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

SUBMISSION DATE / POSTED DATE

10-08-2023 / 15-08-2023

CITATION

Meneguitti Dias, Felipe; Ribeiro, Estela; moreno, ramon; Ribeiro, Adele; samesima, nelson; pastore, carlos; et al. (2023). Artificial Intelligence-Driven Screening System for Rapid Classification of 12-Lead ECG Exams: A Promising Solution for Emergency Room Prioritization. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.23925210.v1

DOI

10.36227/techrxiv.23925210.v1

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

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

FELIPE M. DIAS[1,2], ESTELA RIBEIRO[1,3], RAMON A. MORENO[1], ADELE H. RIBEIRO[1], NELSON SAMESIMA[1,3], CARLOS A. PASTORE[1,3], JOSE E. KRIEGER[1,3], and MARCO A. GUTIERREZ[1,2,3]

[1]Heart Institute (InCor) – Clinics Hospital University of Sao Paulo Medical School (HCFMUSP) – Sao Paulo – SP – Brazil
[2]Polytechnique School, University of Sao Paulo (POLI USP) – Sao Paulo – SP – Brazil
[3]University of Sao Paulo Medical School (FMUSP) – Sao Paulo – SP – Brazil

Corresponding author: Marco A. Gutierrez (e-mail: marco.gutierrez@hc.fm.usp.br).

This work was supported by the Zerbini Foundation, and Foxconn Brazil, as part of the research project "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 like emergency rooms (ERs) can significantly improve the chances of survival for high-risk patients. 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 objective was to accurately classify the exams into three classes: Normal (N), Atrial Fibrillation (AFib), and Other (O). The evaluation of our method resulted in AUROC scores of 95.3%, 99.1%, and 93.3% for N, AFib, and O, respectively. To further validate our approach, we conducted evaluations using the Chinese Physiological Signal Challenge 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 achieved AUROC values of 98.3%, 98.0%, and 97.7% for the classes N, AFib, and O, respectively. Finally, we developed and deployed a system with a trained model within the ER of a tertiary hospital for research purposes. The 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 has resulted in the establishment of 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 Atrial Fibrillation, Artificial Intelligence, ECG, Emergency Room, 12-lead.

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. Moreover, ECGs are readily accessible, non-invasive, and cost-effective. Particularly in emergency units, 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.

Atrial fibrillation (AFib) is the most common form of chronic sustained cardiac arrhythmia [13]–[15], affecting nearly one percent of the global population [16]. 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 [17], [18]. Early detection and intervention of AFib, thereby preventing potential harm, can have a significant impact on healthcare outcomes and associated costs [19].

The 12-lead ECG, interpreted by a trained physician, is the definitive exam for diagnosing AFib [18], [20]. 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 50 years, there have been several attempts to develop computerized ECG interpretation methods [21]. 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, meaning 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 [22]. Additionally, the classification accuracy, especially for arrhythmias, is limited [23].

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 have shown remarkable growth in recent years [24]. These methods offer advantages by eliminating the dependence on specialist-defined features for classification. Instead, they adopt an end-to-end approach where features are automatically extracted from the ECG exam and used for classification. These algorithms have significantly improved the detection of AFib and other cardiac conditions. However, most of these systems rely on digital one-dimensional signals [25].

In hospital settings, ECG exams are typically stored as images or PDF files in the Picture Archiving and Communication System (PACS) [26]. 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 [5], [27], there is still a literature gap regarding the deployment of such methods in clinical environments, with real-time evaluation and appropriate validation. Table 1 provides a summary of recent 12-lead ECG classification methods using different deep learning strategies for classification tasks. As shown, only a minority of studies have utilized 2-D image-based strategies to classify ECG signals.

Using ECG images for diagnosis through deep learning is not a new concept [7], [9], [28]. Recently, researchers in [7] demonstrated that models using ECG images perform comparably, or even better, than those using one-dimensional signals. However, unlike our proposed approach, [9] and [7] used artificially created ECG images to train their models, while [28] used a limited dataset. Furthermore, the suitability of their approaches in hospital environments, such as collecting exams in DICOM format, extracting image and demographic information, image preprocessing, classification, and providing diagnosis feedback to clinicians, was not evaluated.

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, which restricts their applicability. Secondly, algorithms that are specific to particular equipment and training datasets further hinder generalizability. Lastly, the current research landscape often prioritizes improving machine learning model performance, while neglecting crucial considerations of practical applicability.

In this study, we propose 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. We also evaluate the inclusion of demographic information (age, gender, and ethnicity) in the classification system [23]. The classification system considers three classes: Normal (N), Atrial Fibrillation (AFib), and Other cardiac condition (O). To demonstrate the feasibility of our approach in clinical settings, we have developed a screening system specifically designed for implementation in emergency rooms. This system is integrated into the PACS of a tertiary referral hospital, enabling the automatic detection of newly acquired ECG exams within the emergency room. The ECG image exams, along with relevant demographic information, are then 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 emergency room.

