On the impact of ECG data quality for arrhythmia detection using convolutional neural networks and wearable devices

Juan Manuel Sancho

September 13, 2021

Abstract

Cardiovascular diseases are the leading cause of death in the world, accounting for 16% of global deaths, and arrhythmias are usually a symptom and a risk factor. The progress of technologies such as low-cost and low-power sensors, computer networks, microcontrollers, and Internet of Things devices in general, provides many opportunities for mitigating this problem. In this regard, wearable devices can aid in better monitoring, treatment, and medical follow up, potentially preventing life-threatening heart disease; however, these devices usually present limitations on the data they report, such as lower sampling rates, bit depth, and amount of leads recorded, in order to save up on building and energy consumption costs. In combination with Machine Learning techniques, this data can bring important advances in arrhythmia detection. Wearable devices' limitations must be taken into account when working with Machine Learning models, since the performance of said models depends on the quality of the data used to build them. This work studies the impact of data quality reduction on the performance of arrhythmia classification using state-of-the-art deep neural network classification models. Furthermore, a new model that compensates performance loss due to data quality reduction is proposed. In order to achieve this, a dataset with signals at a 500 Hz sampling rate and 32-bit resolution is downgraded to better match the parameters of wearable devices, and the performance on the original and downgraded datasets is compared on classifying 4 types of heart rhythms. The performance on a per-lead and per-class basis is also studied in order to better determine ways to increase overall performance. Results show that indeed the difference in data quality decreases the performance of the model, going from an original accuracy of 95.3% to 93.9% at 100 Hz and 8-bits on a single lead. This accuracy can be increased to 95.4% by combining as little as 2 leads through a majority voter model, and up to 95.8% by using the same model on 4 leads.

Keywords: cardiology, arrhythmia, deep learning, telemedicine.

1 Introduction

According to the World Health Organization (2020), cardiovascular disease is the leading cause of death worldwide; currently, these diseases account for a 16% of deaths globally. Within the spectrum of possible cardiac issues, arrhythmias contribute as a risk factor (Khairy et al., 2006). They introduce a bigger complexity at the time of diagnosis, since many of them do not generate symptoms in its patients (Boriani et al., 2015). Arrhythmias can also be an issue in Covid-19 patients, producing further complications and becoming a risk factor (Babapoor-Farrokhran et al., 2020).

Particularly, the most common type of arrhythmia (Atrial Fibrillation) affects less than 1% of people younger than 60 years old, while it is present in approximately 10% of those older than 60 (Centers for Disease Control and Prevention, 1999). Besides its prevalence, this kind of arrhythmia is related to an increased risk of suffering a stroke, and is in many cases the underlying cause of death (Benjamin et al., 2018).

The most common method for analyzing the heart's electrical and rhythmic patterns, in order to detect possible anomalies, is through an Electrocardiogram (ECG). A traditional ECG has 12 cardiac leads which measure the heart's activity from different planes by placing several electrodes on the chest and limbs and measuring the electrical current between them (Conover, 2002).

Systematic analysis of potential risk patients is still used as one of the ways to detect undiagnosed arrhythmias (Moran et al., 2016). These techniques depend on routine screenings and manual observation of ECG records to confirm if a risk patient has an arrhythmia or not.

In this context, telemecidine and wearable devices can aid in monitoring cardiac patients, allow ing for a more extensive medical follow-up in a less invasive way, and even permitting a better access for people who live far away from health facilities (Wootton et al., 2017). Wearable devices for these patients have the potential to improve their diagnostic and follow-up (Gusev et al., 2017), especially when combining them with techniques for automatic anomaly detection. Since telemedicine devices are designed to report data over longer periods of time than a standard screening, it becomes un feasible for physicians to manually analyze a patient's complete long-form records, which could span days or weeks of monitoring. This makes the addition of automatic detection techniques important for such devices.

Regarding automatic arrhythmia detection, multiple articles have been published about utilizing Machine Learning techniques to work on 12-lead, 2-lead and even 1-lead ECG signals (Acharya et al., 2018; Y. Li et al., 2018; Hannun et al., 2019; Yıldırım et al., 2018; Yildirim et al., 2020).

The efficiency, accuracy and complexity of Machine Learning models is highly dependent on the quality of the training data utilized (N. Gupta et al., 2021). On the other hand, wearable devices are usually designed considering their cost, portability, energy consumption, and data transfer and storage costs (Sodhro et al., 2018; R. Gupta et al., 2014), which leads to lower quality signals being captured by them. From this conflict emerges the need to study the concrete impact of this data quality reduction, as well as ways to compensate for it. This balance of cost and performance is dramatically important in applications related to medicine, where even small performance gains result in potentially more patients receiving adequate care and treatment, which means saving lives.

State-of-the-art arrhythmia classification models usually work on high quality ECG data, and achieve maximum performance using only one or two leads of ECG data (Acharya et al., 2017; Z. Li et al., 2020; Yıldırım et al., 2018; Mousavi et al., 2020). However, in limited data scenarios, as when building and using wearable devices for arrhythmia monitoring, recording and using different leads for classification can be a promising direction. Given that every individual lead provides different information, it is also important to analyze the models' performances on each lead separately, as well as their particular accuracy when detecting each specific arrhythmia class. This provides useful insight such as considering which leads to include when designing a new device, or how to best combine several available leads in order to increase classification performance.

This work aims to first study the degradation of classification performance by using lower quality signals (similar to those provided by wearable devices), and then propose models that compensate for this loss and improve performance. In order to achieve this, a state-of-the-art classifier based on the Convolutional Neural Network (CNN) architecture is trained to detect and classify cardiac

arrhythmias of up to 4 classes; the dataset published by Zheng et al. (2020) will be used for training, which contains over 10,000 12-lead ECG records of cardiac patients recorded at a 500 Hz sampling rate and a 32-bit resolution. Given that the intended use for this model is to utilize the data provided by wearable ECG devices, which usually don't serve the same information as a full traditional ECG recording, the model will also be trained on downsampled and quantized signals as to better approximate the data provided by such devices. As a reference point, this work will utilize Galeno Sys (DataFlow, 2019), a wearable 6-lead (leads I, II, III, aVR, aVF and aVL) ECG device, which measures the heart's activity with a frequency of 100 Hz and a resolution of 8-bits. The information obtained from these proposed models' performances, including complete analysis on a per-lead and per-class basis, will then be used to build combinatory models of the 6 leads (or a subset of them) in order to reach a better classification efficiency. To the best of my knowledge, no other published paper studies the use of 6 leads for arrhythmia classification. Furthermore, no other study provides an analysis on the performance of classification models on each individual lead by arrhythmia classes, which can be useful not only for building classification models, but also for designing and building wearable devices.

While the studied signal downgrade decreases accuracy from 95.3% to 93.9% on the best perform ing individual lead, a soft weighted majority voter model can increase the accuracy of the downgraded signals up to 95.4% utilizing 2 leads, or up to 95.8% utilizing 4 leads.

The remainder of this work is organized as follows: In Section 2, I present a brief review of literature related to this topic and its theoretical background. Section 3 first describes the method ology employed in this work, and later the dataset used for training the various models. Section 4 presents the results, and finally Section 5 introduces the achieved conclusions and discusses this work's findings.

2 Literature Review

2.1 Related Works

The MIT-BIH Arrhythmia Database is probably the most widespread dataset in automatic ECG classification (Moody & Mark, 2001). This database was created in 1980 as the first generally available material of its kind, and is comprised of 48 patients' half-hour excerpts of 2-lead ambulatory ECG recordings, annotated by physicians with the corresponding heartbeat information. These records are digitized with a 360 Hz sampling rate and an 11-bit resolution. While it is a staple of ECG analysis through Machine Learning, technical limitations of the tape recording equipment utilized (Moody & Mark, 2001), imbalanced classes (Shaker et al., 2020) and other issues mark the need for a more up to date database which contemplates these problems. Zheng et al. (2020) present such a database, with over 10,000 records recorded at 500 Hz and 32-bits, which this work utilizes and will be later described in the Data Subsection. Another similar dataset is presented by Wagner et al. (2020), containing 21837 records at 500 Hz and 16-bits, and will also be described further in this work as it will be utilized for final validation of the proposed

models. Several other public datasets of ECG records exist, but many of them are focused on other specific tasks rather than general arrhythmia detection.

A wide array of Machine Learning techniques have been applied to ECG analysis and classifica tion. Many of the works in this regard, however, depend on manual pre-processing of signals and feature extraction prior to the classification phase (Acharya et al., 2007; P lawiak & Acharya, 2020;

De Chazal & Reilly, 2006; Lin, 2008). The need for pre-processing steps increases computational times, while hand-crafted features require specific knowledge in the field (Zhai & Tin, 2018).

More recently, deep learning models have become popular for ECG analysis, proposing end-to end systems which eliminate the need of processing the signal beforehand or analyzing it in order to extract features (Acharya et al., 2018; Y. Li et al., 2018; Hannun et al., 2019). Since each ECG lead is essentially a one-dimensional time series, one-dimensional Convolutional Neural Networks (1D-CNN) possess an ideal structure for these architectures (Kiranyaz et al., 2015). Acharya et al. (2017) developed a 9 layer CNN network which identifies 5 types of heart rhythms based on a single ECG lead; they obtained an accuracy of 94.03%. Z. Li et al. (2020) also aimed for the detection of 5 rhythms, testing both 1 and 2 leads of ECG records with a 31-layer 1D-CNN, obtaining a better accuracy on the 2-lead version (99.38% versus 99.06%). Yıldırım et al. (2018) worked on a single lead, classifying 17 different rhythms (normal sinus rhythm, 15 cardiac arrhythmias and pacemaker rhythm) on 10 second segments with an overall accuracy of 91.33% and a very low classification time per sample (0.015 seconds). All of the aforementioned work utilized the MIT-BIH dataset for training.

Besides 1D-CNNs, long short-term memory networks (LTSM) are another good approach for time series-like data (Hochreiter & Schmidhuber, 1997b). Such algorithms have been used to classify arrhythmias in ECG records (Chang et al., 2021; Yildirim, 2018; Mousavi et al., 2020). The strengths of LTSMs allow them to bridge time lags over 1000 discrete time steps and help to reduce the complexity of backpropagation training through time (Hochreiter & Schmidhuber, 1997a).

In the field of wearable devices, several studies have tried to perform accurate ECG abnormality classification while maintaining efficiency. Y. Li et al. (2018) developed a patient-specific CNN for classification, downsampling the MIT-BIH dataset to a 250 Hz sampling rate and utilizing only Lead II to approximate the data of wearable devices; an accuracy of 96.89% is reached on 5 classes after training a generic CNN on each patient's specific ECGs. Saadatnejad et al. (2019) again utilize a single lead of the MIT-BIH dataset with a patient-specific training phase and a LSTM based architecture, with a previous feature extraction phase, and obtain an accuracy of 99.6% and an F1 score of 0.971 on 7 rhythm classes; they conclude that this architecture has a low enough computational complexity as to be executed on wearable devices directly. Amirshahi and Hashemi (2019) work with the MIT-BIH dataset and spiking neural networks, also using a per-patient training phase, in order to classify 4 heart rhythms, obtaining an accuracy of 97.9% and an F1 score of 0.88, while maintaining low energy consumption. While other works have studied the impact of lower sampling rates and bit depths on several algorithms related to automatic ECG delineation and QRS complex detection (Simon et al., 2007; Ajdaraga & Gusev, 2017), to the best of my knowledge, no research exists of these effects on deep-learning models.

