A Multimodal Deep Learning Approach for Atrial Fibrillation Classification from 12-Lead ECG Recordings

Rafael Laranjeira[1], **Bruno Pimentel**[1], **Thiago Cordeiro**[1], **Álvaro Sobrinho**[2], **Estela Ribeiro**[3], **Felipe Dias**[3], **Marco Antonio Gutierrez**[3]

[1]Instituto de Computação (IC) – Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brasil

thiago@ic.ufal.br

Abstract. *This paper presents a multimodal approach for the automated classification of Atrial Fibrillation (AF). Our approach converts standard 12-lead Electrocardiogram (ECG) images into complementary modalities (i.e., images, time series, and spectrograms) to support neural network analyses. A weighted fusion mechanism combines modality-specific features and adapts their contributions to different classification challenges. We evaluated the approach under a balanced class distribution by downsampling the majority class (Normal Rhythm) to match the number of samples in the minority class (AF). Using the InCor-DB private dataset, the implemented multimodal fusion models achieved an F1 score of 99.28% in the balanced scenario and 88.59% in the imbalanced one. Additional validation on the Zheng-DB public dataset confirmed model generalization, with an F1 Score of 98.13% under balanced conditions. Our results demonstrate the feasibility of combining spectrograms, images, and time series for automated AF classification.*

1. Introduction

Cardiovascular Diseases (CVDs) are especially prevalent in high-income countries, where sedentary lifestyles, poor dietary habits, and aging populations drive their rising incidence. Yet they also pose a health threat in low- and middle-income countries, where limited access to healthcare and preventive measures exacerbates their impact [Geldsetzer and Tisdale 2024]. CVDs include conditions that affect the heart and blood vessels, such as coronary artery disease, stroke, heart failure, and Atrial Fibrillation (AF), which is a health condition characterized by irregular heartbeats.

Electrocardiogram (ECG) results can aid in identifying this condition by revealing characteristic features. It results from disorganized electrical activity in the atria, which produces inconsistent intervals between R waves—unlike the predictable intervals seen in normal sinus rhythm. This combination of features is crucial for diagnosing AF and differentiating it from other arrhythmias, thereby supporting clinicians in making informed risk assessments and treatment decisions. Existing research emphasizes the significance of these ECG characteristics for accurate AF diagnosis, underscoring the crucial need for precise detection in clinical practice [Chousou et al. 2023, Nesheiwat et al. 2024].

Over the past decades, researchers have applied diverse methodologies to detect AF in ECG exams. Previous works have focused on architectures such as Convolutional Recurrent Neural Networks [Zihlmann et al. 2017], hybrid CNNs combined with LSTMs

and shortcut connections [Ping et al. 2020], or frameworks utilizing 1D CNNs with BiLSTMs applied to PPG signals [Aldughayfiq et al. 2023]. While these approaches achieved significant results, they often rely on a single modality or homogeneous data formats.

This paper presents an approach to automated AF classification using multimodal deep learning that processes and integrates multiple ECG representations (i.e., images, spectrograms, and time series) to deliver robust and interpretable AF classification across diverse clinical scenarios. We introduce a specialized pipeline that converts standard 12-lead ECG images into complementary modalities to address the common challenge of heterogeneous ECG data formats in clinical practice.

2. Multimodal Approach

Classifying AF with deep learning requires data preprocessing, model design, and evaluation. We handle multimodal data using images, spectrograms, and time series (Figure 1).

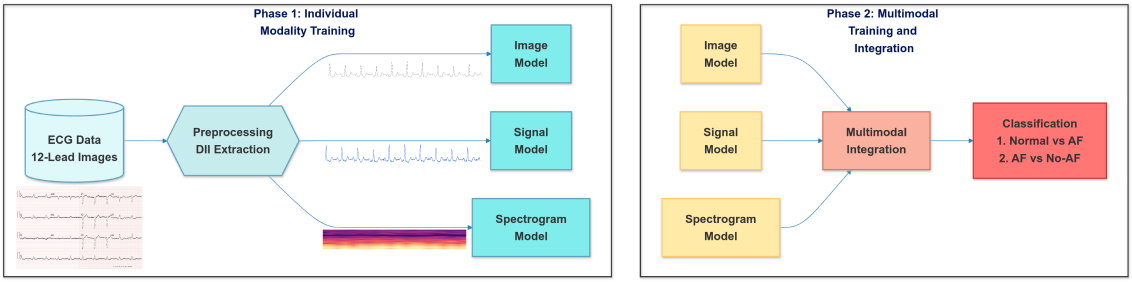

Figure 1. Overview of our approach for ECG-based AF classification.

