Can Deep Learning Models Differentiate Atrial Fibrillation from Atrial Flutter?

Estela Ribeiro[1], **Quenaz Bezerra Soares**[1], **Felipe Meneguitti Dias**[1],
Jose E. Krieger[1], **Marco Antonio Gutierrez**[1]

[1]Heart Institute (InCor) – Clinics Hospital
University of Sao Paulo Medical School (HCFMUSP)
Sao Paulo – SP – Brazil

estela.ribeiro@hc.fm.usp.br, quenaz.soares@hc.fm.usp.br,

f.dias@hc.fm.usp.br, j.krieger@hc.fm.usp.br, marco.gutierrez@incor.usp.br

Abstract. *Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) are prevalent arrhythmias that present similar clinical features, challenging automated ECG differentiation. This study investigates the classification of AFib and AFlut using 12-lead ECGs from the CinC2021 dataset and a private dataset. We evaluated both 1D and 2D deep learning models. For 1D models, LiteVGG-11 demonstrated the highest performance, achieving an Acc of* 77.91*, AUROC of* 87.17*, and F1 score of* 76.59*. For 2D models, the EfficientNet-B2 outperformed other architectures, with an Acc of* 75.20*, AUROC of* 85.50*, and F1 of* 71.59*. Our results show that 1D models outperform 2D ones and that performance varies significantly across datasets, highlighting the difficulty in distinguishing AFib from AFlut.*

1. Introduction

Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) are distinct irregular heart rhythms originating from abnormal activity in the heart's upper chambers, the Atria [Ko Ko et al. 2022]. These conditions pose significant risks, especially for the elderly [Shah et al. 2018]. AFib involves chaotic electrical activity, causing rapid, irregular atrial contractions at 350-500 beats per minute, compromising heart function and raising stroke risk [Thaler 2019, Brundel et al. 2022]. AFlut, often misdiagnosed as AFib, features a single electrical circuit driving atrial contractions at 250-350 beats per minute, disrupting heart function [Thaler 2019]. Early diagnosis and treatment are crucial for managing AFib and AFlut and reducing severe complications such as stroke [Shah et al. 2018].

Subtle or absent symptoms often accompany irregular heart rhythms, including chest pain, dizziness, shortness of breath, fainting, and palpitations [Shah et al. 2018, Brundel et al. 2022], associated with rapid ventricular rate and inadequate diastolic ventricular filling [Ko Ko et al. 2022]. Automated detection systems can significantly aid in promptly and accurately identifying these conditions, improving healthcare efficiency and reducing patient wait times. This is especially beneficial for underprivileged hospitals with limited access to experienced cardiologists, alleviating strain on their healthcare infrastructure.

The electrocardiogram (ECG), an essential tool for diagnosing cardiac issues, is utilized extensively worldwide, with millions of exams conducted annually. ECG involves the measurement of the heart's electrical activity using electrodes affixed to a patient's skin and is considered the gold standard for noninvasive diagnosis of various heart disorders [Thaler 2019]. Clinical assessment of AFib and AFlut predominantly relies on non-invasive 12-lead ECGs where distinct patterns of electrical activity on the ECG signal enable differentiation between these two conditions [Brundel et al. 2022].

On the ECG, AFib is characterized by the absence of P waves, irregular RR intervals, and fibrillatory waves, while AFlut typically displays sawtooth flutter waves [Thaler 2019]. However, despite these distinct patterns, AFlut is often misdiagnosed as AFib due to similar symptoms and AFib's higher prevalence [Ko Ko et al. 2022, Shah et al. 2018, Brundel et al. 2022]. Some studies suggest that AFlut may be misinterpreted as AFib, especially when ventricular activity is highly irregular, causing AFlut to mimic AFib on surface ECGs [Ghafoori et al. 2018]. This misinterpretation can lead to inappropriate treatment, as each condition requires a specific therapeutic approach.

