TÉCNICO LISBOA

Diagnosing pulmonary embolism from electrocardiograms

João Pedro dos Santos Marques

Thesis to obtain the Master of Science Degree in

Bologna Master Degree in Electrical and Computer Engineering

Supervisor: Prof. Arlindo Manuel Limede de Oliveira

Examination Committee

Chairperson: Prof. João Manuel de Freitas Xavier
Supervisor: Prof. Arlindo Manuel Limede de Oliveira
Member of the Committee: Prof. Maria Margarida Campos da Silveira

June 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Agradecimentos

Começo por agradecer aos meus pais, Teresa e Raúl, uma vez que tudo começou com eles. Desde o meu nascimento ao presente término do meu curso, enquanto fomentavam a descoberta de valores, propósitos e gostos, fizeram sempre por me guiar, aconselhar e acompanhar da melhor forma que conseguiram e souberam. Além disso, foram também os principais investidores no financiamento da minha vida académica, permitindo-me a aquisição de vivências múltiplas sem grandes preocupações associadas. Como o meu pai costuma dizer: ”A vida é feita de várias portas abertas que se vão fechando a cada escolha que fazes”. Graças a eles, fui capaz de manter as portas importantes abertas e não me arrepender das que se foram fechando.

Um grande obrigado à minha namorada, Beatriz, a qual foi a minha principal companhia, presente e remota, durante todo o meu percurso universitário, em particular na redação deste trabalho. Não há palavras que descrevam a tua paciência para comigo nos dias em que o sono me faltou e a má-disposição se apoderou do meu corpo. Obrigado pela tua capacidade de descomplicação das minhas dúvidas e hesitações. Conto contigo para me acompanhares em muitas mais aventuras no futuro.

Agradeço muito ao meu orientador, prof. Arlindo, o qual me apoiou de várias maneiras durante todas as etapas desta dissertação e me proporcionou um grande crescimento profissional. Além disso, fez crescer em mim um gosto ainda maior pela área de inteligência artificial. Agradeço também os seus conselhos, paciência, compreensão e disponibilidade total.

Obrigado ao Hospital de Santa Maria, à Dra. Beatriz e ao Dr. Miguel, por me apoiarem na análise clínica deste problema. Um especial obrigado à Dra. Beatriz pela sua simpatia e disponibilidade.

Aos meus amigos e família, obrigado pela companhia e apoio. Sei que não fui fácil de lidar durante a redação deste trabalho. Graças a vós, esta etapa passou a correr.

Agradeço ao Técnico por me fazer desenvolver qualidades, tais como o pensamento crítico, o olhar atento e a resiliência. No entanto, o maior ensinamento que desta casa levo é: ”Ninguém faz nada sozinho”. Isto porque, apesar das aulas e professores me terem apresentado as ferramentas essenciais à minha formação, foi com amigos e colegas da Instituição que as aprendi a utilizar durante as noitadas de preparação de laboratórios, projetos e exames.

Por fim, agradeço ao INESC-ID e aos colegas que lá conheci, os quais me foram aconselhando bastante ao longo deste trabalho.

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.

Resumo

A embolia pulmonar (EP) é uma das causas mais comuns de mortalidade associada a doenças cardiovasculares no mundo. Apesar disso, os seus métodos clínicos de diagnóstico apresentam baixa especificidade e necessitam, muitas vezes, de confirmação através de angiografia pulmonar por tomografia computorizada, a qual traz vários inconvenientes, tais como a exposição de pacientes a radiação, o custo monetário elevado e a impossibilidade de aplicação em vários contextos como o pré-hospitalar. Posto isto, a eletrocardiografia (ECG), sendo uma ferramenta de monitorização do estado cardiovascular do paciente de acesso fácil e rápido, demonstra um grande potencial clínico neste contexto. Com efeito, e tendo em conta os benefícios da aprendizagem profunda em áreas como análise de imagem, processamento de língua natural e processamento de sinal, as redes neuronais podem beneficiar bastante o diagnóstico de EP a partir de ECGs. Para tal, utilizou-se um conjunto de dados extraído da base de dados do Hospital de Santa Maria, constituído por 929 exemplos no total: 261 exemplos positivos e os restantes 668 negativos. Uma vez que, à data da escrita desta dissertação, nenhum estudo abordou o diagnóstico de EP através de redes neuronais utilizando apenas ECGs, recorreu-se a um modelo de diagnóstico de arritmias como base de referência e desenvolveu-se um modelo novo para diagnosticar EP, composto por uma ResNet-18 com uma camada de *self-multihead attention*. Duas versões deste modelo foram desenvolvidas: uma 1D que recebeu os dados originais processados (valores numéricos) dos ECGs e uma 2D que processou os espectrogramas dos ECGs. O desempenho da versão 1D superou tanto o modelo de referência como a versão 2D. Como consequência, esta versão foi escolhida para comparar o seu desempenho com as métricas de previsão clínica recomendadas, apresentando resultados superiores ao nível da especificidade (100%; 95% IC: 94-100), do valor preditivo positivo (100%; 95% IC: 82.35-100.00) e da área sob a curva (0.75; 95% CI: 0.66-0.82), e provando, assim, que as técnicas de aprendizagem profunda podem contribuir para a melhoria da assistência médica no diagnóstico de EP.