II. METHODS

A. DATA SOURCE

We utilized 12-lead ECG exams collected from 2017 to 2020 from the Picture Archiving and Communication System (PACS) of a specialized tertiary referral hospital in Brazil that focuses on cardiology. The exams were obtained from MORTARA™ ELI 250c machines, which digitally captured the ECG signals and transmitted them to the hospital's PACS via a gateway. This gateway automatically converted the

TABLE 1. Summary of recent 12-leads ECG classification methods.

Authors	Dataset	Signal Type	Method
Leur et al. (2020) [4]	Private	1-D	ResNet
Ribeiro et al. (2020) [5]	UFMG-Code [5]	1-D	ResNet
Baek et al. (2021) [6]	Private	1-D	RNN
Sangha et al. (2022) [7]	UFMG-Code [5]	1-D	ResNet
Sangha et al. (2022) [7]	UFMG-Code [5], PTB-XL [8]	2-D	ResNet
Gliner et al. (2020) [9]	CPSC2018 [10]	1-D	CNN
Gliner et al. (2020) [9]	CPSC2018 [10]	2-D	CNN
Vranken et al. (2021) [11]	Private, CPSC2018 [10]	1-D	ResNet
Zhang et al. (2021) [12]	CPSC2018 [10]	1-D	CNN

signals into 2D images in Digital Imaging and Communication in Medicine (DICOM) format. The resulting image was presented as an A4-format chart, incorporating a reference grid with a resolution of 25 mm/s for the time axis and 10 mm/mV for the voltage axis. Subsequently, these images were converted to Portable Network Graphics (PNG) format and subjected to an automated cropping process to eliminate any private information at the top of the image.

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: Normal (N), Atrial Fibrillation (AFib), and ECG abnormalities (O). To build the dataset, we incorporated patient demographic information such as 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 neural network learning process was not influenced by diagnostic uncertainty. After applying these exclusion criteria, the final dataset, referred to as InCor-DB, consisted of 99,746 ECG exams from 64,192 unique patients. It included anonymized 2D image ECG exams, their diagnostic reports (N, O, and AFib), and de-identified patient demographic information (age, gender, and ethnicity). This private dataset complied with all relevant ethical regulations and was approved by the Internal Review Board (IRB) under registration CAAE 45070821.3.0000.0068.

To prevent intra-subject bias, we took meticulous measures to ensure that identical subjects were not included in both the training and test sets. When the training and test sets involved the same subject (intra-subject paradigm), the model captured subject-specific heartbeat characteristics, potentially enhancing test performance due to data leakage.

To demonstrate the generalizability of our proposed method, we tested our model on an external database, the China Physiological Signal Challenge (CPSC). This dataset comprised 6,877 12-lead one-dimensional ECG signals with durations ranging from 6 to 60 seconds. The signals were classified into nine different classes: Normal, Atrial Fibrillation, First-degree atrioventricular block, Left bundle branch block, Right bundle branch block, Premature atrial contraction, Premature ventricular contraction, ST-segment depression, and ST-segment elevation. Similar to InCor-DB, these nine classes were clustered into the N, O, and AFib categories, and patients under 18 years old were excluded. More information about the CPSC dataset can be found in [10]. Table 2 summarizes the datasets used in our study.

B. DEPLOYMENT OF THE MODEL IN THE EMERGENCY UNIT

Furthermore, we deployed our model in the Emergency Unit of a tertiary referral hospital system for research purposes. We developed a system that evaluates every new ECG exam captured in the emergency room and provides a prioritized list of exams to the physicians. Exams classified as atrial fibrillation have higher priority.

Each MORTARA™ ELI equipment in the hospital is connected to a gateway that converts the ECG data from proprietary to DICOM format. This gateway sends each DICOM file to the hospital's PACS. To deploy our model, we first developed a service that processes each new ECG exam from the emergency unit 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 emergency unit. The web page is displayed on a monitor in the emergency room. 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 illustrates the pipeline for deploying our model in a hospital setting and how it integrates within the hospital dataflow infrastructure.

C. DATA PREPROCESSING

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 objective during the preprocessing stage was to optimize the efficiency of our neural network by removing all extraneous and nonessential

TABLE 2. Patients demographic information of our employed datasets.

