	50.53 (1.29)	11.95 (10.38)	59.54 (46.38)	40.65 (46.62)

Table 4: Performance results of seven proposed architectures for one-dimensional input data. Train = InCor-DB dataset. External Validation = CinC 2021 dataset.

Test on InCor-DB dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
LiteVGG-11	**95.50 (0.99)**	**98.33 (0.65)**	**84.77 (3.28)**	**96.07 (1.08)**	**91.70 (2.70)**
LiteResNet-18	95.52 (1.10)	98.28 (0.64)	84.66 (4.19)	96.23 (1.58)	91.26 (3.61)
MobileNet	87.38 (11.42)	97.62 (1.03)	71.31 (14.94)	86.40 (14.02)	93.53 (4.79)
ResNet-50	82.93 (22.41)	96.86 (2.36)	69.01 (19.15)	81.84 (27.50)	88.87 (13.93)
VGG-16	83.88 (24.46)	82.12 (20.69)	56.70 (33.21)	87.72 (29.32)	64.64 (33.53)
DenseNet-121	93.87 (3.76)	97.07 (1.64)	80.80 (7.54)	94.76 (4.54)	88.29 (5.58)
EfficientNet-B2	69.69 (33.03)	73.58 (23.25)	48.56 (33.66)	69.47 (39.74)	72.08 (34.73)
External Validation on CinC 2021 dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
LiteVGG-11	47.56 (0.56)	55.90 (1.01)	31.09 (2.12)	92.78 (1.34)	19.27 (1.70)
LiteResNet-18	46.89 (0.85)	55.90 (0.65)	28.80 (3.44)	93.78 (2.15)	17.56 (2.70)
MobileNet	48.66 (2.88)	53.58 (1.66)	38.37 (12.07)	81.07 (16.31)	28.38 (14.83)
ResNet-50	49.18 (4.77)	55.98 (1.18)	37.95 (17.22)	78.57 (27.71)	30.80 (24.83)
VGG-16	45.40 (6.03)	54.11 (2.79)	24.21 (20.07)	86.28 (28.93)	19.82 (27.44)
DenseNet-121	47.68 (1.21)	57.04 (0.53)	31.57 (5.40)	92.07 (4.47)	19.91 (4.63)
EfficientNet-B2	49.16 (7.30)	52.68 (3.04)	38.59 (24.52)	67.61 (39.09)	37.61 (35.87)

Table 5: Performance results of LiteVGG-11 one-dimensional-based model trained on InCor-DB dataset and with external validation on each individual CinC 2021 dataset.

	Chapman	**CPSC**	**CPSC (extra)**	**GA**	**Ningbo**	**PTB-XL**
Acc	89.15 (0.34)	92.40 (1.54)	78.07 (1.21)	81.10 (1.35)	14.68 (1.55)	92.53 (1.54)
AUROC	89.82 (0.82)	-	78.95 (1.61)	85.92 (0.93)	-	97.42 (0.57)
F1	68.76 (1.79)	-	52.85 (3.89)	65.20 (1.11)	25.58 (2.35)	54.29 (4.83)
Spe	96.46 (0.70)	92.40 (1.54)	88.89 (1.55)	84.09 (2.91)	-	92.43 (1.62)
Se	59.89 (3.57)	-	47.41 (5.69)	71.94 (4.16)	14.68 (1.55)	94.66 (1.56)

3.2 Image-based classification

Table 6 provides the performance results of the five proposed architectures for image-based classification of AFib and AFlut, with AFlut as positive class. We employed PhysioNet CinC 2021 dataset and InCor-DB for training, validation, and testing using 10-fold cross-validation.

Table 6: Performance results of five proposed architectures for image input data.

Architectures	Acc	AUROC	F1	Spe	Se
MobileNet	54.94 (13.94)	76.02 (10.99)	64.16 (5.56)	27.64 (31.50)	92.96 (11.62)
ResNet-50	49.11 (9.51)	72.35 (10.15)	51.51 (19.18)	26.32 (40.88)	80.11 (37.90)
VGG16	46.45 (7.76)	50.0 (0.0)	41.16 (28.40)	30.0 (48.30)	70.0 (48.30)
DenseNet	48.0 (9.89)	64.20 (12.53)	45.65 (24.55)	29.80 (47.84)	72.76 (44.76)
EfficientNet-B2	**75.20 (3.38)**	**85.50 (1.14)**	**71.59 (3.66)**	**74.76 (13.85)**	**75.74 (13.85)**

Similar to the one-dimensional-based classification, in addition to the previous results, we used the following strategy: (i) Train/Validate/Test the models only with the PhysioNet CinC 2021 dataset and then perform an external validation with the InCor-DB dataset (Table 7); and the opposite by (ii) Training, validating and testing the models only with the InCor-DB dataset and then performing an external validation with the PhysioNet CinC 2021 dataset (Table 8).