Yildirim et al. (2020) worked with the newer dataset presented by Zheng et al. (2020), utilizing

a CNN in combination with a long short-term memory (LTSM) unit, in order to predict arrhythmia classes. Their work processes each ECG lead independently in order to measure which one gives the best classification performance. Of the 11 rhythms present in the dataset, two options were tested: 7 classes (eliminating the least common examples in the dataset) and 4 classes (combining down the 11 rythms, as suggested by Zheng et al. on the original dataset's article), obtaining performances of 92.24% and 96.13%, respectively.

2.2 Theoretical background

In this section, important concepts regarding the current work's theoretical background and evalu ation methods are introduced in order to facilitate its understanding.

2.2.1 Electrocardiography

Electrocardiography is a fundamental tool for analyzing the heart's behaviour and monitoring its normal functioning. The cardiac muscle contracts and relaxes through electrical charges that depo larize and repolarize the myocardial cells. This activity can be recorded by electrodes placed on the chest and limbs, and combining them produces the various leads of a full ECG reading (Conover, 2002).

2.2.2 Signal Digitization

When digitizing analog signals, such as the electric current provided by an ECG, the quality of the data can be affected by parameters such as sampling rate, the amount of represented samples per second, and bit depth or quantization, the range of values that can be represented digitally, since a finite number of bits can be used to represent each number (Hammond, 1999).These parameters can affect the performance of different types of algorithms and models applied to the digitized data (Ajdaraga & Gusev, 2017).

2.2.3 Machine Learning Models

Machine Learning is a branch of Artificial Intelligence concerned with automatically extracting mean ingful information from a set of data and being able to learn to perform decision and classification tasks automatically. It is composed of a wide array of techniques that allow for automatic learning to happen, one of them being Neural Networks. Neural Networks are designed in a way as to mimic the behaviour of the human brain, utilizing a structure of layers of neurons which interconnect and update their output process during training.

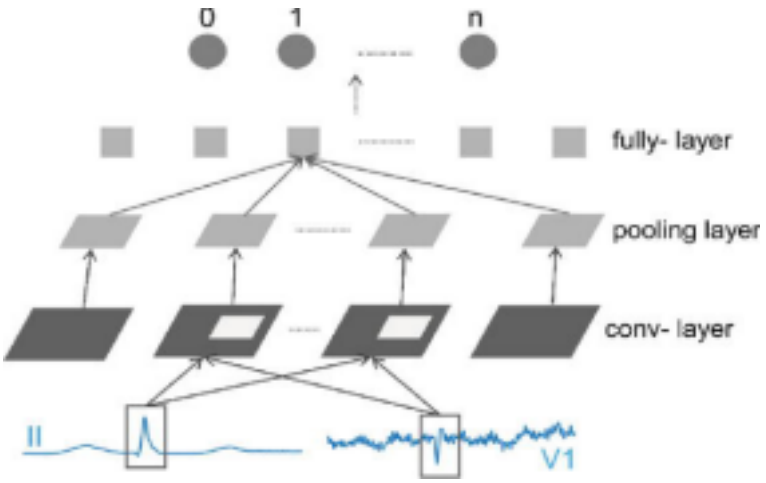
As previously stated, applying some Machine Learning techniques to ECG data requires a pre vious effort of pre-processing and manual feature selection. Given the high dimensionality involved (10 seconds of an ECG record at 500 Hz in 6 channels amount to 30,000 values), some systems rely on dimensionality reduction methods, such as principal component analysis (Castells et al., 2007), to lift some of the computational burden of processing these signals; manual selection of features by experts is an alternative, but it requires time and may be unfeasible for

larger datasets.

Deep Learning is a subset of the Neural Network structure with a bigger depth of layers, giving it an improved capacity to learn, but usually requiring a bigger amount of data for effective training.

In the scope of Deep Learning, the Convolutional Neural Network (CNN) structure is a powerful tool for many Machine Learning tasks. Most CNN architectures are comprised of three main types of layers: convolution layers, where feature maps are convolved with learnable kernels, sub-sampling layers (also called pooling layers), where inputs are downsampled thus reducing their dimensions, and finally classification layers, which are fully connected feed-forward layers at the end of the network that use the preceding feature maps to perform the desired classification (Alom et al., 2018). An example CNN architecture for ECG can be seen on Figure 1.

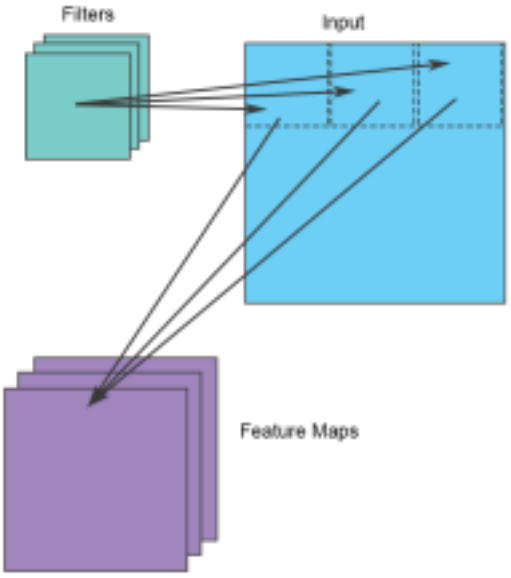
Figure 1: Example CNN architecture for ECG records



Source: Gao (2019)

Convolutional layers (Figure 2) work by running a number of filters of specified sizes through the input, resulting in an activation. The parameters of these filters are adjusted during training, giving them the ability to learn feature maps specific to the training data. In the case of 2D images, these filters run in a pixel-wise manner, while in 1D time-series like data they run in a timestep-wise manner.

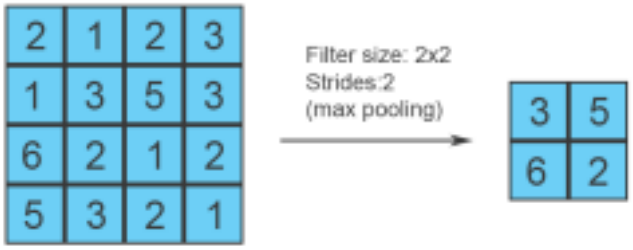
Figure 2: Example of convolutional layer operations on a 2D input



Pooling layers (Figure 3) help to reduce the size of the representations, in order to decrease the amount of necessary parameters in the network. They work by running a fixed size window through

the input by a set number of strides and either returning the average values or the maximum values in said window.

Figure 3: Example of a max pooling operation on a 2D input



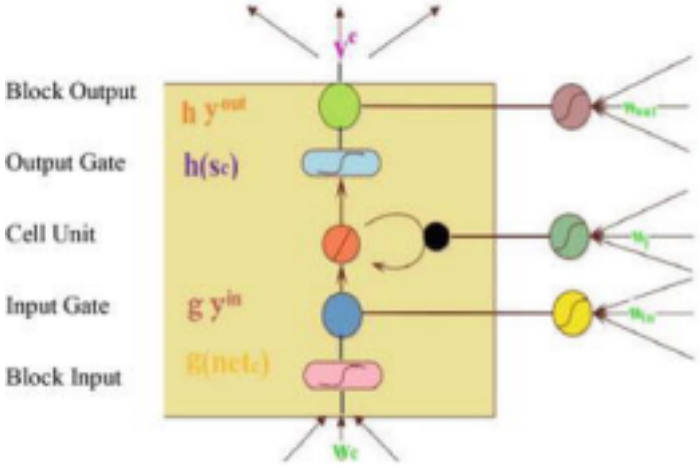
Other advanced techniques such as batch normalization (linearly transforming inputs to have zero mean and unit variance), dropout layers (randomly setting subsets of activations to zero), advanced network initialization methods (such as ”Xavier”, instead of using random initialization) and different activation functions (such as ReLU and Leaky ReLU instead of Sigmoid and TanH) can improve both classification performance and convergence speed (Alom et al., 2018).

CNNs have several advantages in processing larger inputs such as images, since they are highly optimized for learning and extracting abstractions of 2-dimensional features, and are composed of sparse connections and shared weights, minimizing the total amount of parameters in the network (Alom et al., 2018). While CNNs are typically used on image data (both 2D and 3D), their archi tecture allows for higher dimensionality input feature maps while preserving its

abstraction power (Mrazova et al., 2013).

Another deep learning technique which is particularly useful for operating sequences over time is LSTM, a variation of recurrent neural networks (RNNs) first introduced by Hochreiter and Schmid huber (1997b), which store a cell's state and learns when to maintain, use or replace it during a given sequence. ECG signals are a perfect candidate for LSTM models, since these signals are a periodic representation of heart activity through time, and considering the relationship between time-steps is particularly important. An example of a LSTM cell can be observed in Figure 4.

Figure 4: Example of a long short-term memory cell



Source: Gao (2019)

CNNs can be used successfully in combination with LSTMs, taking advantage of the representa tion learning of the former while retaining the sequence processing potential of the latter (Oh et al., 2018).

2.2.4 Model Evaluation

Several metrics can be used to assess the performance of a Machine Learning model. Some

important metrics are based on a multiclass Confusion Matrix (as seen on Table 1). The difference between a multiclass Confusion Matrix and a binary Confusion Matrix is that the former doesn't have the usual classification in True/False Negative/Positive; instead, we can identify the True Positives as the labels that were correctly classified, while the rest is considered missclassified. In this work, the Confusion Matrix is a 4x4 matrix where the main diagonal represents the correctly classified examples, while the rest of the matrix represents the missclassified examples.

Table 1: Multiclass confussion matrix

AFIB		MISSCLASSIFIED (%)	MISSCLASSIFIED (%)	TRUE POSITIV (%)
GSVT		MISSCLASSIFIED (%)	MISSCLASSIFIED (%)	MISSCLAS ED (%
SB				
SR				

	TRUE POSITIVE (%)	MISSCLASSIFIED (%)
AFIB GSVT SB SR	MISSCLASSIFIED (%)	TRUE POSITIVE (%)

From the Confusion Matrix, we can obtain a series of metrics that summarize the performance

of the model on the test set. These metrics are Precision, Recall and F1 Score (M¨uller & Guido, 2016).

Precision (also known as Predictive Positive Value) evaluates how many of the samples predicted as positive are actually positive. It is used when the goal is to limit the amount of false positives. It is calculated as the number of True Positives divided by the sum of True Positives and False Positives (Equation 1).

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

On the other hand, Recall (also known as Sensitivity) measures how many of the actually positive samples are correctly predicted as positive; in this way, Recall is important when trying to minimize False Negatives. It is calculated as the number of True Positives divided by the sum of True Positives and False Negatives (Equation 2).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

A balance between Precision and Recall can be measured utilizing the F1-score, resulting in a trade off between both metrics. F1-score calculates the harmonic mean of Precision and Recall (Equation 3), going from 1 when both measures are perfect to 0 when either one is 0.

$$F_{1score} = 2 \times \text{Precision} \times \text{Recall}$$

$$\text{Precision} + \text{Recall} \quad (3)$$

As previously mentioned, these metrics are designed for binary classification problems, where True/False Positive/Negatives are not ambiguous. In the case of multiclass classification, these metrics are calculated in a per-class binary basis and then averaged. For this work, the averaging technique used is weighted averaging, which computes the mean of the per-class F1-scores, weighted by their support. This measure will be the main evaluation metric of this work, and will be calculated on the test set unseen during the training process.

