IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 September 2023, accepted 25 October 2023, date of publication 30 October 2023, date of current version 6 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328538

RESEARCH ARTICLE

Artificial Intelligence-Driven Screening System for Rapid Image-Based Classification of 12-Lead ECG Exams: A Promising Solution for Emergency Room Prioritization

FELIPE MENEGUITTI DIAS [1,2], ESTELA RIBEIRO [1,3], RAMON ALFREDO MORENO [1], ADÈLE HELENA RIBEIRO [1], NELSON SAMESIMA [1], CARLOS ALBERTO PASTORE [1,3], JOSE EDUARDO KRIEGER [1,3], AND MARCO ANTONIO GUTIERREZ [1,2,3]

[1] Heart Institute (InCor), Clinics Hospital University of São Paulo Medical School (HCFMUSP), São Paulo 05403-000, Brazil
[2] Polytechnique School, University of São Paulo (POLI USP), São Paulo 05508-010, Brazil
[3] Department of Cardiopneumology, University of São Paulo Medical School (FMUSP), São Paulo 01246-903, Brazil

Corresponding author: Marco Antonio Gutierrez (marco.gutierrez@hc.fm.usp.br)

This work was supported in part by the São Paulo Research Foundation (FAPESP) under Grant 2021/12935-0 and Grant 2014/50889-7, in part by Foxconn Brazil, and in part by the Zerbini Foundation as part of the Research Initiative titled “Machine Learning in Cardiovascular Medicine.”

ABSTRACT The electrocardiogram (ECG) serves as a valuable diagnostic tool, providing crucial information about life-threatening cardiac conditions such as atrial fibrillation and myocardial infarction. A prompt and efficient assessment of ECG exams in environments such as Emergency Rooms (ERs) can significantly enhance the chances of survival for high-risk patients. Despite the presence of numerous works on ECG classification, most of these studies have concentrated on one-dimensional ECG signals, which are commonly found in publicly available ECG datasets. Nevertheless, the practical relevance of such methods is limited in hospital settings, where ECG exams are usually stored as images. In this study, we have developed an artificial intelligence-driven screening system specifically designed to analyze 12-lead ECG images. Our proposed method has been trained on an extensive dataset comprising 99,746 12-lead ECG exams collected from the ambulatory section of a tertiary hospital. The primary goal was to precisely classify the exams into three classes: Normal (N), Atrial Fibrillation (AFib), and Other (O). The evaluation of our approach yielded AUROC scores of 93.2%, 99.2%, and 93.1% for N, AFib, and O, respectively. To further validate our approach, we conducted evaluations using the 2018 China Physiological Signal Challenge (CPSC) database. In this evaluation, we achieved AUROC scores of 91.8%, 97.5%, and 70.4% for the classes N, AFib, and O, respectively. Additionally, we assessed our method using 1,074 exams acquired in the ER and obtained AUROC values of 98.3%, 98.0%, and 97.7% for the classes N, AFib, and O, respectively. Furthermore, we developed and deployed a system with a trained model within the ER of a tertiary hospital for research purposes. This system automatically retrieves newly captured ECG chart images from the Picture Archiving and Communication System (PACS) within the ER. These images undergo necessary preprocessing steps and serve as input for our proposed classification method. This comprehensive approach established an efficient and versatile end-to-end framework for ECG classification. The results of our study highlight the potential of leveraging artificial intelligence in the screening of ECG exams, offering a promising solution for the rapid assessment and prioritization of patients in the ER.

INDEX TERMS Artificial intelligence, atrial fibrillation, ECG, ECG image, 12-lead electrocardiogram, emergency room.

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kafiul Islam.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide [1], particularly in low- and middle-income countries, accounting for approximately 80% of these fatalities [2]. Furthermore, CVDs impose a significant economic burden, encompassing both direct costs (e.g., hospitalizations) and indirect costs (e.g., loss of productivity due to incapacity to work) [3]. Therefore, there is a pressing need to develop new approaches for the prevention and early treatment of these diseases.

In this regard, the electrocardiogram (ECG) plays a crucial role in accurately identifying various cardiac conditions, including myocardial infarction and atrial fibrillation (AFib). Moreover, ECGs are readily accessible, non-invasive, and cost-effective. Particularly in Emergency Rooms (ER), their significance is amplified, as prompt screening and diagnosis can significantly enhance the chances of patient survival. Thus, the automated classification of ECG exams in such environments holds the potential to optimize clinical workflow by prioritizing patients in critical conditions.

AFib is the most common form of chronic sustained cardiac arrhythmia [4], [5], [6], affecting nearly one percent of the global population [7]. Its prevalence increases with age [1], and individuals over 65 years old have a fourfold higher prevalence. Moreover, untreated AFib significantly increases the risk of other cardiac conditions, including stroke [8], [9]. Early detection and intervention of AFib, thereby preventing potential harm, can have a significant impact on healthcare outcomes and associated costs [10].

The 12-lead ECG, interpreted by a trained physician, is the definitive exam for diagnosing AFib [9], [11]. Physicians typically extract key characteristics from ECG signals, such as P-wave duration and irregular electrical activity, to identify irregularities. However, visually inspecting the 12-lead ECG to detect irregularities is time-consuming. Over the past 60 years, there have been several attempts to develop computerized ECG interpretation methods [12]. These methods utilize rule-based expert systems that rely on well-known patterns of AFib to provide classification. However, these methods have significant drawbacks. First, the classification algorithms are vendor-specific, which means they can only be used with equipment from the vendor that developed the algorithm. Second, accurately identifying certain key ECG features, such as the QT interval, is challenging [13]. Additionally, the classification accuracy, especially regarding arrhythmias, is limited [14].

On the other hand, the use of deep learning-based tools to enhance the diagnostic capabilities of cardiac arrhythmias in both inpatient and outpatient settings has shown remarkable growth in recent years [15]. These methods offer advantages by eliminating the need for specialist-defined features for classification. Instead, they adopt an end-to-end approach where features are automatically extracted from the ECG exam and employed for classification. These algorithms have substantially enhanced the detection of AFib and other cardiac conditions. Nevertheless, the majority of these systems rely on digital one-dimensional signals [16].

In hospital settings, ECG exams are typically stored as images or PDF files in the Picture Archiving and Communication System (PACS) [17]. Therefore, applying one-dimensional ECG classification methods is not feasible in hospital environments. Although some recent studies have proposed 12-lead classification systems with good performance [18], [19], there is still a literature gap regarding the deployment of such methods in clinical environments with an appropriate validation.

In summary, research studies have a limited impact on clinical practice due to several factors. Firstly, most studies primarily focus on one-dimensional ECG signals, limiting their applicability. Secondly, algorithms that are tailored to specific equipment and training datasets further hinder generalizability. Lastly, the current research landscape prioritizes enhancing machine learning model performance while overlooking critical aspects of practical applicability within clinical settings.

In this study, we introduce a new deep learning-based tool for classifying ECG exams using images from a dataset of 99,746 exams acquired from ambulatory patients at a tertiary referral hospital. The classification system encompasses three classes: Normal (N), Atrial Fibrillation (AFib), and Other cardiac condition (O). To validate the system, we conducted an assessment using both an internal test set and an external validation set. Furthermore, to demonstrate the practical feasibility of our approach in clinical settings, we have developed a screening system specifically designed for implementation in Emergency Rooms (ER). This system has been seamlessly integrated into the PACS of a tertiary referral hospital, enabling the automatic detection of newly acquired ECG exams within the ER. Following detection, the ECG image exams, along with relevant demographic information (age, gender, and ethnicity), are processed by our classification algorithm. We further validate the effectiveness of our method by comparing the algorithm's classification with assessments by a panel of experts using exams obtained through this system. Currently, for research purposes, physicians can access this application through a dedicated screen located in the ER. Additionally, we conducted an evaluation to assess the impact of including demographic information in the classification system [14].

The main contributions of our work include:

- Development of a robust deep learning system designed for the automatic classification of 12-lead ECG exam images into three distinct classes: Normal (N), Atrial Fibrillation (Afib), and Other cardiac condition (O);
- Training and assessment of the proposed system conducted on images of ambulatory ECG examinations acquired from a specialized tertiary referral hospital with a distinct focus on cardiology.
- Rigorous validation of our system's performance through evaluation on an external dataset, extending the reach of our research beyond the initial training dataset;
- Deployment of the proposed ECG classification system into a real clinical setting (ER) and evaluation of the system by comparing its results against a committee of experts (cardiologists);

- Detailed investigation into the significance of demographic information in the classification of N, AFib, and O classes.