In the first phase, individual-modality training processes raw ECG data through preprocessing and DII lead extraction, yielding three parallel inputs: the ECG image, the processed signal, and the spectrogram. A dedicated model handles each input. In the second phase, multimodal training and integration are performed by retraining the individual models and combining them through a multimodal integration layer. This step enables the final classification of normal ECGs versus AF.

2.1. Database Description, Preprocessing, and Validation

We relied on a private, image-based dataset of 12-lead ECG examinations, known as InCor-DB, collected between 2017 and 2020 [Dias et al. 2023]. The dataset contains over 100,000 exams in PNG format, each with a resolution of 3,122 × 1,671 pixels, and includes detailed diagnostic reports to ensure accurate labeling and analysis. As an initial screening step, we identified and removed examinations with overlapping leads to avoid compromising signal quality and interpretation. As an additional validation step, we also used a publicly available dataset of 10,646 12-lead ECG examinations (Zheng-DB) [Zheng et al. 2020]. We converted the signals, originally in time-series format, to images to match the processing applied to the private dataset. We considered only AF and typical rhythm cases from a subset of exams.

The extraction of the DII lead signal from ECG examinations used a specialized multi-stage digital signal processing pipeline to ensure accurate digitization. The process began with image segmentation, isolating the DII lead region through vertical pixel-density analysis. The system generates a density profile in which high pixel concentrations highlight signal traces, while low-density regions mark the spaces between leads.

Signal separation was performed using peak detection with a minimum threshold of 150 pixels and a separation distance of 300 pixels. The algorithm identifies points of minimum pixel density between peaks as optimal cut points. Signals were extracted column-wise by computing the average position of detected pixels to generate a sequence of coordinate pairs.

To remove artifacts, the extracted signal was subjected to DBSCAN clustering (radius 50, min 5 points) and outlier filtering to ensure continuity. Initial preprocessing applied bilateral filtering (kernel size = 9, $\sigma_{\text{color}} = 15$, $\sigma_{\text{space}} = 15$) to reduce noise while preserving edges. This was followed by Otsu's adaptive thresholding for binarization and morphological dilation with a 2×2 elliptical kernel. Quality control validated signal continuity and temporal relationships, flagging failures for manual review. Finally, coordinates were converted to time-series format and amplitude-normalized.

Time-series processing begins by extracting Lead II for AF-detection relevance. Signals are standardized to 300 Hz for consistency with the InCor-DB dataset. Baseline wander is removed via a 201-sample moving average filter, while noise is suppressed using Savitzky-Golay filtering (window = 15, order = 3). Finally, the amplitude is normalized, and the signal length is fixed to 3,010 samples by cropping or padding.

The image pipeline standardizes ECG traces to 80 × 320 pixels at 300 DPI to maintain aspect ratio and efficiency. Images are converted to grayscale, and the intensities are normalized to the 0–1 range. Grid lines and axes are removed, and signal traces are enhanced to maximize contrast.

Spectrograms are generated using STFT with a Hann window (512 samples, 75% overlap) to balance time-frequency resolution [Jeon et al. 2020]. The frequency range is restricted to 0.5–50 Hz to highlight clinical bands. Following percentile-based contrast enhancement, spectrograms are converted to grayscale, normalized, and resized to 80 × 320 pixels to match the image representation methodology.

We achieved class balance by downsampling the majority class. To ensure robustness, the experiment employed 5-fold cross-validation across four different random seeds. This decision was motivated by the sensitivity to random initialization observed in preliminary experiments, in which model performance varied considerably with the seed. By combining multiple seeds with k-fold cross-validation, the evaluation provides a more comprehensive assessment of model stability and performance reliability, which is particularly important given the stochastic nature of neural network training.

The InCor-DB dataset comprises 16,894 samples with a perfectly balanced distribution (8,447 samples per class, AF and Normal), achieved by downsampling the majority class. In contrast, the Zheng-DB dataset reflects a natural distribution, with 413 AF samples and 1,366 Normal samples, enabling evaluation of the model's performance under imbalanced conditions. We utilize StratifiedGroupKFold with a custom split function to maintain class distribution and prevent patient data leakage. This mechanism achieves three key goals: stratification to maintain class balance, patient-level grouping to keep all records from a single patient within a single split, and reproducible shuffling with a fixed seed.

The implementation divides the dataset into five folds. By ensuring that no patient's data appears in both the training and test sets within the same fold, we prevent the

model from memorizing patient-specific patterns, thereby forcing it to learn generalizable ECG features. Within each fold, the data is further divided into training and validation subsets without patient overlap. For efficient processing, each fold's multi-modal data is serialized and stored in TFRecord format.