Table 7: Performance results of five proposed architectures for image input data. Train = CinC 2021 dataset. External Validation = InCor-DB dataset.

Test on CinC 2021 dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
MobileNet	56.11 (11.53)	67.49 (12.28)	52.94 (31.86)	50.82 (45.52)	59.66 (42.85)
ResNet-50	52.13 (11.56)	68.60 (9.97)	43.22 (35.07)	55.27 (48.75)	50.45 (47.51)
VGG-16	52.25 (11.91)	50.0 (0.0)	45.68 (39.31)	40.0 (51.63)	60.0 (51.63)
DenseNet-121	51.0 (13.32)	65.02 (15.25)	37.50 (39.27)	60.81 (46.79)	44.91 (48.58)
EfficientNet-B2	**63.42 (3.73)**	**65.08 (10.09)**	**74.88 (6.0)**	**18.31 (29.54)**	**91.85 (18.04)**
External Validation on InCor-DB dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
MobileNet	45.53 (32.33)	65.98 (20.96)	19.31 (13.19)	42.65 (43.18)	63.64 (46.31)
ResNet-50	52.15 (36.67)	65.88 (14.41)	21.29 (13.43)	51.72 (49.46)	54.90 (44.87)
VGG-16	42.74 (37.45)	50.0 (0.0)	14.49 (12.47)	40.0 (51.63)	60.0 (51.63)
DenseNet-121	52.58 (36.16)	61.18 (16.95)	11.99 (12.43)	53.61 (49.75)	46.12 (49.20)
EfficientNet-B2	20.84 (14.84)	68.79 (14.27)	25.53 (2.91)	8.77 (18.41)	96.63 (7.70)

Table 8: Performance results of five proposed architectures for image input data. Train = InCor-DB dataset. External Validation = CinC 2021 dataset.

Test on InCor-DB dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
MobileNet	91.19 (6.74)	97.71 (1.84)	74.92 (11.39)	91.57 (9.11)	87.54 (17.27)
ResNet-50	72.37 (32.93)	92.51 (15.03)	60.81 (24.52)	69.39 (39.36)	91.99 (14.96)
VGG-16	71.19 (31.06)	50.0 (0.0)	4.39 (9.27)	80.0 (42.16)	20.0 (42.16)
DenseNet-121	40.81 (35.60)	85.79 (19.38)	34.17 (25.11)	33.42 (87.86)	87.86 (30.17)
EfficientNet-B2	**94.08 (2.59)**	**97.03 (2.41)**	**80.11 (8.57)**	**95.06 (3.46)**	**87.76 (11.09)**
External Validation on CinC 2021 dataset					
Architectures	**Acc**	**AUROC**	**F1**	**Spe**	**Se**
MobileNet	82.66 (8.96)	84.96 (3.19)	46.97 (12.03)	85.71 (13.11)	61.58 (25.22)
ResNet-50	51.71 (5.52)	52.51 (5.87)	48.30 (19.72)	60.56 (39.63)	46.18 (33.05)
VGG-16	43.09 (9.71)	50.0 (0.0)	15.23 (32.11)	80.0 (42.16)	20.0 (42.16)
DenseNet-121	55.70 (8.45)	53.58 (13.65)	59.41 (27.17)	29.04 (42.42)	72.37 (40.12)
EfficientNet-B2	47.26 (2.06)	54.64 (2.94)	32.28 (8.62)	89.13 (7.40)	21.06 (7.10)

Table 9 shows the results of our best model for image-based classification, i.e., the EfficientNet-B2 model, trained only with the InCor-DB dataset and perform an external validation with each individual datasets of CinC 2021 databases.

4 Discussion

In the present study, we used different approaches to assess if Deep Convolutional Neural Network models are capable to differentiate Atrial Fibrillation from Atrial Flutter. We are the first to present an assessment of AFib and AFlut

Table 9: Performance results of EfficientNet-B2 image-based model trained on InCor-DB dataset and with external validation on each individual CinC 2021 dataset.