		InCor-DB	CPSC	Emergency Room
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%)

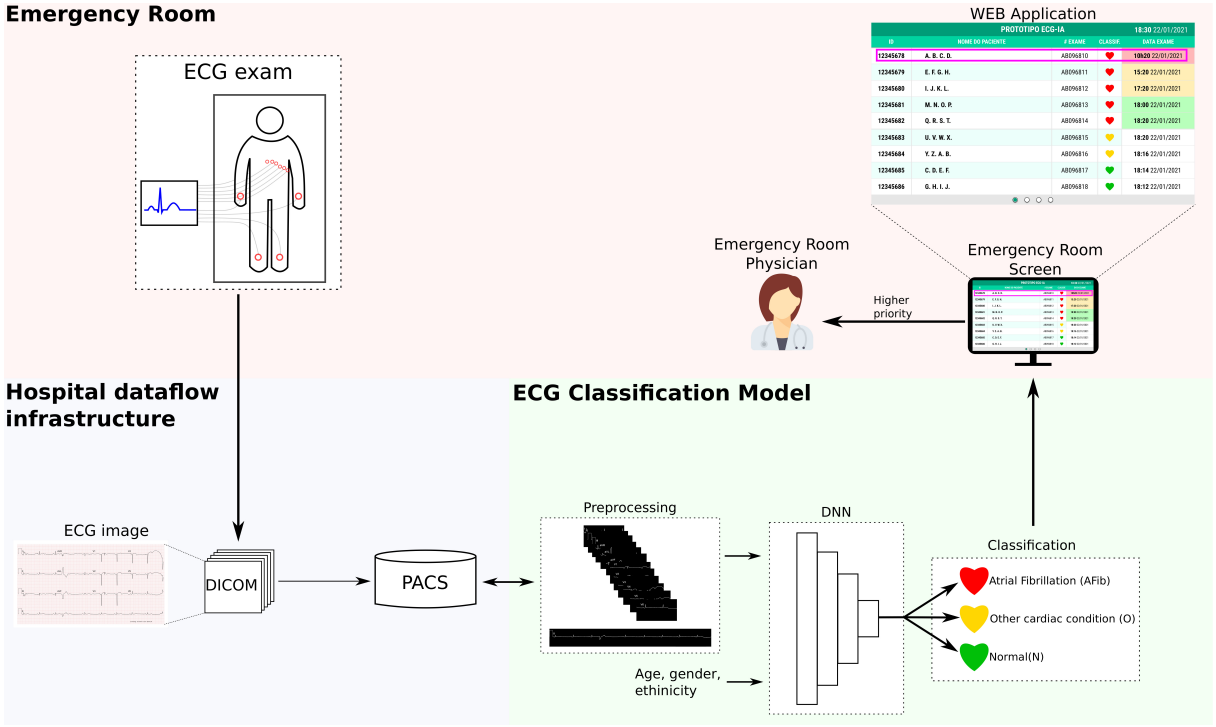


FIGURE 1. Diagram of the integration of the proposed method within the hospital dataflow infrastructure.

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 salt and pepper noise. To eliminate the noise, a morphological erosion operation followed by dilation was performed. Afterward, each lead, including the 10-second DII lead, was separated individually. To decrease the computational complexity, the images were resized to 30% of their original size. As a result, the short lead images (DI, DII, DIII, avR, avL, avF, V1, V2, V3, V4, V5, and V6) became 144 x 224 pixels, while the long lead image (10-second DII) became 141 x 898 pixels. The short lead images were stacked to create a 3D volume, which, along with the long lead image, served as the input for our proposed neural network architecture. Figure 2 illustrates the preprocessing steps involved in our methodology.

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 Afro-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 deep neural network with three inputs: 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). Each lead is individually cropped from preprocessed ECG

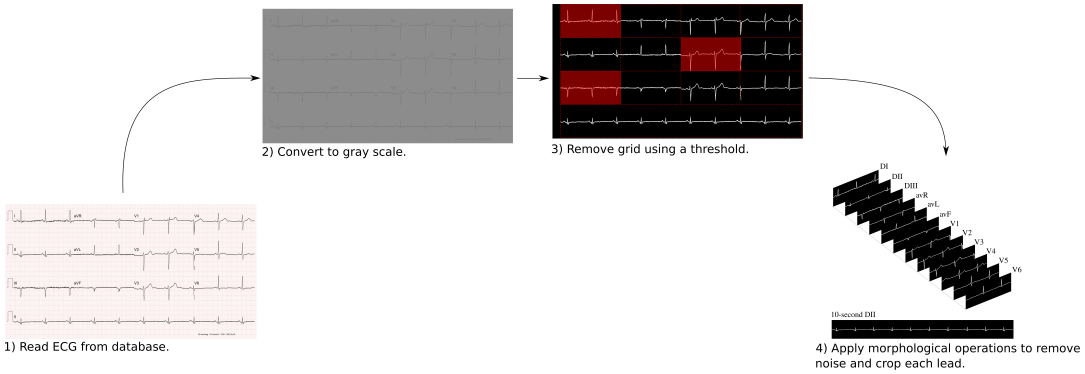


FIGURE 2. Diagram with the preprocessing steps.

images. The short leads are then packed into a 3D volume of size 144 x 224 x 12 and fed into a stack of 3D convolutions. Simultaneously, the long lead is processed by a stack of 2D convolutions. 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 3x3, 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 pool size of 2x3. 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 2x2.