2.2. Architecture for the Experiment

2.2.1. Image-Based Architecture

The image modality architecture adopts a residual neural network design optimized for ECG-based AF detection using image data. It balances feature extraction with computational efficiency through four sequential convolutional blocks, each employing a dual-path structure. The residual block consists of two paths. In the main path, a Conv2D layer with a (3×3) kernel, He initialization, and L2 regularization ($\lambda = 0.001$) is followed by batch normalization with a momentum of 0.99 and a ReLU activation. A second Conv2D layer, maintaining the same number of filters, is then applied, followed by an additional batch normalization step. In the residual path, a 1×1 convolution ensures dimensionality matching, and its output is added to the main path. The block concludes with a final ReLU activation applied to the combined output.

The filter progression expands systematically across the blocks ($16 \rightarrow 32 \rightarrow 64 \rightarrow 128$) to manage the growth of feature space. We perform a spatial reduction using a 2×2 max-pooling operation after each block, and apply a dropout with a rate of 0.3 for regularization. The classification head begins with a dual-stream pooling stage, which combines global average pooling and Global Max Pooling before concatenation. A dense layer with 128 units, L2 regularization ($\lambda = 0.01$), batch normalization, and ReLU activation follows, accompanied by dropout at a rate of 0.3. Another dense layer with 64 units and identical regularization is then applied, and the architecture concludes with a final sigmoid activation layer for classification.

The optimized image model uses a four-block residual network with filter expansion from 16 to 128. This structure enables hierarchical learning of both subtle wave morphologies and broader rhythmic patterns. The main path employs two 3×3 convolutional layers with batch normalization for feature extraction, while the residual path uses a 1×1 convolution for identity mapping. This design preserves signal integrity while allowing the network to learn complex representations. A dual-stream spatial reduction strategy—combining max and average pooling—is applied after each block. This preserves both critical peak information (e.g., R waves) and overall morphology. To prevent overfitting, dropout (0.3) is applied between blocks. The classification head comprises parallel global average and max-pooling streams, followed by dense layers with 128 and 64 units, respectively. This effectively captures multiscale patterns, ranging from individual wave components to global rhythm characteristics.

2.2.2. Spectrogram-Based Architecture

This architecture adopts an inception-inspired design optimized for analyzing time–frequency representations of ECG signals in AF classification. This multi-scale strategy captures both localized frequency transitions and broader spectro-temporal patterns characteristic of arrhythmias, while maintaining computational efficiency through

strategic filter expansion ($32 \rightarrow 64 \rightarrow 128$) and the use of bottleneck layers. The model uses an Inception-inspired architecture optimized for ECG time–frequency analysis. Processing begins with **batch normalization** and a 7×7 convolution (32 filters) to capture broad initial spectro-temporal patterns.

The core comprises three sequential Inception modules with progressively larger filters ($32 \rightarrow 64 \rightarrow 128$). Each module features four parallel paths to capture multi-scale features: 1×1 convolution: Efficient channel-wise transformation; 3×3 & 5×5 paths: Bottlenecked (1×1 pre-layer) to capture transitions at various scales; Max-pooling path: 1×1 projection to preserve salient frequency components.

Independent batch normalization and ReLU activations ensure stable learning before feature concatenation. This synthesis is vital for detecting AF signatures across different scales. The classification head employs global average pooling, a 256-unit dense layer, and a dropout rate of 0.5 to mitigate overfitting. A final sigmoid layer produces the output, effectively distilling complex frequency-domain patterns for cardiac rhythm differentiation.

2.2.3. Time Series Architecture

The hybrid ConvLSTM design integrates convolutional and recurrent components for sequential AF detection. It features two stages: 1D convolutions for local morphology extraction and dual bidirectional LSTMs for modeling complex rhythm patterns. This approach captures multi-scale temporal relationships efficiently for multimodal integration. Temporal feature extraction begins with a 1D convolutional layer (32 filters, kernel size =3, stride=2), followed by batch normalization and ReLU. A second 1D convolution (64 filters, stride=2) with batch normalization further stabilizes training.

Sequential processing employs two bidirectional LSTM layers (64 and 32 units) with a dropout rate of 0.3. The classification head comprises a 32-unit dense layer (ReLU), a 0.2 dropout, and a 16-unit dense layer, followed by a sigmoid activation for binary classification.

The time-series model employs a hybrid ConvLSTM architecture for sequential ECG analysis. Temporal feature extraction uses two Conv1D layers (32- and 64-filter) with a stride of 2, enabling efficient downsampling while preserving morphological features such as P-waves, QRS complexes, and T-waves. Batch normalization and ReLU activation stabilize this stage and improve computational efficiency.

Sequential processing employs a dual BiLSTM structure to capture complex temporal dependencies. A 64-unit Bidirectional LSTM layer with sequence retention analyzes forward and backward relationships, followed by a 32-unit BiLSTM layer to refine temporal representations. To prevent overfitting, dropout layers (0.3) are integrated between LSTM components.