	Chapman	**CPSC**	**CPSC (extra)**	**GA**	**Ningbo**	**PTB-XL**
Acc	85.47 (2.48)	87.93 (8.05)	77.2 (3.35)	80.61 (3.44)	16.95 (6.51)	87.85 (8.93)
AUROC	83.44 (3.86)	-	77.32 (3.23)	83.69 (2.70)	-	96.45 (0.74)
F1	60.29 (8.24)	-	54.46 (7.22)	61.90 (7.94)	28.51 (9.70)	46.16 (13.49)
Spe	93.63 (5.10)	87.93 (8.05)	86.50 (8.28)	85.32 (8.12)	-	87.68 (9.57)
Se	56.85 (14.19)	-	52.03 (13.73)	66.16 (15.92)	16.95 (6.51)	91.50 (7.52)

specific discrimination based on end-to-end CNNs, as far as we know, demonstrating that it is possible to reasonably differentiate these two diagnoses with CNN models.

Our main findings can be summarized as follows, and will be further discussed bellow: (1) When using all databases (CinC 2021 and InCor-DB), only models based on one-dimensional data were able to discriminate AFib from AFlut, with reasonable performance; (2) Models based on two-dimensional data had poor performance, except the EfficientNet-B2; (3) Regarding the available datasets, models based only on the PhysioNet CinC Challenge 2021 databases failed to discriminate the study classes, resulting in metrics close to chance level, whereas models based only on InCor-DB private dataset were successfully able to separate the classes; (4) We argue that before deciding to combine AFib and AFlut exams into one class for further analysis, it is important to assess the classes' separability on the study dataset. As we see in this work, for the InCor-DB dataset, these two classes are clearly separable, however, this is not the case for the PhysioNet CinC Challenge 2021 databases; Furthermore, (6) Regarding the datasets on CinC 2021, caution must be taken when using the Ningbo dataset, since our results demonstrate that most of the exams labeled as AFlut are predicted as AFib by our models.

4.1 Models based on both CinC2021 and InCor-DB dataset

For one-dimensional models, the EfficientNet-B2 architecture failed to learn the problem, with performance near chance level. On the other side, the LiteVGG-11 architecture achieved the best results in terms of F1-score. All image-based models performed poorly, except for the EfficientNet-B2 architecture, that presented results similar to the one-dimensional models.

Other works proposed to discriminate AFib and AFlut using datasets from MIT-BIH Wang and Wu [2022]. However, as pointed by Jekova et al. (2022) Jekova et al. [2022] these datasets contain long-term Holter recordings with a limited number of subjects. These recordings fail to represent the diversity of arrhythmias, compared to the CinC 2021 databases and the InCor-DB dataset that have many subjects and a higher number of exams.

4.2 Models Based on CinC2021 dataset, with external validation on InCor-DB dataset

Our first hypothesis was that the discrimination of AFib and AFlut for Deep Learning models were not trivial, given that both classes are frequently combined into one in some research Jekova et al. [2022] since these two diagnoses share similar characteristics. Results on the PhysioNet CinC Challenge 2021 databases indicated that our image-based proposed networks were not able to separate AFib from AFlut, given that most metrics used were close to chance level. Table 7 support this evaluation on CinC test set, and even the EfficientNet-B2 architecture (with better performance compared to the others) shows that the model is probably assigning the exams to a single class, as can be seen by the specificity and sensibility metrics. In contrast, our one-dimensional-based models (Table 3 performed reasonably with LiteVGG-11, LiteResNet-18 and MobileNet architectures.

When looking at the data distribution over the individual datasets from the PhysioNet CinC Challenge 2021 database, given the unbalanced data, it is possible that the models are learning to separate datasets instead of learning the problem of AFib and AFlut discrimination. As described in Table 1, the CinC database contains 5,238 AFib exams and 8,373 AFlut exams. Considering that most of the AFlut samples come from the Ningbo First Hospital 12-lead ECG dataset, one could reasonably think that the models are discriminating the exams by Ningbo dataset vs all.

4.3 Models based on InCor-DB dataset, with external validation on CinC2021 dataset

When using the InCor-DB private dataset (Tables 4 and 8), even with the unbalanced classes, our models were able to achieve great performance, with Acc and AUROC above 90%. In this case, when performing external validation on CinC dataset, for most of the models, the networks were unable to generalize. Considering the aforementioned findings, it appears to be appropriate to combine AFib and AFlut labeled exams in subsequent experiments for ECG classification

when using the PhysioNet CinC Challenge 2021 databases, given the models performance being close to chance level. This is not the case when using the InCor-DB dataset, where the two classes are clearly separable.