Palavras-chave

Embolia pulmonar (EP), eletrocardiogramas (ECG), aprendizagem profunda, ResNet, ECG's espectrogramas

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, residual networks, self-multihead attention, ECG's spectrograms

Contents

List of Tables

List of Figures

Abbreviations

1D	1-dimensional
2D	2-dimensional
ANN	Artificial Neural Networks
AUC	Area under the curve
CI	Confidence Interval
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
CT Scan	Computed Tomography Scan
DD	D-dimer
DFT	Discrete Fourier Transform
DNN	Deep Neural Networks
DVT	Deep Vein Thrombosis
ECG	Electrocardiogram
FN	False Negative
FP	False Positive
HSM	Hospital de Santa Maria
INESC-ID	Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa
LSTM	Long short-term memory
nECGs	Novel Eletrochardiographic Score
NPV	Negative Predictive Value
PE	Pulmonary Embolism
PEGeD	Pulmonary Embolism Graduated d-Dimer

PPV	Positive Predictive Value
RBBB	Right Bundle Branch Block
ReLU	Rectified Linear Unit
ResNet	Residual Network
STFT	Short-time Fourier Transform
TA	Test Accuracy
Tanh	Hyperbolic Tangent
TN	True Negative
TP	True Positive
ROC	Receiver operating characteristics
WHO	World Health Organization

1

Introduction

Contents

1.1 Motivation

Pulmonary Embolism (PE) is the third most common cause of cardiovascular myopathy death worldwide, just after stroke and heart attack [1]. It represents 2% to 5% of all causes of out-of-hospital cardiac arrest and is associated with a highly unfavorable prognosis. According to the World Health Organization (WHO), PE caused an average of almost 40,000 deaths in Europe every year between 2013 and 2015[1]. 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, most commonly by a thrombus or, more rarely, from the embolization of other materials into the pulmonary circulation, such as fat, air, or tumor cells [2]. When caused by a thrombus, it can be a consequence of a Deep Vein Thrombosis (DVT). DVT is a medical condition characterized by the formation of a blood clot within a deep vein in the body (usually developed in the lower leg, thigh, and pelvis). In this situation, when a part of the clot breaks off and travels through the bloodstream reaching the lungs, PE may develop. If the clot is small and treatable, people can recover from it with no more than some lung damage. However, if it is large, it can block the blood flow and be fatal[2].

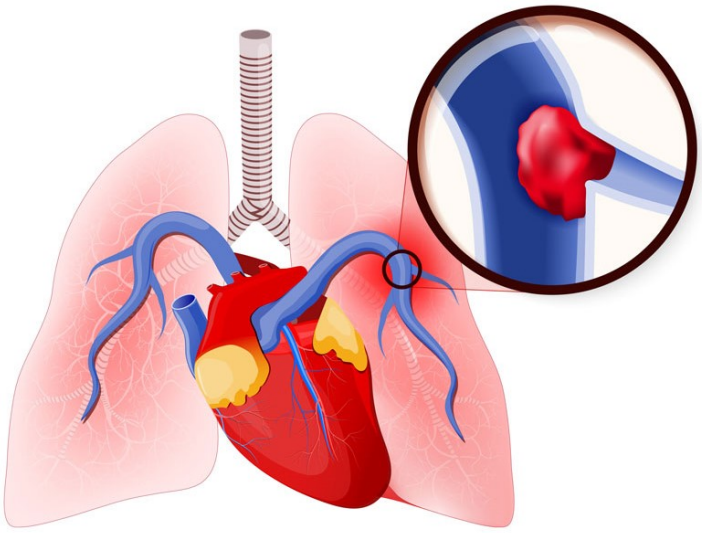


Figure 1.1: A representative drawing of a pulmonary embolism [1].