The classification head performs progressive dimensionality reduction using two dense layers (32 and 16 units) with ReLU activation. A final 0.2 dropout layer regularizes the model, allowing it to synthesize learned temporal features for effective cardiac condition identification.

2.2.4. Multimodal Fusion Architecture

The multimodal architecture integrates features from three ECG representations—images, spectrograms, and time series—using an adaptive weighting mechanism (Figure 2). This approach enables the model to automatically learn the optimal contribution of each modality during training, enhancing robustness in atrial fibrillation detection.

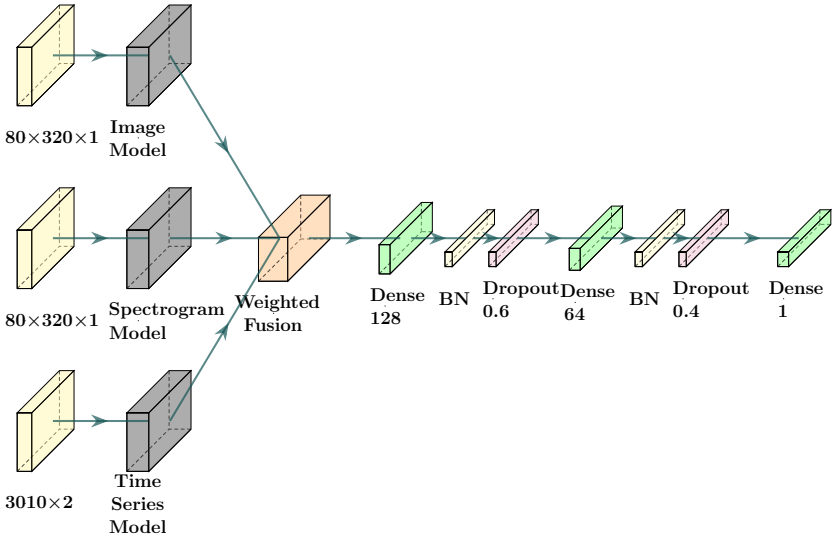

Figure 2. Overview of the multimodal fusion architecture.

We implemented the fusion process through a custom trainable layer that performs a weighted sum of the features extracted before each modality's classification head. The weights are initialized equally (approximately 0.33 per modality) and are updated during training via backpropagation to highlight the most informative modalities for the task. A softmax function constrains the weights to sum to 1, making their contributions directly interpretable as percentages of the final decision. This design enables resource-efficient integration by requiring only three additional parameters to be learned, while effectively capturing the complementary strengths of the different modalities.

We implemented the trainable weighted fusion mechanism as a custom TensorFlow layer that dynamically combines the extracted features from each modality. It uses a softmax-normalized weight vector, initialized with equal values ($[0.33, 0.33, 0.34]$), which adapts during training through backpropagation to optimize the contribution of each modality.

The fusion mechanism combines features through a weighted sum, with each modality's contribution determined by its learned importance, and is defined as:

$$\text{Fused Features} = w_{\text{image}} \cdot x_{\text{image}} + w_{\text{spec}} \cdot x_{\text{spec}} + w_{\text{ts}} \cdot x_{\text{ts}} \tag{1}$$

where w_i represents the learned weight for modality i, and x_i is the feature vector from that modality. The weights are normalized using softmax:

$$w_i = \frac{\exp(w_i^{\text{raw}})}{\sum_{j=1}^{n} \exp(w_j^{\text{raw}})} \tag{2}$$

where w_i^{raw} represents the trainable weight factor before softmax normalization, ensuring all weights sum to 1 and remain positive.

The classification head consists of two fully connected layers with 128 and 64 units, respectively. Each layer applies batch normalization to improve convergence, ReLU activation to introduce non-linearity, and dropout regularization with rates of 0.6 and 0.4 to prevent overfitting. The architecture concludes with an output layer composed of a single neuron with a sigmoid activation function for binary classification.

2.2.5. Training Configuration and Pipeline

We implemented the models in Python 3.12.6 using key libraries, including TensorFlow 2.17.0, Keras 3.5.0, Scikit-learn 1.5.1, NumPy 1.26.4, SciPy 1.14.1, and Pandas 1.2.2. For visualization and analysis, the environment included Matplotlib 3.9.2 and Seaborn 0.13.2, together with interpretability tools SHAP 0.46.0 and LIME 0.2.0.1. We run all experiments in a CUDA 12 environment with NVIDIA cuDNN 8.9.7.29, the NVIDIA CUDA Toolkit 12.3, and NVIDIA TensorRT 10.4.0, which provides GPU acceleration and optimization. The training infrastructure used an NVIDIA GeForce RTX 4070 GPU paired with an AMD Ryzen 5 7600 CPU and 32 GB of DDR5 RAM.