If we take our best model in terms of F1-score, given the significant class imbalance in the dataset, the F1-score offers a more cautious assessment of the model's performance. Thus, the MobileNet was selected for one-dimensional models and EfficientNet-B2 for image-based models. We trained the models only with the InCor-DB dataset and perform an external validation with each individual datasets of CinC 2021 databases, we can see that for most of the individual databases the trained models performed reasonably, except for the CPSC and the Ningbo datasets (Table 9). For CPSC dataset, most of the exams are classified as AFib, which is expected since all the exams are labeled as AFib. However, on Ningbo dataset, considering AFlut as true label, we see that for our model most of the exams labeled as AFlut are predicted as AFib. It seems unreasonable to think that there are 7,615 samples of AFlut and none of AFib in this dataset. Therefore, we argue that these results may be related to the criteria used by cardiologists from different nations to diagnose Atrial Fibrillation and Atrial Flutter. Since the diagnoses made on the InCor-DB dataset came from a joint of cardiologists from the same hospital, they should adopt the same criteria for these diagnoses. In the worst case, one could say that these exams are mislabeled or, even worst, misdiagnosed.

Ethics statement

This research was approved by the Internal Review Board (IRB) of the Heart Institute (InCor) – Clinics Hospital University of Sao Paulo Medical School (HCFMUSP), registration CAAE 45070821.3.0000.0068, as part of the Machine Learning in Cardiovascular Medicine Project. Since this is a secondary analysis of anonymized data stored in the InCor-DB, informed consent was not required by the Research Ethics Committee for the present study.

Authors contributions statement

E.R. Conceptualization, Methodology, Implementation and Writting. Q.B.S Conceptualization, Methodology, Implementation and Writting. F.M.D. Conceptualization, Methodology, Implementation and Writting. J.E.K. and M.A.G. Supervision and Review. All authors analyzed the results and revised critically the manuscript. All authors read and approved the submitted manuscript.

Competing interests

The authors declare no competing interests.

Data availability

Publicly available datasets were analyzed in this study. This data can be found in the PhysioNet databank: https://physionet.org/content/challenge-2021/1.0.2/. Private InCor-DB dataset will be made available upon reasonable request to the authors.

Code availability

Codes will be made available on request.

Acknowledgements

This study was financially supported in part by São Paulo Research Foundation (FAPESP) – grant n° 2021/12935-0, the Foxconn Brazil, and the Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

References

Fransisco G. Cosío. Atrial flutter, typical and atypical: A review. *Arrhythmia and Electrophysiology Review*, 6(2): 55–62, 2017. doi:10.15420/aer.2017:5:2.

Nway Lea Ko Ko, Anila Sriramoju, Banveet Kaurb Khetarpal, and Komandoora Srivathsan. Atypical atrial flutter: review of mechanisms, advances in mapping and ablation outcomes. *Current Opinion in Cardiology*, 37(1):36–45, 2022. doi:10.1097/HCO.0000000000000924.

Syed Raza Shah, Sue-Wei Luu, Matthew Calestino, John David, and Christopher Bray. Management of atrial fibrillation-flutter: uptodate guideline paper on the current evidence. *Journal of Community Hospital Internal Medicine Perspectives*, 8(5):269–275, 2018. doi:10.1080/20009666.2018.1514932.

Malcolm S. Thaler. *The Only EKG Book You'll Ever Need*. Philadelphia :Wolters Kluwer Health/Lippincott Williams and Wilkins, 2019.

Bianca J. J. M. Brundel, Xun Ai, Mellanie True Hills, Myrthe F. Kuipers, Gregory Y. H. Lip, and Natasja M. S. de Groot. Atrial fibrillation. *Nature Reviews Disease Primers*, 21(8), 2022. doi:10.1038/s41572-022-00347-9.

Elyar Ghafoori, Nathan Angel, Derek J. Dosdall, Rob S. MacLeod, and Ravi Ranjan. Atrial fibrillation observed on surface ecg can be atrial flutter or atrial tachycardia. *Journal of Electrocardiology*, 51(6, Supplement):S67–S71, 2018. ISSN 0022-0736. doi:10.1016/j.jelectrocard.2018.07.010.