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 3x3x3. The pool size of the max-pooling layers in the first two blocks is 2x2x2, in the third block it is 3x2x2, and in the last three blocks it is 1x2x2.

The outputs of the 2D and 3D branches are concatenated with the demographic information and then passed through a dense layer with 16 units. Finally, a dense layer with 3 units and a sigmoid activation function is employed for classification into three classes: Normal (N), Atrial Fibrillation (AF), and Other diseases (O). Gender, age, and ethnicity information are significant factors in clinical practice for the diagnosis of cardiovascular diseases [29]. Therefore, this information is incorporated into the network. The proposed architecture for ECG classification is illustrated in Figure 3.

E. NETWORK TRAINING

We built our neural network using the Keras API (version 2.4.3) with the TensorFlow backend (version 2.3.0) in Python (version 3.6.8). For training, we utilized the Adam optimizer with default parameters to minimize the cross-entropy. A batch size of 64 and a maximum of 100 epochs were employed.

To prevent overfitting, we implemented an early stopping callback with a patience of seven epochs. This means that if the model does not improve on the validation dataset for seven consecutive epochs, training is stopped. 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 the */dev/shm* partition. This step accelerates batch construction during training and reduces training time. Training and evaluation on the InCor-Db dataset took approximately 4 hours.

III. RESULTS

A. EXPERIMENTAL SETUP

We used the InCor-DB dataset, acquired retrospectively from a tertiary referral hospital, which consists of 99,746 ambulatory exams from 64,192 different patients. External validation was performed using the CPSC dataset. Additionally, our method was validated using 1,074 exams captured during a one-month period in the emergency room of a tertiary referral hospital. For all datasets, we considered three classes for classification: Normal (N), Atrial Fibrillation (AFib), and Other cardiac conditions (O). To train the proposed network, we divided the InCor-DB into training, validation, and testing sets, with proportions of 60%, 20%, and 20% respectively. This division followed an inter-patient protocol, ensuring that exams of a patient did not appear in different splits. In order to compare our results with the existing literature, we employed five classification metrics Sensitivity (Se), Specificity (Spe), F1-score (F1), Area Under the Receiving Operating Curve (AUROC), and Accuracy (Acc). In the upcoming sections, we will present the results obtained for the testing set of the InCor-DB, the CPSC dataset, and the evaluation of our deployed method in the emergency unit of a tertiary hospital over a one-month period. In the latter case, we compared the outcomes obtained by our proposed method against those of a committee comprising three certified cardiologists with a minimum of 5 years of experience. Furthermore, we

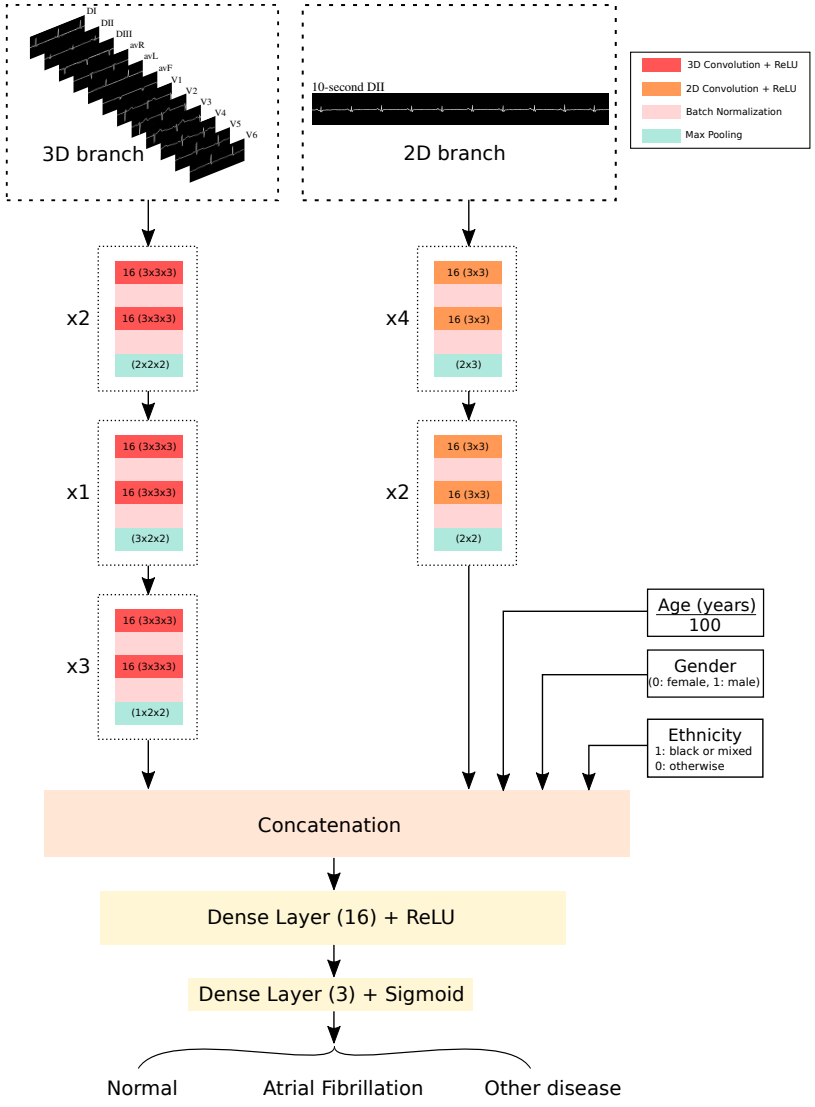


FIGURE 3. Proposed network architecture.

investigated whether the inclusion of demographic variables could enhance the performance of our model.