When splitting up training and validation sets, cross validation can be used in order to more properly measure the predictive performance of a model. One particular cross validation technique is random subsampling, which generates K data splits at random and uses each one for training, and then averages the performance of all the trained models, thus obtaining a better error estimate. As well as this, the splits can be stratified in order to maintain the original class distribution, avoiding training sets with unbalanced classes. This work will utilize these technique for all the different models, reporting this average performance.

3 Metodology

3.1 Methodology

The current work's methodology was structured as follows: first, the architecture and results of Yildirim et al. (2020) were reproduced on single leads at the dataset's original 500 Hz frequency and 32-bit resolution. Afterwards, since the objective of this work is first to study the potential performance loss with lower quality samples, the original signals were downsampled and quantized and the same architecture was utilized in order to compare performances. The per-lead and per-class

performances of both versions of the model was studied as well in order to determine if particular leads provide better results, or if any particular arrhythmia classes are harder to identify. Afterwards, in order to test if the performance loss can be compensated, several combinatory models were tried on the downgraded signals to see if there is an improvement over single lead performance. A second state-of-the-art model based on ECG spectrogram images was then tested on the data to compare it to the original model based on raw ECG signals. Finally, the proposed models were tested on a different dataset in order to better estimate their performance on a different data distribution, further approximating real-life performance.

3.1.1 Single original leads

Original signals were used to train the model presented by Yildirim et al.. Table 2 presents this architecture as proposed by its original authors.

Table 2: Reference architecture proposed by Yildirim et al.

Layer Type	Parameters	Output Shape
Conv1D	Filters=64, Size=21, Strides= 1	1 453 × 64
MaxPooling1D	Pool size=2	226 × 64
Batch Norm	-	226 × 64

LeakyReLU Alpha=0.1 226 × 64
Dropout Rate=0.3 226 × 64
Conv1D Filters=64, Size=7, Strides= 1 220 × 64
MaxPooling1D Pool size=2 110 × 64
Batch Norm - 110 × 64
Conv1D Filters=128, Size=5, Strides= 1 106 × 128
MaxPooling1D Pool size=2 53 × 128
Conv1D Filters=256, Size=13, Strides= 1 41 × 256
Conv1D Filters=512, Size=7, Strides= 1 35 × 512
Dropout Rate=0.3 35 × 512
Conv1D Filters=256, Size=9, Strides= 1 27 × 256
MaxPooling1D Pool size=2 13 × 256
LSTM Unit=128, Return Sequences=True 13 × 128
Flatten - 1664
Dense Units=64, Activation=ReLU 64
Dense Units=4, Activation=Softmax 4

Each lead's model (leads I, II, III, aVR, aVF and aVL) was trained 10 times in order to obtain average performances and reduce the impact of stochasticity in the results, varying the random split of train and validation sets (80% and 20% of the previously split training set, respectively). The resulting models were further evaluated on the separate test set.

All the models were trained for 25 epochs using a batch size of 64 and Adam optimizer with a learning rate of 0.001. The loss function will be categorical cross-entropy.

Since there appears to be some small overfitting on 25 epochs of training as pointed out by Yildirim et al. (2020), model checkpointing wasset up in order to also preserve the model on the epoch that performs best on the validation set. Performance was evaluated on said best models.

3.1.2 Single downsampled leads

Afterwards, the original signals were downsampled to 100 Hz and quantized to 8-bits, repeating the training of the original leads. The downsampling was performed through Fast Fourier Transform (Cooley et al., 1969). The same architecture was utilized, only modifying the strides of the first convolutional layer in order to preserve the shape of the feature maps propagated into the deeper layers, since the input dimensions are altered during downsampling. The performance of each lead and its downgraded counterpart were then compared to measure the impact of this modification.

On both cases (original and downsampled single leads), performance was also evaluated on a per lead and per-class basis, in order to determine if specific leads classify different arrhythmias better than others, or if any lead in particular is generally better at classification.

3.1.3 Combined leads

After defining the best performances on each lead, several combinatory models (which combine the input of all 6 leads) were developed in order to determine if predictive performance can be improved on a multi-lead scale. These models utilize the downsampled and quantized versions of

the data, as it is the reference point for this work.

3.1.3.1 Majority Voter model

First, a majority voter model was implemented, which takes the prediction of each lead's model and decides classification based on the most predicted class across all leads. Three versions were tested: a hard voter (only sum the absolute classes predicted by each model), a soft voter (summing over each model's predicted probabilities of class predictions) and a weighted soft voter (where the class probabilities predicted are scaled by each model's class F1 scores). These models were tested using all possible combinations of leads in order to determine if any particular combination results in better performance.

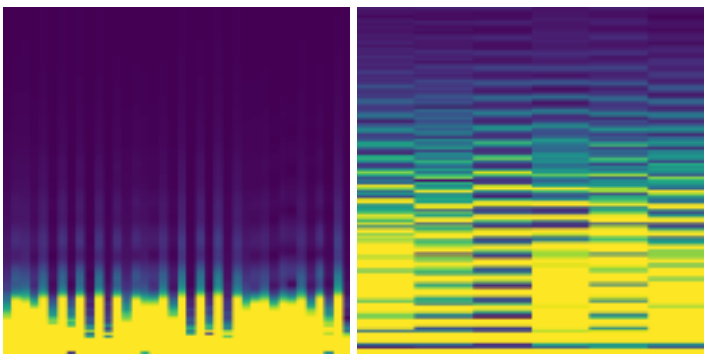
3.1.3.2 Combined model

Afterwards, an new combined model was implemented, removing the final dense and output layer from each lead's best models, and combining them with a new fully connected and output layer. This new model was re-trained on the training data, only updating the weights of the fully connected layers (leaving the original models intact, preserving their previously trained feature extraction and LSTM layers). Several iterations were trained in order to reduce stochasticity and obtain an average performance. This training was performed over 10 epochs with a learning rate of 0.0001 as the ensemble model tends to encounter overfitting earlier than the original trainings.

3.1.4 Alternative spectrogram-based 2d-CNN model

As an alternative approach to compare against the proposed model, a new architecture is tested based on spectrogram images of the ECG records rather than using the raw signal. Based on the proposed model by Huang et al. (2019), raw ECG signals are converted to 256x256 pixels spectrogram images through short-time Fourier transform. Examples of resulting images can be seen on Figure 5.

Figure 5: Example of resulting spectrogram images



(a) Original signal (b) Downsampled/quantized signal

As the input of this model are images, 2-dimensional convolutions are utilized instead of the 1-dimensional convolutions used in the previous models.

The model's architecture was first manually tweaked until an acceptable baseline performance was reached, and then Hyperband Optimization (L. Li et al., 2017) was used to select the best hyperparameters for the architecture (number of convolutional filters, filter size, learning rate and number of units in the fully connected layer). An overview of the resulting architecture can be seen on Figure 6, while the full details can be seen on Table 3.

Figure 6: Overview of the spectrogram-based model's architecture

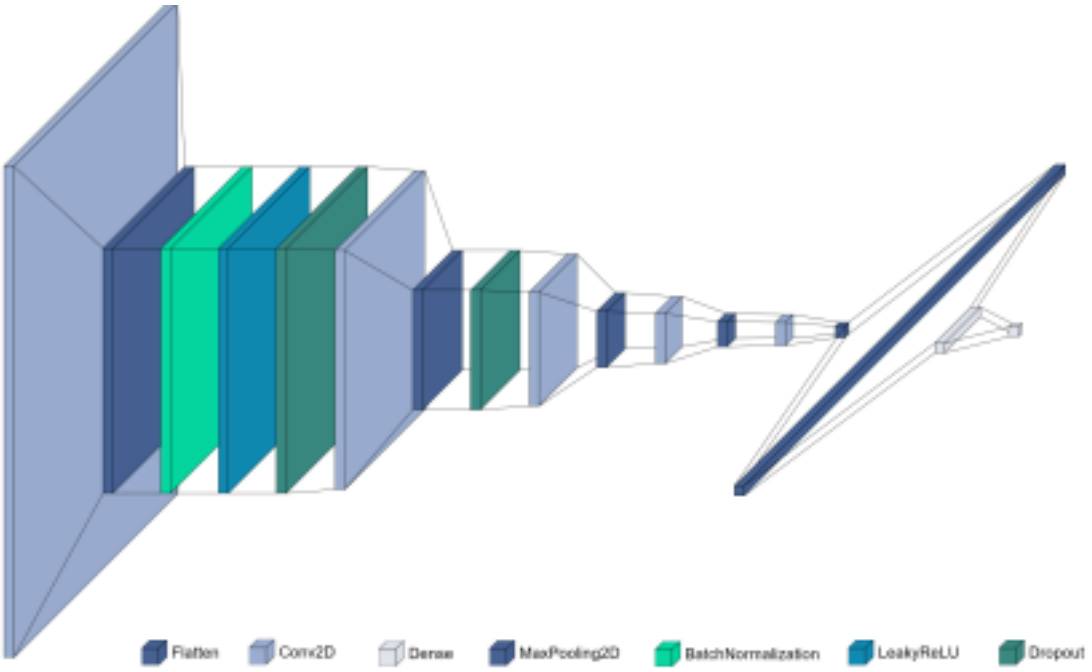


Table 3: Details of the spectrogram-based model

Layer type	Parameters	Output shape
Conv2D	Filters=56, Size=(2,2),Strides=1	(255, 255, 56)
Max Pooling 2D	Pool size = (2,2)	(127, 127, 56)
Batch Normalization	-	(127, 127, 56)
Leaky ReLu	Alpha=0.1	(127, 127, 56)
Dropout	Rate=0.3	(127, 127, 56)
Conv2D	Filters=40, Size=(4,4),Strides=1	(124, 124, 40)
Max Pooling 2D	Pool size = (2,2)	(62, 62, 40)
Dropout	Rate=0.3	(62, 62, 40)
Conv2D	Filters=64, Size=(4,4),Strides=1	(59, 59, 64)
Max Pooling 2D	Pool size = (2,2)	(29, 29, 64)
Conv2D	Filters=32, Size=(4,4),Strides=1	(26, 26, 32)
Max Pooling 2D	Pool size = (2,2)	(13, 13, 32)
Conv2D	Filters=56, Size=(2,2),Strides=1	(12, 12, 56)
Max Pooling 2D	Pool size = (2,2)	(6, 6, 56)
Flatten	-	2016
Dense	Units = 56, Activation=ReLu	56
Dense (Output)	Units = 4, Ativation=Softmax	4

3.1.5 Final test on a new dataset

In order to further measure this work's models' abilities to generalize on unseen data and better approximate real-life performance, a new dataset is used for a final test pass. Beyond the

evaluation on the hold-out test set, which comes from the same distribution as the training set, using an out of-domain evaluation dataset for real-world applications is the best indicator of how well a model is truly generalizing the problem, instead of only overfitting to the characteristics of the original training dataset (Monarch, 2021).

The dataset presented by Wagner et al. (2020) contains 21837 ECG records of 10 seconds of length obtained from 18885 patients, annotated by up to two cardiologists identifying rhythm classes and other heart conditions. This dataset was prepared for this test by fusing the arrhythmia classes as defined by the original dataset (Table 7), and randomly subsampling an amount of each class' records (in order to obtain an even class distribution). The resulting test set composition can be seen on Table 4.