We applied the CosineDecayRestarts scheduler from TensorFlow to both the single-modality and multimodal models, with an initial learning rate of 1×10^{-3}. For the single-modality models, the first_decay_steps parameter was set to 4000; for the multimodal model, it was set to 2000. We manually tuned these values based on an analysis of the models' convergence behavior.

We also used the AdamW optimizer for training, a variant of Adam that incorporates decoupled weight decay to improve generalization. We configured the optimizer with a learning rate controlled by the CosineDecayRestarts scheduler, a weight decay of 0.001, and gradient clipping with a norm of 1.0. Additional parameters included $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-7}$, and the AMSGrad variant enabled (amsgrad=True).

We utilized TensorFlow's LossScaleOptimizer for mixed-precision training to improve numerical stability and GPU performance with a batch size of 128 for all models. Binary cross-entropy loss was employed to distinguish standard ECG signals from AF.

As the dataset is balanced, class weights were unnecessary during training from TFRecord files. Training ran for up to 100 epochs, using early stopping with a patience of 8 epochs based on the validation F1 score, which also governed model checkpointing.

To evaluate model effectiveness, we monitored binary accuracy, PR AUC, and ROC AUC, along with precision, recall, specificity, and the F1 score. We also recorded the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to provide a complete view of classification performance.

3. Experimental Results

We evaluated four approaches: image-based analysis, spectrogram analysis, time-series analysis, and a multimodal fusion approach. We tested each model with four random seeds (42, 73, 99, 122) to ensure robust evaluation and assess the stability of the results.

We employed a 5-fold stratified group K-fold cross-validation strategy, repeated with four different random seeds. Table 1 presents the dataset statistics.

Table 1. Dataset statistics across random seeds.

Seed	Split	Samples/Fold	Normal (%)	AF (%)	Class Ratio
3*42	Train	10836 ± 26	49.91 ± 0.35	50.09 ± 0.35	1.004 ± 0.014
	Val	2680 ± 26	50.38 ± 1.39	49.62 ± 1.39	0.986 ± 0.056
	Test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
3*73	Train	10833 ± 21	49.77 ± 0.31	50.23 ± 0.31	1.009 ± 0.012
	Val	2682 ± 22	50.93 ± 1.24	49.07 ± 1.24	0.964 ± 0.049
	Test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
3*99	Train	10808 ± 31	50.04 ± 0.15	49.96 ± 0.15	0.998 ± 0.006
	Val	2707 ± 32	49.83 ± 0.59	50.17 ± 0.59	1.007 ± 0.024
	Test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000
3*122	Train	10826 ± 23	49.99 ± 0.21	50.01 ± 0.21	1.000 ± 0.009
	Val	2689 ± 22	50.04 ± 0.87	49.96 ± 0.87	0.999 ± 0.035
	Test	3379 ± 1	50.00 ± 0.00	50.00 ± 0.00	1.000 ± 0.000

Values are reported as mean $\pm$ standard deviation across folds.
Class ratio is calculated as AF / Normal.

The results reveal a moderate positive correlation between the image and spectrogram modalities ($r = 0.21$), while both exhibit strong negative correlations with the time-series modality ($r = -0.71$ and $r = -0.84$, respectively). These findings suggest that the model distributes importance differently across modalities, with image and spectrogram features often complementing each other, whereas time-series features contribute in a contrasting manner. This balance indicates that the fusion mechanism leverages complementary information from heterogeneous ECG representations to enhance classification robustness.

Figure 3 presents the scatterplot matrix of the normalized weights assigned to all modalities. The diagonal histograms show the distribution of weights for each modality, while the off-diagonal plots illustrate the pairwise relationships between modalities. The results highlight the narrow distributions and consistent weighting across runs, with moderate correlations observed between image and spectrogram weights, and a clearer trade-off between time-series weights and the other modalities.

Table 2 reports the average performance metrics (mean ± standard deviation across four seeds) for the image, spectrogram, time-series, and multimodal models. The multimodal model consistently achieves the strongest overall performance, with the highest accuracies (0.993), recall (0.993), and F1 score (0.993), while exhibiting low variance across seeds, indicating stable training behavior. Among the single-modality models, the image-based approach performs best across most metrics, slightly outperforming the time-series and spectrogram models. Although all models achieve high ROC and PR AUC values, the multimodal fusion approach shows a clear advantage by integrating complementary information across heterogeneous ECG representations.