B. PERFORMANCE ON INCOR-DB DATASET

Using 20% of the InCor-DB dataset for testing, we achieved the following results for the AFib class: 94.5% sensitivity (Se), 98.4% specificity (Spe), 90.3% F1-score (F1), 99.1% Area Under the Receiving Operating Curve (AUROC), and 98.0% accuracy (Acc). Notably, our model exhibited a high sensitivity value for detecting AFib, providing compelling evidence for its successful applicability in 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 3.

TABLE 3. Performance of ECG classification in the InCor-Db dataset test set.

	Sen	Spe	F1	AUROC	Acc
Normal	86.4	90.8	81.0	95.3	89.7
AFib	94.5	98.4	90.3	99.1	98.0
Others	89.3	84.5	91.5	93.3	88.0
Average	*90.0*	*91.2*	*87.6*	*95.9*	*91.9*

C. EXTERNAL VALIDATION ON CPSC DATASET

The division of the training, validation, and test sets within the InCor-DB dataset was performed randomly to ensure that patients did not overlap across different sets. However, it is important to consider that all ECG exams in this 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: sensitivity (Se) of 88.6%, specificity (Spe) of 97.8%, F1-score of 89.2%, area under the receiver operating characteristic curve (AUROC) of 97.5%, and accuracy (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 4 provides a summary of the results obtained for the CPSC dataset.

TABLE 4. 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*

D. DEPLOYED MODEL INTO THE EMERGENCY UNIT

To validate our model in the emergency unit, we followed a specific procedure. Initially, 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 ensure accuracy, 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 one of the following classes: N, AFib, or O. In cases where the two cardiologists disagreed on the classification 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 emergency unit, 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. Throughout the one-month evaluation, a total of 1,074 valid exams were collected. The results obtained from this evaluation are presented in Table 5, showcasing the outcomes observed in the emergency unit.

TABLE 5. Performance of ECG classification on the Emergency Unit.

	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*

E. 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 6, 7, and 8, indicate that the inclusion of these demographic variables did not improve the performance of our classification model.

IV. DISCUSSION

We have successfully developed an externally validated automated diagnosis tool capable of accurately identifying rhythm disorders from ECG images. The tool demonstrated high discriminatory power across multiple test sets, effectively distinguishing between the proposed classes. Moreover, it exhibited robust generalization across an external dataset.

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 our methodology into the emergency unit of a tertiary referral hospital for research purposes. The primary objective of this system is not to replace physicians or provide definitive diagnoses for patients, but rather to serve as a classification/prioritization tool. Its purpose is to identify patients requiring immediate care and initiate appropriate diagnostic measures. By implementing this system, the efficiency of screening services in Emergency Units can be significantly enhanced, given the high volume of daily patients. Additionally, it can be utilized to identify problematic exams, such as those involving lead swaps.

Given the increased cardiovascular risk associated with atrial fibrillation, prioritizing patients with this condition in emergency units is highly desirable. Therefore, we developed a system that utilizes our ECG classification methodology to provide physicians in the emergency unit with a prioritized list of exams, giving higher priority to exams classified as atrial fibrillation.

In clinical settings, the availability of one-dimensional signals may be limited, as ECG devices commonly store signals as images. Consequently, the significance of image-based ECG classification systems is widely recognized. Our model exhibited exceptional performance on a validation dataset obtained from an emergency unit, which was subsequently validated by cardiologists.

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 is characterized by an increased amplitude of the QRS complex in leads V1 and V6, among other indicators. To emulate the physician's approach and explore the interdependence between leads, our proposed network architecture incorporates a 3D stack of short leads, facilitating the detection of abnormalities. Additionally, the separate utilization of the 2D network on the 10-second DII lead enables the identification of irregularities in heartbeat