[1] Visit https://www.prnewswire.co.uk/news-releases/isth-data-from-the-world-health-organization-mortality-database-reveal-fewer-deaths-from-pulmonary-embolism-in-the-european-region-over-the-past-15-years-840655850.html for details. Accessed in March 2023

[2] Visit https://www.cdc.gov/ncbddd/dvt/facts.html for details. Accessed in May 2023.

As in many other diseases, the faster PE is noticed, the better consequences can be prevented by treating such as anticoagulants, thrombolytics, and even clot removal. However, diagnosing PE may be challenging because the symptoms are like many other diseases, such as heart attack and pneumonia. To find blood clots, based on each patient's risk, some tests are often done, such as a chest X-Ray, Computed Tomography Scan (CT Scan) and Electrocardiogram (ECG)[3]. Although CT Scan 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 undesirable). Because of that, electrocardiograms 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 [3, 4]. There are multiple types of scores and various types of PE severity. Therefore, diagnosing PE 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 all of these factors, building a model which can perform an analysis of an ECG, inferring from lots of different features if a patient has PE or not, may be beneficial. I believe such a tool can support medical decisions and improve diagnostic accuracy (e.g., if used together with some electrocardiographic score performed by the doctor), enhancing, as a consequence, pulmonary embolism prevention and its faster treatment. I present this decision support system built using neural network technology.

1.2 Objectives

This dissertation focuses on improving the diagnosis of PE using Deep Neural Networks (DNN)s. Currently, the commonly used clinical methods for predicting the disease have high sensitivity and low specificity [3]. Because of that, cardiologists ask for something more specific, which can increase the certainty with which a diagnosis is made.

Numerous studies have focused on utilizing ECGs for diagnosing various diseases [5]. However, exploring ECG for diagnosing PE remains relatively limited. Thus, it is crucial to emphasize the significance of leveraging existing Deep Learning technologies and tailoring them to this specific context of ECG analysis for PE diagnosis.

The main objectives of this work can be summarized as follows:

- Train multiple DNNs architectures over data related to patients with PE;
- Evaluate the performance of each architecture, comparing it with other clinical approaches;
- Find the best way to look over the dataset, perform data augmentation and feed it to the DNNs;

[3] Visit https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/pulmonary-embolism/diagnosis.html for details. Accessed in May 2023

- Discuss the achieved results.

Note that the achieved results in this thesis prove that, despite the problem's difficulty, it is possible to improve PE diagnosis current results by the use of Deep Learning.

1.3 Contributions

The original contributions of this thesis are presented as follows:

- New model's architecture for PE diagnosis; All the code and documentation is publicly available at:
  - https://github.com/Sargazzo/PE-diagnosis-from-ecgs-Deep-Learning
- Review over ECG analysis using deep learning;
- Review over ECG preprocessing, data management, and dealing with imbalanced datasets.

The contributions of this thesis have been partially published in the "Portuguese Journal of Cardiology" [6].

1.4 Thesis Outline

This dissertation comprises a sequence of chapters that cover the essential steps undertaken in this research.

Chapter 2 introduces the necessary knowledge about deep learning theory and PE diagnosis.

Chapter 3 covers the related work, starting from exploring clinical approaches to PE and leading up to previous deep learning research on the field on the subject.

Chapter 4 centers around the dataset used in this study and provides detailed insights into its acquisition and management processes.

Chapter 5 offers insights into how the related work influenced the design of the network, delving further into the thought behind each component and the chosen approach.

Chapter 6 presents crucial information about the experiments, including the experimental setup, where the models were trained, the training methodology employed, and the achieved results. Furthermore, it compares the proposed model and multiple clinical approaches.

Chapter 7 presents some final remarks and suggestions for improvements on this subject for the future.

2

Background

Contents

This chapter provides the essential background on deep learning theory, electrocardiography, clinical approaches, and essential metrics for evaluating the performance of models in this context.

2.1 Pulmonary Embolism Diagnosis and Detection

As mentioned earlier, diagnosing PE can be difficult due to symptoms shared with various other diseases. However, in an optimal scenario, its diagnosis and detection follows some steps[1]:

1. First, the person may experience symptoms such as chest pain (most common), sudden cough that can produce blood or bloody mucus, dizziness... After experiencing this, the person should go to a doctor;

2. Secondly, his/hers doctor may discuss his/hers medical history and the symptoms of the patient, make them perform a physical exam, and order several tests such as an electrocardiogram and/or a CT scan;

3. Finally, by gathering and analyzing all the results obtained, the doctor will take his/hers conclusions and decide whether to diagnose or not the patient with PE (the diagnostics' result is binary: either the patient has the disease or he has not).