II. RELATED WORK

ECG classification is a research topic that dates back 60 years, with pioneering works like Caceres et al. [20], which relied on designing features based on clinical knowledge of the ECG. Subsequently, the advent of openly accessible datasets, such as the MIT-BIH dataset [21], provided an opportunity for various research groups to delve deeper into this area. However, a notable limitation of the MIT-BIH dataset is that it contains only 2 ECG leads, whereas typical ECG exams are recorded with 12 leads. Additionally, it contains a limited number of patients (48). Using a dataset with a limited number of leads and patients may hinder the full generalization of the research findings to real-world scenarios.

Despite the limitations of this dataset, numerous recent studies have proposed various approaches for ECG classification. In this context, Marinho et al. [22] presented a feature extraction-based approach for ECG heartbeat classification. They evaluated various feature extraction techniques, including Fourier, Goertzel, Higher-Order Statistics, and Structural Co-Occurrence Matrix. The study employed four different classifiers: Support Vector Machine, Multi-Layer Perceptron, Bayesian, and Optimum-Path Forest. The authors reported achieving accuracies above 90% with low computational complexity. Additionally, in a recent study by Houssein et al. [23], signal descriptors were extracted based on one-dimensional local binary pattern (LBP), wavelet, higher-order statistical, and morphological information. They employed a support vector machine classifier to categorize ECG heartbeats into five classes, achieving an accuracy rate exceeding 98%.

The availability of large-scale 12-lead ECG datasets, such as the 2018 China Physiological Signal Challenge (CPSC) dataset [24], has enabled the exploration of data-hungry approaches like deep learning. In this context, Ribeiro et al. [18] introduced a convolutional neural network (CNN) inspired by the ResNet architecture to classify 12-lead ECGs. Their study employed the most extensive ECG exam dataset to date (CODE dataset), comprising six classes: 1st-degree AV block (1DAVB), Right bundle branch block (RBBB), Left bundle branch block (LBBB), Sinus bradycardia (SBC), AFib, and Sinus tachycardia (STC). Their model outperformed cardiology resident physicians, achieving F1 scores exceeding 80% for all classes. Additionally, Che et al. [25] introduced a hybrid approach that combines a transformer network with a CNN architecture. Through the utilization of a link constraint to enhance the discriminative power of ECG embedding vectors, they achieved favorable classification results. Their research used the CPSC dataset [24] and focused on classifying ECG exams into nine different classes, including AFib, 1DAVB, LBBB, RBBB, Premature atrial contraction (PAC), Premature ventricular contraction (PVC), ST-segment depression (STSD), and ST-segment elevation (STSE), obtaining a mean F1 score of 78.6%. Similarly, Dong et al. [26] used a depth-wise separable convolutional network, along with vision transforms, for ECG classification, also using the CPSC dataset. Their approach demonstrated even higher effectiveness, achieving a mean F1 score of 82.9%.

ECG classification based on 12-lead ECG signals has demonstrated practicability and clinical relevance, making it a preferred choice over approaches relying on the MIT-BIH dataset. Nonetheless, it is essential to take into account that a substantial number of medical facilities store 12-lead ECG examinations in image format within their data repositories [17]. Consequently, the feasibility of employing signal-centric ECG techniques within hospital settings becomes constrained. Recent works have addressed this limitation by proposing methods that classify ECG exams using their image representations as input. For instance, Gliner et al. [27] introduced both one-dimensional and image-based ECG classification systems employing CNNs in both approaches. They achieved accuracies exceeding 90% for all cardiac conditions considered, with a 98% accuracy for detecting AFib. For image-based ECG classification, they mapped the ECG signals to a blank ECG chart. Using the CPSC 2018 dataset, they obtained an AUC of 96% for AFib classification, similar to the performance obtained using one-dimensional signals (98%). Similarly, Sangha et al. [28] also proposed an image-based ECG classification system. They trained their model with the CODE data [18] using an EfficientNet architecture and demonstrated that models using ECG images perform comparably to those using one-dimensional ECG signal models. The approach proposed in [29], on the other hand, used a very small dataset for training, so the generalization power is also compromised.

Table 1 presents a summary of recent studies on ECG classification.

III. METHODS

A. DATA SOURCE

We employed 12-lead ECG exams collected between 2017 and 2020 from ambulatory patients visiting a specialized tertiary referral hospital in Brazil, with a primary focus on cardiology. These examinations were acquired from MORTARA™ ELI 250c electrocardiograph systems, which digitally recorded the ECG signals. Additionally, the system applied a series of filters to each recorded ECG signal to ensure the retrieval of high-quality data. The process involves applying a low-pass filter with adjustable cut-off frequencies (40Hz, 150Hz, and 300Hz) along with a baseline filter. This filtering procedure effectively eliminates both baseline wander and high-frequency noise, including high-frequency electromagnetic interference. Moreover, to mitigate the impact of powerline interference, the MORTARA™ ELI 250c electrocardiograph system incorporates a notch filter tuned to the power line frequency. This notch filter effectively

TABLE 1. Overview of recent ECG classification studies.

Work	Year	Dataset	Leadset	ECG Input Type	Method
Marinho et al. [22]	2019	MIT-BIH	Single lead	Signal (1D)	Features + Classifier
Houssein et al. [23]	2021	MIT-BIH	Single lead	Signal (1D)	Features + Classifier
Ribeiro et al. [18]	2020	CODE	12 lead	Signal (1D)	ResNet
Che et al. [25]	2021	CPSC	12 lead	Signal (1D)	CNN + Transformer
Dong et al. [26]	2023	CPSC	12 lead	Signal (1D)	CNN + Transformer
Gliner et al. [27]	2020	CPSC	12 lead	Signal (1D)	CNN
Sangha et al. [28]	2022	CODE	12 lead	Signal (1D)	EfficientNet
Gliner et al. [27]	2020	CPSC	12 lead	Image (2D)	CNN
Sangha et al. [28]	2022	CODE	12 lead	Image (2D)	EfficientNet
Hao et al. [29]	2020	Private	12 lead	Image (2D)	ResNet

attenuates powerline noise, thereby contributing to the reduction of signal artifacts and enhancing the overall quality of the recorded ECG data.

The ECG signals are transmitted to the hospital's PACS through a dedicated gateway, ensuring integration and accessibility of the recorded data within the medical facility. The gateway seamlessly converts the ECG signals into 2D images in the widely used Digital Imaging and Communications in Medicine (DICOM) format, with dimensions of 3,320 × 2,219 pixels, facilitating easy storage, retrieval, and analysis of the data within the medical system. The resultant image is displayed in the form of an A4-format chart, featuring a reference grid with time axis resolution of 25 mm/s and a voltage axis resolution of 10 mm/mV. For the specific objectives of this investigation, these images were transformed into the Portable Network Graphics (PNG) format. Furthermore, an automated cropping process was applied, resulting in the dimensions of 3,122 × 1,671 pixels. This cropping step was designed to eliminate any sensitive or confidential information located at the upper section of the image as illustrated in Figure S1 of the Supplementary Material.

Each ECG exam was accompanied by a diagnostic report in structured text format. Exams with the same diagnosis shared the same diagnostic text. The dataset consisted of 52 different diagnoses, which were categorized into three classes: N, AFib, and O. To construct the dataset, we integrated patient demographic details, including age, ethnicity, and gender. Patients with pacemakers or under 18 years of age were excluded from the study due to different diagnostic criteria used for evaluating their ECG exams. Additionally, exams without an associated diagnosis or with ambiguous diagnoses, such as "ECG may present first-degree atrioventricular block" were disregarded to ensure the CNN learning process was not influenced by diagnostic uncertainty. After implementing these exclusion criteria, the resulting dataset, which we refer to as InCorDB, comprised a total of 99,746 ECG exams collected from 64,192 unique patients. It included anonymized 2D image ECG exams, de-identified patient demographic information, and their diagnostic reports (N, AFib, and O). This private dataset complied with all pertinent ethical regulations and received approval from the Institutional Review Board (IRB) under registration number CAAE 45070821.3.0000.0068.

To demonstrate the generalizability of our proposed method, we conducted testing on an external database (CPSC) [24]. This dataset comprised 6,877 12-lead one-dimensional ECG signals with durations ranging from 6 to 60 seconds. These signals were classified into nine different classes: N, AFib, 1DAVB, LBBB, RBBB, PAC, PVC, STSD, and STSE. Similar to the approach taken with InCor-DB, we regrouped these nine classes into the categories N, AFib, and O. Furthermore, patients under 18 years of age were excluded from this dataset. Table 2 summarizes the datasets utilized in our study.

B. DEPLOYMENT OF THE MODEL IN THE EMERGENCY ROOM

We deployed our model in the ER of a tertiary referral hospital system for research purposes. We established a system to assess each newly acquired ECG exam within the ER and generated a prioritized list of exams for the attending physicians. Exams classified as AFib are given a higher priority within this system.