TABLE 6. Demographic variable analysis for Atrial Fibrillation class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	94.532 ± 1.158	98.521 ± 0.212	89.088 ± 1.174	99.050 ± 0.239	98.213 ± 0.164
(1) All demographic variables	95.172 ± 1.180	98.488 ± 0.183	89.300 ± 0.873	99.088 ± 0.083	98.255 ± 0.175
(2) Gender and Age	94.785 ± 1.176	98.631 ± 0.203	89.775 ± 1.283	99.038 ± 0.213	98.325 ± 0.183
(3) Gender and Ethnicity	95.318 ± 0.835	98.484 ± 0.232	89.363 ± 1.316	99.163 ± 0.192	98.238 ± 0.200
(4) Age and Ethnicity	94.797 ± 1.521	98.533 ± 0.198	89.313 ± 0.890	99.100 ± 0.169	98.238 ± 0.141
(5) Gender	94.940 ± 0.845	98.572 ± 0.120	89.575 ± 1.042	99.025 ± 0.260	98.288 ± 0.141
(6) Age	95.154 ± 1.107	98.545 ± 0.237	89.575 ± 1.491	99.175 ± 0.175	98.275 ± 0.225
(7) Ethnicity	95.185 ± 0.936	98.492 ± 0.140	89.313 ± 1.093	99.113 ± 0.125	98.225 ± 0.175

TABLE 7. Demographic variable analysis for Normal class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	84.707 ± 3.195	87.458 ± 1.198	71.525 ± 0.822	93.113 ± 0.569	86.938 ± 0.444
(1) All demographic variables	85.207 ± 1.697	87.093 ± 0.634	71.300 ± 0.355	93.188 ± 0.327	86.725 ± 0.282
(2) Gender and Age	86.411 ± 1.923	86.397 ± 0.724	71.100 ± 0.727	93.238 ± 0.453	86.413 ± 0.458
(3) Gender and Ethnicity	86.917 ± 3.149	86.318 ± 1.623	71.250 ± 0.758	93.275 ± 0.480	86.425 ± 0.789
(4) Age and Ethnicity	85.415 ± 2.297	87.197 ± 0.655	71.538 ± 0.761	93.350 ± 0.518	86.850 ± 0.220
(5) Gender	87.436 ± 2.247	86.267 ± 0.859	72.225 ± 2.081	93.463 ± 0.518	86.500 ± 0.447
(6) Age	86.601 ± 1.669	86.654 ± 0.605	71.525 ± 0.526	93.375 ± 0.276	86.663 ± 0.283
(7) Ethnicity	87.097 ± 1.072	86.309 ± 0.514	71.338 ± 0.825	93.338 ± 0.220	86.463 ± 0.297

TABLE 8. Demographic variable analysis for Others class.

Configuration	Sensitivity	Specificity	F1-score	AUROC	Accuracy
(0) No demographic variable	84.137 ± 1.368	87.785 ± 2.206	89.188 ± 0.503	92.900 ± 0.407	85.150 ± 0.450
(1) All demographic variables	83.696 ± 0.625	88.347 ± 1.184	89.013 ± 0.290	92.938 ± 0.311	84.950 ± 0.307
(2) Gender and Age	83.158 ± 0.922	89.025 ± 1.510	88.825 ± 0.369	93.050 ± 0.351	84.725 ± 0.406
(3) Gender and Ethnicity	82.848 ± 1.939	89.522 ± 2.271	88.713 ± 0.788	93.038 ± 0.350	84.650 ± 0.886
(4) Age and Ethnicity	83.834 ± 0.685	88.496 ± 1.716	86.125 ± 0.212	93.138 ± 0.507	85.100 ± 0.177
(5) Gender	82.943 ± 0.865	89.774 ± 1.706	88.838 ± 0.346	93.175 ± 0.413	84.788 ± 0.356
(6) Age	83.239 ± 0.579	89.279 ± 1.109	88.900 ± 0.283	93.100 ± 0.262	84.875 ± 0.243
(7) Ethnicity	82.887 ± 0.559	89.612 ± 0.864	88.775 ± 0.249	93.075 ± 0.243	84.725 ± 0.260

rhythm. This architecture enhances the versatility of the proposed model, enabling its application across different ECG configurations

The assumption underlying the test set is that it adequately represents the data encountered in various contexts, thereby enabling future generalizability. However, when deploying the system, there is a risk that clinicians may overestimate the model's accuracy, potentially leading to patient harm if the proposed system fails to exhibit robust generalization. It is well-known that performance usually drops when models are tested on other datasets [30], [31]. Despite the CPSC dataset comprising data collected from hospitals in China, our model still achieved favorable results. This successful performance on a distinct dataset demonstrates the model's capacity to generalize effectively.

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. Moreover, it is important to acknowledge that patterns learned from datasets employing complex machine learning algorithms may not inherently convey precise and easily understandable knowledge. In the context of AFib classification, it becomes crucial to determine whether the prediction of "AFib" is linked to relevant clinical information rather than unrelated characteristics that happen to correlate with the predicted class.