Steven A. Hicks, Jonas L. Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary M. Maleckar, Michael A. Riegler, and Jørgen K. Kanters. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific Reports*, 11(1):10949, 2021. doi:10.1038/s41598-021-90285-5.

Felix K. Wegner, Lucas Plagwitz, Christian Doldi, Kevin Willy, Julian Wolfes, Sarah Sandmann, Julian Varghese, and Lars Eckardt. Machine learning in the detection and management of atrial fibrillation. *Clinical Research in Cardiology*, 111(9):1010–1017, 2022. doi:10.1007/s00392-022-02012-3.

Marija D. Ivanovic, Vladimir Atanasoski, Alexei Shvilkin, Ljupco Hadzievski, and Aleksandra Maluckov. Deep learning approach for highly specific atrial fibrillation and flutter detection based on rr intervals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1780–1783, 2019. doi:10.1109/EMBC.2019.8856806.

Felipe M Dias, Nelson Samesima, Adele Ribeiro, Ramon A Moreno, Carlos A Pastore, Jose E Krieger, and Marco A Gutierrez. 2d image-based atrial fibrillation classification. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021. doi:10.23919/CinC53138.2021.9662735.

Bambang Tutuko, Siti Nurmaini, Alexander Edo Tondas, Muhammad Naufal Rachmatullah, Annisa Darmawahyuni, Ria Esafri, Firdaus Firdaus, and Ade Iriani Sapitri. Afibnet: an inplementation of atrial fibrillation detection with convolutional neural network. *BMC Med Inform Decis Mak*, 21:216, 2021. doi:10.1186/s12911-021-01571-1.

Irena Jekova, Ivaylo Christov, and Vessela Krasteva. Atrioventricular synchronization for detection of atrial fibrillation and flutter in one to twelve ecg leads using a dense neural network classifier. *Sensors*, 22(16), 2022. ISSN 1424-8220. doi:10.3390/s22166071.

George B. Moody and Roger G. Mark. A new method for detecting atrial fibrillation using r-r intervals. *Computers in Cardiology*, 10:227–230, 1983. URL `https://physionet.org/content/afdb/1.0.0/`.

A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000. URL `https://physionet.org/`.

George B. Moody and Roger G. Mark. The impact of the mit-bih arrhythmia database. *IEEE Eng in Med and Biol*, 20 (3):45–50, 2000. URL `https://physionet.org/content/mitdb/1.0.0/`.

Philip de Chazal, M. O'Dwyer, and R.B. Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51(7):1196–1206, 2004. doi:10.1109/TBME.2004.827359.

Eduardo José da S. Luz, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. Ecg-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127:144–164, 2016. ISSN 0169-2607. doi:10.1016/j.cmpb.2015.12.008.

Monika Butkuvienė, Andrius Petrėnas, Andrius Sološenko, Alba Martín-Yebra, Vaidotas Marozas, and Leif Sörnmo. Considerations on performance evaluation of atrial fibrillation detectors. *IEEE Transactions on Biomedical Engineering*, 68(11):3250–3260, 2021. doi:10.1109/TBME.2021.3067698.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Quenas B. Soares, Rosangela Monteiro, Fábio B. Jatene, and Marco A. Gutierrez. A lightweight unidimensional deep learning model for atrial fibrillation detection. In *2022 Computing in Cardiology (CinC)*, pages 1–4, 2022.

MA Reyna, N Sadr, EA Perez Alday, A Gu, AJ Shah, C Robichaux, AB Rad, A Elola, S Seyedi, S Ansari, H Ghanbari, Q Li, A Sharma, and Clifford GD. Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021. *Computing in Cardiology*, 48:1–4, 2021. URL `https://moody-challenge.physionet.org/2021/`.

MA Reyna, N Sadr, EA Perez Alday, A Gu, AJ Shah, C Robichaux, AB Rad, A Elola, S Seyedi, S Ansari, H Ghanbari, Q Li, A Sharma, and Clifford GD. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *Physiol. Meas*, 2022. URL `https://moody-challenge.physionet.org/2021/`.

Jibin Wang and Xiaotai Wu. A deep learning refinement strategy based on efficient channel attention for atrial fibrillation and atrial flutter signals identification. *Applied Soft Computing*, 130:109552, 2022. ISSN 1568-4946. doi:10.1016/j.asoc.2022.109552.