The ER also employs MORTARA™ ELI 250 electrocardiograph systems that are seamlessly integrated into the PACS. Similarly to the InCorDB, the ECG exams obtained within the ER underwent cropping to the dimensions 3,122 × 1,671 to safeguard patient information from exposure.

To deploy our model, we first developed a service that processes each new ECG exam from the ER sent to the PACS. It classifies the ECG and saves the information in a database. This information is then exposed through a REST service to a web client, which provides visual feedback to the clinical staff in the ER. The web page is displayed on a monitor in the ER. Patients are listed following a prioritization protocol: exams classified as Atrial Fibrillation (AFib) have the highest priority, followed by Other diseases (O), and then Normal (N). Additionally, within the same classification, more recent exams have lower priority. On the web page, higher priority results in a higher position in the spreadsheet.

Figure 1 shows our model deployment pipeline in a hospital setting, seamlessly integrated into the hospital's dataflow infrastructure. The web application depicted in this figure presents the actual interface accessible to doctors in the emergency room for research purposes.

TABLE 2. Demographic information of patients in our employed datasets.

		InCor-DB	CPSC	ER
Demography	Male: N (%)	51,778 (51.9%)	3,622 (53.9%)	602 (56.0%)
	Age: Mean (SD)	60.2 (16.6)	61.2 (17.9)	61.6 (14.7)
	Ethnicity: N (%)	85,395 (85.6%)	-	697 (62.9%)
Diagnosis	Normal: N (%)	19,282 (19.3%)	821 (12.2%)	139 (12.9%)
	FA: N (%)	9,017 (9.0%)	1,219 (18.1%)	141 (13.1%)
	Other: N (%)	80,464 (80.7%)	4684 (69.7%)	794 (73.9%)

To validate our model in the ER, we gathered ECG exams from this unit over a period of one month. Subsequently, our model was employed to predict the classification of each ECG exam. To provide the ground-truth for the ECGs, two cardiologists were provided with the same set of exams, along with relevant information such as gender, age, and ethnicity. Each cardiologist independently assigned each exam to the following classes: N, AFib, or O. In cases where the two cardiologists disagreed on the diagnosis of an exam, a third cardiologist was consulted to determine the final label. During our analysis, we excluded exams conducted on patients below 18 years of age. However, due to the unavailability of reports for the exams conducted in the ER, we were unable to exclude exams from patients with pacemakers. Consequently, we requested the cardiologists to determine whether each exam belonged to a pacemaker user or not. It is worth mentioning that the cardiologists are board-certified with a minimum of 5 years of experience. Throughout the one-month evaluation, a total of 1,074 valid exams were collected.

C. DATA PREPROCESSING

We used images of ECG exams as input for our model (Figure S1 of the Supplemental Material). A significant portion of the information in an ECG image, such as color information, holds limited relevance for diagnostic purposes. Additionally, considering the consistent grid scale used in our ECG exams (25 mm/s on the x-axis and 0.5 mV/mm on the y-axis), the grid-related information is non-informative. Therefore, our primary goal during the preprocessing stage was to optimize the efficiency of our CNN by removing all nonessential information.

The initial step involved converting the 2D ECG images into grayscale. Subsequently, a threshold filter was applied to remove the reference grid, but this process introduced noise. To eliminate the noise, a morphological erosion operation followed by dilation was performed. Afterward, each lead, including the 10-second DII lead, was individually cropped from the ECG image. To decrease the computational complexity, the images were resized to 30% of their original size. Consequently, the short-lead images (DI, DII, DIII, avR, avL, avF, V1, V2, V3, V4, V5, and V6) were resized to dimensions of 144 × 224 pixels, while the long-lead image (10-second DII) was adjusted to dimensions of 141 × 898 pixels. The short-lead images were combined to form a 3D volume, which, along with the long-lead image, constituted the input for our proposed CNN architecture. Figure 2 outlines the preprocessing steps involved in our methodology: 1) ECG image input, 2) Transformation to grayscale and grid removal, 3) Individual lead cropping and resizing, 4) Preprocessing output: 3D stack of short leads and the 10-second DII lead.

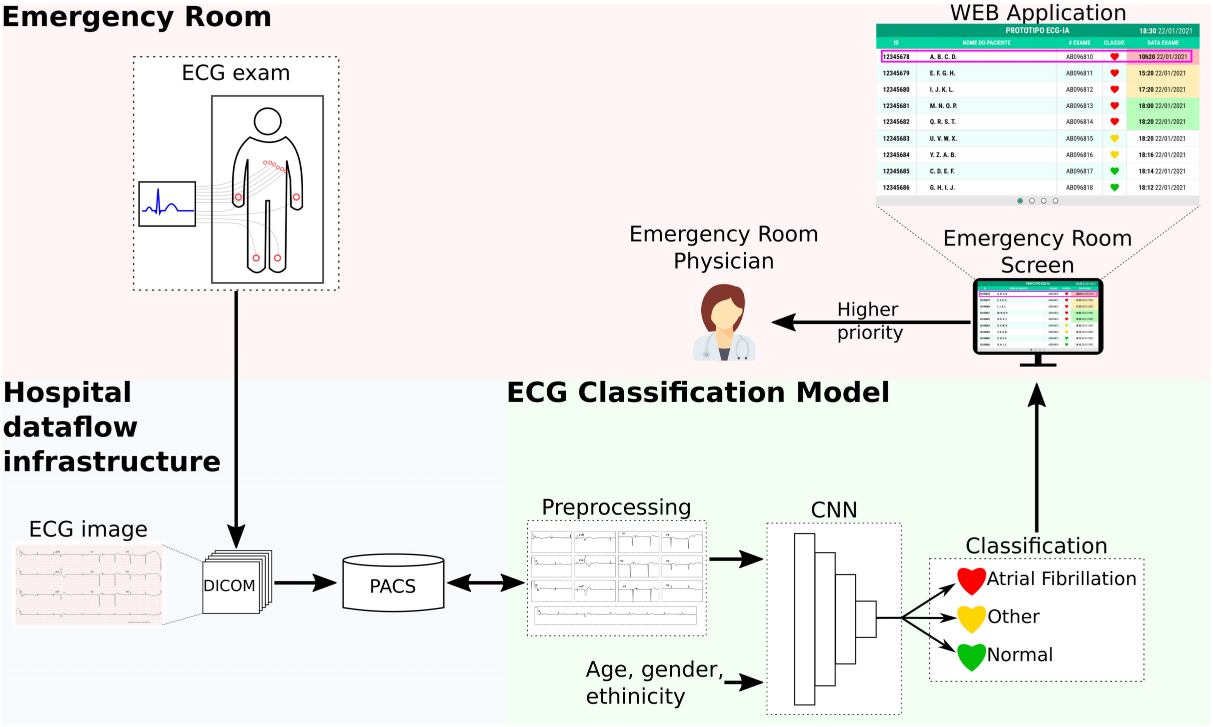
To test the network using the CPSC dataset, it was necessary to transform the one-dimensional signals into corresponding image representations. To achieve this, a MORTARA ECG image template without any signal was used as the background, onto which the signals were superimposed. This image-based representation required leads with a minimum length of 10 seconds. However, some signals in the CPSC dataset did not meet this requirement. To address this challenge, the insufficiently long signals were padded by replicating the initial segment until they reached a duration of 10 seconds.

The demographic information data also underwent preprocessing procedures. Gender information was mapped, designating male and female patients as 1 and 0, respectively. A similar approach was taken for ethnicity, assigning a value of 0 to patients identified as African American or mixed, and a value of 1 to others. To normalize age, the actual age was divided by 100.

D. PROPOSED NETWORK ARCHITECTURE

The proposed approach is a CNN with three branches: a stack of short leads (DI, DII, DIII, avR, avL, avF, V1, V2, V3, V4, V5, and V6), a long lead (10-second DII), and demographic information (age, gender, and ethnicity). Every lead is independently extracted through cropping from the preprocessed ECG images. Subsequently, the short leads are assembled into a 3D volume with dimensions of 144 × 224 × 12, which is then input into a series of 3D convolution layers. In parallel, the long lead undergoes processing through a set of 2D convolution layers. The outputs of these two branches are concatenated with the demographic information and connected to a fully connected layer, followed by a classification output layer.