In the field of one-dimensional signals, the SHAP method was utilized by [12] to visualize significant segments of ECG signals. This approach proved helpful in identifying AFib and other cardiac conditions, aligning with standard ECG interpretation and the Grad-CAM method for image-based ECG classification has also been used [7]. 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 [32], [33], future research will focus on enhancing the interpretability of ECG classification models. Furthermore, while incorporating demographic information into automatic ECG classification systems has been suggested [23], our findings indicate that the inclusion of these variables did not improve results for the classes examined in this study (Atrial Fibrillation, Normal, and Other), as shown in Tables 6, 7, and 8. 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 conclusion, we have successfully developed a robust

and versatile artificial intelligence image-based ECG classification system. This system has been integrated into an end-to-end framework, enabling its utilization in emergency room settings for screening 12-lead ECG exams.

ACKNOWLEDGEMENTS

This study was financially supported by the Foxconn Brazil and the Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

ETHICS STATEMENT

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

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] A. Seki and M.C. Fishbein. Chapter 2 - age-related cardiovascular changes and diseases. In L. Maximilian Buja and Jagdish Butany, editors, Cardiovascular Pathology (Fourth Edition), pages 57–83. Academic Press, San Diego, fourth edition edition, 2016.

[2] World Health Organization et al. Global status report on noncommunicable diseases 2014. Number WHO/NMH/NVI/15.1. World Health Organization, 2014.

[3] Alessandra de Sá Earp Siqueira, Aristarco Gonçalves de Siqueira-Filho, and Marcelo Gerardin Poirot Land. Analysis of the economic impact of cardiovascular diseases in the last five years in brazil. Arquivos brasileiros de cardiologia, 109:39–46, 2017.

[4] Rutger R van de Leur, Lennart J Blom, Efstratios Gavves, Irene E Hof, Jeroen F van der Heijden, Nick C Clappers, Pieter A Doevendans, Rutger J Hassink, and René van Es. Automatic triage of 12-lead ecgs using deep convolutional neural networks. Journal of the American Heart Association, 9(10):e015138, 2020.

[5] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. Nature communications, 11(1):1–9, 2020.

[6] Yong-Soo Baek, Sang-Chul Lee, Wonik Choi, and Dae-Hyeok Kim. A new deep learning algorithm of 12-lead electrocardiogram for identifying atrial fibrillation during sinus rhythm. Scientific Reports, 11:12818, 2021.

[7] Veer Sangha, Bobak J. Mortazavi, Adrian D. Haimovich, Antônio H. Ribeiro, Cynthia A. Brandt, Daniel L. Jacoby, Wade L. Schulz, Harlan M. Krumholz, Antonio Luiz P. Ribeiro, and Rohan Khera. Automated multilabel diagnosis on electrocardiographic images and signals. Nature Communications, 13(1583), 2022.

[8] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. Scientific data, 7(1):1–15, 2020.

[9] Vadim Gliner, Noam Keidar, Vladimir Makarov, Arutyun I Avetisyan, Assaf Schuster, and Yael Yaniv. Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms. Scientific Reports, 10(1):1–12, 2020.

[10] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. Journal of Medical Imaging and Health Informatics, 8(7):1368–1373, 2018.

[11] Jeroen F Vranken, Rutger R van de Leur, Deepak K Gupta, Luis E Juarez Orozco, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Sadaf Gulshad, and René van Es. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. European Heart Journal - Digital Health, 2(3):401–415, 05 2021.

[12] Dongdong Zhang, Samuel Yang, Xiaohui Yuan, and Ping Zhang. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. iScience, 24(4):102373, 2021.

[13] Felipe Atienza and Omer Berenfeld. 42 - dominant frequency and the mechanisms of initiation and maintenance of atrial fibrillation. In Douglas P. Zipes and José Jalife, editors, Cardiac Electrophysiology: From Cell to Bedside (Sixth Edition), pages 419–432. W.B. Saunders, Philadelphia, sixth edition edition, 2014.

[14] Bianca J. J. M. Brundel, Xun Ai, Mellanie True Hills, Myrthe F. Kuipers, Gregory Y. H. Lip, and Natasja M. S. de Groot. Atrial fibrillation. Nature Reviews Disease Primers, 21(8), 2022.

[15] Bartłomiej Król-Józaga. Atrial fibrillation detection using convolutional neural networks on 2-dimensional representation of ecg signal. Biomedical Signal Processing and Control, 74:103470, 2022.

[16] Kathrin Hahne, Gerold Mönnig, and Alexander Samol. Atrial fibrillation and silent stroke: links, risks, and challenges. Vascular Health and Risk Management, 12:65, 2016.

[17] Emma Svennberg, Johan Engdahl, Faris Al-Khalili, Leif Friberg, Viveka Frykman, and Mårten Rosenqvist. Mass screening for untreated atrial fibrillation: the strokestop study. Circulation, 131(25):2176–2184, 2015.

[18] P Kirchhoff, S Benussi, D Kotecha, A Ahlsson, D Atar, B Casadei, et al. 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 Cardiothorac Surg, 50(5):e1–e88, 2016.