Table 4: Composition of the final test set

Class	Total examples (%)
AFIB	637 (25.0)
GSVT	637 (25.0)
SB	637 (25.0)
SR	637 (25.0)
All	2548 (100)

As the records in this dataset are recorded both in 500 Hz and 100 Hz, no manual downsampling is necessary for the downsampled models, although quantization to an 8-bit resolution is still required.

3.1.6 Model evaluation

The comparison between models and leads was done through several metrics in terms of their pre dictive performance, as previously defined in the Theoretical Background section.

In addition to the general metrics of each model, performance on individual leads was analyzed in order to determine if there is any difference in their classification capabilities. Per-class performance is also presented to see if any particular class is more difficult to classify.

Finally, since the aim of this work is medical diagnosis, the kind of errors made by the classification process should also be considered. In medical diagnosis, there are two types of errors: Type I, where a patient is diagnosed with a disease and is healthy, and Type II, where a patient has a disease but is diagnosed as healthy. While both account for a misdiagnosis, Type II errors are widely considered more harmful, as the patient might miss the opportunity for a cure or treatment, endangering their health (Cummins & Hazinski, 2000).

In this regard, a recall measure will be created in order to compare the correct classification as arrhythmia against normal rhythm. For this case, True Positives are the examples correctly labeled as any arrhythmia kind, while False Negatives are all the arrhythmia examples identified as normal. This measure will be higher as less examples are wrongly classified as ”healthy”, reaching 1 when no arrhythmia sample is classified as a normal rhythm.

3.2 Data

This work uses a dataset created under the auspices of Chapman University and Shaoxing People's Hospital (Zheng et al., 2020). It is composed of 10 second measurements of 12-lead ECG records of 10,646 patients. Each segment was manually labeled by a licensed physician in order to identify the presence of any arrhythmias or other cardiac issues, and then validated by a second physician.

The dataset consists of 5,956 male patients and 4,690 female patients; among them, 17% present a normal cardial rhythm (Sinus Rhythm), while the rest possesses at least one anomaly. Among these anomalies, 12 rhythms are recognized (1 normal and 11 anomalous). The distribution of these rhythms is presented in Table 5.

Table 5: Rhythm types and distribution in the dataset

Acronym Name Full Name Frequency (%)

SB Sinus Bradycardia 3,889 (36.53) SR Sinus Rhythm 1,826 (17.15) AFIB Atrial Fibrillation 1,780 (16.72) ST Sinus Tachycardia 1,568 (14.73) AF Atrial Flutter 445 (4.18) SI Sinus Irregularity 399 (3.75) SVT Supraventricular Tachycardia 587 (5.51) AT Atrial Tachycardia 121 (1.14) AVNRT Atrioventricular Node Reentrant Tachycardia 16 (0.15) AVRT Atrioventricular Reentrant Tachycardia 8 (0.07) SAAWR Sinus Atrium to Atrial Wandering Rhythm 7 (0.07) All All 10,646 (100)

Each record contains 10 seconds of every ECG lead, sampled at a rate of 500 Hz and a resolution of 32 bits. There are two versions of each record: one is the original recording with noise, while the

other is a version de-noised through a Butterworth low pass filter, LOESS curve fitting and non local means (Zheng et al., 2020); the difference can be seen in Figure 7. Apart from the rhythm labels, each record contains other useful information such as gender, age, QRS duration and others (Table 6).

Figure 7: Comparison of original and de-noised signal on Lead I

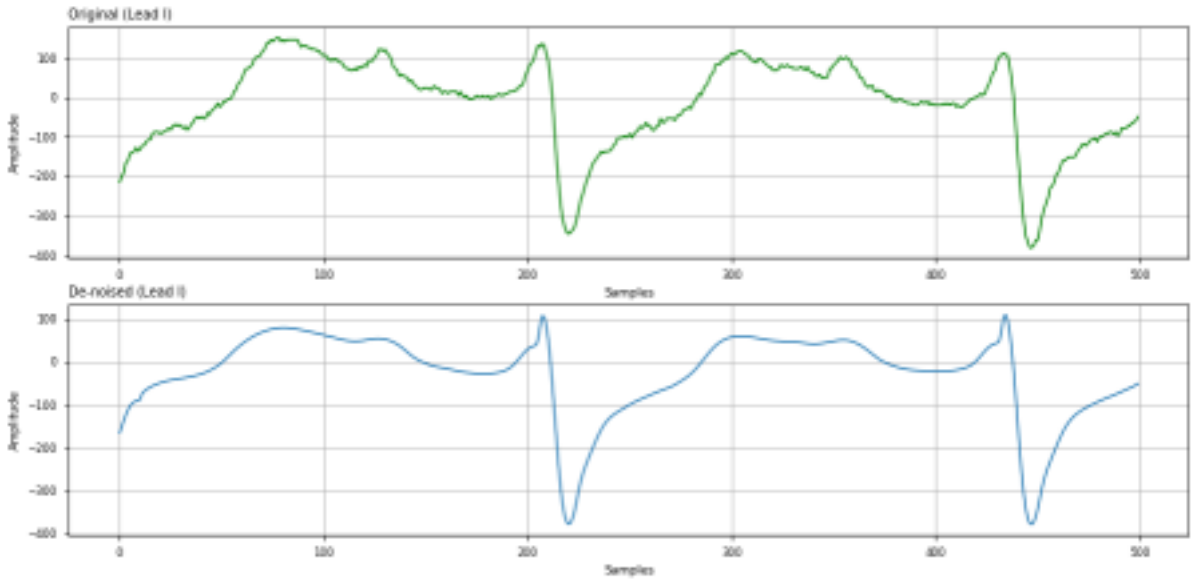


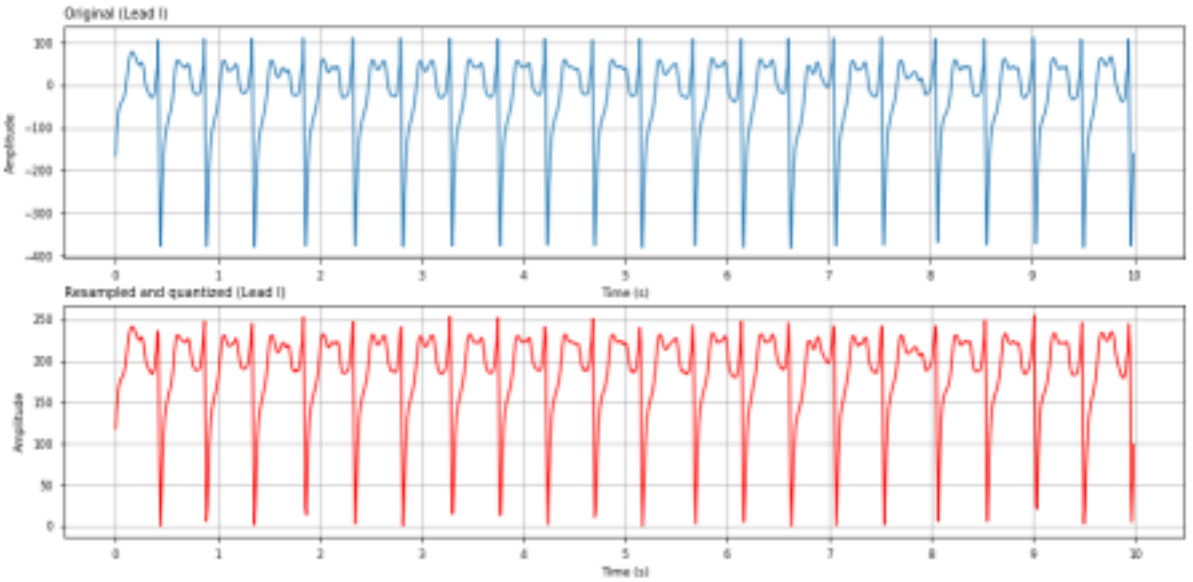
Table 6: Complete list of attributes for each record

Attributes	Type	Value Range	Description
FileName	String		ECG data file name (unique ID)
Rhythm	String		Rhythm Label
Beat	String		Other conditions Label
PatientAge	Numeric	0–999	Age
Gender	String	MALE/FEMALE	Gender
VentricularRate	Numeric	0–999	Ventricular rate in BPM
AtrialRate	Numeric	0–999	Atrial rate in BPM
QRSDuration	Numeric	0–999	QRS duration in msec
QTInterval	Numeric	0–999	QT interval in msec
QTCorrected	Numeric	0–999	Corrected QT interval in msec
RAxis	Numeric	-179~180	R axis
TAxis	Numeric	-179~181	T axis
QRSCount	Numeric	0–254	QRS count
QOnset	Numeric	16 Bit Unsigned	Q onset (In samples)
QOffset	Numeric	17 Bit Unsigned	Q offset (In samples)
TOffset	Numeric	18 Bit Unsigned	T offset (In samples)

Since the reference point of this work the ECG data provided by the Galeno Sys device, which has a sampling rate of 100 Hz, the dataset needs to be downsampled from its original 500 Hz rate. While some literature concludes that sampling rates as low as 62.5 Hz do not impact the ability to perform automatic ECG delineation (Simon et al., 2007), there is evidence supporting the fact that accuracy

of such algorithms is unsatisfactory beyond 120 Hz (Ajdaraga & Gusev, 2017). Furthermore, the device has a resolution of 8 bits, which Ajdaraga and Gusev (2017) conclude is insufficient, at least for some QRS detection algorithms. Since the device works at 100 Hz and 8-bits, these variables will be tested to measure their impact on performance. Figure 8 presents the comparison of an original signal and its downgraded and quantized counterpart.

Figure 8: Comparison of original and downsampled and quantized signal



As previously seen on Table 5, some arrhythmia classes have very few examples, which could impact the model's ability to generalize their classification. For this reason, and as suggested by Zheng et al. (2020), the classes are merged into 4 superclasses (AFIB,GSVT,SB,SR), following the recommendations of cardiologists. An additional 58 records are discarded because they are either empty or contain incomplete data. A test set of 20% of the records is separated in order to measure the models' ability to generalize outside of the training set. This test set is stratified in order to preserve the proportion of classes of the entire dataset. Table 7 presents the fused classes and their distribution.

Table 7: Resulting classes after merging and splitting train-test sets

Original Classes	Merged Class	Total Examples (%)	Training set (%)	Test set (%)
AFIB, AF	AFIB	2218 (20.94)	1774 (20.94)	444 (20.94)
SVT, AT, SAAWR, ST, AVNRT, AVRT	GSVT	2260 (21.34)	1808 (21.34)	452 (21.34)
SB	SB	3888 (36.72)	3110 (36.72)	778 (36.72)
SR, SI	SR	2222 (20.98)	1778 (20.98)	444 (20.98)
All	All	10588 (100)	8470 (100)	2118 (100)

4 Results

4.1 Impact of data quality reduction on classification performance

Before detailing the per-lead and per-class performances on the original and downgraded signals, it is important to first address the general performance loss across all leads. An overview of the average performance across all leads on the original and downsampled signals can be seen on Table 8

Table 8: Overview of performance loss on downsampled leads

Model	Accuracy	Validation Accuracy	Test Accuracy	F1 Score (Test)
Original Signals	0.965 (0.008)	0.943 (0.011)	0.923 (0.013)	0.923 (0.01)
Downsampled Signals	0.941 (0.018)	0.944 (0.01)	0.927 (0.01)	0.927 (0.01)
Performance Loss	2.4%	2.0%	1.7%	1.7%

Note: Standard deviation in parentheses.