In the 2D branch, we employ four consecutive convolutional blocks. Each block consists of a 2D convolution layer with 16 filters of size 3 × 3, followed by a batch normalization layer. Another 2D convolution layer with the same configuration is added, followed by another batch normalization layer, and finally, a max-pooling layer with a



PROTOTIPO ECG-IA 18:30 22/01/2021

ID	NOME DO PACIENTE	# EXAME	CLASSIF.	DATA EXAME
12345678	A. B. C. D.	AB096810	❤️	10h20 22/01/2021
12345679	E. F. G. H.	AB096811	❤️	15:20 22/01/2021
12345680	I. J. K. L.	AB096812	❤️	17:20 22/01/2021
12345681	M. N. O. P.	AB096813	❤️	18:00 22/01/2021
12345682	Q. R. S. T.	AB096814	❤️	18:20 22/01/2021
12345683	U. V. W. X.	AB096815	💛	18:20 22/01/2021
12345684	Y. Z. A. B.	AB096816	💛	18:16 22/01/2021
12345685	C. D. E. F.	AB096817	💚	18:14 22/01/2021
12345686	G. H. I. J.	AB096818	💚	18:12 22/01/2021

FIGURE 1. Diagram of the integration of the proposed method within the hospital dataflow infrastructure.

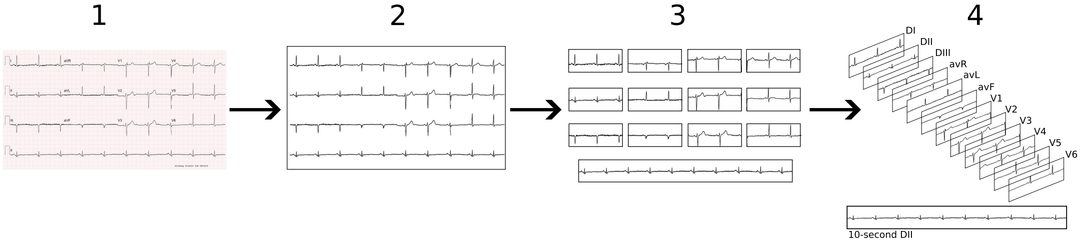


FIGURE 2. Diagram with the preprocessing steps.

pool size of 2 × 3. Two additional convolutional blocks with the same layout are stacked. However, the pool size of the max-pooling layer in these blocks is reduced to 2 × 2. This 2D branch was built upon previous work [30].

Table 3 provides an overview of the layers within the 2D branch, including the employed number of filters, kernel sizes, and the parameter count for each layer. In total, the number of parameters of this branch amounts to 26,448.

The 3D branch consists of six convolutional blocks. Each block contains a 3D convolutional layer, followed by a batch normalization layer, another 3D convolutional layer, another batch normalization layer, and finally a max-pooling layer. Each 3D convolutional layer uses 16 filters of size 3 × 3 × 3. The pool size of the max-pooling layers in the first two blocks is 2 × 2 × 2, in the third block it is 3 × 2 × 2, and in the last three blocks it is 1 × 2 × 2.

Table 4 presents the layers in the 3D branch, along with the number of filters, kernel sizes, and the corresponding number of parameters for each layer. In total, the number of parameters for this branch amounts to 77,424.

The outputs of the 2D and 3D branches are concatenated with the demographic information, and this concatenated data is subsequently fed through a dense layer featuring 16 units. Finally, a dense layer with 3 units and a sigmoid activation function is employed for classification into three classes: N, AFib, and O using a multi-label classification setup, i.e., two classes can be classified as true simultaneously. Gender, age, and ethnicity information are significant factors in clinical practice for the diagnosis of cardiovascular diseases [31]. Therefore, this information is incorporated into the network. Table 5 summarizes the proposed model, demonstrating the concatenation of the outputs from the 2D and 3D branches with the demographic information (age, gender, ethnicity). In total, our model contains 106,547 parameters. The proposed architecture for ECG classification is illustrated in Figure 3.

Our proposed deep learning model presents a novel multi-branch approach to ECG classification, offering several distinct advantages. Firstly, in the 2D branch, we leverage the DII lead as input, which is usually used for arrhythmia classification. By focusing on this lead, this branch achieves better performance in detecting and classifying arrhythmia, enhancing its diagnostic capabilities for these specific diseases. Moreover, our model incorporates a 3D branch, where we utilize a stack of all ECG leads. This integration allows us to capture inter-lead information and morphological changes occurring in multiple leads simultaneously. This improved ability to analyze inter-lead dynamics empowers our model to achieve accurate and robust predictions, making it highly suitable for diagnosing conditions that are characterized by intricate multi-lead patterns. Also, we have integrated demographic information (age, gender, and ethnicity) that have been identified in the literature as important factors in ECG classification [14]. Lastly, our model's compact size is a significant advantage compared to conventional image classification architectures, such as ResNet and VGG, which usually have millions of parameters. With just 106,547 parameters, our model ensures faster training and more efficient inference. Furthermore, the reduced parameter count mitigates the risk of overfitting, enhancing the model's generalization capabilities.

TABLE 3. Summary of the 2D branch of the proposed model.

Layer	Output size	Filters	Param
Input_2D	141x898x1	-	-
ConvBlock_1	8x11x16	CONV: [16 3x3]×2 POOL: 2x3 ×4	16,912
ConvBlock_2	2x2x16	CONV: [16 3x3]×2 POOL: 2x2 ×2	9,536
Flatten	64	-	-

TABLE 4. Summary of the 3D branch of the proposed model.

Layer	Output size	Filters	Param
Input_3D	12x144x224x1	-	-
ConvBlock_1	3x36x56x16	CONV: [16 3x3x3]×2 POOL: 2x2x2 ×2	21,488
ConvBlock_2	1x18x28x16	CONV: [16 3x3x3]×2 POOL: 3x2x2 ×1	13,984
ConvBlock_3	1x2x3x16	CONV: [16 3x3x3]×2 POOL: 1x2x2 ×3	41,952
Flatten	64	-	-

TABLE 5. Summary of the proposed CNN model.

Layer	Output size	Param
Input_2D	141x898x1	-
Branch_2D	96	26,448
Input_3D	12x144x224x1	-
Branch_3D	96	77,424
Input_demographic	3	-
Concatenate (2D, 3D, demographic)	163	-
Dense_1+ReLU	16	2,624
Dense_2+Sigmoid	3	51

E. DATA SPLIT

The InCor-DB dataset was employed both for the training and comprehensive evaluation of our proposed methodology. As an initial step, a subset comprising 20% of the dataset was isolated to serve as a test set, meticulously safeguarded from any involvement during the entirety of the training phase. This test set was solely used to evaluate the performance of our system and was not involved in any decision-making process during training. To avoid any patient overlap between different data splits, we took great care during the separation process, ensuring that exams from a patient could not appear in both the test set and any other splits. This critical step aims to prevent over-optimistic results that may not accurately represent real-world scenarios, as previous studies have highlighted [32].

For the remaining 80% of the dataset (referred to as the ''work dataset''), we employed two distinct data splitting strategies. Firstly, to train our ECG classification system for N, AFib, and O classes, we used a train-validation split, allocating 75% of the work dataset for training and 25% for validation. Secondly, to assess the significance of demographic information on the performance of the ECG classification system, we adopted an 8-fold cross-validation approach. We chose the K-fold strategy because it allows for better comparison between different CNN setups, with eight different setups being compared in this case. In both strategies, we ensured that exams from the same patient remained within the same split, guaranteeing that no patient's data was spread across multiple splits.

F. NETWORK TRAINING

We conducted the network training process in the following manner. Firstly, we pre-processed each image in our dataset as explained in Section III-C. These pre-processed images and the demographic information (age, gender, ethnicity), were stored as HDF5 and CSV files, respectively. The diagnostic reports for each ECG exam were mapped to the three classes considered in this work: N, AFib, and O. It is worth noting that an ECG exam could belong to both the AFib and O classes, making it a multi-label problem. Next, we performed our data split, as described in Section III-E, where 20% of the dataset was reserved as an internal test set, while the remaining 80% was used for training and validation of the model. The proposed network, as outlined in Section III-D, consists of three branches: a 2D branch (receiving the DII long lead), a 3D branch (receiving a

