Deep Learning Approach for Detection of Atrial Fibrillation and Atrial Flutter Based on ECG Images

Estela Ribeiro[1,2], Felipe M. Dias[1,3], Quenaz B. Soares[1,3], Jose E. Krieger[1,2], Marco A. Gutierrez[1,2,3]

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

[2]University of Sao Paulo Medical School (FMUSP)
Sao Paulo – SP – Brazil

[3]Polytechnique School, University of Sao Paulo (POLI USP)
Sao Paulo – SP – Brazil

`{estela.ribeiro, f.dias, quenaz.soares, j.krieger}@hc.fm.usp.br`

`marco.gutierrez@incor.usp.br`

Abstract. *This study explores the application of image-based deep learning techniques to distinguish between Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) using images of standard 12-lead ECG exams from a private database. By implementing a MobileNet Convolutional Neural Network architecture, we achieve a high classification performance, with an accuracy of 95.6%, AUROC of 97.6%, F1-score of 83.2%, specificity of 99.6%, and sensitivity of 72.7%. We also applied explainable methods, such as Grad-CAM and LIME, to try to interpret the model's decision-making process and identify significant regions within the ECG images that contribute to the classification. Our results demonstrate the potential of image-based deep learning approaches for accurate and reliable discrimination between AFib and AFlut, paving the way for enhanced diagnostic capabilities in clinical settings.*

1. Introduction

Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) are heart conditions characterized by irregular heart rhythms at the Atria [Cosío 2017, Ko Ko et al. 2022], being particularly dangerous for the elderly population [Shah et al. 2018]. Both conditions are associated with an increased risk of stroke and other heart-related complications [Brundel et al. 2022]. Thus, the development and validation of AFib and AFlut automatic detection methods may collaborate to speed up its diagnosis on triage to provide the necessary treatment.

The use of non-invasive 12-lead electrocardiogram (ECG) is the standard method for the clinical diagnosis of AFib and AFlut [Brundel et al. 2022, Cosío 2017] since these two heart conditions can be distinguished by the patterns of electrical activity registered on an ECG exam. AFib is characterized by a chaotic electrical activity at the atria, the absence of P waves, irregular RR intervals, and the presence of fibrillatory waves, whereas AFlut commonly presents a sawtooth flutter wave pattern [Thaler 2019].

There have been several efforts at developing automated ECG interpretation methods that use traditional machine learning approaches and rule-based expert systems to provide a classification of well-known cardiac condition patterns. Deep learning methods have recently been shown to be more effective than traditional methods for automatic electrocardiogram (ECG) analysis [Hicks et al. 2021, Wegner et al. 2022]. Regarding AFib and AFlut classification methods, only a limited number of works propose to differentiate these two rhythms, using one-dimensional signals with long records of two-lead ECGs exams [Ivanovic et al. 2019], mainly using the MIT-BIH Atrial Fibrillation [Moody and Mark 1983, Goldberger et al. 2000] and MIT-BIH Arrhythmia [Moody and Mark 1983, Goldberger et al. 2000] datasets. However, these studies usually use a limited number of leads with few subjects, which could hinder the classification.

In this study, we developed a Convolutional Neural Network based on 12-lead ECG image records to differentiate AFib from AFlut using data from a private database acquired from ambulatory patients of a tertiary referral hospital. To the best of our knowledge, this is the first study to present an AFib/AFlut classification model based on an end-to-end convolutional neural network with 12-lead ECG exams.

2. Methods

2.1. Dataset

We used a private 12-lead ECG dataset collected from 2017 to 2020 from the Picture Archiving and Communication System (PACS) of a tertiary referral hospital in Brazil specialized in cardiology (Heart Institute Hospital), acquired from Mortara™ ELI 250c machines, with 52 different clinical diagnoses of cardiac abnormalities [Dias et al. 2021]. We only considered exams that included the specific diagnostic labels of AFib and AFlut arrhythmia. Exams with different diagnostic annotations were not considered in this work. Table 1 summarizes the entire number of ECG records available in our private dataset.

Table 1. Number of selected 12-lead ECG records from our private database.

ECG records (total)	Atrial Fibrillation	Atrial Flutter
9,528	8,219	1,309

2.2. Data preprocessing

Our preprocessing steps are based on previous works [Dias et al. 2021] for Atrial Fibrillation classification. The standard 12-lead ECG raw one-dimensional signals of limb leads (I, II, III, aVR, aVL, and aVF) and chest leads (V1 to V6) had a frequency of 500 Hz and a fixed length of 10 seconds. It was necessary to map the one-dimensional raw signals to their image counterparts by employing a MORTARA ECG image template without any signal as background and drawing the signals onto this image (original dimension 1671x3122x3). 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 (resized dimension 501x936).