Figure 4 compares the cross-validation performance of the image, spectrogram, time-series, and multimodal models across four random seeds (42, 73, 99, and 122). The results show that the multimodal model consistently achieves the highest F1 scores, demonstrating superior stability and robustness compared to the single-modality approaches. Among the individual modalities, the image-based model generally performs

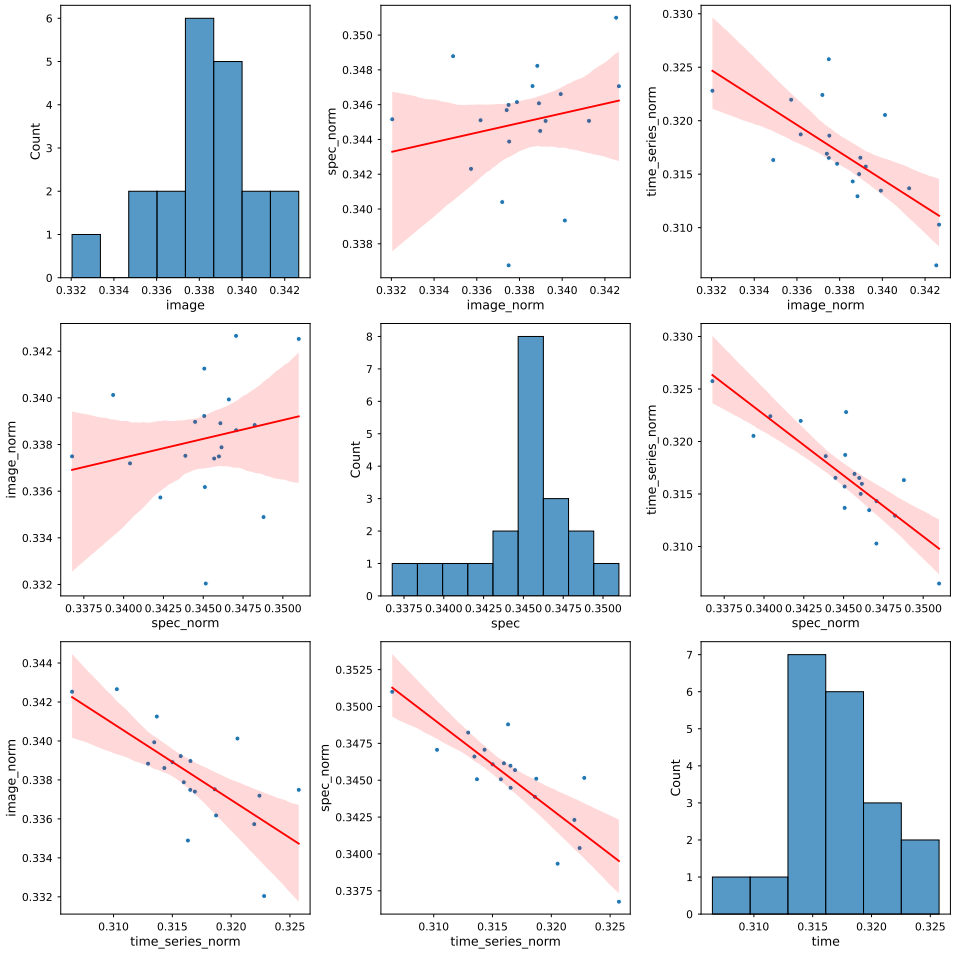

Figure 3. Analysis of modality fusion weights across five cross-validation folds and four seeds. The diagonal shows the weight distributions for the image, spectrogram, and time-series modalities. Off-diagonal plots display pairwise correlations between modalities with fitted regression lines (red).

Table 2. Average performance across modalities (mean ± std over four seeds).

Metric	Image	Spec	Time Series	Multimodal
Accuracy	0.991 ± 0.001	0.979 ± 0.003	0.982 ± 0.001	0.993 ± 0.000
Precision	0.993 ± 0.001	0.978 ± 0.003	0.979 ± 0.002	0.992 ± 0.001
Recall	0.988 ± 0.002	0.979 ± 0.003	0.985 ± 0.001	0.993 ± 0.001
F1 Score	0.991 ± 0.001	0.979 ± 0.003	0.982 ± 0.002	0.993 ± 0.000
Specificity	0.993 ± 0.001	0.978 ± 0.003	0.979 ± 0.002	0.992 ± 0.001
ROC AUC	0.999 ± 0.000	0.996 ± 0.002	0.994 ± 0.001	0.998 ± 0.001
PR AUC	0.999 ± 0.000	0.995 ± 0.002	0.993 ± 0.001	0.997 ± 0.001

best, followed by the time-series and spectrogram models. These findings highlight the advantage of multimodal fusion, which leverages complementary features across heterogeneous ECG representations to improve classification accuracy.