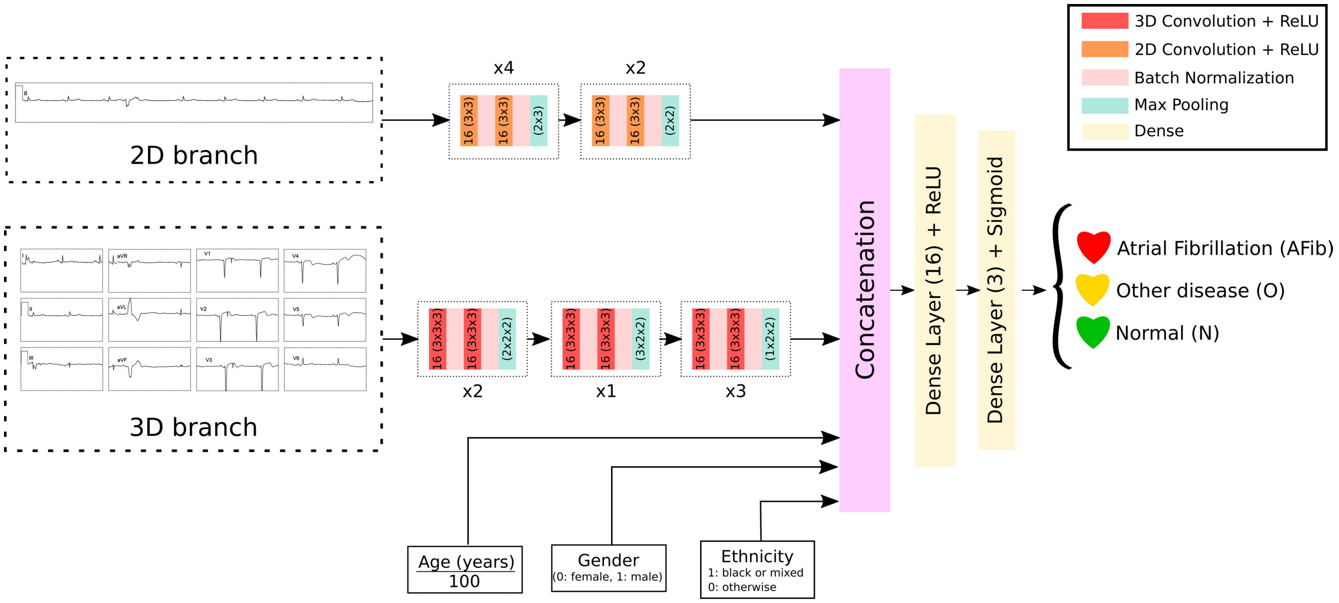


FIGURE 3. Proposed network architecture.

stack of the 12 short leads), and a demographic branch (receiving age, gender, and ethnicity information). The demographic information was also preprocessed as described in Section III-C. For the training process, we utilized the Adam optimizer [33] with a learning rate of 0.001, set 30 epochs for training and used a batch size of 64. Additionally, we implemented an early stopping callback with patience of 9, which automatically halts the training if the validation loss does not decrease within 9 epochs, thus preventing overfitting to the training dataset. Other employed hyperparameters can be found in Table 6. The model was trained using binary cross-entropy loss, and we applied weights to samples from minority classes to address the imbalanced dataset issue. Once the training was completed, we evaluated the model on the internal test set.

We built our CNN using the Keras API (version 2.4.3) with the TensorFlow backend (version 2.3.0) in Python (version 3.6.8). The training was conducted on a computer server equipped with four 16 GB V100 GPUs, 128 GB of RAM, and 16 4 GHz CPUs. The entire training dataset, approximately 4 GB in size, was directly transferred to the computer RAM through a dedicated partition. This step accelerates batch construction during training and reduces training time. Training and evaluation on the InCor-Db dataset took approximately 4 hours.

G. EXPERIMENTAL SETUP

Internal and external validation procedures were executed utilizing the InCor-DB and CPSC datasets, respectively. Furthermore, the validation of our method encompassed the utilization of 1,074 examinations acquired over a duration of one month within the ER setting of a tertiary referral hospital. Across all datasets, a classification scheme was employed, encompassing the categories of N, AFib, and O. To facilitate a comprehensive benchmarking against prior research, five essential classification metrics were adopted, including Sensitivity (Se), Specificity (Spe), F1-score (F1), Area Under the Receiving Operating Curve (AUROC), and Accuracy (Acc).

TABLE 6. Hyperparameters employed in the experiments.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1.00E-03
Loss-function	Binary Cross-Entropy
Batch Size	64
Epochs	30
Early Stopping Patience	9

In the upcoming sections, we will discuss the results from analyzing the test sets in the InCor-DB and CPSC datasets, along with evaluating our method's effectiveness in the Emergency Room of a tertiary hospital over a month. In this case, we compared our method's results with those of three experienced cardiologists, each with at least five years of experience. We also investigated the impact of including demographic variables on our model's performance.

IV. RESULTS

A. PERFORMANCE ON INCOR-DB DATASET

Using 20% of the InCor-DB dataset for testing, we achieved the following results for the AFib class: Se 94.5%, Spe 98.4%, F1 90.3%, AUROC 99.1%, and Acc 98.0%. Notably, our model exhibited a high sensitivity value for detecting AFib, providing compelling evidence for its successful applicability for screening purposes. Additionally, commendable results were obtained for the Normal and Other classes, with an AUC exceeding 90%. For a comprehensive overview of the results obtained on this dataset, please refer to Table 7. Also, in the Supplemental Material, we provide the confusion matrices and Receiving Operating Curves for the three considered classes in Figures S2 and S3, respectively.

We employed the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [34] to provide insights regarding the interpretability of our network's predictions. To do this, we applied Grad-CAM to one ECG sample from each of the three classes. Since our model has multiple inputs, we performed this interpretability analysis separately for each input branch (2D and 3D). Afterward, we combined the results into a single image for each example. The interpretability findings for ECG samples in the Normal (N), Atrial Fibrillation (AFib), and Other (O) classes can be found in Figures S4, S5, and S6 in Section IV of the Supplementary Material.

TABLE 7. Performance of ECG classification in the InCor-Db dataset test set.

	Sen	Spe	F1	AUROC	Acc
Normal	88.5	84.9	70.2	93.2	85.6
AFib	94.4	98.5	89.2	99.2	98.1
Others	81.4	90.3	88.0	93.1	83.8
Average	*88.1*	*91.2*	*82.5*	*95.2*	*89.2*

Figure 4 displays the training progress of our network, presenting the loss value and mean AUROC value over the epochs for both training and validation sets.

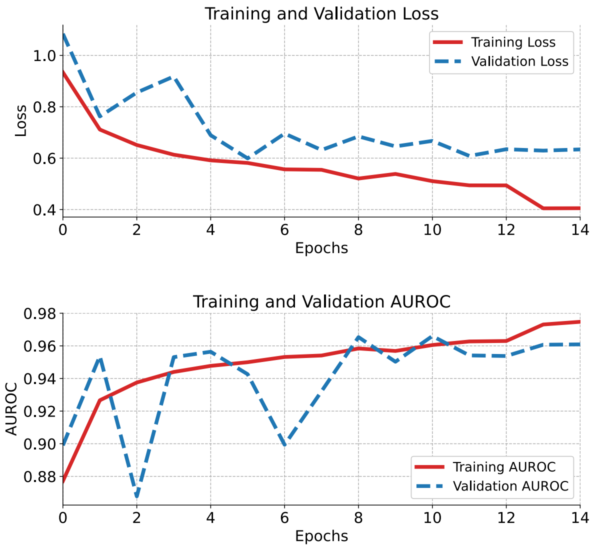


FIGURE 4. Loss and ROC obtained during training for both the train and validation datasets.

B. EXTERNAL VALIDATION ON CPSC DATASET

It is important to consider that all ECG exams in the InCor-DB dataset were captured using the same equipment, which introduces the possibility of potential bias. To demonstrate the generalizability of our results, we applied our trained network to the ECG data from the CPSC dataset. For Atrial Fibrillation, our method achieved the following performance metrics on the CPSC dataset: Se of 88.6%, Spe of 97.8%, F1 of 89.2%, AUROC of 97.5%, and Acc of 96.2%. Despite the CPSC dataset comprising patients from a different hospital, country, equipment, and using 1D signals instead, our method still achieved comparable results to our internal dataset (InCor-DB). Table 8 provides a summary of the results obtained for the CPSC dataset. The Supplemental Material provides additional results regarding the confusion matrices and Receiving Operating Curves obtained in this dataset in Figure S2 and S3, respectively.

TABLE 8. Performance of ECG classification on the external validation set in the CPSC dataset.

	Sen	Spe	F1	AUROC	Acc
Normal	92.5	77.3	54.4	91.8	79.3
AFib	88.6	97.8	89.2	97.5	96.2
Others	71.9	62.8	77.3	70.4	69.4
Average	*84.3*	*79.3*	*73.6*	*86.5*	*81.6*

Table 9 presents a comparison of our achieved results for AFib classification in the CPSC dataset with other relevant recent studies from the literature. Regardless of whether they used image or signal inputs, these works have been evaluated on the same CPSC dataset.