The creation of a good performing autonomous tool for supporting the evaluation of the presence of PE in a particular patient from ECG is relevant due to the ease of performing an ECG and its availability in a hospital or a prehospital context, while it does not make patients go through radiation exposure like a CT Scan does.

2.2 Electrocardiography

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 [7, 8].

Electrocardiography is especially important in diagnosing abnormal origins of cardiac activation and conduction abnormalities [9]. Note that by analyzing data from several ECG electrodes, a graphical representation of the heart's electrical activity can be acquired. Consequently, each lead's ECG recording is different in shape since each one is recording the heart's electrical activity from a different perspective.

[1] Visit https://utswmed.org/conditions-treatments/acute-pulmonary-embolism/ for details. Accessed in May 2023.

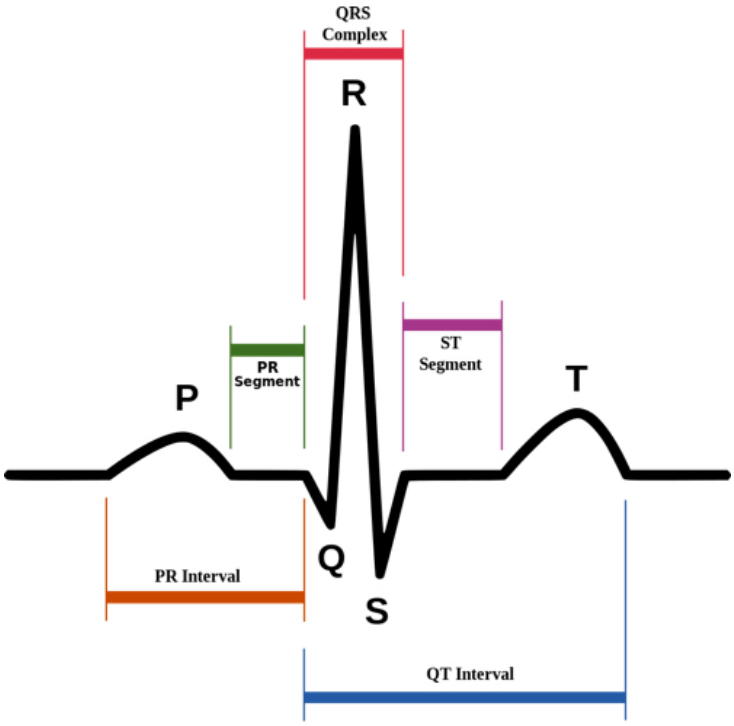


Figure 2.1: The ECG multiple components[2].

The ECG has multiple components that can be identified:

- P waves - represent atrial depolarisation;
- PR interval - begins at the start of the P wave and ends at the beginning of the Q wave;
- QRS complex - represents depolarisation of the ventricles;
- ST segment - starts at the end of the S wave and ends at the beginning of the T wave;
- T wave - represents ventricular repolarisation;
- RR interval - begins at the peak of one R wave and ends at the peak of the next R wave;
- QT interval - begins at the start of the QRS complex and finishes at the end of the T wave.

Since a single ECG rhythm strip only contains information recorded from one perspective of the heart, it is worth creating a configuration of multiple rhythm strips containing information recorded from many different perspectives. One of the most used configurations is the 12-lead ECG. As the name suggests, 12 different electrodes are placed all over the patient to record multiple waveforms of different sensitivity according to fig.2.2. According to their placement, these leads have the following identifiers: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, and V6. The 12-lead ECG makes it possible to analyze the

[2]Retrieved from https://geekymedics.com/understanding-an-ecg/. Accessed in April 2023

major shape of QRS complexes, the ventricular and atrial rate, the precordial lead, the limb lead, and others [8].

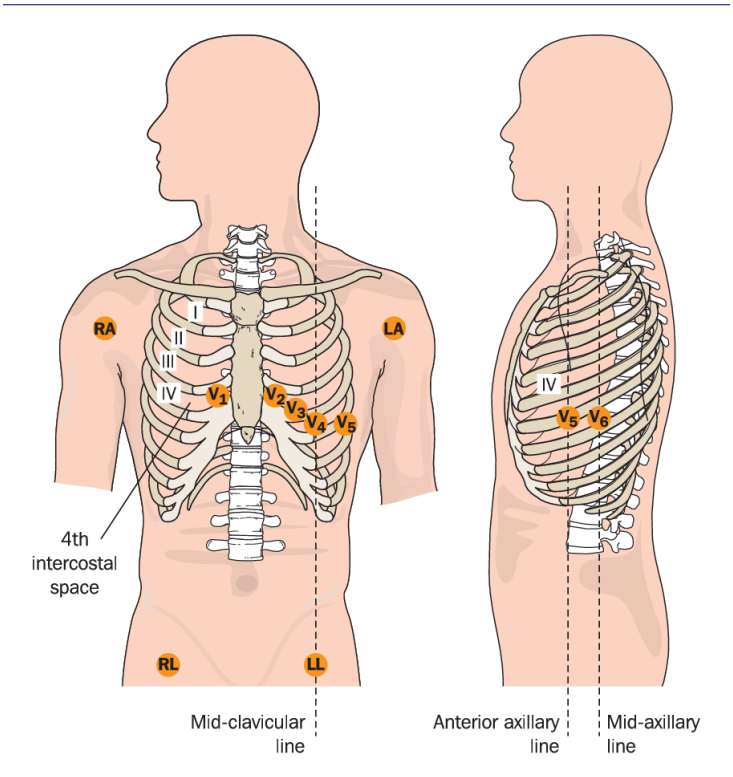


Figure 2.2: The ECG 12-lead placement on a patient [8].

Previous studies that applied ECG scoring systems for acute PE diagnosis, mostly based on 12-lead ECG, demonstrated a reasonable performance (especially in sensitivity terms) but performed poorly in specificity terms. Moreover, their application is time-consuming, requires expertise in ECG interpretation, and is not routinely implemented in clinical practice.

2.3 Relevant performance metrics

To infer if a model or test/exam has a good performance in diagnosing the disease, there are some important terms and metrics widely used [10]:

- True Positive (TP): the number of cases correctly identified as positive for disease;
- False Positive (FP): the number of cases incorrectly identified as positive for disease;
- True Negative (TN): the number of cases correctly identified as negative for disease;
- False Negative (FN): the number of cases incorrectly identified as negative for disease;

- Sensitivity: percentage of people that were correctly identified with the disease;
- Specificity: percentage of people that were correctly identified without the disease;
- Positive Predictive Value (PPV): percentage of the cases giving positive test results which are truly positive for disease;
- Negative Predictive Value (NPV): percentage of the cases giving negative test results which are truly negative for the disease.
- Receiver operating characteristics (ROC) - graphical representation used to assess the performance of a binary classification model. It evaluates the trade-off between the true positive rate (sensitivity) on the y-axis and the false positive rate (1 - specificity) on the x-axis.
- Area under the curve (AUC) - Metric to quantify the performance of a classification model. It measures the area under the ROC curve, ranging between 0 and 1, where a higher value indicates better discrimination ability.

We will also use some Machine Learning related metrics that translate the metrics above and help the improvement and focus of the model:

- Accuracy - Represents the percentage of the correct outputs of the model over all its outputs:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{2.1}$$

- Precision - Represents the same as the PPV:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.2}$$

- Recall - Represents the same as sensitivity:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.3}$$

- F1 score - represents the harmonic mean of precision and sensitivity:

$$F1score = \frac{2 \times precision \times recall}{precision + recall} \tag{2.4}$$

2.4 Deep Learning

Deep learning is a subfield of machine learning that uses Artificial Neural Networks (ANN) to model and solve complex problems. An ANN uses multiple layers of neurons to create a hierarchical representation of the input data, allowing the network to learn increasingly abstract and more complex features as the data flows through the layers. This section revises important core concepts regarding the subject.

2.4.1 Learning Through Training

The main objective of a machine learning model is to generate a meaningful output from a given input. Considering a model f, this happens through a process called training where the aim of f is to minimize a loss function $\mathcal{L}(\hat{y}, y)$, where $\hat{y}$ is a produced output of $f(x)$, x being an input of the model. Suppose a significant number of examples (x, y) is given to the model. In that case, it should be able to learn the task by following, for example, the stochastic gradient descent algorithm [11]. This is possible by updating each model's parameter, θ, for each example or set of multiple instances (mini-batch) by back-propagation. The steps for performing the back-propagation algorithm and updating the weights θ are the following:

1. Input an input or a set of inputs to the model;
2. Perform forward propagation;
3. Compute the loss between the outputs y and the targets $\hat{y}$, $\mathcal{L}(\hat{y}, y)$;
4. Perform backpropagation and get the gradient loss;
5. Update the model's parameters θ.

In the end, the model's performance is evaluated using two disjoint sets of the dataset: the test set and the training set. The training set, as the name implies, is used for training the model and performing the back-propagation algorithm, while the test set is reserved to evaluate the model's ability to perform and generalize well.

It is also essential for the model to use an activation function since it enables it to learn more complex patterns in the input data. It consists of a non-linear operation applied at the end of each given network layer, giving the model the capability of approximating non-linear functions. Otherwise, it could only learn linear transformations, restricting its classification ability.

Multiple non-linear activation functions can be used, and their choice depends on the problem we are facing, but the most common ones are Rectified Linear Unit (ReLU), Sigmoid, and Hyperbolic Tangent (Tanh) [12].

2.4.2 Loss Functions and Focal Loss

As said in the previous segment, the goal of a loss function is to measure 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 real one.

There are many different loss functions. The choice of which we should use depends on the type of task we are working on. For example, in classification tasks, such as the diagnosis of a condition where the goal is to predict if a patient either has a disease or not (either belongs to one class or another), a common loss function used is the cross-entropy loss [13]. This loss function quantifies the difference between two probability distributions: predicted and real/true.

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i), \text{for n classes} \tag{2.5}$$

In the equation 2.5, the cross-entropy calculation is formally described with t_i being the actual label and p_i the softmax probability for the i^{th} class.

Although cross-entropy is one of the most used loss functions, it does not consider dataset imbalance. Therefore, when faced with this situation, it will make the model benefit the most common class in the dataset rather than the others, leading to poor performance in detecting the minority class. For instance, if we have two classes, a and b, and class a is four times more common than class b, guessing all examples that belong to class a will lead to a performance way superior when compared to guessing all instances belong to class b.

A loss function known as Focal Loss was introduced to mitigate the class imbalance in machine learning tasks [14]. It is designed to give more importance to complex or misclassified samples by applying a modulating term to the cross-entropy loss - it is a dynamically scaled cross-entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. 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 samples [14] as presented in equation 2.6.

$$FL_{p_t} = -(1 - p_t)^\gamma log(p_t) \tag{2.6}$$

Another term that can be added to the equation 2.6 to better handle the class imbalance problem in binary classification is a hyperparameter α_t.

$$\alpha_t = \begin{cases} \alpha, if\ class = 1 \\ 1 - \alpha, otherwise \end{cases} \tag{2.7}$$

Using this α_t presented in the equation 2.7, 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 2.8.

$$FL_{p_t} = -\alpha_t(1-p_t)^{\gamma}log(p_t) \tag{2.8}$$

2.4.3 Convolutional Neural Networks

Perhaps the earliest precursor of the Convolutional Neural Network (CNN) was the Neocognitron proposed by Fukushima [15], which presented concepts such as feature extraction, pooling layers, and using convolution in a neural network. However, the CNN name originated with the design of a model called LeNet developed by LeCun et al. [16]. This work demonstrated that a CNN model, which gathers simpler features into progressively more complicated ones, can be successfully used for handwritten character recognition (this was tested using the MNIST database of handwritten digits), being largely developed between 1989 and 1998. Nevertheless, it was only in 2012 that this architecture became one of the most well-known techniques applied to image recognition when Krizhevsky et al. built one deep CNN named AlexNet [17] that was, for the first time, more successful than traditional hand-crafted feature learning on the ImageNet Dataset.

A CNN consists of a class of neural networks that specializes in processing data that has a grid-like topology, such as an image[3]. An image is seen digitally as a binary representation of visual data where each pixel is represented by a value, which denotes its brightness and color in the image grid. The goal of the CNN is to predict a specific output by interpreting this input image. This is done in 3 different steps: the convolutional layer, the pooling layer, and the fully connected layer.

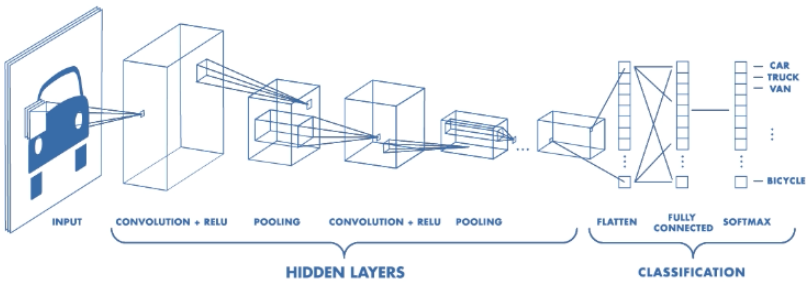


Figure 2.3: The architecture of a Convolutional Neural Network[4].

The convolutional layer is the core component of the CNN and carries its central portion of the

[3]Visit https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939 for details. Accessed in May 2023.

[4]Retrieved from https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html. Accessed in May 2023.

computational load. In this layer, a dot product is performed between two matrices: one with a set of learnable parameters (kernel) and the other with a portion of the receptive field. Note that the kernel is smaller than the input in space. The kernel will slide across the input image during the forward pass, from left to right and top to bottom, generating a new-sized output containing the kernel's response at each spatial position of the image (activation map). Notice that this operation leverages three critical ideas: sparse interaction, parameter sharing, and equivariant representation.

The pooling layer can be interpreted as summarizing some defined output's locations by compacting each one into just one point. This helps reduce the representation's spatial size and, consequently, decreases the required amount of computation and weights while providing some translation invariance. This summary replacement is achieved by using pooling functions such as the L2 norm of the rectangular neighborhood, the weighted average of the neighborhood...

The fully connected layer is formed by several layers where, as the name suggests, the neurons have full connectivity from one to another with all of the preceding and succeeding layers. This operation facilitates the mapping of the input to the output representation.

Finally, an output derived from this analysis is expected after making an input image pass through all these layers. In addition, since data is generally not linearly separable, there are some non-linearity layers (activation functions) placed after the convolutional layer, allowing the network to learn more complex mappings from input to output [18].

2.4.4 Residual Neural Networks

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 [19]. This problem appears when DNNs cannot propagate useful gradient information from the model output back to the layers near its input. Nevertheless, in 2015 the Residual Network (ResNet) [20] appeared to address this problem by introducing the residual block. Thanks to this addition, deeper networks than the ones used previously became easier to optimize and could gain accuracy from considerably increased depth [20].

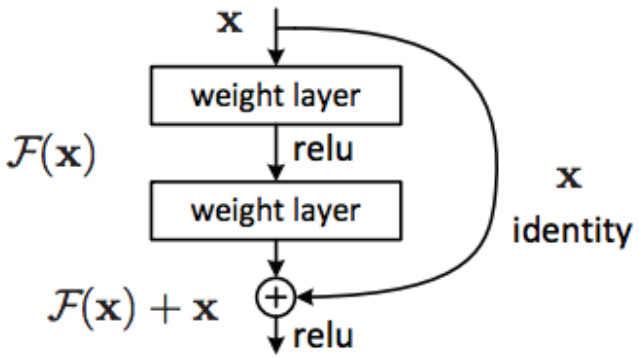


Figure 2.4: The residual block [20].

The residual block adds a new connection to the network called skip connection/identity mapping. It consists of connecting activations of a layer to further layers by skipping some of them in between, enabling information to flow directly through a network layer rather than being passed through a series of nonlinear transformations.

For the identity mapping to have the same dimensions as the output it is added to (notice that these outputs come from convolutions that usually change the input's dimensions), it is multiplied by a linear projection to expand the channels of shortcut to match the residual [20], allowing them to be combined.

All in all, adding the identity mapping [20] to deep CNN has empirically shown to increase performance, achieving state-of-the-art performance on a wide range of computer vision tasks such as image classification, object detection, segmentation, and others, while making the training of the network more accessible.

2.4.5 Transformers and Self-Attention

The transformer was first presented to the world in 2017 in the paper "Attention is All you need" [21] and quickly revolutionized machine learning by outperforming several state-of-the-art architectures, mainly in natural language processing. One of the most significant innovations of this model is the use of attention mechanisms as its primary way of routing information.

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 [21].

Formally, self-attention is applied to an input x of length n and dimension d by following some steps[5].

[5] Visit https://towardsdatascience.com/demystifying-efficient-self-attention-b3de61b9b0fb for details. Accessed

At first, the input x is projected through 3 trainable weight matrices, WQ, WK, WV, outputting 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 2.9.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.9}$$

By applying equation 2.9, a model can focus on the important features by establishing relations between every item of the input (every item looks at all the remaining ones), building a global self-attention mechanism.

The performance of this mechanism can even be improved by introducing multiheaded attention [21]. This concept consists of projecting matrices Q, K and V h times, using different learned projection matrices H (heads) each time. Then, equation 2.9 is applied to each of these h projections in parallel, producing h outputs that are concatenated and projected again to build the final result (all the process illustrated in fig.2.5).