As seen in these results, there is a performance loss when training the model on the 100 Hz and 8-bit signals. Further details can be found on the following subsections.

4.2 Single leads

4.2.1 Original signals

Each individual lead was trained 10 times on the original data at 500 Hz and 32-bit resolution with varying train/validation splits. Table 9 presents the overall average metrics while Figure 9 presents

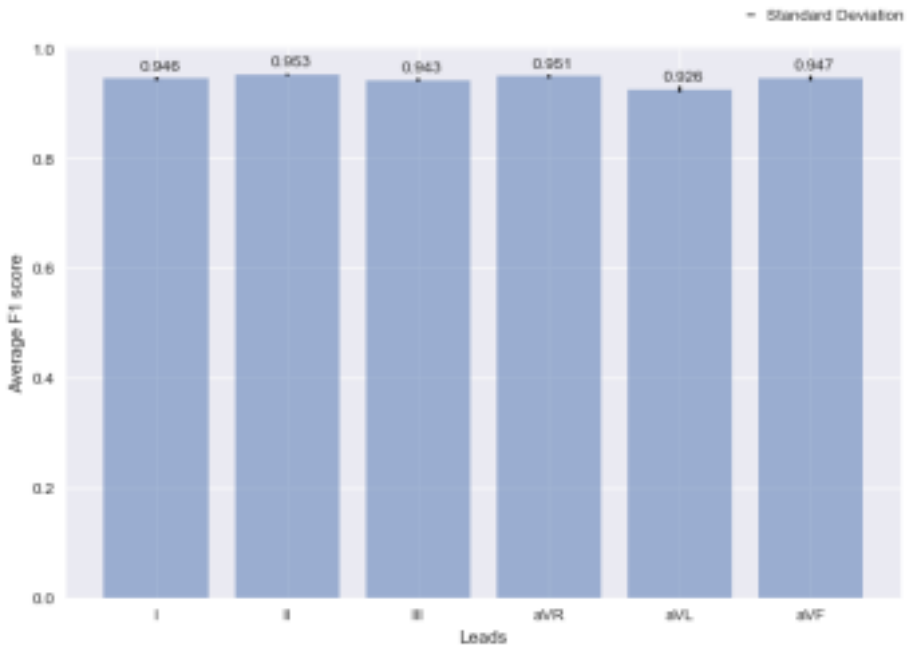
the average F1 score (calculated on the test set) of each with its corresponding standard deviation.

Table 9: Average model results on single original leads

Lead	Accuracy	Validation Accuracy	Test Accuracy	F1 Score (Test)
Lead I	0.96 (0.0108)	Lead aVR	0.971 (0.0067)	(0.0026) 0.951 (0.003) 0.943 (0.0038) 0.953 (0.0033) 0.923 (0.0062) 0.943 (0.0065) 0.947 (0.0033) 0.953 (0.0019) 0.943 (0.004) 0.95	(0.0032) 0.926 (0.0056) 0.947 (0.0045) 0.946 (0.0033) 0.953 (0.0019) 0.943 (0.004) 0.951 (0.0032) 0.926 (0.0058) 0.947 (0.0047)
Lead II	0.975 (0.007)	Lead aVL	0.953 (0.0161)		
Lead III	0.967 (0.0081)	Lead aVF	0.962 (0.0067) 0.943		

Note: Standard deviation in parentheses. 17

Figure 9: Average F1 score on each original lead



As seen on Table 9, every lead provides a good performance of over 0.92 F1 score, with lead aVL performing the worst.

For further analysis, per-class average F1 scores are presented on Table 10 and Figure 10, while the Confusion Matrices of each lead's best model are presented on Figure 11.

Table 10: Average Class F1 scores across each original lead

Class	Lead I	Lead II	Lead III	Lead aVR	Lead aVL	Lead aVF

AFIB	0.902 (0.0094)	SR	0.954 (0.0033)	0.957 (0.0028)	(0.0067)	0.853 (0.0172)	(0.013)
GSVT	0.919 (0.0062)		0.92 (0.0037)	0.898 (0.0087)	0.911 (0.0065)	0.906 (0.0058)	0.921 (0.0073)
SB	0.983 (0.0021)		0.926 (0.0052)	0.918 (0.0062)	0.917 (0.0056)	0.978 (0.0027)	0.983 (0.0026)
			0.986 (0.0016)	0.984 (0.0028)	0.985 (0.0025)	0.928 (0.0061)	0.953 (0.0047)
				0.942	0.964 (0.0028)	0.903	

Note: Standard deviation in parentheses. 18

Figure 10: Average F1 score of each class per original lead

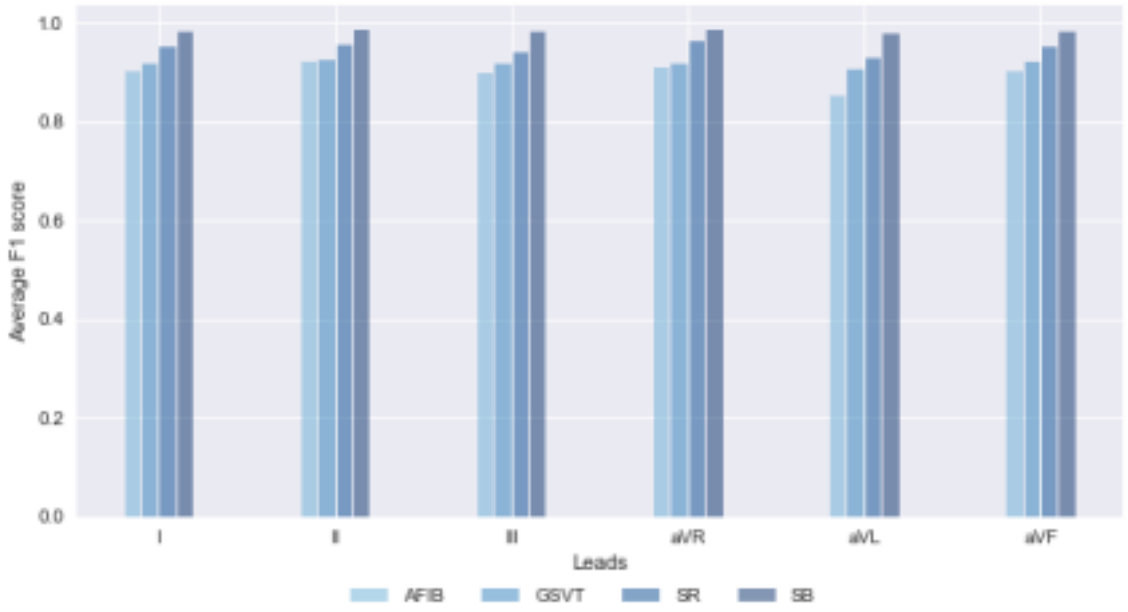
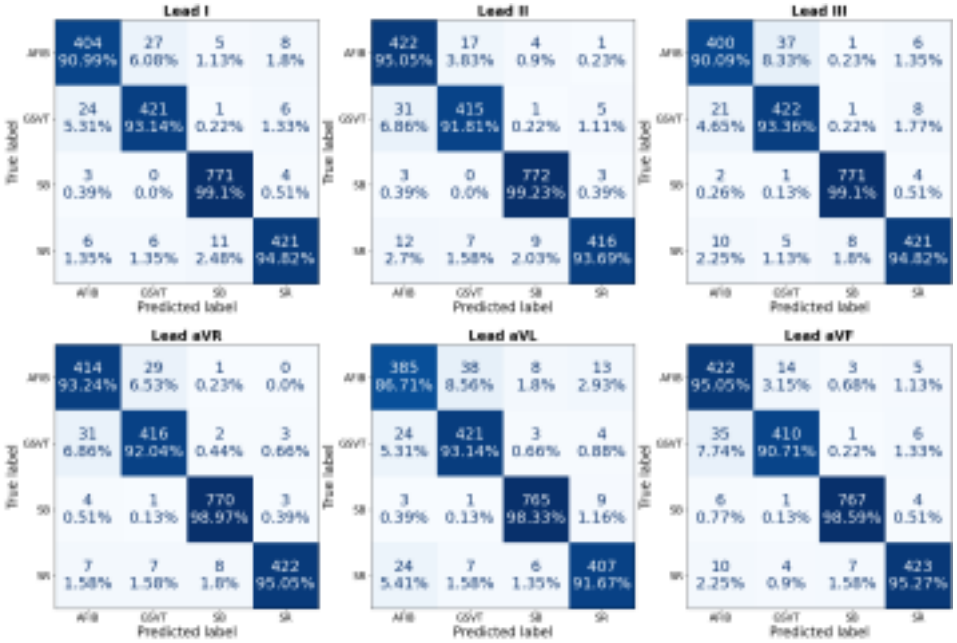


Figure 11: Confusion matrix of each original lead's best model



Class Sinus Bradycardia (SB) performs the best in all leads, followed by Sinus Rhythm (SR), Gen eral Supraventricular Tachycardia (GSVT) and finally Atrial Fibrillation (AFIB). Lead II presents the best F1 score for AFIB, GSVT and SB, while lead aVR has the best score for SR.

Finally, the average Arrhythmia Recall metric is presented on Table 11.

Table 11: Average Arrhythmia Recall across original leads

Lead	Arrhythmia Recall
I	0.989 (0.0017)
II	0.991 (0.0022)
III	0.989 (0.0032)
aVR	0.994 (0.0023)
aVL	0.984 (0.0035)
aVF	0.99 (0.0036)

Note: Standard deviation in parentheses.

Lead aVR presents the smallest number of arrhythmias wrongly classified as a normal rhythm.

4.2.2 Downsampled signals

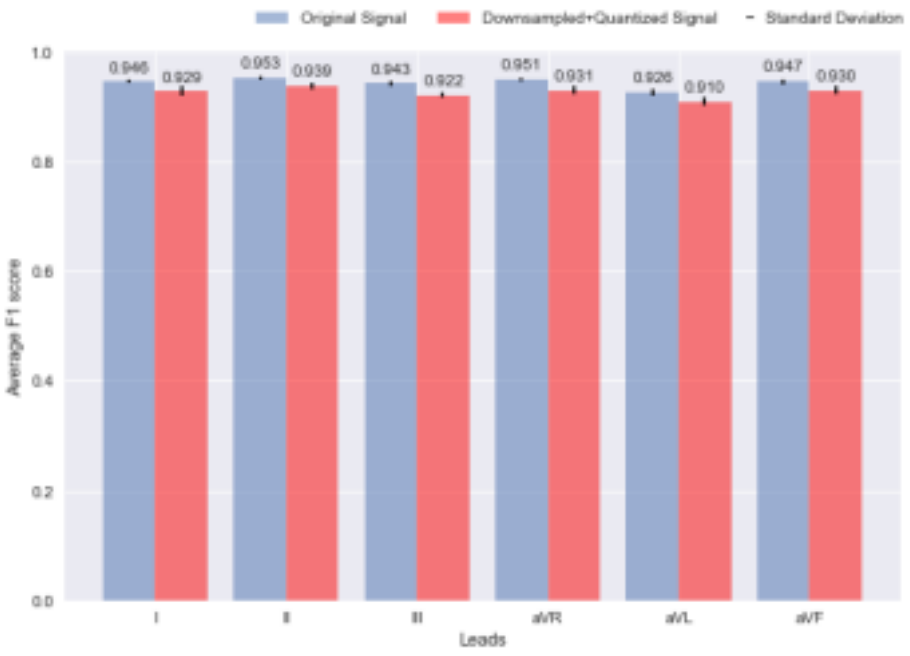
Next, new models were trained with data downsampled to 100 Hz and quantized to an 8-bit resolu tion. Each lead's models were once again trained 10 times to obtain an average performance. Table 12 presents the overall average metrics of the downsampled models, while Figure 17 presents the comparison of average F1 scores between the original and the downsampled signals.