C. MODEL DEPLOYMENT INTO THE EMERGENCY ROOM

We also evaluated our models using the 1,074 ECG exams obtained from the ER during a one-month period. Our model's predictions were compared to the label obtained through a committee of cardiologists. The results obtained from this evaluation are presented in Table 10. Figures S2 and S3 of the Supplemental Material show, respectively, the confusion matrices and Receiving Operating Curves for the classes N, AFib, and O obtained in this dataset.

D. ANALYSIS OF DEMOGRAPHIC VARIABLES

We conducted an analysis to investigate whether the inclusion of demographic variables, namely gender, age, and ethnicity, would enhance the performance of the model. We employed an 8-fold cross-validation approach, training the model with various configurations: (0) No demographic variable; (1) All demographic variables; (2) Gender and Age; (3) Gender and Ethnicity; (4) Age and Ethnicity; (5) Gender; (6) Age; and (7) Ethnicity. However, our findings, presented in Tables 11, 12, and 13, indicate that the inclusion of these demographic variables did not improve the performance of our classification model.

V. DISCUSSION

We have successfully developed an externally validated automated diagnosis tool that accurately detects rhythm disorders from ECG images. This tool displayed strong discriminatory capabilities across various test sets, effectively discerning

TABLE 9. Comparison with the literature regarding AFib classification in the CPSC dataset.

Work	ECG Input Type	Sensitivity	Specificity	F1	AUROC	Accuracy
Gliner et al [27]	Image (2D)	-	-	-	96	-
Gliner et al [27]	Signal (1D)	94.1	99.1	95.4	98	98
Che et al [25]	Signal (1D)	-	-	85.8	98.6	-
Dong et al [26]	Signal (1D)	-	-	92.4	98.65	-
Proposed	Image (2D)	88.6	97.8	89.2	97.5	96.2

TABLE 10. Performance of ECG classification on the ER.

	Sen	Spe	F1	AUROC	Acc
Normal	86.3	97.3	84.5	98.3	95.9
AFib	88.7	96.1	82.8	98.0	95.2
Others	95.6	85.6	96.5	97.7	94.1
Average	*90.2*	*93.0*	*87.9*	*98.0*	*95.0*

between the specified classes. Furthermore, it demonstrated resilient generalization when applied to an external dataset.

The use of ECG images instead of ECG one-dimensional signals offers several advantages, particularly in hospital environments where ECG devices commonly store signals as images. This compatibility simplifies the integration of image-based methods into clinical practice, making them more practical for real-world applications. However, it is important to note that ECG images come with larger input sizes, which can lead to an increase in the computational complexity of CNNs. Additionally, to accommodate these larger inputs, ECG images often require resizing before being used as inputs to the CNN. This resizing process may introduce the risk of information loss.

Although ECG images typically require larger CNNs for accurate classification due to their size, we introduced a relatively compact network with just 106,547 parameters in this study. In contrast to other image-based ECG studies that employ multi-million parameter networks, such as VGG, our approach employs a notably smaller neural network. This provides better generalizability and faster inference times. Furthermore, the visualization of abnormal changes in various leads simultaneously by physicians plays a crucial role in identifying several diseases during ECG examinations. For example, left ventricle enlargement can be identified by specific indicators, including an elevated amplitude of the QRS complex in leads V1 and V6, among others. In an effort to replicate the methodology of physicians and investigate the interplay between leads, our proposed network architecture integrates a 3D stack comprising the short leads. This design allows for the detection of abnormalities. Furthermore, the isolated use of the 2D network on the 10-second DII lead aids in identifying irregularities in heart rhythm. Such an architecture enhances the adaptability of the proposed model, making it suitable for deployment across various ECG configurations.

The test set is designed to represent data from a variety of situations, aiming to ensure its relevance in future scenarios. Yet, many recognize that models can underperform on new datasets [35], [36]. Nonetheless, when applying automated classification systems in medical settings, they must be consistently reliable. Therefore, to validate our model in a different setting, we tested it with the CPSC dataset, derived from ECG tests in Chinese hospitals. Our model obtained good results, emphasizing its strong capability to adapt to different datasets.

The results we obtained from the CPSC dataset allow us to assess the effectiveness of our proposed approach concerning previous studies in the field. The comparison is particularly centered on the AFib class and is presented comprehensively in Table 9. Our approach, utilizing two-dimensional data (2D Images), demonstrated metrics that are slightly lower but still comparable to those achieved by [26] and [27], both of which use one-dimensional ECG signals as input. Notably, our F1-score surpasses that of [25], even though our AUROC and Acc are slightly lower. Furthermore, we conducted a comparative analysis of our method with the image-based ECG classification approach proposed by [27], and our model achieved a higher AUROC score. It's important to note that the models we compared against were trained on the CPSC dataset, whereas our model underwent exclusive external testing, without any exposure to ECG examples from the CPSC dataset during training. In summary, despite external evaluation on the CPSC dataset, our approach delivers results comparable to the literature employing one-dimensional ECG signals while also outperforming other existing image-based ECG classification approaches.

Nonetheless, while there is an extensive body of literature on ECG classification, the implementation of such systems in hospital settings to improve medical care remains limited. Despite numerous studies reporting exceptional results, the clinical validation and real-world impact on healthcare are still unknown. To address this gap, in addition to proposing a new method for ECG classification, we integrated and evaluated our methodology into the ER of a tertiary referral hospital for research purposes. The primary goal of this system is not to replace physicians or provide definitive diagnoses for patients, but rather to serve as a classification/prioritization assistant tool. Its purpose is to assist in identifying patients who require immediate care. By implementing this system, the efficiency of screening services in ERs can be significantly enhanced, given the high volume of daily patients. Given the increased cardiovascular risk associated with AFib, prioritizing patients with this

TABLE 11. Demographic variable analysis for Atrial Fibrillation class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	94.5 ± 1.2	98.5 ± 0.2	89.1 ± 1.2	99.0 ± 0.2	98.2 ± 0.2
(1) All demographic variables	95.2 ± 1.2	98.5 ± 0.2	89.3 ± 0.9	99.1 ± 0.1	98.2 ± 0.2
(2) Gender and Age	94.7 ± 1.2	98.6 ± 0.2	89.8 ± 1.3	99.0 ± 0.2	98.3 ± 0.2
(3) Gender and Ethnicity	95.3 ± 0.8	98.5 ± 0.2	89.4 ± 1.3	99.2 ± 0.2	98.2 ± 0.2
(4) Age and Ethnicity	94.8 ± 1.5	98.5 ± 0.2	89.3 ± 0.9	99.1 ± 0.2	98.2 ± 0.1
(5) Gender	94.9 ± 0.8	98.6 ± 0.1	89.6 ± 1.0	99.0 ± 0.3	98.3 ± 0.1
(6) Age	95.1 ± 1.1	98.5 ± 0.2	89.6 ± 1.5	99.2 ± 0.2	98.3 ± 0.2
(7) Ethnicity	95.2 ± 0.9	98.5 ± 0.1	89.3 ± 1.1	99.1 ± 0.1	98.2 ± 0.2

TABLE 12. Demographic variable analysis for Normal class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	84.7 ± 3.2	87.5 ± 1.2	71.5 ± 0.8	93.1 ± 0.6	86.9 ± 0.4
(1) All demographic variables	85.2 ± 1.7	87.1 ± 0.6	71.3 ± 0.4	93.2 ± 0.3	86.7 ± 0.3
(2) Gender and Age	86.4 ± 1.9	86.4 ± 0.7	71.1 ± 0.7	93.2 ± 0.5	86.4 ± 0.5
(3) Gender and Ethnicity	86.9 ± 3.1	86.3 ± 1.6	71.2 ± 0.8	93.3 ± 0.5	86.4 ± 0.8
(4) Age and Ethnicity	85.4 ± 2.3	87.2 ± 0.7	71.5 ± 0.8	93.3 ± 0.5	86.8 ± 0.2
(5) Gender	87.4 ± 2.2	86.3 ± 0.9	72.2 ± 2.1	93.5 ± 0.5	86.5 ± 0.4
(6) Age	86.6 ± 1.7	86.6 ± 0.6	71.5 ± 0.5	93.4 ± 0.3	86.7 ± 0.3
(7) Ethnicity	87.1 ± 1.1	86.3 ± 0.5	71.3 ± 0.8	93.3 ± 0.2	86.5 ± 0.3

TABLE 13. Demographic variable analysis for Others class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	84.1 ± 1.4	87.8 ± 2.2	89.2 ± 0.5	92.9 ± 0.4	85.1 ± 0.4
(1) All demographic variables	83.7 ± 0.6	88.3 ± 1.2	89.0 ± 0.3	92.9 ± 0.3	84.9 ± 0.3
(2) Gender and Age	83.2 ± 0.9	89.0 ± 1.5	88.8 ± 0.4	93.0 ± 0.3	84.7 ± 0.4
(3) Gender and Ethnicity	82.8 ± 1.9	89.5 ± 2.3	88.7 ± 0.8	93.0 ± 0.3	84.6 ± 0.9
(4) Age and Ethnicity	83.8 ± 0.7	88.5 ± 1.7	86.1 ± 0.2	93.1 ± 0.5	85.1 ± 0.2
(5) Gender	82.9 ± 0.9	89.8 ± 1.7	88.8 ± 0.3	93.2 ± 0.4	84.8 ± 0.4
(6) Age	83.2 ± 0.6	89.3 ± 1.1	88.9 ± 0.3	93.1 ± 0.3	84.9 ± 0.2
(7) Ethnicity	82.9 ± 0.6	89.6 ± 0.9	88.8 ± 0.2	93.1 ± 0.2	84.7 ± 0.3

condition in ERs is highly desirable. Therefore, we developed a system that utilizes our ECG classification methodology to provide physicians in the ER with a prioritized list of exams, giving higher priority to exams classified as AFib.