Over the past decades, the research community has increasingly focused on automating AFib detection, with Deep Learning (DL) emerging as an effective technique for ECG analysis [Hicks et al. 2021, Wegner et al. 2022]. Studies consistently show high accuracy in detecting AFib compared to non-AFib classes [Ivanovic et al. 2019, Dias et al. 2021, Tutuko et al. 2021, Jekova et al. 2022, Dias et al. 2023], with some proposing merging AFib and AFlut into a single class for classification [Jekova et al. 2022]. However, distinguishing between AFib and AFlut has received limited attention, and existing studies have produced unsatisfactory results [Ivanovic et al. 2019].

Most studies differentiating between AFib and AFlut typically use datasets such as 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], featuring extended records of two-lead one-dimensional ECGs. These studies adopt a classification approach for ECG signals, categorizing them into AFib, AFlut, and Normal Sinus Rhythm, often with limited subject pools. Consequently, they frequently had to partition this data into smaller segments for analysis. While employing the same subject for both training and testing sets may yield more precise results due to intra-subject heartbeat interdependence, caution is necessary, as relying solely on intra-subject paradigms could lead to overly optimistic and biased classifications [de Chazal et al. 2004, Butkuvienė et al. 2021].

In clinical environments, ECG exams are often stored in Picture Archiving and Communication Systems (PACS) as images. PACS are widely used in healthcare for storing and managing medical imaging data, such as X-rays, MRIs, CT scans, and ECGs, in a standardized and accessible manner. This system helps streamline the workflow for clinicians, allowing them to retrieve and review images efficiently. ECG exams are typically stored in PACS as images rather than as one-dimensional (1D) signals. This practice presents a challenge when attempting to extract the original 1D signals from these stored images, making it a non-trivial task. Consequently, instead of developing models that classify various heart diseases using 1D ECG signals, it is more practical and applicable to focus on creating image-based models. These models can directly utilize the ECG images stored in PACS, aligning better with the existing clinical workflows and data storage practices.

Furthermore, most studies classifying ECGs rely on 1D signals [Dias et al. 2023]. However, in clinical practice, physicians diagnose by visually examining and interpreting 12-lead ECGs exams. Thus, we hypothesize that 2D (image-based 12 lead ECG exams) DL models designed for AFib and AFlut discrimination may outperform one-dimensional models. Additionally, considering the common occurrence of misdiagnoses between these conditions, we also anticipate sub-optimal performance from DL models.

In this study, our aim is to investigate the effectiveness of employing 12-lead ECGs to differentiate between AFib and AFlut, utilizing either one-dimensional signals or traditional 12-lead ECG images, with a binary classification approach. We utilized data from the six largest PhysioNet Cinc Challenge 2021 (CinC2021) databases, along with a private database sourced from ambulatory patients at a tertiary referral hospital. We explored two types of input data: images (2D) and one-dimensional (1D) signals, with the objective of determining which yields better performance. This approach distinguishes our study from other state-of-the-art DL-based ECG classification research. To conduct our experiments, we evaluated the performance of various different Convolutional Neural Network (CNN) architectures for image-based and one-dimensional-based input data. To the best of our knowledge, this study represents the first report on the assessment of AFib and AFlut discrimination using end-to-end CNNs.

2. Methods

In this section, we describe the dataset, the preprocessing steps, and the deep neural networks architecture used for binary classification of ECG signals. 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).

2.1. Datasets

Tabela 1. Number of selected 12-lead ECG exams from the six CinC2021 datasets and the InCor-DB private database.

Datasets	ECGs (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 (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
Private InCor-DB 12-Lead ECG	9,528	8,219	1,309
Total	*23,139*	*13,457*	*9,682*

We utilized the PhysioNet CinC Challenge 2021 (CinC2021) databases [Reyna et al. 2021, Reyna et al. 2022], which offer a repository of standard 12-lead ECGs covering 30 cardiac abnormality diagnoses. It comprises 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.