Figure 5 illustrates the trade-off between average performance (F1 score) and computational cost (training time in seconds) for the image, spectrogram, time-series, and multimodal models. The multimodal model achieves the highest F1 score but at the expense of the longest training time (approximately 1167.8s), reflecting the added complex-

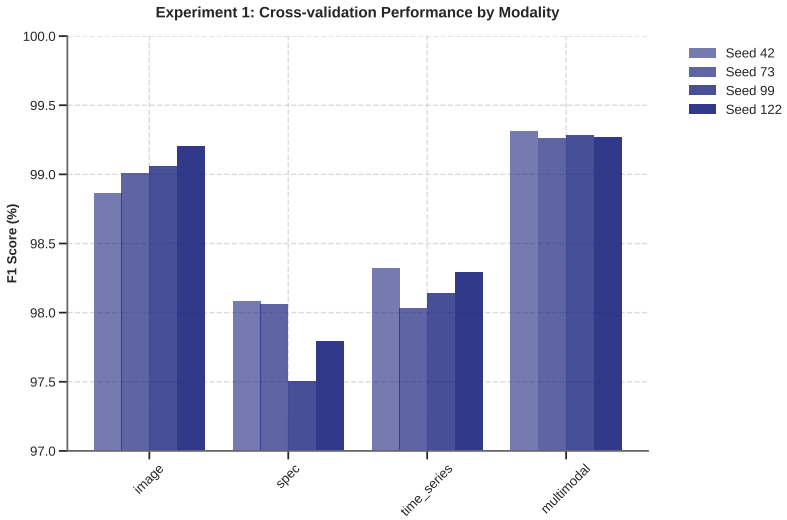

Figure 4. F1 score comparison across different modalities.

ity of integrating multiple modalities. In contrast, the image-based and time-series models achieve competitive performance while incurring substantially lower training costs (around 605.2 and 819.5 seconds, respectively). In comparison, the spectrogram model exhibits slightly lower performance but remains the most efficient in terms of training time (281.0 seconds). These results highlight the balance between accuracy and computational efficiency when selecting model architectures for ECG classification.

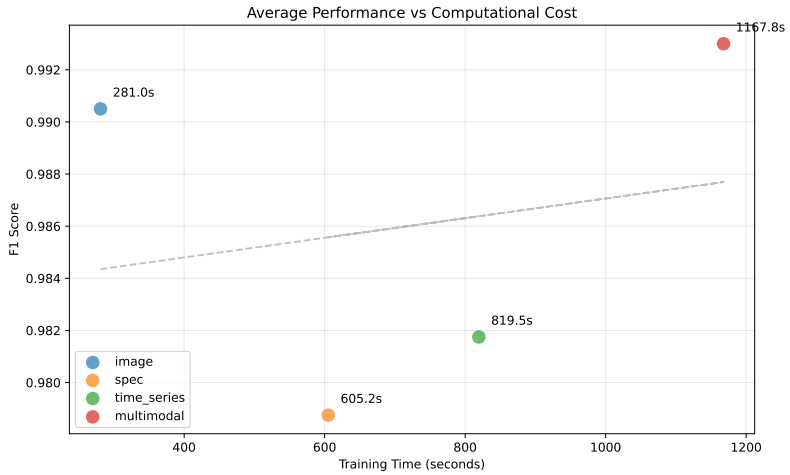

Figure 5. Average performance vs cost comparison across different modalities.

Figure 6 compares the performance of the image, spectrogram, time-series, and multimodal models between cross-validation and external validation. The results show a consistent decrease in F1 score across all modalities when moving from cross-validation to external validation, reflecting the expected generalization gap. Despite this reduction, the multimodal model continues to achieve the highest F1 scores, followed by the image-based and time-series models, with the spectrogram model performing slightly lower. These findings highlight both the robustness of the multimodal approach and the importance of external validation for assessing real-world generalizability.

Figure 7 shows the distribution of learned fusion weights for all modalities across

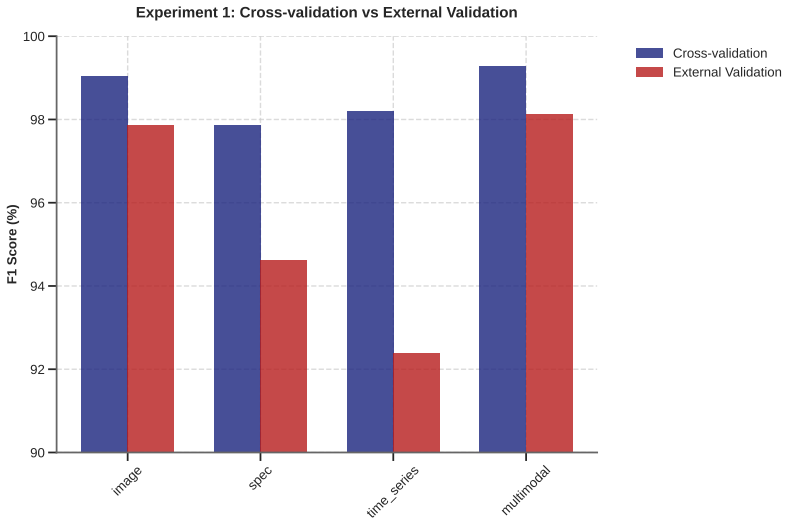

Figure 6. Cross-validation and external validation F1 scores across all modalities.

the four random seeds. The results indicate that the time-series modality consistently receives the highest weight, followed by the spectrogram and image modalities, with only minor variations across seeds. This pattern suggests that the fusion mechanism prioritizes temporal information while still leveraging complementary contributions from spectral and image-based features, ensuring robust integration of multimodal data.