An inherent limitation of current AFib detection algorithms is their primary focus on distinguishing AFib from normal rhythms, disregarding other types of arrhythmias or cardiac conditions within different categories. In this study, we specifically defined three classes: Normal, AFib, and Others. As a result, our investigation solely revolves around identifying these three classes. We are building a curated ECG data set to expand the scope by including a larger number of classes. Furthermore, CNNs tailored for image inputs often require larger computational resources. To manage complexity, we resized ECG exam images before inputting them into our network, though this could potentially lead to data loss affecting classification. Our approach's advantages and drawbacks are summarized in Table 14, offering a comprehensive overview of our methodology.

In the domain of one-dimensional signals, the SHAP method was employed by [37] to visualize significant segments of ECG signals. This approach proved beneficial in identifying AFib and other cardiac conditions, aligning with standard ECG interpretation. Additionally, the Grad-CAM method has also been utilized for image-based ECG classification in [28]. Unfortunately, the latter study failed to establish clear connections between the interpretations and specific cardiac conditions. Given the increasing concerns surrounding the use of black-box systems in critical domains like medicine [38], [39], concerns regarding interpretability are particularly noteworthy.

We attempted to apply the Grad-CAM technique to gain insights into the decision process of our CNN using three ECG image examples, one for each class (N, AFib, and O). We present our obtained results in Section IV of the Supplemental Material (Figures S4, S5, and S6). In Figure S4, which showcases the Grad-CAM results for an ECG exam classified as class N, it is evident that the CNN primarily directs its attention to the P waves of the ECG heartbeats, with a particular focus on the DII long lead (the bottom lead). The presence or absence of the P-wave is a crucial marker for atrial fibrillation detection, suggesting that the network might be leveraging this information to rule out the possibility of AFib. However, the interpretability on the short leads does not reveal a specific pattern that can be precisely correlated with clinical information. For the ECG exam with AFib in Figure S5, it seems that the network is focusing its attention on the QRS complex, especially in the short leads. This could indicate that the network is detecting irregular rhythms, another marker of AFib. However, the interpretability on the

TABLE 14. Advantages and drawbacks of the proposed method.

Aspect	Advantages	Drawbacks
ECG images	Better integration into hospital environment	Increase in CNN size Information loss
Proposed network	Smaller compared to other image-based approaches	Generally larger than one-dimensional ECG approaches
Training with real-world ECG images	Better representation of real-world scenarios	-
Emergency Room evaluation	Provides a better grasp on the real-world performance	-
Number of classes	-	Only three classes (N, AFib, O)

DII long lead does not provide any interpretable information regarding AFib detection. In Figure S6, which represents an ECG exam diagnosed as class O, the network's attention appears to be on the QRS complex of the short leads, particularly leads V4, V5, and V6. These QRS complexes exhibit an enlarged size potentially prompting the network to pay more attention to this region, which could serve as an indicator of an abnormality. However, the DII long lead's interpretability results show sparse attention to different regions, encompassing the P-wave, QRS complex, and T-wave, making it challenging to draw any interpretable conclusion based on this lead.

It is important to note that while Grad-CAM is a popular technique for interpretability, it still does not provide a clear and straightforward interpretation of the CNN's decisions. The results of Grad-CAM require users to "interpret" them, making the process somewhat convoluted as "interpreting the interpretability technique". Moreover, since we need to configure the activation layer to evaluate in Grad-CAM, different choices may lead to distinct results and interpretations for the same ECG example. This inherent variability limits the reliability of the interpretability method. Given these limitations, we acknowledge the need for future studies on specific interpretability methods tailored towards ECG exams. Addressing these shortcomings would be beneficial to obtain more transparent and clinically meaningful insights from deep learning models applied to ECG data.

Furthermore, while incorporating demographic information into automatic ECG classification systems has been suggested [14], our findings indicate that the inclusion of these variables did not improve results for the classes examined in this study, as shown in Tables 11, 12, and 13. Nevertheless, we acknowledge that demographic variables may hold importance for new classes, and we plan to conduct further investigations in this area in future studies.

In summary, we have successfully developed a robust and versatile artificial intelligence image-based ECG classification system. This system has been seamlessly integrated into an end-to-end framework, making it readily applicable in ER settings for the screening of 12-lead ECG exams.

ETHICS STATEMENT

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

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] A. Seki and M. C. Fishbein, "Age-related cardiovascular changes and diseases," in *Cardiovascular Pathology*, L. M. Buja and J. Butany, Eds. New York, NY, USA: Academic, 2016, pp. 57–83.

[2] *Global Status Report on Noncommunicable Diseases 2014*, World Health Organization, Geneva, Switzerland, 2014.

[3] A. D. S. E. Siqueira, A. G. D. Siqueira-Filho, and M. G. P. Land, "Analysis of the economic impact of cardiovascular diseases in the last five years in Brazil," *Arquivos Brasileiros Cardiologia*, vol. 109, pp. 39–46, Jan. 2017.

[4] F. Atienza and O. Berenfeld, "Dominant frequency and the mechanisms of initiation and maintenance of atrial fibrillation," in *Cardiac Electrophysiology: From Cell to Bedside*, P. Douglas and Z. J. Jalife, Ed. Philadelphia, PA, USA: W. B. Saunders, 2014, pp. 419–432.

[5] B. J. J. M. Brundel, X. Ai, M. T. Hills, M. F. Kuipers, G. Y. H. Lip, and N. M. S. de Groot, "Atrial fibrillation," *Nature Rev. Disease Primers*, vol. 8, pp. 1–21, 2022, doi: 10.1038/s41572-022-00347-9.

[6] B. Król-Józaga, "Atrial fibrillation detection using convolutional neural networks on 2-dimensional representation of ECG signal," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103470.

[7] A. Samol, K. Hahne, and G. Mönnig, "Atrial fibrillation and silent stroke: Links, risks, and challenges," *Vascular Health Risk Manage.*, vol. 12, p. 65, Mar. 2016.

[8] E. Svennberg, J. Engdahl, F. Al-Khalili, L. Friberg, V. Frykman, and M. Rosenqvist, "Mass screening for untreated atrial fibrillation: The STROKESTOP study," *Circulation*, vol. 131, no. 25, pp. 2176–2184, Jun. 2015.

[9] P. Kirchhoff, S. Benussi, D. Kotecha, A. Ahlsson, D. Atar, and B. Casadei, "ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS: The task force for the management of atrial fibrillation of the European society of cardiology (ESC) developed with the special contribution of the European heart rhythm association (EHRA) of the ESC, endorsed by the European stroke organization (ESO)," *Eur. J. Cardio-Thoracic Surg.*, vol. 50, no. 5, pp. e1–e88, 2016.

[10] G. Hindricks, T. Potpara, N. Dagres, E. Arbelo, J. J. Bax, and C. Blomström-Lundqvist, "2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European association for cardio-thoracic surgery (EACTS): The task force for the diagnosis and management of atrial fibrillation of the European society of cardiology (ESC) developed with the special contribution of the European heart rhythm association (EHRA) of the ESC," *Eur. Heart J.*, vol. 42, no. 5, pp. 373–498, Aug. 2020.

[11] K. Harris, D. Edwards, and J. Mant, "How can we best detect atrial fibrillation?" *J. Roy. College Physicians Edinburgh*, vol. 42, no. 18, pp. 5–22, 2012.

[12] P. W. Macfarlane and J. Kennedy, "Automated ECG interpretation—A brief history from high expectations to deepest networks," *Hearts*, vol. 2, no. 4, pp. 433–448, Sep. 2021.

[13] H. Smulyan, "The computerized ECG: Friend and foe," *Amer. J. Med.*, vol. 132, no. 2, pp. 153–160, Feb. 2019.