Table 12: Average model results on single downsampled leads

Lead	Accuracy	Validation Accuracy	Test Accuracy	F1 Score (Test)
Lead I	0.927 (0.0152)	Lead aVR	0.958 (0.0182)	(0.0074) 0.94 (0.0051) 0.911 (0.0055) 0.93 (0.0077) 0.904 (0.0037) 0.927 (0.0046) 0.929 (0.0083) 0.939 (0.0046) 0.922 (0.0046) 0.931
Lead II	0.953 (0.0099)	Lead aVL	0.915 (0.0149)	(0.0064) 0.911 (0.006) 0.93 (0.0073) 0.929 (0.0086) 0.939 (0.0047) 0.922 (0.0045) 0.931 (0.0063) 0.91 (0.0064) 0.93 (0.0073)
Lead III	0.938 (0.0189)	Lead aVF	0.956 (0.0137) 0.926	

Note: Standard deviation in parentheses. 20

Figure 12: Comparison of average F1 scores between each original and downsampled



lead

As seen on Figure 17, every lead has worse performance when trained on the downsampled data. The average F1 performance loss across all leads is 1.74%.

Per-class average F1 scores are presented on Table 13 and Figure 13. The Confusion Matrices of each lead's best model are presented on Figure 14.

Table 13: Average Class F1 scores across each downsample lead

Class	Lead I	Lead II	Lead III	Lead aVR	Lead aVL	Lead aVF
AFIB	0.855 (0.0233)	0.932 (0.0112)	0.945 (0.008)	0.868 (0.0145)	0.812 (0.0142)	0.905 (0.0097)
GSVT	0.91 (0.0065)	0.883 (0.0119)	0.844 (0.0094)	0.903 (0.006)	0.898 (0.0064)	0.936 (0.0073)
SB	0.98 (0.003)	0.914 (0.0087)	0.901 (0.0096)	0.979 (0.0036)	0.972 (0.0025)	
		0.981 (0.0024)	0.978 (0.0027)	0.939 (0.0082)	0.913 (0.0086)	
			0.924		0.864	

Note: Standard deviation in parentheses. 21

Figure 13: Average F1 score of each class per downsampled lead

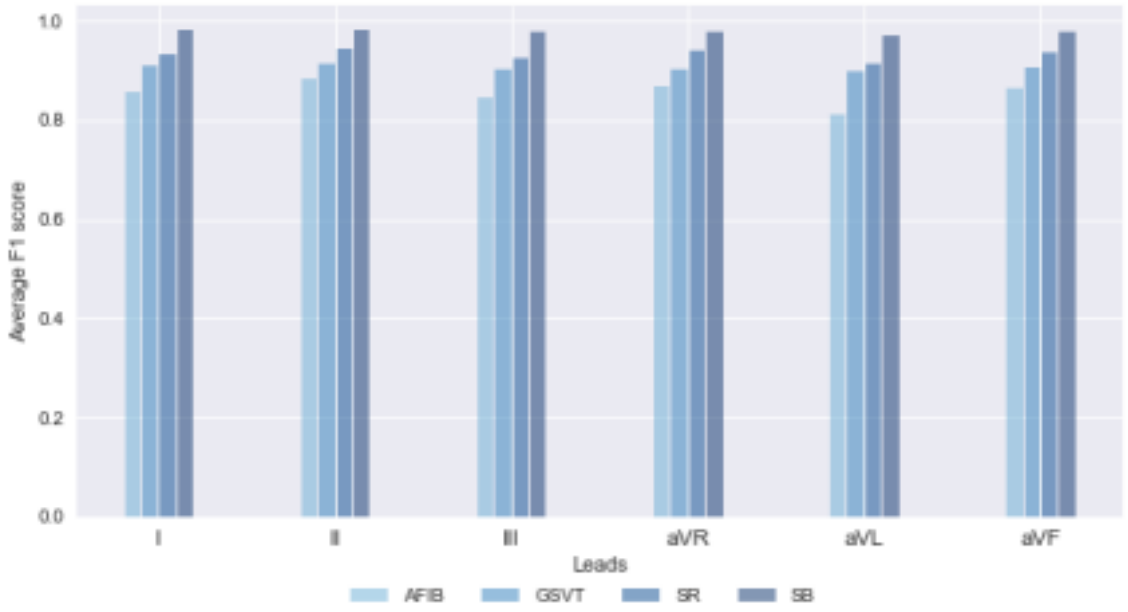
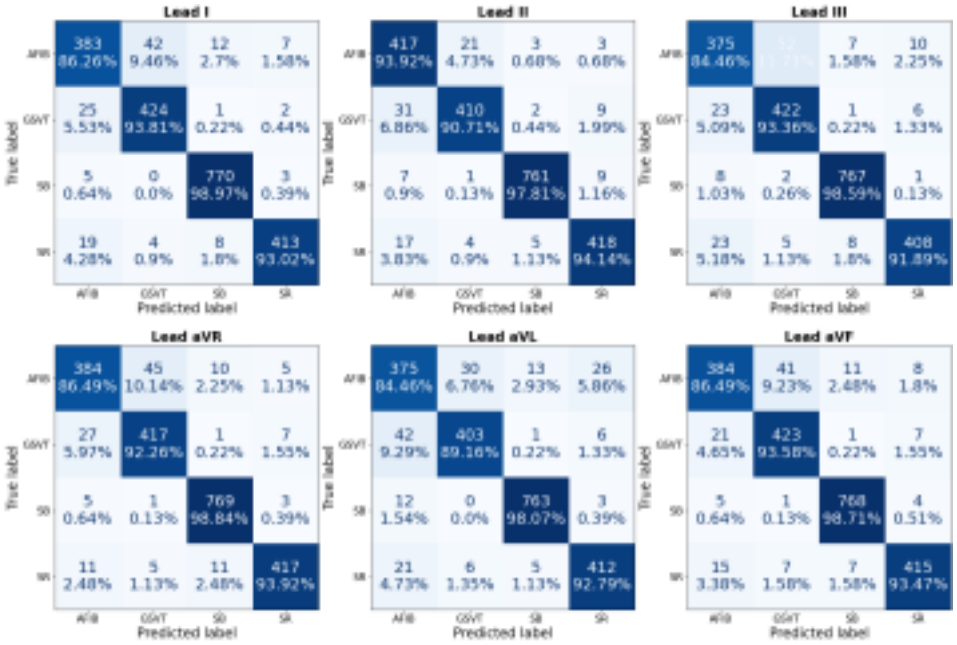


Figure 14: Confusion matrix of each downsampled lead's best model



Finally, downsampled leads' average Arrhythmia Recall is presented on Table 14

Table 14: Average Arrhythmia Recall across downsampled leads

Lead	Arrhythmia Recall
I	0.983 (0.0071)
II	0.989 (0.0061)
III	0.982 (0.0066)
aVR	0.988 (0.0053)
aVL	0.979 (0.0031)
aVF	0.986 (0.0044)

Note: Standard deviation in parentheses.

As seen on these results, the models trained on the downsampled data perform worse than those trained on the original signals.

4.3 Combined leads

4.3.1 Majority voter

Three majority voters were implemented for the downsampled data's best models, combining all 6 leads: a hard majority voter, a soft majority voter and a weighted soft majority voter. Resulting accuracy and F1 scores of each majority voter model can be seen on Table 15.

Table 15: Accuracy and F1 scores for majority voter models

Model	Accuracy	F1 Score
Hard Voter	0.94475	0.94471
Soft Voter	0.95467	0.95463
Weighted Soft Voter	0.95561	0.95555

The hard voter version performs worse than the best single lead performance on downsampled data (Lead II), which obtained an F1 score of 0.947, while both soft majority voters performed better (with the weighted version performing marginally better). The latter two also perform better than the F1 scores of each model utilized to make the individual predictions (as seen on Table 16)

Table 16: F1 scores for majority voter's selected models

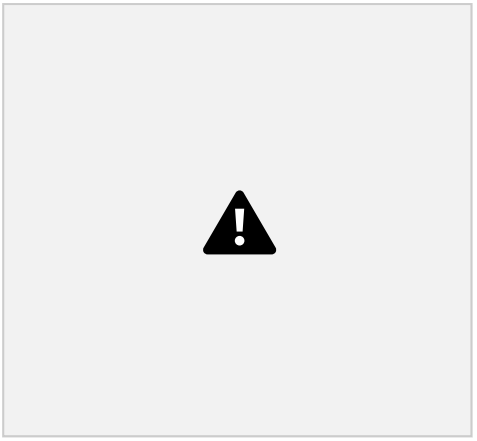
Lead	F1 Score (test)
I	0.9447
II	0.947
III	0.9345
aVR	0.9417
aVL	0.9255
aVF	0.9283

The per-class F1 scores for the soft weighted majority voter are presented on Table 17, while its Confusion Matrix is found on Figure 15.

Table 17: Per class F1 scores for the soft weighted majority voter

Class	F1 Score (test)
AFIB	0.916
GSVT	0.935
SB	0.987
SR	0.961

Figure 15: Confusion matrix of the soft weighted majority voter



Finally, the Arrhythmia Recall metric for the soft weighted majority voter is 0.993, which out performs the best average Arrhythmia Recall for single lead models (Lead II, with an Arrhythmia Recall of 0.989).

Furthermore, the soft weighted majority voter scheme was tested on all possible lead combinations in order to see if combining fewer leads produced better results. All the combinations and their respective scores can be seen on Table 18.

Table 18: All possible combinations for the soft weighted majority voter's F1 scores

Leads	F1 Score (test)
I-II-aVL-aVF	0.958
I-II-aVF	0.957
I-II-aVR-aVL-aVF	0.957
I-II-aVR-aVF	0.957
I-II-III-aVR	0.956
I-II-III-aVR-aVL-aVF	0.956
I-II-III-aVR-aVF	0.956
I-II-III-aVR-aVL	0.955
II-III-aVR-aVL	0.955
I-II-III-aVL-aVF	0.955
I-II-III-aVF	0.955
I-II-III	0.955
I-II-III-aVL	0.955
II-III-aVR-aVL-aVF	0.955
I-aVR-aVL-aVF	0.955
II-aVR-aVL-aVF	0.954
II-III-aVR	0.954
I-III-aVR-aVF	0.954
I-III-aVR-aVL	0.954
I-II	0.954
I-III-aVR	0.954
II-aVL	0.954
I-II-aVR-aVL	0.953
II-III-aVR-aVF	0.953
I-II-aVL	0.953
II-aVR-aVL	0.953
I-aVR-aVL	0.952
II-aVL-aVF	0.952
I-III-aVR-aVL-aVF	0.952

Leads	F1 Score (test)
I-II-aVR	0.952
I-aVR-aVF	0.952
III-aVR	0.952
I-aVR	0.951

II-aVR 0.951	I-III-aVL-aVF 0.948
II-III-aVL 0.95	aVR-aVF 0.948
II-aVR-aVF 0.95	aVR-aVL-aVF 0.948
III-aVR-aVL 0.95	I-III 0.948
III-aVR-aVF 0.95	I-aVL 0.947
II-III-aVL-aVF 0.95	I-III-aVL 0.947
II-III-aVF 0.949	aVR-aVL 0.947
II-aVF 0.949	I-aVL-aVF 0.947
I-III-aVF 0.949	III-aVL-aVF 0.941
I-aVF 0.949	III-aVL 0.939
III-aVR-aVL-aVF 0.949	III-aVF 0.939
II-III 0.948	aVL-aVF 0.936

The all leads version ranks 6th in performance, with versions of only 3, 4 or 5 leads obtaining better F1 scores. The highest performance (using leads I, II, aVL and AVF) perform 0.2% better than on all 6 leads. While this difference between combinations is small and could be attributed to stochasticity, performance gains can be obtained with as little as 2 leads (against using the best single lead available).