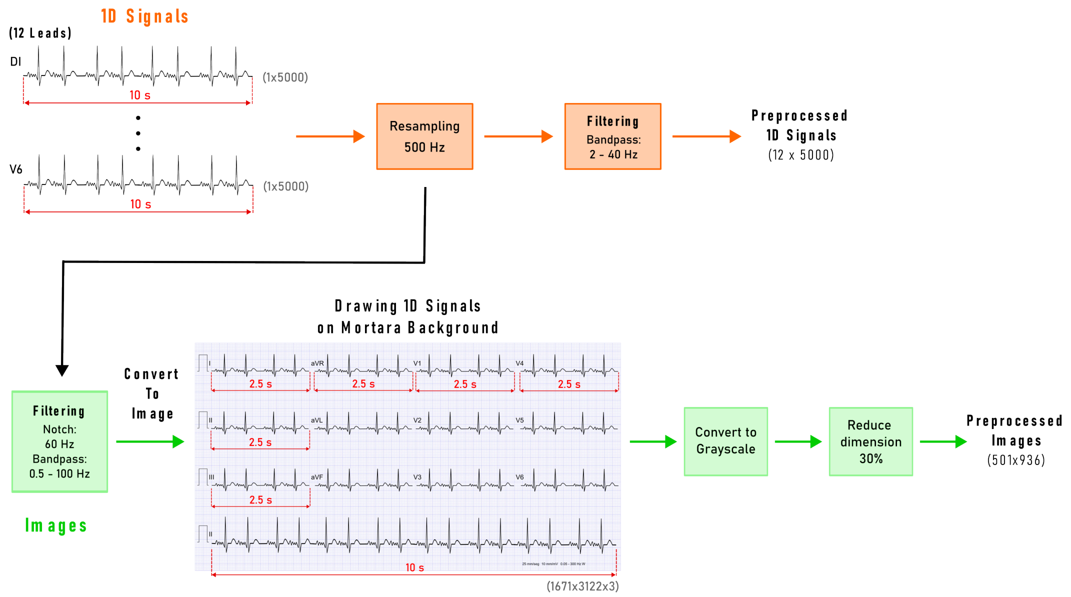

Figura 1. Preprocessing Steps for 1D and 2D ECG signals.

Typically, 12-lead ECGs present 10 s of recorded signals. To avoid losing arrhythmic morphologies present in long-term ECG labels, we selected datasets with recordings of approximately 10 s. Therefore, we excluded PTB and INCART datasets due to longer recordings exceeding 10 s. We also incorporated a private dataset of 12-lead ECGs, denoted as InCor-DB, comprising data collected between 2017 and 2020 [Dias et al. 2023]. This dataset was sourced from the Picture Archiving and Communication System (PACS) of a specialized tertiary referral hospital in Brazil with a focus on cardiology, namely Heart Institute Hospital. Data were acquired using MORTARA TM ELI 250c machines, encompassing 52 distinct clinical diagnoses related to cardiac abnormalities. It is important to note that this private dataset fully adheres to all pertinent ethical regulations and was approved by the Institutional Review Board (IRB).

This study aimed to analyze patient data diagnosed with AFib and AFlut arrhythmia, excluding records with different diagnostic annotations. We utilized class weight estimation techniques to handle dataset imbalance. Table 1 outlines the number of ECGs in the six largest CinC2021 databases and the InCor-DB dataset.

2.2. Data preprocessing

The standard 12-lead ECG raw signals were resampled to 500 Hz and standardized to a length of 10 seconds. This entailed either truncating longer signals to the initial 10 seconds or zero-padding shorter signals to achieve the desired duration. Our preprocessing consists of two phases: one for 1D signals and the other for 2D signals, which are essentially images. Figure 1 displays our preprocessing approach.

For 1D signals, we applied a Butterworth bandpass filter with a frequency range of 2-40 Hz, maintaining the original sampling rate of 500 Hz. To create the dataset for 2D signals, we converted the 1D raw signals from the original datasets into images using the MORTARA ECG image template, with the signals drawn onto this background. The original image dimensions were 1671x3122x3. We opted for the MORTARA template to mimic how physicians would typically encounter ECG exams. Prior to conversion, the signals underwent filtering with a 60 Hz notch filter and a 0.5–100 Hz bandpass Butterworth filter. We then converted the images to grayscale and resized them to 30% of their original dimensions (resulting in a 501x936 grayscale image) to reduce computational complexity.

2.3. Deep learning models

2.3.1. One-dimensional Classification

We employed seven 1D CNN architectures to assess the performance of AFib and AFlut classification in 1D data: (i) LiteVGG-11 [Soares et al. 2022]; (ii) LiteResNet-18; (iii) MobileNet; (iv) ResNet-50; (v) VGG-16; (vi) DenseNet-121; and (vii) EfficientNet-B2. For the traditional CNNs, we adapted the 2D convolutions to 1D convolutions. In the case of Lite models (LiteVGG-11 and LiteResNet-18), we implemented a lightweight CNN proposed by Quenaz et al. (2022) [Soares et al. 2022]. These Lite models deliver comparable performance to their original counterparts, while demanding fewer computing resources. Their approach incorporates depth-wise separable 1D convolution layers (DWConv), a reduced number of filters, a global average pooling for flattening, and fewer units in the dense layers. We retained the fully connected layers of the original models, only modifying the replacement of the last layer with a single output using a sigmoid activation. Each model underwent training for over up to 120 epochs with a batch size of 64. To mitigate overfitting, we incorporated an early stopping callback with a patience of seven epochs. This means that if the model does not improve in the validation dataset for seven consecutive epochs, the training process is stopped.

2.3.2. Image Classification

To assess the performance of image-based classification of AFib and AFlut, we used five traditional and widely used 2D CNNs: (i) MobileNet; (ii) ResNet-50; (iii) VGG16; (iv) Densenet-121; and (v) EfficientNetB2. The fully connected layers comprised a customized 3-layer perceptron with dropout regularization set at 30%, ReLU activation function in intermediate layers, and a sigmoid function in the final layer. Each model underwent training for 30 epochs, utilizing a batch size of 8. Similar to the 1D classification, we implemented an early stopping callback with a patience of seven epochs to prevent overfitting.

2.4. Performance Evaluation

We conducted a 10-fold cross-validation for all experiments, and we present the results in the following format: mean (std). To prevent data leakage, we ensured that exams from the same patient did not appear in different partitions of the cross-validation protocol. In order to evaluate the performance of the employed models, we considered five distinct metrics, including: Sensitivity (Se), Specificity (Spe), F1-score (F1), Area Under Receiver Operating Characteristic curve (AUROC) and Accuracy (Acc). Given the significant class imbalance within the dataset, we defined our best model based on the F1-score.

2.5. Experimental Setup

We conducted experiments to assess the performance and generalizability of our models utilizing distinct approaches: 1D and image-based ECGs. To evaluate the effectiveness of our models, we employed various evaluation setups. Our goal was to assess the overall performance of the models and their ability to generalize to external datasets. The experiments were carried out for both the 1D and image-based ECGs, utilizing the setups described in Table 2.

Tabela 2. Summary of experimental setups. Each setup was applied to both 1D and image-based ECGs.

Setup	Train / Validation / Test (10-fold)	External Validation
1	CinC2021 and InCor-DB	None
2	CinC2021	InCor-DB
3	InCor-DB	CinC2021
4	InCor-DB	Chapman, CPSC, GA, Ningbo, PTB-XL

3. Results

3.1. One-dimensional-based classification

Table 3 displays performance results for seven proposed architectures, considering AFlut as the positive class. We employed **Setup 1**, utilizing both the CinC2021 and InCor-DB datasets for training, validation, and testing.

Tabela 3. Performance results of seven proposed architectures for 1D 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 these results, we adopted different strategies: (**Setup 2**) Training/validating/testing exclusively with CinC2021 followed by external validation with InCor-DB (Table 4); and the opposite (**Setup 3**) training with InCor-DB and external validation using CinC2021 (Table 5).

Moreover, Figure 2 presents the accuracy results for the top-performing 1D-based classification model, LiteVGG-11, trained exclusively on InCor-DB and validated externally on individual CinC2021 datasets (**Setup 4**).

3.2. Image-based classification

Table 6 showcases performance results for five proposed architectures in image-based AFib and AFlut classification. We employed **Setup 1**, utilizing both CinC2021 and InCor-DB for training, validation, and testing.

Similar to 1D-based classification, we implemented the following strategies: (**Setup 2**) Training/validating/testing exclusively with CinC2021, followed by external validation with InCor-DB (Table 7); and the opposite (**Setup 3**) training with InCor-DB, externally validating using CinC2021 (Table 8).

Tabela 4. Performance results of seven proposed architectures for one-dimensional input data. Train = CinC2021 dataset. External Validation = InCor-DB dataset.

Test on CinC2021 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)

Tabela 5. Performance results of seven proposed architectures for one-dimensional input data. Train = InCor-DB dataset. External Validation = CinC2021 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 CinC2021 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)

Tabela 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)**

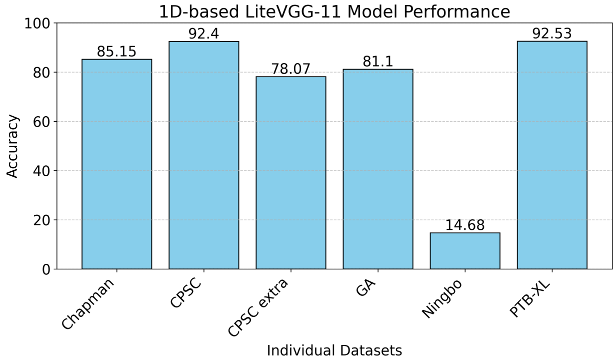

Figura 2. Performance of LiteVGG-11 1D-based model trained on InCor-DB dataset and with external validation on each individual CinC2021 dataset.

Tabela 7. Performance results of five proposed architectures for image input data. Train = CinC2021 dataset. External Validation = InCor-DB dataset.

			Test on CinC2021 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)

Tabela 8. Performance results of five proposed architectures for image input data. Train = InCor-DB dataset. External Validation = CinC2021 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 CinC2021 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)

Figure 3 depicts accuracy results for our top-performing image-based classification model, EfficientNet-B2. This model was solely trained on InCor-DB and externally validated on individual CinC2021 datasets (**Setup 4**).

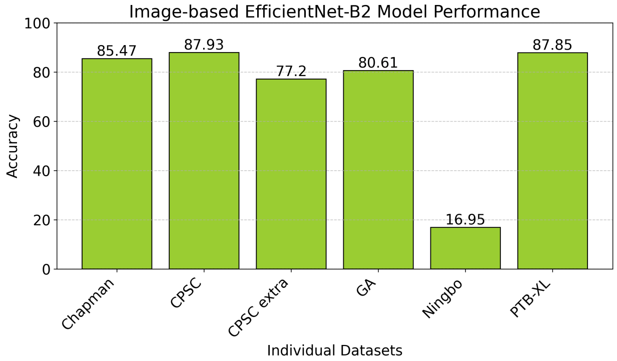

Figura 3. Performance of EfficientNet-B2 image-based model trained on InCor-DB dataset and with external validation on each individual CinC2021 dataset.

4. Discussion

In our current study, we employed different approaches to evaluate the potential of CNN models in distinguishing between AFib and AFlut. As far as we know, we are the first to present an assessment of AFib and AFlut specific discrimination using end-to-end CNNs, demonstrating the feasibility of reasonably distinguishing these two diagnoses.

Our primary findings can be summarized as follows and will be further discussed below: (1) When utilizing all available databases (CinC2021 and InCor-DB), models based on 1D data achieved the capacity to discriminate between AFib and AFlut, exhibiting reasonable performance; (2) Models based on 2D data demonstrated poor performance, with the exception of the EfficientNet-B2 model; (3) Concerning the available datasets, models trained solely on the CinC2021 databases struggle to differentiate the study classes, resulting in metrics that closely resemble chance levels. Conversely, models exclusively based on the InCor-DB private dataset successfully separated the classes; (4) We emphasize the significance of evaluating the separability of classes within the study dataset before contemplating the combination of AFib and AFlut exams into a single class for further analysis. In our research, we observed a clear differentiation between these two classes in the InCor-DB dataset, but this was not evident in the case of the CinC2021 databases; Additionally, (5) Concerning the CinC2021 datasets, we advise exercising caution when using the Ningbo dataset. Our results indicate that a majority of the exams labeled as AFlut are predicted as AFib by our models.