2.3. Deep Learning Model

To evaluate the performance of image-based Atrial Fibrillation and Atrial Flutter classification, we used a traditional and widely used 2-dimensional Convolutional Neural Network named MobileNet [Howard et al. 2017]. The fully connected layers were a customized 3-layer perceptron with a dropout rate of 30%, a ReLU activation function in intermediate layers, and a sigmoid function in the last layer. Our model was trained over 30 epochs using a batch size of 8 samples. Additionally, to avoid overfitting, an early stopping callback to terminate the model training process after seven epochs was also added. Furthermore, we estimate class weights to deal with the unbalanced datasets.

2.4. Performance Evaluation

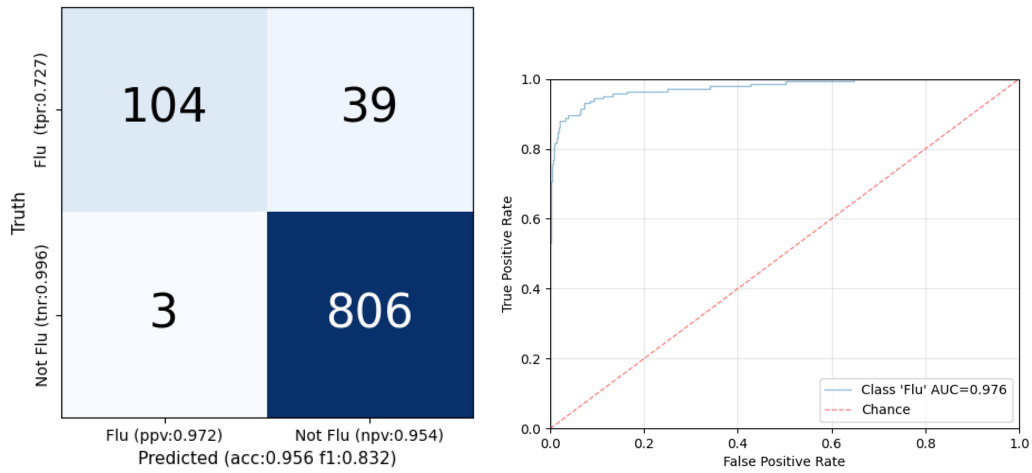
We performed an 80% train, 10% validation, and 10% test split. To prevent data leakage, we ensured that exams from the same patient does not exist in different divisions. To evaluate the employed models, we addressed five statistical metrics, including Sensitivity (Se), Specificity (Spe), F1-score (F1), Area Under Operational Receipt Curve (AUROC), and Accuracy (Acc). All experiments were performed using a Foxconn High-Performance Computer (HPC) M100-NHI with a 4 GPU cluster of 16 GB NVIDIA Tesla V100 cards. The methodology was implemented using the Python framework (version 3.6.8) and Keras/TensorFlow (version 2.3.0).

2.5. Explainable methods

We selected two well-known methods to understand the predictions of our model. The Gradient-weighted Class Activation Mapping (Grad-CAM) method [Selvaraju et al. 2017], creates a heatmap for a certain class label, allocating significant regions of the ECG input image to the prediction, and creating a coarse localization map using the gradient information flowing into the final convolutional layer. The Local Interpretable Model-Agnostic Explanations (LIME) method [Ribeiro et al. 2016] uses a simpler interpretable surrogate model (e.g. linear regression), applying perturbations on the ECG input. Furthermore, it weights the perturbed images in accordance with how similar they are to the original image, and utilizes this knowledge to train an interpretable model that imitates the behavior of the original model for a certain ECG input image. In our work, we used a linear regression as the surrogate model, and segmented the images with a quickshift segmentation algorithm with 1000 perturbations. Additionally, for the LIME method, visualizations produce a map with the most positively significant areas highlighted in green and the least positively significant areas highlighted in red, whereas for the Grad-CAM method, the most significant pixels are highlighted in red and the least significant in blue.