4.3.2 Combined model

The combined model was created using the same best models also selected for the majority voter. It was trained 10 times to obtain average performances. The results can be seen on Table 19.

Table 19: Average accuracy and F1 scores of the ensemble model

Accuracy (Test set)	F1 Score
0.946 (0.0052)	0.945 (0.0056)

Note: Standard deviation in parentheses. 25

This new model performs better than the best average performance on a single lead (Lead II, with an F1 of 0.939), while it performs very similarly to the soft majority voter models (which obtained an F1 score of 0.9484 and 0.9488 on the regular and weighted variants, respectively) on average. Per-class average F1 scores of this model are presented on Table 20, while the Confusion Matrix for the best ensemble model is presented on Figure 16.

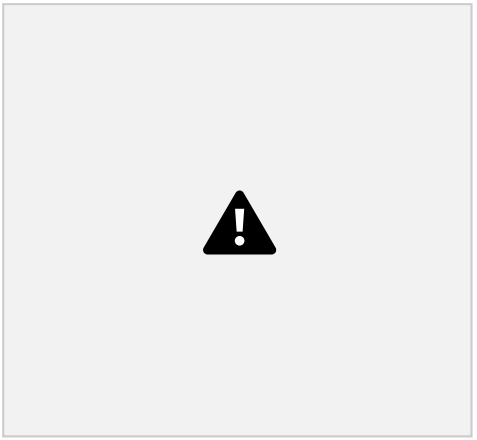
Table 20: Per class average F1 scores for the ensemble models

Class	Average F1 score
AFIB	0.8983 (0.009)
GSVT	0.9268 (0.0055)
SB	0.9823 (0.0029)

SR	0.9436 (0.007)

Note: Standard deviation in parentheses.

Figure 16: Confusion matrix of the best ensemble model



The ensemble model has, on average, worse per-class performances than the soft majority voter.

Finally, the average Arrhythmia Recall metric for the ensemble model is 0.990 (with a standard deviation of 0.0026). This average Arrhythmia Recall achieved was lower than the one reached by the soft weighted majority voter.

4.4 Spectrogram-based model

This model was also trained for 25 epochs, saving the model on the epochs with the best validation accuracy as some overfitting occurred at different points. The full results for both the original, and the downsampled and quantized signals can be seen on Table 21 and Figure 17.

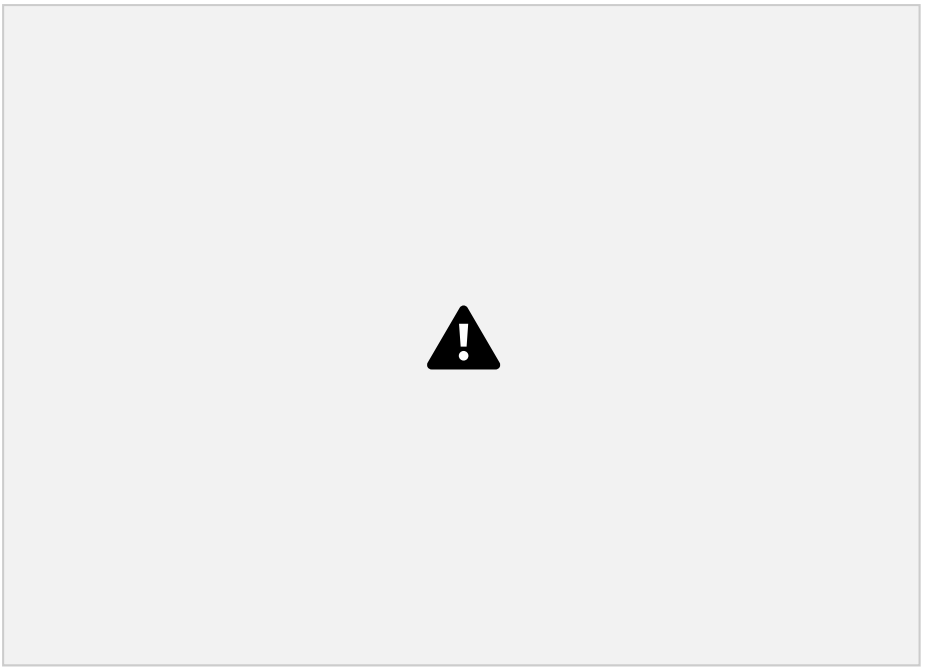
Table 21: Average F1 scores of the spectrogram-based model on original and downsampled signals

Lead	Accuracy (Original)	F1 Score (Original)	Accuracy (Downsampled)	F1 Score (Downsampled)
Lead I	0.892 (0.0056)	0.891 (0.006)	0.9 (0.0041)	0.897 (0.0052)
Lead II	0.901 (0.0032)	0.9 (0.0041)	0.897 (0.0052)	0.905 (0.0053)
Lead III	0.898 (0.0049)	0.897 (0.0052)	0.905 (0.0053)	0.889 (0.0044)
Lead aVR	0.906 (0.0049)	0.905 (0.0053)	0.889 (0.0044)	0.9 (0.0043)
Lead aVL	0.89 (0.0041)	0.889 (0.0044)	0.9 (0.0043)	0.91 (0.0065)
Lead aVF	0.901 (0.0042)	0.9 (0.0043)	0.91 (0.0065)	0.919 (0.005)
				0.911 (0.0041)
				0.917 (0.0022)
				0.897 (0.0041)
				0.908 (0.0056)
				0.909 (0.007)
				0.918 (0.005)
				0.911 (0.0045)
				0.916 (0.0024)
				0.897 (0.0043)
				0.907 (0.0059)

Note: Standard deviation in parentheses.

Figure 17: Comparison of average F1 scores between each original and downsampled lead on the

spectrogram-based model



As seen on the results, this model performs worse on average than the proposed model by 4.73% on the original signals and 1.74% on the downsampled and quantized signals. One interesting result is that the downsampled signals, unlike on the proposed model, perform better than the original signals. This may be due to a noise reduction effect resulting from the quantization of the signal, which results in less redundant information on the produced images.

4.5 Final test on a new dataset

This evaluation is performed on the final trained models (training on the entire training/validation set used previously for random splitting); the reported original performance is measured on the

originally separated test set (not used during final training). Arrhythmia Precision is measured in the same way as the Arrhythmia Recall, considering all arrhythmia classes as positive and normal rhythms as negative. The full results can be seen on Table 22.

Table 22: Models' performance on the final test set

Model Accuracy F1 Score Arrhythmia Recall Arrhythmia Precision Single-1D CNN

Downsampled Signal (Lead I) 0.901 (0.976) 0.901 (0.947) 0.980 (0.985) 0.935 (0.988) Multiple-1D CNN

Soft weighted majority voter 0.912 (0.955) 0.911 (0.955) 0.982 (0.993) 0.941 (0.986) Combined 0.899 (0.946) 0.899 (0.945) 0.899 (0.949) 0.934 (0.983)

Spectrogram model

Downsampled Signal (Lead II) 0.829 (0.920) 0.827 (0.922) 0.986 (0.975) 0.886 (0.974) Note: Original dataset scores in parentheses.

As seen on these results, while the models perform worse on the new test data, they are still able to generalize well and maintain a high recall rate, meaning that the amount of false negatives stays low. This loss of performance may be due to the different distribution compared to the

training set and differences among datasets (namely, the analog/digital conversion on the new dataset uses different parameters, and the records aren't subjected to de-noising as the original dataset is). The soft weighted majority voter on all 6 leads again performs best, with a F1 score loss of 0.043 compared to the original dataset.

Furthermore, the per-lead performance of the 1D-CNN proposed model on downsampled data differs in respect of the original dataset. Table 23 presents the per-lead analysis on this new dataset.

Table 23: Per-lead results on the 1D-CNN model on the final test set

Lead	Accuracy	F1 Score	Arrhythmia Recall	Arrhythmia Precision
Lead I	0.901	0.901	0.980	0.935
Lead II	0.885	0.884	0.968	0.939
Lead III	0.868	0.867	0.984	0.928
Lead aVR	0.892	0.891	0.975	0.932
Lead aVL	0.869	0.867	0.970	0.929
Lead aVF	0.872	0.871	0.970	0.938

In this newer dataset, while all leads obtain an acceptable performance, Lead II is no longer the best-performing one, with Lead I and Lead aVR surpassing it. This difference may indicate that some leads are not universally better for classification, or it could be due to differences in the methods these datasets have when recording each lead.

In order to further test the performance of the voter model on the new dataset, and considering that per-lead performance differs from the original dataset, all majority voter combinations are tested again on this new dataset. These results can be seen on Table 24.

Table 24: All possible combinations for the soft weighted majority voter's F1 scores on the new test set

Leads	F1 Score (test)
I-II-III-aVR-aVL	0.9147
I-II-III-aVR	0.9135
I-II-aVR	0.9134
I-III-aVR-aVL	0.9127
I-II-aVR-aVL	0.9123
I-III-aVR-aVL-aVF	0.912
I-II-III-aVR-aVL-aVF	0.9117
I-aVR-aVF	0.9116
I-III-aVF	0.9116
I-II-III	0.9116
I-II-III-aVR-aVF	0.9112
I-III-aVR	0.9111
I-III-aVR-aVF	0.9109
I-II-III-aVL	0.9109
I-II-aVR-aVL-aVF	0.9102
I-aVR-aVL	0.9099
I-II	0.9099
I-aVR-aVL-aVF	0.9093
II-III-aVR-aVL	0.9089
I-II-aVR-aVF	0.9088
II-III-aVR	0.9085
I-aVR	0.9083
II-III-aVR-aVL-aVF	0.9082
I-II-aVL	0.908
II-aVR-aVL	0.9077
I-II-III-aVL-aVF	0.907
III-aVR-aVF	0.9063
I-II-III-aVF	0.9062
II-aVR	0.9057
III-aVR	0.9057
aVR-aVF	0.9055
I-II-aVL-aVF	0.9054
III-aVR-aVL-aVF	0.9053
I-aVL	0.905
II-III-aVR-aVF	0.905
I-aVF	0.9049
I-III	0.9047
II-aVR-aVL-aVF	0.9046
III-aVR-aVL	0.9036
I-III-aVL-aVF	0.9033
II-aVR-aVF	0.9031
I-II-aVF	0.9031
aVR-aVL	0.903
aVR-aVL-aVF	0.9023
II-III-aVL	0.9004
I-III-aVL	0.9
I-aVL-aVF	0.8998
II-III-aVF	0.8992
II-III-aVL-aVF	0.8986
II-aVL-aVF	0.8959
II-aVL	0.8957
II-aVF	0.8953
II-III	

0.8952 III-aVF 0.8912 aVL-aVF 0.8873 III-aVL
III-aVL-aVF 0.889 0.882

While the best performing combinations differ from the ones in the original dataset, there is still a gain in performance to be obtained when combining as little as 2 leads, with versions of 3, 4, and 5 leads obtaining the best scores.