4.1. Models based on both CinC2021 and InCor-DB dataset

In 1D models (Table 3), EfficientNet-B2 struggled to address the problem, performing close to chance level, while LiteVGG-11 had the best performance. Moreover, among image-based models (Table 6), performance was generally poor, except for EfficientNet-B2, which exhibited results similar to 1D models. Previous research aimed at distinguishing AFib and AFlut using MIT-BIH datasets [Wang and Wu 2022] that have limitations due to a limited number of subjects in long-term Holter recordings [Jekova et al. 2022]. These recordings failed in representing arrhythmia diversity compared to CinC2021 databases and the InCor-DB dataset, which feature more subjects and exams.

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

The results from the CinC2021 databases indicated that our proposed image-based networks struggled to differentiate AFib from AFlut, with most evaluation metrics hovering near chance level. Table 7 supports this assessment, particularly for the CinC2021 test set. Even the EfficientNet-B2 architecture, while showing better performance compared to others, seemed to assign exams predominantly to one class, as indicated by specificity and sensitivity metrics. In contrast, our 1D-based models (Table 4) performed reasonably well, particularly LiteVGG-11, LiteResNet-18, and MobileNet architectures.

Upon analyzing data distribution across individual datasets within the CinC2021 database, we observed that the imbalanced data can potentially lead our models to learn to differentiate datasets rather than addressing the primary task of discriminating AFib and AFlut. As shown in Table 1, the CinC2021 database comprises 5,238 AFib exams and 8,373 AFlut exams. Given that a significant portion of AFlut samples originates from the Ningbo First Hospital 12-lead ECG dataset, it's reasonable that our models are distinguishing exams based on their origin rather than their underlying arrhythmia type.

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

When utilizing the InCor-DB private dataset (Tables 5 and 8), despite class imbalance, our models achieved outstanding performance, with Acc and AUROC scores exceeding 90%. However, during external validation on the CinC2021 dataset, our models struggled to generalize. Given these findings, it appears appropriate to merge AFib and AFlut labeled exams in future ECG classification experiments using the CinC2021 databases, as our models' performance approaches chance level, contrasting with the clear discriminability of the two classes in the InCor-DB dataset.

Considering the significant class imbalance, we evaluated model performance based on F1-score, selecting MobileNet for 1D models and EfficientNet-B2 for image-based models as our top-performing models. These models were exclusively trained on the InCor-DB dataset and validated externally with each dataset from the CinC2021 databases. In most cases, our trained models performed well, except for the CPSC and Ningbo datasets (Table 3). For the CPSC dataset, most exams were predicted as AFib, which aligns with all exams being labeled as such. However, regarding the Ningbo dataset, where AFlut is the designated label for all exams, our model predominantly predicted AFib for exams labeled as AFlut. This discrepancy may be attributed to potential variations in diagnostic criteria among cardiologists from different nations when distinguishing AFib and AFlut. Conversely, diagnoses in the InCor-DB dataset originated from cardiologists within the same hospital, likely following consistent diagnostic criteria. In the worst-case scenario, one might argue that exams in the Ningbo dataset could be mislabeled or, more concerning, misdiagnosed.

5. Conclusion

In this work, we explored the potential of CNN models to distinguish between AFib and AFlut based on ECG data. Contrary to our initial expectations, our findings suggest that 1D models generally outperformed image-based models in this discrimination task. This discrepancy underscores the complexity of translating clinical intuition into computational models and highlights the importance of empirical validation in machine learning research. Specifically, our analysis demonstrated that 1D models exhibited superior performance, particularly when trained on the InCor-DB dataset. However, model performance decreased when validated on CinC2021 dataset. Additionally, we emphasized the importance of careful dataset selection and evaluation, as well as consistency in exam labeling. While models trained on InCor-DB achieved high accuracy, there were discrepancies in model predictions for the Ningbo dataset, highlighting the need for standardized diagnostic criteria.

Acknowledgements

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

Referências

Brundel, B. J. J. M., Ai, X., Hills, M. T., Kuipers, M. F., Lip, G. Y. H., and de Groot, N. M. S. (2022). Atrial fibrillation. *Nature Reviews Disease Primers*, 21(8).

Butkuvienė, M., Petrėnas, A., Sološenko, A., Martín-Yebra, A., Marozas, V., and Sörnmo, L. (2021). Considerations on performance evaluation of atrial fibrillation detectors. *IEEE Transactions on Biomedical Engineering*, 68(11):3250–3260.