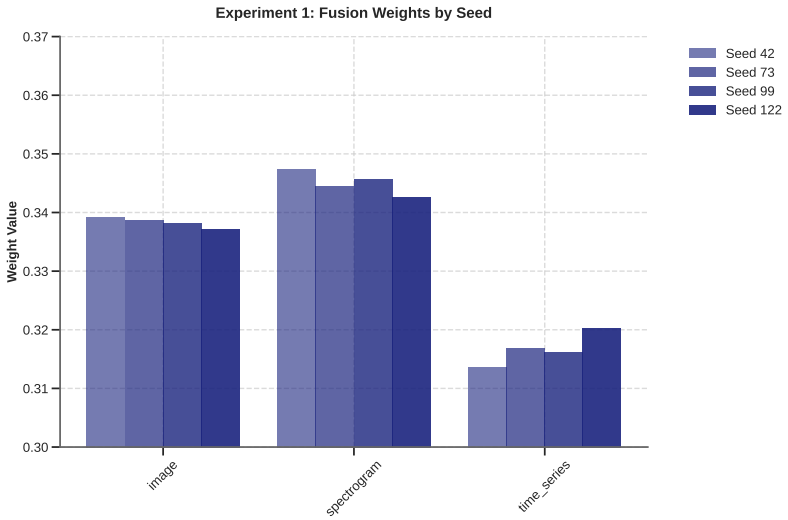

Figure 7. Distribution of fusion weights across different seeds.

4. Discussion and Conclusions

The experimental results demonstrate the effectiveness of combining multiple ECG representations for Atrial Fibrillation (AF) classification. Under controlled conditions, the multimodal approach outperformed single-modality models, achieving an F1 score of 0.9928 ± 0.0002 and maintaining stability across random seeds. The balanced fusion weights (image: 0.3382 ± 0.0025, spectrogram: 0.3450 ± 0.0033, time series: 0.3167 ± 0.0045) indicate that each modality contributed complementary information for distinguishing AF from normal rhythms.

Furthermore, the multimodal model demonstrated strong generalization on the external validation dataset (F1 score = 0.9813 ± 0.0036), highlighting its robustness to dataset-specific variations and its potential for practical clinical applications across diverse healthcare settings.

Despite these promising results, the computational cost of processing multiple representations warrants future optimization. Future research should focus on refining the signal extraction pipeline with advanced denoising techniques, exploring more adaptive fusion mechanisms, and conducting multi-center validations to evaluate performance across different patient populations and clinical protocols. Finally, preprocessing optimizations, such as reducing input dimensionality, are needed to prevent the model from learning non-diagnostic boundary features.

References

Aldughayfiq, B., Ashfaq, F., Jhanjhi, N., and Humayun, M. (2023). A deep learning approach for atrial fibrillation classification using multi-feature time series data from ecg and ppg. *Diagnostics*, 13(14):2442.

Chousou, P. A., Chattopadhyay, R., Tsampasian, V., Vassiliou, V. S., and Pugh, P. J. (2023). Electrocardiographic predictors of atrial fibrillation. *Medical Sciences*, 11(2):30.

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*.

Geldsetzer, P. and Tisdale (2024). The prevalence of cardiovascular disease risk factors among adults living in extreme poverty. *Nature Human Behaviour*, 8:903–916.

Jeon, H., Jung, Y., Lee, S., and Jung, Y. (2020). Area-efficient short-time fourier transform processor for time–frequency analysis of non-stationary signals. *Applied Sciences*, 10(20):7208.

Nesheiwat, Z., Goyal, A., and Jagtap, M. (2024). *Atrial Fibrillation*. StatPearls Publishing, Treasure Island (FL), Updated 2023 Apr 26 edition. StatPearls [Internet].

Ping, Y., Chen, C., Wu, L., Wang, Y., and Shu, M. (2020). Automatic detection of atrial fibrillation based on cnn-lstm and shortcut connection. In *Healthcare*, volume 8, page 139. MDPI.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., and Rakovski, C. (2020). A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):48.

Zihlmann, M., Perekrestenko, D., and Tschannen, M. (2017). Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE.