Diagnosing pulmonary embolism from electrocardiograms

João Pedro dos Santos Marques
joao.p.d.s.marques@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2023

Abstract

Pulmonary Embolism (PE) is a significant cause of cardiovascular-related deaths worldwide, often posing diagnostic challenges due to the lack of specificity in clinical presentation. In addition, the diagnosis frequently requires confirmation by computed tomography pulmonary angiography (CTPA), which has limitations such as radiation exposure, cost, and availability. On the other hand, Electrocardiography (ECG) analysis holds promise as a reliable and easily accessible tool for monitoring cardiovascular health. Given the constant advances we have seen in deep learning applied to domains such as image analysis, natural language processing, and signal processing, neural networks have the potential to play a crucial role in PE diagnosis from ECGs. In this study, we used a dataset obtained from the Hospital de Santa Maria database, consisting of 929 examples with 261 positive and 668 negative cases. As no previous studies, at the time of this dissertation, focused on using neural networks fed only by ECGs for PE diagnosis, we employed an arrhythmia network architecture as the baseline model and developed a novel architecture for PE diagnosis based on a ResNet-18 model enhanced with a self-multihead attention layer. Two versions of the model were created: a 1D version that processed raw data and a 2D version fed with ECG spectrograms. The performance of the 1D version exceeded that of the baseline and the spectrogram version. Evaluation against guideline-recommended clinical prediction rules demonstrated improved specificity (100%; 95% CI: 94-100), positive predictive value (100%; 95% CI: 82.35-100.00), and area under the curve (AUC) of 0.75 (95% CI: 0.66-0.82), demonstrating that deep learning can contribute to improving medical assistance in the diagnosis of PE.
Keywords: Pulmonary embolism (PE), electrocardiograms (ECG), deep learning, self-multihead attention, spectrograms

1. Introduction

PE is the third most common cause of cardiovascular myopathy death worldwide, just after stroke and heart attack [6]. Some improvements have been made during the past years to decrease its mortality. Nevertheless, it is still very high in Eastern Europe (from 10% up to 30% depending on how early it is diagnosed) and presents an increasing trend in some low and middle-income countries. Furthermore, as the age of the specific population under observation increases, there is a corresponding increase in the likelihood of both developing PE and experiencing fatal outcomes.

PE occurs when the flow of blood in the pulmonary artery or its branches is disrupted. However, diagnosing PE can be challenging due to similar symptoms to other diseases, such as heart attack. Although CT Scans and X-Ray are good test tools for diagnosing this disease, not all hospitals have access to it, and they expose patients to radiation (which is not desirable). Because of that, ECGs are seen as a more accessible and healthy alternative to infer if a patient has or not a pulmonary embolism by applying some electrocardiographic score [4, 26]. There are multiple types of scores and various types of PE severity. Therefore, diagnosing it from an ECG may not be straightforward. Furthermore, some good performing scores for detecting the disease have some too complex criteria for a doctor to apply to every patient.

Considering these factors, a model that can perform an ECG analysis to predict PE and assist medical decisions was built using neural network technology. The contributions of this dissertation have been partially published in the ”Portuguese Journal of Cardiology” [24].

2. Background

The ECG is a low-cost, rapid, and widely available test that cardiologists and non-cardiologists have used for decades. It records the heart’s electrical activity from different angles to identify and locate

pathologies. For that, electrodes are placed on different parts of a patient's limbs and chest to record the electrical activity[2, 11]. One of the most used configurations is the 12-lead ECG, where, as the name suggests, 12 different electrodes are placed all over the patient to record multiple waveforms of different sensitivities. Multiple clinical approaches based on ECGs make it possible to diagnose PE from it and numerous important metrics are employed to evaluate each method's performance in this context, such as sensitivity, specificity, positive predictive value (PPV), and area under the ROC curve (AUC).

2.1. Deep neural networks

There are several different deep learning architectures, but two of the most popular ones are the convolutional neural networks (CNN) [12] and the residual neural networks (ResNet) [8]. Starting with the CNN, it can gather simpler features from a chosen input into progressively more complicated ones. This way, it can get important non-obvious findings from the input and use them to evaluate a desired task. Many layers are commonly added to a CNN to increase the network's performance. However, the deeper a network like this gets, the more difficult it is to train it due to the vanishing gradient problem [17]. ResNets appeared to address this problem by introducing the residual block, which helps the network better track information. Thanks to this addition, deeper networks than the ones used previously became easier to optimize and could gain accuracy from considerably increased depth [8] and be trained easier.

2.2. Focal loss

A Loss function measures how well a machine learning model performs on a given task by computing a scalar value to represent how distant a predicted output is from the actual output. A standard loss function used is the cross-entropy loss [9]. This loss function quantifies the difference between two probability distributions: predicted and real/true.

Although cross-entropy is one of the most used loss functions, it does not consider dataset imbalance. To mitigate this problem in machine learning tasks, a loss function known as Focal Loss was introduced[14]. It is designed to give more importance to complex or misclassified samples by applying a modulating term to the cross-entropy loss. Formally, it adds a factor $(1 - p_t)^\gamma$ to the standard cross-entropy criterion. Moreover, setting $\gamma > 0$, being γ a tunable focusing parameter, reduces the relative loss for well-classified examples ($p_t > 0.5$), putting more focus on hard, misclassified examples.

Another term that can be added to better handle the class imbalance problem in binary classification is a hyperparameter α_t.

$$\alpha_t = \begin{cases} \alpha, if\ class = 1 \\ 1 - \alpha, otherwise \end{cases} \quad (1)$$

Using this α_t presented in the equation 1, it is possible to handle and control the weight/importance of each sample depending on the class it belongs to. Thus, the alpha form of the Focal Loss is defined as shown in equation 2.

$$FL_{p_t} = -\alpha_t (1 - p_t)^\gamma log(p_t) \quad (2)$$

2.3. Self-attention

Attention is a technique that enables the model to emphasize crucial parts of the input while deeming less relevant portions as less significant. This approach facilitates the establishment of better long-range contextual relationships. Various types of attention exist, each with its own utilization methods. One such type is self-attention, wherein the input focuses on itself rather than being related to a separate target sequence: in self-attention, the input and target sequences are identical [25].

Formally, self-attention is applied to an input x of length n and dimension d by following some steps [5]. At first, the input x is projected through 3 trainable weight matrices, WQ, WK, WV, producing 3 matrices: Q (queries), K (keys) and V (values) of dimensions $n \times d$. Then, a score is computed using these vectors in the form $Attention(Q, K, V) = Score(Q, K)V$. Applying a scaling factor leads to more stable gradients, and as the most commonly used score's function is the softmax, the scaled-dot product attention is achieved as in equation 3.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

The performance of this mechanism can even be improved by introducing multiheaded attention [25]. This concept consists of projecting matrices Q, K and V h times, using different learned projection matrices H (heads) each time.

As a result, the model's capability expands to cover a broader range of the input while allowing the attention mechanism to generate multiple representations for each input section. This enables the extraction of information that would otherwise be hard to get with a single attention head.

3. Related Work

3.1. Clinical scores

Most guideline-recommended clinical prediction rules consist of a sheet containing a list of metrics where each finding in the patient values some points. At the end of the diagnosis, all the points

are summed, and a score is obtained. How the amount of points is related to the presence of PE depends on the chosen scoring method. Still, typically the score increases with the severity and/or likelihood of PE. Some of the most relevant clinical scores are the following:

- The Daniel score [4] is based on ECGs findings and is easy to apply. It has high specificity but low sensitivity overall.
- The Novel Electrocardiographic Score (nECG) [26] is also based on ECG findings. Although it is a score challenging to perform, since its metrics are complex, it outperforms the Daniel score and most state-of-the-art clinical approaches.
- The Wells and the Revised Geneva scores differ from the previous two because their main metrics are not based on ECG's findings. The Wells Score uses subjective characteristics, while the Revised Geneva Score is based entirely on objective variables, but their overall performance is similar. Moreover, the Wells score is usually more sensitive, while the Revised Geneva score is more specific.

The metrics and correspondent points for the Daniel, nECG, Revised Geneva, and Wells score are presented in the thesis.

A comparison between the performance of these 4 approaches in the same context is exposed in table 1.

Scores	Sensitivity	Specificity	PPV	NPV	TA
nECG	**98.3**%	72.7%	83.1%	**97%**	**87.5**%
Daniel	20%	**88.6%**	70.6%	44.8%	49%
Wells	51.7%	86.4%	**83.8%**	56.7%	66.3%
Geneva	63.3%	45.5%	61.3	47.6%	55.8%

Table 1: The sensitivity, specificity, test accuracy, and predictive values of the different investigated methods when they were applied together in a similar test example[26].

3.2. Deep Learning Approaches

Applying deep learning to ECGs to interpret and diagnose cardiovascular diseases is gaining much interest. However, at the time of this research, no work was found relating the PE diagnosis of this condition using solely ECGs as input (most of the models found use CT Scan images [27]). The most relevant work found was the one developed by Sulaiman S. Somani et al. [22], where a 1D DNN extracts the main features from ECGs that are then used as input to a fusion model. This fusion model is fed not only by these ECG features but also by CT Scan findings and clinical data. Because of that, this thesis had to search for inspiration in models relating to other diseases diagnosis using only or almost only ECG findings to accomplish its goals.

In 2020, Al-Zaiti et al. [1] presented various results obtained by applying multiple machine learning-based methods for predicting underlying acute myocardial ischemia in patients with chest pain.

Many studies showed that ECG diagnosis accuracy using DNNs improves when residual blocks are added to the network's architecture compared to using CNNs alone. Because of that, based on the ResNet architecture [8], Wenxiao Jia et al. built a 1D 34-layer residual network to address the Cardiology Challenge 2020 [10]. As a result, it outperformed multiple other methods, such as CNN-based ones, and achieved a high and stable performance score measured by challenge metric, proving it can easily be used as an assisting tool for cardiologists.

In 2021, Chao Che et al. [3] proposed an end-to-end deep learning framework combining a CNN and a transformer network for ECG signal processing and arrhythmia classification. The proposal's main goal was to help cardiologists perform assisted diagnosis of heart disease and improve the efficiency of healthcare delivery. The model's performance was measured using the F1 score, achieving a score of 78,6% on this metric. Moreover, this work proved the relevance of utilizing a transformer and/or attention mechanisms to capture temporal continuity relations of the input data.

4. Data Management and Pre-processing

4.1. Data acquisition and selection

The dataset provided by HSM was highly imbalanced, containing 1014 ECG examples: 293 belonging to the positive class (with PE) and the remaining 721 belonging to the negative class. The acquisition was made through ECG devices using Dotlogic technologies. These devices apply a digital low pass filter of 40 Hz at -3 dB to all the waveforms measured, being the results then saved in a pdf file like the one shown in fig.1.

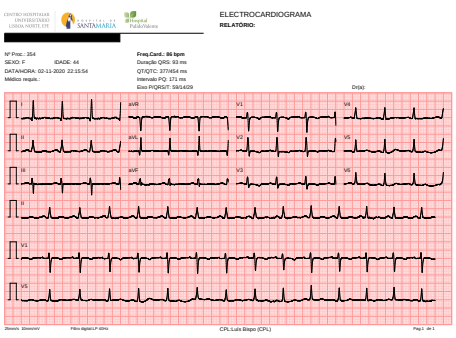

Figure 1: Acquired ECG example.

Some provided ECG examples arranged the in-

formation differently and contained hard-to-use information such as lead attaching noise. Because of that, some of the examples were disposed of. Consequently, the final useful dataset was composed of 929 samples in total, 261 belonging to the positive class and 668 to the negative class. Afterward, it was split into two independent datasets: a training set containing 826 examples (222 of the positive class and 604 of the negative one) and a test set containing 103 samples (38 of the positive class and 65 of the negative one).

The images extracted from the pdf files were excessively large at 2161 x 1121 pixels and lost a significant amount of information when downsampling or size reduction techniques were employed. So, we used raw data instead. In this work, we refer to raw data as an ECG format where instead of an image, the data is presented as numeric values corresponding to the measurements made by the ECG machine at a given sampling rate. A tool from Dotlogic (the company responsible for the ECG devices and software) was provided, which could extract this information and return it in a standard format XML file as presented in fig.2.

Figure 2: Raw ECG data .xml file example.

All the dataset's examples contained 12 leads with a length of 10 seconds each and were acquired at a fixed sampling rate of 500 Hz, corresponding to a total of 5000 data points per lead. Each lead presents the measured values ordered by acquisition, and the amplitude unit is 0.00000250 V, which means a value of 100 in the XML file corresponds to a measurement of 0.250 mV.

4.2. Pre-processing

Pre-processing was applied to make it easier for a network to use and learn from the dataset.

First, a threshold was set to prevent big outliers or spike values. This value was discussed with HSM's specialists and set to 4mV. Secondly, a median filter could be applied to reduce the baseline drift a lot of ECGs have. However, it was not needed on this dataset since, by the acquisition of the Dotlogic equipment, there was no relevant baseline drift. Thirdly, to ease some network operations and data augmentation techniques such as shifting, just 8,192 seconds of the signal (corresponding to 4096 data points on the raw data file) were used. This segment could start at any point from second 0 to second 1,808, and its starting point in time would be set the same for all the 12 leads on each example. Last but not least, normalization was applied. The most beneficial and commonly used techniques for that purpose are the Z-score normalization and the Min-Max normalization [7].

$$Zscore = \frac{x - mean(x)}{std(x)} \tag{4}$$

$$MinMax = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

4.3. Data augmentation

To address our training set, which was highly imbalanced and small, we used data augmentation techniques to artificially increase the size and diversity of the dataset [19]. Data augmentation involves applying various transformations to the original data, introducing variations, and improving the developed models' robustness, generalization, and accuracy. There are multiple data augmentation techniques on ECGs [16], but the main ones are shift, flip, scale, random drop, section drop, lead drop, sin sum, and square pulse sum.

4.4. Spectrograms

In some cases, one approach that may improve ECG's analysis is to use spectrograms [28, 13]. A spectrogram is a visual representation of the frequency content of a signal over time. It is a two-dimensional plot where the x-axis represents time, the y-axis represents frequency, and the color/intensity of each plot's point represents the magnitude of the frequency component at that correspondent segment in time. To create a spectrogram from an ECG, first, the signal is broken down into small, overlapping time windows. Then, to which one of these a DFT is obtained using the STFT method and, therefore, the frequency component is acquired for each segment in time [21]. Finally, we plot the magnitudes of these frequency components over time, resulting in a spectrogram, which can be defined as the squared magnitude of STFT as expressed in 6.

$$S(t,f) = \left| \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f} \right|^2 dt \tag{6}$$

where $S(t,f)$ is the time-frequency representation, $x(\tau)$ is the input signal and $w(t)$ is the observation window [21]. The choice of $w(t)$ can signif-

icantly impact the quality of the spectrogram obtained. The window size determines the time resolution of the spectrogram, while the window type affects the frequency resolution and the amount of spectral leakage that occurs. In general, the window size should be chosen to balance the need for good time resolution with the desire for good frequency resolution (a smaller window size will provide better time resolution but poorer frequency resolution). The choice of the window type will depend on the context we are in, but the most common ones are the Hamming window, the Blackman window, and the Tukey window (this last one is widely used in ECG's analysis).

Since the frequencies that make up the ECG are found in the low-frequency range (mostly below 50 Hz), it is usual to use a log-spectrogram to help to emphasize the lower frequency components, balance the magnitudes across the plot and reduce the impact of high-amplitude outliers, while improving its visual perception [28].

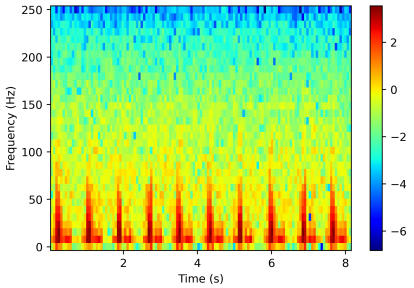

Figure 3: Log-spectrogram example of an ECG.

5. Development of the proposed model

5.1. Small 1D-DNN with residual blocks

The first chosen approach applied to the PE diagnosis case scenario was based on the work developed by Ribeiro et al. in 2020 [20], where a DNN model was trained in a dataset with more than 2 million labeled exams and outperformed resident medical doctors in recognizing 6 types of abnormalities in 12-lead ECG recordings. Specifically, it achieved F1 scores above 80% and specificity over 99%. The DNN architecture is exposed in fig.4. The model architecture is based on a small standard residual network [8], yet it adopts a slight modification to the residual blocks (it uses 4 identical modified residual blocks).

Although this work was tested on a considerably larger dataset than the one this thesis was presented with (about 2000 times larger), the promising results achieved, the simple network architecture, and the detail given by the authors relative to the training and learning process, made this idea a valuable approach. Thus, the results of this network, when applied to this thesis dataset, were used as the baseline.

5.2. Attention-enhanced residual network

From the baseline defined in the previous section, much thought was put into how a neural network can better interpret ECG findings. The first improvement that came to mind was Wenxiao Jia et al.'s approach [10] of adapting standard residual networks [8] of multiple depths to 1D format to get the network to learn more information from deeper features and by having in mind that usually networks for ECG diagnosis benefit from residual blocks. After testing several standard ResNet formats, the ResNet-18 architecture was chosen.

Beyond the capability of extracting complex relevant features from ECGs, it was also important to focus on temporal relations between segments of the input. The ECG raw data on the dataset comprises an extensive sequence of 4096 samples per lead. Thus the 1D-ResNet is not able to capture long-range context by itself. For that, the network needed, for example, a self-attention mechanism to enable a global interpretation of the input sequence.

The combination of these two powerful techniques proved to be worthy in the context of ECG analysis by the work developed by Hasan and Young in 2021 [18] through the development of a new deep learning model they called "HeartNet", whose goal was the automatic diagnosis of several heart-related diseases (mainly arrhythmia) based only on ECGs. This model uses a multi-head attention component on top of a CNN to capture long-range dependencies and temporal information between all the embeddings of the input samples. Inspired by all of this information, a new model was developed within the scope of this thesis: an attention-enhanced residual network.

This model comprises two main components: a ResNet-18, whose output is used as an input of a multi-head self-attention layer. The ResNet is responsible for acquiring deep, complex features, and the attention layer focus on the important information and relations between each embedding segment. In the end, the output of the attention layer is used as an input to a linear, fully connected block instead of a global average pooling as in the HeartNet case. This final linear layer is responsible for generating the model's prediction. The network's architecture and input data flowchart is shown in fig.5.

A deeper analysis of the way the ECG data flows through this network is the following: First, the input ECG is preprocessed using the procedures presented in section 4 and the Z-score normalization (equation 4). After this process, the ECG comprises 12 leads of 4096 samples each. From the network's point of view, this is a signal of length

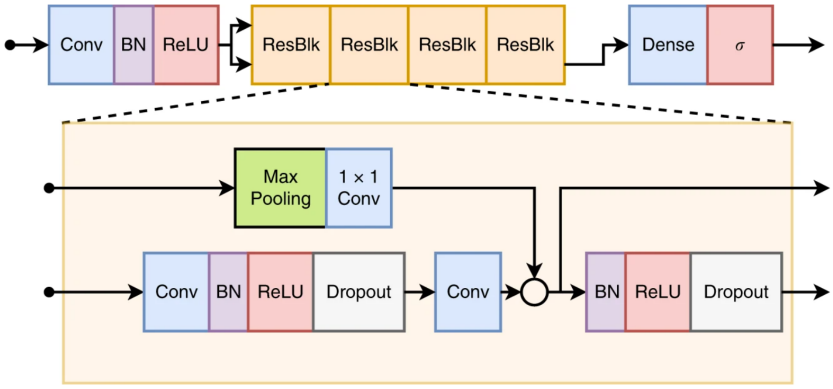

Figure 4: The small 1D-DNN residual neural network architecture [20].

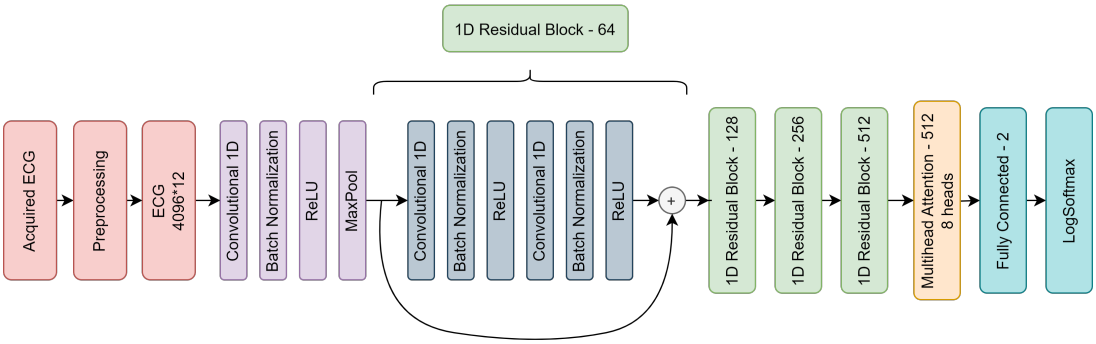

Figure 5: Developed model architecture.

4096 with 12 different channels. After that, the input goes through the entry block (purple zone on fig.5) where a convolutional kernel of size 17 and downsampling by a factor of 4 are applied, resulting in an embedding of length 1024 with 64 channels. From here, 4 residual blocks (green boxes on fig.5) are applied twice each to the samples, doubling the number of channels when passing to the following residual block and downsampling by a factor of 2 (just the first residual block pair does not apply downsampling neither doubles the number of channels). Before being fed to the multi-head self-attention layer with 4 heads, the ECG embedding has a length of 128 and 512 channels. Afterward, its output is finally fed into the classifier (fully connected layer), followed by a LogSoftmax layer. Dropout is used as a regularization method on the classifier. Notice that there are 2 outputs because the classes are one-hot encoded: class 0 represents a patient without PE, and class 1 represents a patient with it.

5.3. Spectrogram based model

Another approach is to utilize ECG's log-spectrogram representation as input data to the DNN. Notice that spectrograms have two dimensions, which makes it possible for pre-trained 2D networks to be used in this case. Nevertheless, pre-trained models such as standard ResNet demonstrated poor results when applied to this dataset, probably because they had not seen many spectrograms before.

In the research developed by Martin Zihlmann et al. [28], a special network called CRNN is applied to an ECG's spectrogram dataset. This architecture follows the same idea presented by the model developed in section 5.2, being composed by a CNN for deep, complex feature extraction and an LSTM block [23] for capturing long-term temporal dependencies. Since the architecture presented in fig.5 has all the benefits brought by this CRNN, adding even others, such as the skip connections, to understand if there is a benefit of using spectrograms over raw data ECGs, the attention-enhanced residual network was modified so it could be fed with 2D samples. Consequently, the samples flow inside the network the same way as presented in section 5.2, but the ResNet-18 performs 2D operations instead of 1D ones.

The main differences happen before the ECG is fed into the network. At first, the input ECG is preprocessed using the procedures presented in chapter 4, and the Min-Max normalization (equation 5) is applied following the recommendations of the work developed by Hongzu Li et al. [13]. Afterward, the spectrogram is built following the steps presented in section 4.4. In this work, the spectrograms are made using a Tukey window with a length of 64 samples, 50% overlap, and a shape parameter of 0.25 (standard recommendation in

the literature). Finally, a log function is applied over the resulting spectrogram intensity values to emphasize the samples' features.

6. Experimental Results

6.1. Setup of the experiments

The experiments were focused on three different models explained in section 5: the small 1D-DNN (section 5.1), the 1D-attention-enhanced residual network (section 5.2) and the spectrogram network (section 5.3). These were developed in PyTorch and trained using an NVIDIA 32GB V100S installed in a DELL PowerEdge C41402 server at INESC-ID.

All these models were trained on the PE training set presented in section 4. This dataset was split into a training set (90%) and a validation set (10%) to perform independent validation. The test set was kept out of the training routine and used to evaluate generalization and overall performance. The training set was loaded in the CSV format provided by the Dotlogic software. The extraction of each lead content, preprocessing, and data augmentation were done, in that order, on each batch before the samples were fed to the network. The pipeline was the same in the spectrogram network case, but a log-spectrogram was built from each lead. After acquiring the 12 log-spectrogram (one per lead), the sample is fed to the network as an image with 12 channels. Both pipelines are illustrated in fig.6 and fig.7.

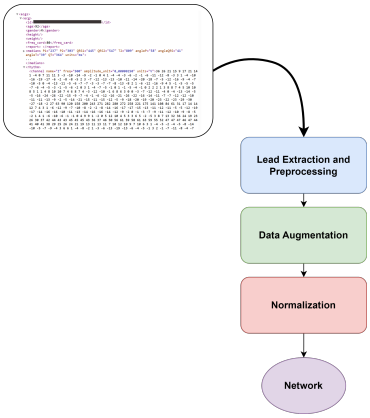

Figure 6: Raw data pipeline before going into the network.

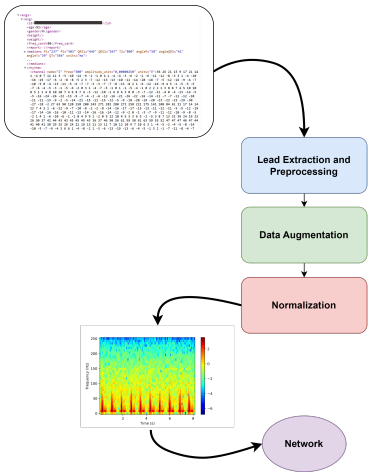

Figure 7: Spectrograms pipeline before going into the network.

During the data augmentation process, a shift operation was consistently performed on each sample. Three possible shifts were implemented: starting the sample at 0 seconds, 0.904 seconds, or 1.808 seconds. Each shift had an equal probability of approximately 33.3%. In addition to this operation, one of the data augmentation techniques mentioned in section 4.3 could be applied with an equal probability of $\frac{1}{n_{aug}+1}$, where n_{aug} represents the number of available augmentations. It is important to note that the selected type of shift and the accompanying data augmentation technique remain consistent across all leads within a single sample.

The most important metric considered to compare performances on the test set was the PPV vs. recall tradeoff (maximum confidence when identifying a positive patient while catching the maximum possible number of positive PE patients). Nevertheless, other metrics were used to help the evaluation, such as specificity, F1-score, and overall accuracy.

Regarding hyperparameters and training routines, all models were trained several times during 50 epochs to perform hyperparameter tuning. Those were chosen among the following options: kernel size of the entry layer - odd numbers from 3 to 31, batch size - [4,8,16,32,64,128], initial learning rate - [0.01,0.005,0.002,0.001,0.0005,...,0.00001], optimizers - [SGD,Adam,AdamW], dropout probability - [0,0.2,0.5,0.8].

6.2. DNN's achieved results

After tuning the hyperparameters and determining the final training conditions for each model, the selected values were used to train each network for 100 epochs. All weights were initialized following the Xavier uniform distribution, and biases were assigned 0 value. Each model's state was saved every time PPV and recall were at least 0.6 and

0.3, respectively. From these models' states, the best one for each model was chosen considering all the metrics mentioned before. The best state for each model was achieved using a kernel size in the first layer of 17, batch size of 32, the AdamW optimizer [15], and dropout probability of 0.5.The chosen learning rate was different between models. Regarding the focal loss, the values defined were $\gamma = 2$ and $\alpha = [0.3, 0.7]$. The training and validation loss are exposed in the thesis.

Finally, the best models had their performance evaluated on the test set, whose results are revealed in the table 2. The 1D attention-enhanced residual network performed the best among these three models, surpassing the baseline.

Table 2: Developed models' best results on the test set.

Models	Small 1D-DNN	1D-Attention ResNet	2D-Attention ResNet
Learning rate	0.001	0.0002	0.0001
Sensitivity, %	30.77	48,72	46.15
Specificity, %	95.31	96.88	81.25
PPV, %	80.00	90.48	60.00
NPV, %	69.32	75.61	71.23
F1-score, %	44.45	63.34	52.17
Accuracy, %	70.87	78.64	67.96

Regarding the spectrogram analysis, we noticed it was more difficult for the network to extract ECG findings associated with PE than when dealing with raw data only. This probably means temporal-related features are more meaningful when diagnosing PE than frequency-related ones, so it does not pay off to trade some temporal features for more information on the spectrum.

6.3. Comparison between AI and clinical approaches

Once the final developed model was trained and set (the 1D-ResNet enhanced with attention), it was compared against the guideline-recommended clinical prediction rules for PE. Those results are compiled in the table 3.

Notice that the AUC value is around 0.75, which is superior to all the clinical approaches' scores acquired on this metric. The ROC is represented in fig.8.

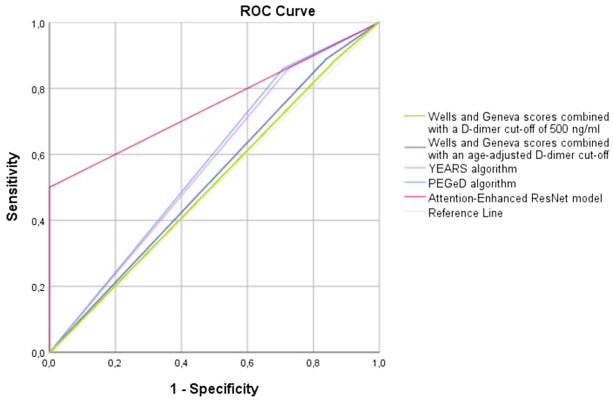

Figure 8: ROC curve demonstrating the diagnostic performance of different decision rules to predict pulmonary embolism..