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

Dias, F. M., Ribeiro, E., Moreno, R. A., Ribeiro, A. H., Samesima, N., Pastore, C. A., Krieger, J. E., and Gutierrez, M. A. (2023). Artificial intelligence-driven screening system for rapid image-based classification of 12-lead ecg exams: A promising solution for emergency room prioritization. *IEEE Access*.

Dias, F. M., Samesima, N., Ribeiro, A., Moreno, R. A., Pastore, C. A., Krieger, J. E., and Gutierrez, M. A. (2021). 2d image-based atrial fibrillation classification. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4.

Ghafoori, E., Angel, N., Dosdall, D. J., MacLeod, R. S., and Ranjan, R. (2018). Atrial fibrillation observed on surface ecg can be atrial flutter or atrial tachycardia. *Journal of Electrocardiology*, 51(6, Supplement):S67–S71.

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

Hicks, S. A., Isaksen, J. L., Thambawita, V., Ghouse, J., Ahlberg, G., Linneberg, A., Grarup, N., Strümke, I., Ellervik, C., Olesen, M. S., Hansen, T., Graff, C., Holstein-Rathlou, N.-H., Halvorsen, P., Maleckar, M. M., Riegler, M. A., and Kanters, J. K. (2021). Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific Reports*, 11(1):10949.

Ivanovic, M. D., Atanasoski, V., Shvilkin, A., Hadzievski, L., and Maluckov, A. (2019). 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.

Jekova, I., Christov, I., and Krasteva, V. (2022). Atrioventricular synchronization for detection of atrial fibrillation and flutter in one to twelve ecg leads using a dense neural network classifier. *Sensors*, 22(16).

Ko Ko, N. L., Sriramoju, A., Khetarpal, B. K., and Srivathsan, K. (2022). Atypical atrial flutter: review of mechanisms, advances in mapping and ablation outcomes. *Current Opinion in Cardiology*, 37(1):36–45.

Moody, G. B. and Mark, R. G. (1983). A new method for detecting atrial fibrillation using r-r intervals. *Computers in Cardiology*, 10:227–230.

Moody, G. B. and Mark, R. G. (2000). The impact of the mit-bih arrhythmia database. *IEEE Eng in Med and Biol*, 20(3):45–50.

Reyna, M., Sadr, N., Perez Alday, E., Gu, A., Shah, A., Robichaux, C., Rad, A., Elola, A., Seyedi, S., Ansari, S., Ghanbari, H., Li, Q., Sharma, A., and GD, C. (2021). Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021. *Computing in Cardiology*, 48:1–4.

Reyna, M., Sadr, N., Perez Alday, E., Gu, A., Shah, A., Robichaux, C., Rad, A., Elola, A., Seyedi, S., Ansari, S., Ghanbari, H., Li, Q., Sharma, A., and GD, C. (2022). Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *Physiol. Meas.*

Shah, S. R., Luu, S.-W., Calestino, M., David, J., and Bray, C. (2018). Management of atrial fibrillation-flutter: uptodate guideline paper on the current evidence. *Journal of Community Hospital Internal Medicine Perspectives*, 8(5):269–275.

Soares, Q. B., Monteiro, R., Jatene, F. B., and Gutierrez, M. A. (2022). A lightweight unidimensional deep learning model for atrial fibrillation detection. In *2022 Computing in Cardiology (CinC)*, pages 1–4.

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

Tutuko, B., Nurmaini, S., Tondas, A. E., Rachmatullah, M. N., Darmawahyuni, A., Esafri, R., Firdaus, F., and Sapitri, A. I. (2021). Afibnet: an inplementation of atrial fibrillation detection with convolutional neural network. *BMC Med Inform Decis Mak*, 21:216.

Wang, J. and Wu, X. (2022). A deep learning refinement strategy based on efficient channel attention for atrial fibrillation and atrial flutter signals identification. *Applied Soft Computing*, 130:109552.

Wegner, F. K., Plagwitz, L., Doldi, C., Willy, K., Wolfes, J., Sandmann, S., Varghese, J., and Eckardt, L. (2022). Machine learning in the detection and management of atrial fibrillation. *Clinical Research in Cardiology*, 111(9):1010–1017.