[14] J. Schläpfer and H. Wellens, "Computer-interpreted electrocardiograms: Benefits and limitations," *J. Amer. College Cardiol.*, vol. 70, no. 9, pp. 1183–1192, 2017.

[15] R. K. Sevakula, W. M. Au-Yeung, J. P. Singh, E. K. Heist, E. M. Isselbacher, and A. A. Armoundas, "State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system," *J. Amer. Heart Assoc.*, vol. 9, no. 4, Feb. 2020, Art. no. e013924.

[16] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Rev. Cardiol.*, vol. 18, no. 7, pp. 465–478, Jul. 2021.

[17] R. Sassi, R. R. Bond, A. Cairns, D. D. Finlay, D. Guldenring, G. Libretti, L. Isola, M. Vaglio, R. Poeta, M. Campana, C. Cuccia, and F. Badilini, "PDF–ECG in clinical practice: A model for long–term preservation of digital 12–lead ECG data," *J. Electrocardiol.*, vol. 50, no. 6, pp. 776–780, Nov. 2017.

[18] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, Apr. 2020.

[19] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

[20] C. A. Caceres, C. A. Steinberg, S. Abraham, W. J. Carbery, J. M. Mcbride, W. E. Tolles, and A. E. Rikli, "Computer extraction of electrocardiographic parameters," *Circulation*, vol. 25, no. 2, pp. 356–362, Feb. 1962.

[21] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, Mar. 2001.

[22] L. B. Marinho, N. D. M. M. Nascimento, J. W. M. Souza, M. V. Gurgel, P. P. R. Filho, and V. H. C. de Albuquerque, "A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification," *Future Gener. Comput. Syst.*, vol. 97, pp. 564–577, Aug. 2019.

[23] E. H. Houssein, I. E. Ibrahim, N. Neggaz, M. Hassaballah, and Y. M. Wazery, "An efficient ECG arrhythmia classification method based on manta ray foraging optimization," *Expert Syst. Appl.*, vol. 181, Nov. 2021, Art. no. 115131.

[24] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, and E. N. Yin Kwee, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *J. Med. Imag. Health Informat.*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018.

[25] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ECG signal processing and arrhythmia classification," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, pp. 1–13, Dec. 2021.

[26] Y. Dong, M. Zhang, L. Qiu, L. Wang, and Y. Yu, "An arrhythmia classification model based on vision transformer with deformable attention," *Micromachines*, vol. 14, no. 6, p. 1155, May 2023.

[27] V. Gliner, N. Keidar, V. Makarov, A. I. Avetisyan, A. Schuster, and Y. Yaniv, "Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Oct. 2020.

[28] V. Sangha, B. J. Mortazavi, A. D. Haimovich, A. H. Ribeiro, C. A. Brandt, D. L. Jacoby, W. L. Schulz, H. M. Krumholz, A. L. P. Ribeiro, and R. Khera, "Automated multilabel diagnosis on electrocardiographic images and signals," *Nature Commun.*, vol. 13, no. 1, p. 1583, Mar. 2022.

[29] P. Hao, X. Gao, Z. Li, J. Zhang, F. Wu, and C. Bai, "Multi-branch fusion network for myocardial infarction screening from 12-lead ECG images," *Comput. Methods Programs Biomed.*, vol. 184, Feb. 2020, Art. no. 105286.

[30] F. M. Dias, N. Samesima, A. Ribeiro, R. A. Moreno, C. A. Pastore, and J. E. Krieger, "2D image-based atrial fibrillation classification," in *Proc. Comput. Cardiol. (CinC)*, vol. 48, 2021, pp. 1–4.

[31] P. W. Macfarlane, I. A. Katibi, S. T. Hamde, D. Singh, E. Clark, B. Devine, B. G. Francq, S. Lloyd, and V. Kumar, "Racial differences in the ECG—Selected aspects," *J. Electrocardiol.*, vol. 47, no. 6, pp. 809–814, Nov. 2014.

[32] E. J. D. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Comput. Methods Programs Biomed.*, vol. 127, pp. 144–164, Apr. 2016.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[35] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Syst. Appl.*, vol. 115, pp. 465–473, Jan. 2019.

[36] M. Butkuviene, A. Petrenas, A. Sološenko, A. Martín-Yebra, V. Marozas, and L. Sörnmo, "Considerations on performance evaluation of atrial fibrillation detectors," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 11, pp. 3250–3260, Nov. 2021.

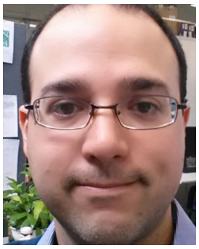
[37] D. Zhang, S. Yang, X. Yuan, and P. Zhang, "Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram," *iScience*, vol. 24, no. 4, Apr. 2021, Art. no. 102373.

[38] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[39] S. Kundu, "AI in medicine must be explainable," *Nature Med.*, vol. 27, no. 8, p. 1328, Aug. 2021.

FELIPE MENEGUITTI DIAS received the B.Sc. and M.Sc. degrees from the Federal University of Juiz de Fora (UFJF), in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in biomedical engineering with the University of São Paulo (USP), working with machine learning applications in electrocardiogram and photoplethysmogram biomedical signals. He is also a Researcher with the Heart Institute (Incor-HCFMUSP). His research interests include biomedical signal processing, machine learning, and compressive sensing.

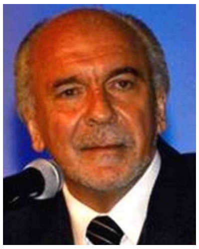
ESTELA RIBEIRO received the B.Sc. degree in mechanical engineering from the FSA University Center, São Paulo, Brazil, in 2015, and the M.Sc. and Ph.D. degrees in electrical engineering from the FEI University Center, São Paulo, in 2017 and 2020, respectively. She is currently a Researcher with the Laboratory of Biomedical Informatics, Heart Institute, Clinics Hospital, University of São Paulo Medical School. Her research interests include pattern recognition, cognitive perception, biomedical signal processing, and machine learning.



RAMON ALFREDO MORENO received the B.Sc. and Ph.D. degrees in electrical engineering from the University of São Paulo, São Paulo, Brazil, in 1998 and 2005, respectively. He is currently a Researcher with the Heart Institute (InCor), São Paulo. His current research interests include developing and implementing models for contextual visualization of medical images and open standards, such as common object request broker architecture (CORBA), digital imaging and communications in medicine (DICOM), java programming, picture archiving and communication systems (PACS), and open source software.

ADÈLE HELENA RIBEIRO received the B.Sc. degree in applied mathematics and the M.Sc. and Ph.D. degrees in computer science from the University of São Paulo, in 2011, 2014, and 2018, respectively. She was a Post-Doctoral Fellow at the Heart Institute until 2019 and she currently with Dominik Heider's Research Group, Philipps-Universität Marburg, Germany. She contributed in the beginning of this research with the development of machine learning methods. Her research interests include the development and application of machine learning and AI tools equipped with causal and counterfactual reasoning for more fair, explainable, scalable, reliable, and personalized decision-making.

NELSON SAMESIMA received the M.D. and Ph.D. degrees. He is currently a Medical Supervisor with the Resting Electrocardiography Clinical Unit, Heart Institute (InCor). His research interests include cardiac arrhythmias, electrophysiology, electrocardiography, cardiology, and surface electrocardiographic mapping/BSPM.



CARLOS ALBERTO PASTORE received the M.D. and Ph.D. degrees. He is currently a Professor with the University of São Paulo Medical School and the Director of the Electrocardiography Unit, Heart Institute (InCor). His research interests include mapping of cardiac electrical activity using multiple lead electrocardiography, high-resolution electrocardiogram, microvolt t-wave alternans (mTWA), and ECG in cardiac electrical diseases.

JOSE EDUARDO KRIEGER received the M.D. and Ph.D. degrees. He is currently a Professor in genetics and molecular medicine with the University of São Paulo Medical School and the Director of the Laboratory of Genetics & Molecular Cardiology, Heart Institute (InCor). His research interests include the genetic determinants of cardiovascular diseases to improve health management algorithms and to the development of novel therapeutics.

MARCO ANTONIO GUTIERREZ received the B.Eng. and D.Sc. degrees in electrical engineering from the University of São Paulo, Brazil, in 1985 and 1996, respectively. He has been with the Heart Institute, University of São Paulo, Brazil, since 1986, where he is currently the Head of the Biomedical Informatics Laboratory and the Informatics Division. He is also an Assistant Professor with the Polytechnic School, since 1997, and School of Medicine, since 2004, University of São Paulo. His research interests include biomedical image and signal processing and health information systems.