Although Daniel's score was applied to the test set, it was not feasible to determine a specific threshold for distinguishing positive from negative examples due to the similarity in the score's median between patients with and without PE. Consequently, no results from the Daniel score were included in this research.

Despite the developed DNN exhibiting lower sensitivity compared to clinical approaches, it compensates for this with high specificity and PPV. This indicates that when the DNN predicts a patient belongs to class 1, it provides a high level of confidence that the patient has PE.

7. Discussion

The management of PE in emergency assistance and pre-hospital situations is critical to decreasing the mortality rate. Given the potential benefit of treatment approaches when it is known someone has the disease, every effort should be made to diagnose PE, especially acute PE, quickly and accurately. In the scope of this thesis, a review of ECG analysis using deep learning was performed with a special focus on PE analysis. As a result, a 1D Attention-enhanced ResNet for PE detection using 12-lead ECG was developed, which achieved high specificity and PPV on a test set where multiple current clinical approaches do not come close. This is important for two main reasons. Firstly, high specificity means people who do not have the disease will be less likely considered a false positive, and, on the other hand, a high PPV gives confidence to a prediction of a positive patient to PE. Secondly, multiple exams, such as the D-dimer blood test for PE, have extremely high sensitivity for PE but very low specificity. Because of that, applying this model to a patient subjected to a D-Dimer test highly assists medical decisions toward a good diagnosis (we get a joint test with high specificity and high sensitivity). Unfortunately, at the time of this research, to the author's knowledge, no previously deep learning model was developed for PE diagnosis using only ECGs. This can probably be explained by the lack of labeled ECG examples positive to PE and by the fact that, unlike other conditions, such as some arrhythmias where just one ECG waveform is needed to detect the anomalies, PE needs a longer sequence in time.

Given the small and highly imbalanced dataset provided for this work, extracting relevant ECG findings from the samples and, consequently, achieving even better results than the ones accomplished was extremely difficult. One reason that can explain the lack of positive samples is the little importance many hospitals give to labeling. Information is power, and although there are many pos-

Table 3: Diagnostic accuracy of Wells and Geneva scores combined with a fixed and an age-adjusted cut-off, YEARS algorithm and PEGeD algorithm to predict pulmonary embolism.

Metrics	Wells score +DD threshold of 500 ng/mL	Geneva score+DD threshold of 500 ng/mL	Wells score+age-ajusted DD cut-off	Geneva score+age-ajusted DD cut-off	YEARS algorithm	PEGeD algorithm	Attention-Enhanced ResNet model
Sensitivity, % (95% CI)	89.47 [75.20-97.06]	89.47 [75.20-97.06]	89.47 [75.20-97.06]	89.47 [75.20-97.06]	86.84 [71.97-95.59]	86.84 [71.97-95.59]	50.00 [33.38-66.62]
Specificity, % (95% CI)	12.31 [5.47-22.82]	12.31 [5.47-22.82]	18.46 [9.92-30.03]	18.46 [9.92-30.03]	29.23 [18.60-41.83]	30.77 [19.91-43.45]	100 [94.48-100.00]
PPV, % (95% CI)	37.36 [27.44-48.13]	37.36 [27.44-48.13]	39.08 [28.79-50.13]	39.08 [28.79-50.13]	41.77 [30.77-53.41]	42.31 [31.19-54.02]	100 [82.35-100.00]
NPV, % (95% CI)	66.67 [34.89-90.08]	66.67 [34.89-90.08]	75.00 [47.62-92.73]	75.00 [47.62-92.73]	79.17 [57.85-92.87]	80.00 [59.30-93.17]	77.38 [66.95-85.80]
AUC, % (95% CI)	0.51 [0.39-0.63]	0.51 [0.39-0.63]	0.54 [0.43-0.65]	0.54 [0.43-0.65]	0.58 [0.47-0.69]	0.59 [0.48-0.70]	0.75 [0.64-0.86]

itive to PE ECGs examples on the HSM database, most of them are not labeled since cardiologists, after analyzing each sample, do not save their interpretation and, as a consequence, it is tough to search for a significant number of positive examples.

All in all, and despite all the limitations, this work can potentially assist medical decisions on PE diagnosis.

8. Future Work

PE is difficult to diagnose since its findings and symptoms are similar to many other conditions, such as heart attack. Because of that, external sample validation with more positive PE samples and different condition samples is critical to substantiate the results.

The analysis would be very beneficial to be made directly from pictures taken at an ECG or, at least, from the pdf files the ECG software provides. This way, a tool based on a model like the one here developed could be accessible from any hospital. Furthermore, it would not depend on the Dotlogic technology (not every hospital uses the Dotlogic software).

Finally, keeping the network's performance while analyzing fewer leads would be interesting. For example, when diagnosing, doctors do not look at all the leads but only at 4 or 5 that contain the most relevant information. Imagining a farfetched scenario where only 1 lead would be sufficient to diagnose PE with a considerable performance, a model like the one we developed could receive a single lead ECG that could be acquired, for instance, from a smartphone.