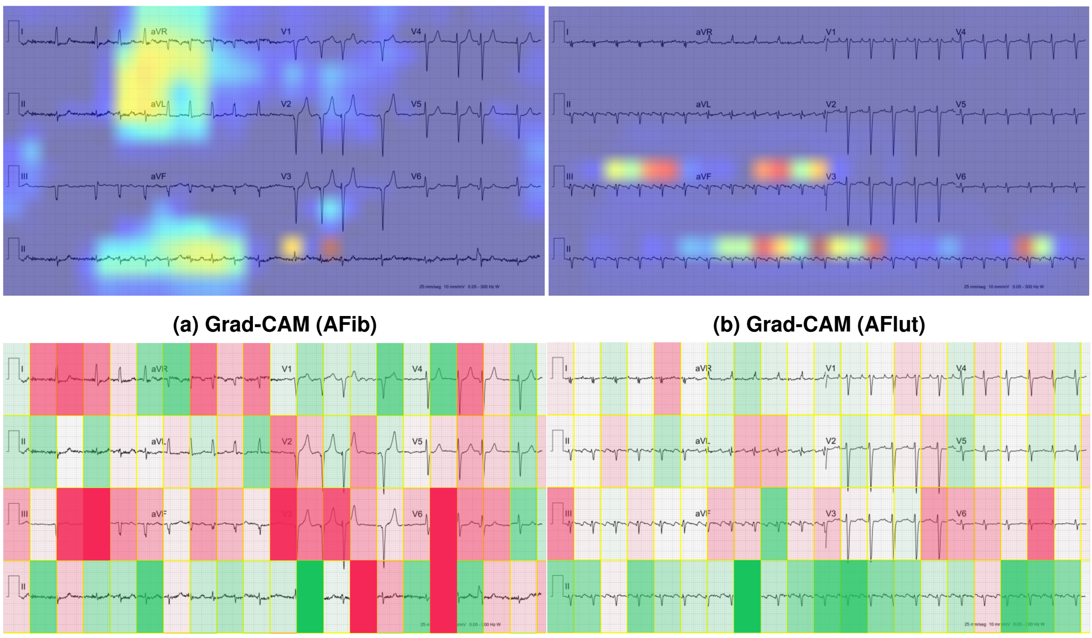
3. Results

Our model performance achieved Acc 95.6%, AUROC 97.6%, F1 83.2%, Spe 99.6%, and Se 72.7%, with AFlut as a positive class. Figure 1 shows the confusion matrix and the AUROC curve for the test split. Figure 2 displays the results of Grad-CAM and LIME methods of an exam labeled Atrial Fibrillation (a) and (c) and an exam labeled Atrial Flutter (b) and (d).



(a) (b)

Figure 1. MobileNet performance on our private dataset test split: (a) The confusion matrix; (b) The AUROC curve.



(a) Grad-CAM (AFib) (b) Grad-CAM (AFlut)

(c) LIME (AFib) (d) LIME (AFlut)

Figure 2. MobileNet performance on our private dataset test split: (a) Grad-CAM of a AFib example; (b) Grad-CAM of a AFlut example; (c) LIME of a AFib example; (d) LIME of a AFlut example.

4. Discussion

In this study, we demonstrated the effectiveness of an image-based deep learning approach for the classification of Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) using standard 12-lead ECG records. We utilized a MobileNet Convolutional Neural Network architecture, which provided a high classification performance. These results indicate that our model is capable of accurately distinguishing between AFib and AFlut based on ECG images.

Our approach differs from previous works, which mostly focused on one-dimensional signals and a limited number of leads. The utilization of a private database with a diverse range of patients and clinical diagnoses also contributed to the robustness of our model.

The data preprocessing steps, including the conversion of raw ECG signals into images and the application of filtering techniques, proved to be effective in maintaining the essential features of the ECG signals while reducing computational complexity. Additionally, the use of class weights in the training process helped to address the class imbalance issue, which could have otherwise affected the model's performance.

In Figure 2, we applied two interpretation methods – Grad-CAM and LIME – to examine electrocardiogram (ECG) examples of Atrial Fibrillation (AFib) and Atrial Flutter (AFlut). For both cases, Grad-CAM and LIME attributed the highest importance to the DII long lead, which was expected, as both AFib and AFlut are rhythm disorders and this lead contains the most significant information about rhythm. Nevertheless, in both instances, we were unable to derive a valid interpretation for the highlighted regions. Consequently, further research is needed to enhance the interpretability capabilities of the proposed model.

Despite the promising results, there are some limitations to our study. Firstly, our dataset was acquired from a single hospital, which may limit the generalizability of the results to other populations and settings. Secondly, our model may not perform as well for other types of arrhythmias or in the presence of noisy ECG signals. Future research should focus on expanding the dataset to include a more diverse patient population and investigating the performance of the model for various types of arrhythmias.

5. Conclusion

In this work, we presented an image-based deep-learning approach for the classification of Atrial Fibrillation (AFib) and Atrial Flutter (AFlut) using images of standard 12-lead ECG exams. The implementation of the MobileNet Convolutional Neural Network architecture demonstrated high classification performance. Furthemore, integration of explainable methods, such as Grad-CAM and LIME, allowed us to identify significant regions within the ECG images, that contribute to the classification. Despite the limitations related to the dataset, our results demonstrate the potential of image-based deep-learning approaches for accurate and reliable discrimination between AFib and AFlut.

Ethics Statement

This research was approved by the Internal Review Board (IRB), registration CAAE 45070821.3.0000.0068, as part of the Machine Learning in Cardiovascular Medicine Project.

Acknowledgements

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

References

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).

Cosío, F. G. (2017). Atrial flutter, Typical and Atypical: A review. *Arrhythmia Electrophysiology Review*, 6(2):55–62.

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.

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.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

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.

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.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. arXiv: 1602.04938.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

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.

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

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.