5 Discussion and Conclusions

In this work, the impact of performing arrhythmia classification on data similar to that provided by wearable devices, with lower signal quality, was tested, while also proposing methods which attenuate that impact. Using the ECG measurement parameters of the reference Galeno Sys (DataFlow, 2019) device (100 Hz sampling rate and 8-bits resolution), a CNN model was trained on a public dataset with over 10,000 ECG records in order to detect normal rhythms and 3 arrhythmia classes.

Firstly, the tests confirmed that there is indeed a classification performance loss when utilizing the lower quality signals, downgrading maximum classification accuracy from 95.3% to 93.9%. In this scenario, it is no longer enough to only consider the classification provided by a single lead's signal in order to diagnose patients. While these differences in performance may seem small, given the application to the medical diagnosis field, this difference could result in patients not receiving adequate care.

Secondly, it was observed that different leads provide different performances, in some cases with varying classification accuracies on different arrhythmia classes. This knowledge can help build better models by taking into consideration each lead's strengths and weaknesses.

Thirdly, by considering the previously mentioned results, leads can be combined in a way which allows to regain the lost performance on the lower quality signals. It was shown that by just combining leads I and II on a majority voter model, accuracy can be increased to 95.4%, and by combining leads I, II, aVL, and aVF, accuracy can be further increased to 95.8%. This information could be taken into consideration when designing wearable devices, as the reduced signal parameters and number of leads would aid in their simplification, and construction and operational costs, while still retaining good diagnostic performance. These results therefore suggest that the limited data captured in 10-second ECG segments at a 100 Hz sampling rate and an 8-bit resolution of only two leads provide a sufficient classification performance while retaining efficiency for a wearable device to process and transmit to a cloud server.

An alternative 2 dimensional CNN-based approach, which utilizes spectrogram images of ECG records, was also tested. While this model performed worse than the previous model based on raw ECG signals (reaching up to 91.9% accuracy on a single lead), further analysis might be beneficial, since the image-based CNN is a simpler and more lightweight model (231,060 parameters versus 2,916,228 parameters on the previous model). This simpler model would allow to obtain savings in cloud computational costs.

The techniques presented on this work could be applied on different datasets in order to determine if certain leads are universally better for classification on average, or if this difference is dataset and device specific. The two datasets tested in this work have different best performing

leads, but further testing is required in order to conclude if this is an effect of the particular examples each dataset has and the methods used for data recording. Another possible direction for further research is to use patient's clinical information, such as gender and age, in order to improve the predictive capabilities of these models.

References

Acharya, U. R., Fujita, H., Oh, S. L., Raghavendra, U., Tan, J. H., Adam, M., . . . Hagiwara, Y. (2018). Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. *Future Generation Computer Systems*, *79* , 952–959.

Acharya, U. R., Krishnan, S. M., Spaan, J. A., & Suri, J. S. (2007). *Advances in cardiac signal processing*. Springer.

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, *89* , 389–396.

Ajdaraga, E., & Gusev, M. (2017). Analysis of sampling frequency and resolution in ecg signals. In *2017 25th telecommunication forum (telfor)* (pp. 1–4).

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., . . . Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164* .

Amirshahi, A., & Hashemi, M. (2019). Ecg classification algorithm based on stdp and r-stdp neural networks for real-time monitoring on ultra low-power personal wearable devices. *IEEE*

transactions on biomedical circuits and systems, *13* (6), 1483–1493.

Babapoor-Farrokhran, S., Rasekhi, R. T., Gill, D., Babapoor, S., & Amanullah, A. (2020). Arrhyth mia in covid-19. *SN Comprehensive Clinical Medicine*, 1–6.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., . . . others (2018). Heart disease and stroke statistics—2018 update: a report from the american heart association. *Circulation*.

Boriani, G., Valzania, C., Biffi, M., Diemberger, I., Ziacchi, M., & Martignani, C. (2015). Asymp tomatic lone atrial fibrillation-how can we detect the arrhythmia? *Current pharmaceutical design*, *21* (5), 659–666.

Castells, F., Laguna, P., S¨ornmo, L., Bollmann, A., & Roig, J. M. (2007). Principal component analysis in ecg signal processing. *EURASIP Journal on Advances in Signal Processing*, *2007* , 1–21.

Centers for Disease Control and Prevention. (1999). *Public health and aging: Atrial fibrillation as a contributing cause of death and medicare hospitalization.* Retrieved 2021-01-12, from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5207a2.htm

Chang, K.-C., Hsieh, P.-H., Wu, M.-Y., Wang, Y.-C., Chen, J.-Y., Tsai, F.-J., . . . Huang, T.-C. (2021). Usefulness of machine learning-based detection and classification of cardiac arrhythmias with 12-lead electrocardiograms. *Canadian Journal of Cardiology*, *37* (1), 94–104.

Conover, M. B. (2002). *Understanding electrocardiography*. Elsevier Health Sciences. Cooley, J. W., Lewis, P. A., & Welch, P. D. (1969). The fast fourier transform and its applications. *IEEE Transactions on Education*, *12* (1), 27–34.

Cummins, R. O., & Hazinski, M. F. (2000). Guidelines based on fear of type ii (false-negative) errors: why we dropped the pulse check for lay rescuers. *Circulation*, *102* (suppl 1), I–377.
DataFlow. (2019). *Galeno sys.* Retrieved 2021-01-12, from https://galenosys.com.uy/ De Chazal, P., & Reilly, R. B. (2006). A patient-adapting heartbeat classifier using ecg morphology and heartbeat interval features. *IEEE transactions on biomedical engineering*, *53* (12), 2535– 2543.
Gao, X. (2019). Diagnosing abnormal electrocardiogram (ecg) via deep learning. IntechOpen.
Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., . . . others (2021). Data quality for machine learning tasks. In *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining* (pp. 4040–4041).
Gupta, R., Gamad, R., & Bansod, P. (2014). Telemedicine: A brief analysis. *Cogent Engineering*, *1* (1), 966459.
Gusev, M., Stojmenski, A., & Guseva, A. (2017). Ecgalert: A heart attack alerting system. In *International conference on ict innovations* (pp. 27–36).
Hammond, J. (1999). Fundamentals of signal processing. In *Modal analysis and testing* (pp. 35–52). Springer.
Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, *25* (1), 65–69.
Hochreiter, S., & Schmidhuber, J. (1997a). Long short-term memory. *Neural computation*, *9* (8), 1735–1780.
Hochreiter, S., & Schmidhuber, J. (1997b). Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, 473–479.
Huang, J., Chen, B., Yao, B., & He, W. (2019). Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. *IEEE Access*, *7* , 92871–92880.

Khairy, P., Dore, A., Talajic, M., Dubuc, M., Poirier, N., Roy, D., & Mercier, L.-A. (2006). Ar rhythmias in adult congenital heart disease. *Expert review of cardiovascular therapy*, *4* (1), 83–95.
Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, *63* (3), 664–675.
Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18* (1), 6765–6816.
Li, Y., Pang, Y., Wang, J., & Li, X. (2018). Patient-specific ecg classification by deeper cnn from generic to dedicated. *Neurocomputing*, *314* , 336–346.
Li, Z., Zhou, D., Wan, L., Li, J., & Mou, W. (2020). Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram. *Journal of electrocardiology*, *58* , 105–112.
Lin, C.-H. (2008). Frequency-domain features for ecg beat discrimination using grey relational analysis-based classifier. *Computers & Mathematics with Applications*, *55* (4), 680–690. Monarch, R. M. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered ai*. Simon and Schuster.
Moody, G. B., & Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, *20* (3), 45–50.
Moran, P. S., Teljeur, C., Ryan, M., & Smith, S. M. (2016). Systematic screening for the detection

of atrial fibrillation. *Cochrane Database of Systematic Reviews*(6).

Mousavi, S., Fotoohinasab, A., & Afghah, F. (2020). Single-modal and multi-modal false arrhythmia alarm reduction using attention-based convolutional and recurrent neural networks. *PloS one*, *15* (1), e0226990.

Mrazova, I., Pihera, J., & Veleminska, J. (2013). Can n-dimensional convolutional neural networks distinguish men and women better than humans do? In *The 2013 international joint conference on neural networks (ijcnn)* (pp. 1–8).

M¨uller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: a guide for data scientists*. " O'Reilly Media, Inc.".

Oh, S. L., Ng, E. Y., San Tan, R., & Acharya, U. R. (2018). Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in biology and medicine*, *102* , 278–287.

P lawiak, P., & Acharya, U. R. (2020). Novel deep genetic ensemble of classifiers for arrhythmia detection using ecg signals. *Neural Computing and Applications*, *32* (15), 11137–11161.

Saadatnejad, S., Oveisi, M., & Hashemi, M. (2019). Lstm-based ecg classification for continuous monitoring on personal wearable devices. *IEEE journal of biomedical and health informatics*, *24* (2), 515–523.

Shaker, A. M., Tantawi, M., Shedeed, H. A., & Tolba, M. F. (2020). Generalization of convolutional neural networks for ecg classification using generative adversarial networks. *IEEE Access*, *8* , 35592–35605.

Simon, F., Martinez, J. P., Laguna, P., van Grinsven, B., Rutten, C., & Houben, R. (2007). Impact of sampling rate reduction on automatic ecg delineation. In *2007 29th annual international conference of the ieee engineering in medicine and biology society* (pp. 2587–2590).

Sodhro, A. H., Sangaiah, A. K., Sodhro, G. H., Lohano, S., & Pirbhulal, S. (2018). An energy efficient algorithm for wearable electrocardiogram signal processing in ubiquitous healthcare applications. *Sensors*, *18* (3), 923.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, *7* (1), 1–15.

Wootton, R., Craig, J., & Patterson, V. (2017). *Introduction to telemedicine*. CRC Press. World Health Organization. (2020). *Who reveals leading causes of death and disability worldwide: 2000-2019.* Retrieved 2021-01-12, from https://www.who.int/news/item/09-12-2020-who -reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 Yildirim, O. (2018). A novel wavelet sequence based on deep bidirectional lstm network model for ¨ ecg signal classification. *Computers in biology and medicine*, *96* , 189–202.

Yıldırım, O., P lawiak, P., Tan, R.-S., & Acharya, U. R. (2018). Arrhythmia detection using deep con- ¨ volutional neural network with long duration ecg signals. *Computers in biology and medicine*, *102* , 411–420.

Yildirim, O., Talo, M., Ciaccio, E. J., San Tan, R., & Acharya, U. R. (2020). Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ecg records. *Computer methods and programs in biomedicine*, *197* , 105740.

Zhai, X., & Tin, C. (2018). Automated ecg classification using dual heartbeat coupling based on convolutional neural network. *IEEE Access*, *6* , 27465–27472.

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., & Rakovski, C. (2020). A 12-lead

electrocardio gram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, *7* (1), 1–8.