[19] Gerhard Hindricks, Tatjana Potpara, Nikolaos Dagres, Elena Arbelo, Jeroen J Bax, Carina Blomström-Lundqvist, Giuseppe Boriani, Manuel Castella, Gheorghe-Andrei Dan, Polychronis E Dilaveris, Laurent Fauchier, Gerasimos Filippatos, Jonathan M Kalman, Mark La Meir, Deirdre A Lane, Jean-Pierre Lebeau, Maddalena Lettino, Gregory Y H Lip, Fausto J Pinto, G Neil Thomas, Marco Valgimigli, Isabelle C Van Gelder, Bart P Van Putte, Caroline L Watkins, and ESC Scientific Document Group. 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. European Heart Journal, 42(5):373–498, 08 2020.

[20] K Harris, D Edwards, and J Mant. How can we best detect atrial fibrillation? The journal of the Royal College of Physicians of Edinburgh, 42:5–22, 2012.

[21] Peter W Macfarlane and Julie Kennedy. Automated ecg interpretation—a brief history from high expectations to deepest networks. Hearts, 2(4):433–448, 2021.

[22] Harold Smulyan. The computerized ecg: friend and foe. The American journal of medicine, 132(2):153–160, 2019.

[23] Jürg Schläpfer and Hein J Wellens. Computer-interpreted electrocardiograms: benefits and limitations. Journal of the American College of Cardiology, 70(9):1183–1192, 2017.

[24] Rahul Kumar Sevakula, Wan-Tai M Au-Yeung, Jagmeet P Singh, E Kevin Heist, Eric M Isselbacher, and Antonis A Armoundas. State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. Journal of the American Heart Association, 9(4):e013924, 2020.

[25] Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology, 18(7):465–478, 2021.

[26] Roberto Sassi, Raymond R Bond, Andrew Cairns, Dewar D Finlay, Daniel Guldenring, Guido Libretti, Lamberto Isola, Martino Vaglio, Roberto Poeta, Marco Campana, et al. Pdf–ecg in clinical practice: A model for long–term preservation of digital 12–lead ecg data. Journal of electrocardiology, 50(6):776–780, 2017.

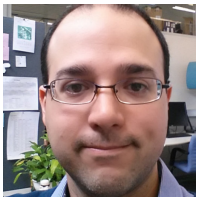
[27] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine, 25(1):65–69, 2019.

[28] Pengyi Hao, Xiang Gao, Zhihe Li, Jinglin Zhang, Fuli Wu, and Cong Bai. Multi-branch fusion network for myocardial infarction screening from 12-lead ecg images. Computer methods and programs in biomedicine, 184:105286, 2020.

[29] PW Macfarlane, IA Katibi, ST Hamde, D Singh, E Clark, B Devine, BG Francq, S Lloyd, and V Kumar. Racial differences in the ecg—selected aspects. Journal of electrocardiology, 47(6):809–814, 2014.
[30] Rasmus S. Andersen, Abdolrahman Peimankar, and Sadasivan Puthusserypady. A deep learning approach for real-time detection of atrial fibrillation. Expert Systems with Applications, 115:465–473, 2019.
[31] Monika Butkuvienė, Andrius Petrėnas, Andrius Sološenko, Alba Martín-Yebra, Vaidotas Marozas, and Leif Sörnmo. Considerations on performance evaluation of atrial fibrillation detectors. IEEE Transactions on Biomedical Engineering, 68(11):3250–3260, 2021.
[32] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215, 2019.
[33] Shinjini Kundu. AI in medicine must be explainable. Nature Medicine, 27(8):1328–1328, 2021.

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

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



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 at the Heart Institute (InCor), Sao Paulo, Brazil. His current research interests include developing and implementing models for contextual visualization of medical images, open standards such as common object request broker architecture (CORBA), and digital imaging and communications in medicine (DICOM), Java programming, picture archiving and communication systems (PACS), and open source software.

ÀDELE HELENA RIBEIRO received the B.Sc. in Applied Mathematics from the University of Sao Paulo in 2011, where she also obtained M.Sc and Ph.D. in Computer Science in 2014 and 2018, respectively. Her research interests are focused on 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. She is currently a postdoc in Dominik Heider's research group at Philipps-Universität Marburg, Germany.

NELSON SAMESIMA , MD, PhD, is a medical supervisor of the Resting Electrocardiography Clinical Unit at the Heart Institute (InCor). His research interests include cardiac arrhythmias, electrophysiology, electrocardiography, cardiology, and surface electrocardiographic mapping/BSPM.

CARLOS ALBERTO PASTORE , MD, PhD, is a Professor at the University of Sao Paulo Medical School and Director of the Electrocardiography Unit at the 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 , MD, PhD, is Professor of Genetics and Molecular Medicine at the University of Sao Paulo Medical School and Director of the Laboratory of Genetics & Molecular Cardiology at the Heart Institute (InCor). His research interests are focused on the genetic determinants of cardiovascular diseases to improve health management algorithms and to the development of novel therapeutics.