I expect such developments in deep learning and clinical practice could enhance the improvement and confidence of the PE diagnosis.

9. Acknowledgement

This work was partly supported by Center for Responsible AI - Application number: C645008882-00000055 and national funds through FCT - Fundação para a Ciência e Tecnologia, under project UIDB/50021/2020.

References

[1] S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. Frisch, C. Martin-Gill, R. Gregg, S. Saba, C. Callaway, and E. Sejdić. Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nature Communications*, 11(1):3966, Aug. 2020.

[2] M. AlGhatrif and J. Lindsay. A brief review: history to understand fundamentals of electrocardiography. *Journal of Community Hospital Internal Medicine Perspectives*, 2(1):14383, Jan. 2012.

[3] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin. Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21(1):184, Dec. 2021.

[4] K. R. Daniel, D. M. Courtney, and J. A. Kline. Assessment of Cardiac Stress From Massive Pulmonary Embolism With 12-Lead ECG. *Chest*, 120(2):474–481, Aug. 2001.

[5] T. v. Dongen. Demystifying efficient self-attention, Nov. 2022.

[6] E.-O. Essien, P. Rali, and S. C. Mathai. Pulmonary Embolism. *The Medical Clinics of North America*, 103(3):549–564, May 2019.

[7] W. Hao and K. Jingsu. Investigating deep learning benchmarks for electrocardiography signal processing. *arXiv pre-print 2204.04420*, 2022.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] K. Janocha and W. M. Czarnecki. On Loss Functions for Deep Neural Networks in Classification. *Schedae Informaticae*, 25, 2017.

[10] W. Jia, X. Xu, X. Xu, Y. Sun, and X. Liu. Automatic detection and classification of 12-lead ecgs using a deep neural network. In *Computing in Cardiology*, pages 1–4, 2020.

[11] E. Khan. Clinical skills: the physiological basis and interpretation of the ECG. *British Journal of Nursing*, 13(8):440–446, Apr. 2004.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998.

[13] H. Li and P. Boulanger. Structural Anomalies Detection from Electrocardiogram (ECG) with Spectrogram and Handcrafted Features. *Sensors*, 22(7):2467, Mar. 2022.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, Feb. 2020.

[15] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. *arXiv pre-print 1711.05101*, Jan. 2019.

[16] N. Nonaka and J. Seita. Randecg: Data augmentation for deep neural network based ecg classification. In *Advances in Artificial Intelligence*, pages 178–189, Cham, 2022. Springer International Publishing.

[17] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training Recurrent Neural Networks. In *30th International Conference on Machine Learning, ICML 2013*, volume 28.

[18] T. H. Rafi and Y. Woong Ko. Heartnet: Self multihead attention mechanism via convolutional network with adversarial data synthesis for ecg-based arrhythmia classification. *IEEE Access*, 10:100501–100512, 2022.

[19] A. Raghu, D. Shanmugam, E. Pomerantsev, J. Guttag, and C. M. Stultz. Data augmentation for electrocardiograms. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 282–310. PMLR, 07–08 Apr 2022.

[20] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, Apr. 2020.

[21] E. F. Shair, S. A. Ahmad, A. R. Abdullah, M. H. Marhaban, and S. B. M. Tamrin. Selection of Spectrogram's Best Window Size in EMG Signal During Core Lifting Task. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1-16):81–85.

[22] S. S. Somani, H. Honarvar, S. Narula, I. Landi, S. Lee, Y. Khachatoorian, A. Rehmani, A. Kim, J. K. De Freitas, S. Teng, S. Jaladanki, A. Kumar, A. Russak, S. P. Zhao, R. Freeman, M. A. Levin, G. N. Nadkarni, A. C. Kagen, E. Argulian, and B. S. Glicksberg. Development of a machine learning model using electrocardiogram signals to improve acute pulmonary embolism screening. *European Heart Journal - Digital Health*, 3(1):56–66, Mar. 2022.

[23] R. C. Staudemeyer and E. R. Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks. *arXiv pre-print 1909.09586*, 2019.

[24] B. Valente Silva, J. Marques, M. Nobre Menezes, A. L. Oliveira, and F. J. Pinto. Artificial intelligence-based diagnosis of acute pulmonary embolism: Development of a machine learning model using 12-lead electrocardiogram. *Revista Portuguesa de Cardiologia*, 2023.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] A. Vereckei, A. Simon, G. Szénási, G. Katona, L. Hankó, M. Krix, V. B. Szőke, V. Baracsi Botos, Z. Járai, and T. Masszi. Usefulness of a Novel Electrocardiographic Score to Estimate the Pre-Test Probability of Acute Pulmonary Embolism. *The American Journal of Cardiology*, 130:143–151, Sept. 2020.

[27] S. Vijayachitra, K. Prabhu, M. Abarana, A. Deepa, and L. Loga Priya. Deep Learning Technique-Based Pulmonary Embolism (PE) Diagnosis. In *Advances in Electrical and Computer Technologies*, pages 695–702, 2022.

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