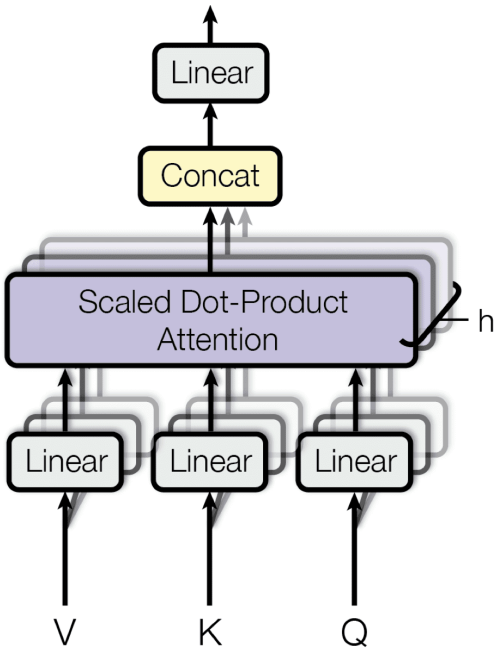


Figure 2.5: Multi-head attention [21].

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

in May 2023

extraction of information that would otherwise be hard to get with a single attention head.

3

Related Work

Contents

This chapter covers the related work, starting from exploring clinical approaches to PE and leading up to previous deep learning research on the field on the subject.

3.1 Clinical Approaches

This section contains the most relevant clinical' score metrics for acute pulmonary embolism diagnosis.

3.1.1 Daniel score

This score was proposed by Daniel L. R. et al. [3], and although it was proposed in 2001, it is still used due to its simplicity. In fact, Hospital de Santa Maria (HSM) makes use of this score while diagnosing PE instead of more recent ones since it is easier to apply.

The metrics used are presented in appendix A. For each identified metric, the corresponding value is summed up, and by doing that, a final resulting score is obtained in the end. Because of that, this derived ECG score increases with the severity of pulmonary hypertension from PE. Any score $\geqslant 10$ highly suggests severe pulmonary hypertension from PE [3].

In a case study made by HSM related to PE diagnosis, 98% specificity was achieved using this score, which is similar to the results achieved by the score's creators [3], where at a cutoff of 10 points, the ECG score was 23.5% (95% Confidence Interval (CI), 16 to 31%) sensitive and 97.7% (95% CI, 96 to 99%) specific for the recognition of severe pulmonary hypertension secondary to PE.

3.1.2 Wells and Revised Geneva scores

Two guideline recommend clinical prediction rules for PE widely used to assess the probability of a patient having PE based on various clinical factors by assigning points are the Wells and the Revised Geneva Scores [22]. These methods differ from the Daniel score because their 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 (both score sheets are presented in Appendix A). Although their overall accuracy is similar, the Wells score is usually more sensitive, while the Revised Geneva score is more specific.

In both cases, a point is assigned to each variable detected, and the final score will indicate the likelihood of PE. By that, these two methods estimate the pre-test probability of PE, which helps guide further diagnostic workup. Based on this final value, clinicians can determine whether additional tests or further clinical evaluation are necessary to confirm or rule out the presence of this disease. For instance, a high pretest probability can lead to further imaging. In contrast, a low probability can lead to

a D-dimer (DD) blood test[1]. All in all, these scoring systems aid in risk stratification and assist in making informed decisions regarding the appropriate management and treatment of patients.

3.1.3 Novel Electrocardiographic Score

The work developed by András Vereckei et al. [4], motivated by the lack of standardized prediction rules in empiric clinical judgment based on ECG, derived a Novel Eletrochardiographic Score (nECGs). The score usage is similar to the Daniel one [3]. However, it uses additional and different selected criteria, some of them being more detailed. These criteria are based on the reflection of important components of the pathomechanism of acute PE, mainly transmural right ventricular ischemia[2], right ventricular dilation[3], acute pulmonary arterial hypertension[4] and right-sided intraventricular conduction disturbances[5] (due to right ventricular ischemia, dilation, and increased right ventricular wall tension) [4]. The score was tested in patients with and without the Right Bundle Branch Block (RBBB) pattern[6], being slightly different for each of the two cases (both score' sheets are available in the appendix A, where fig.A.2 shows the nECGs and fig.A.3 shows a modified version of the nECGs that was used in patients with RBBB pattern). The RBBB is an ECG finding that can be noticed in the ECG leads as ST depression and/or T-wave inversion, specifically in leads V1-3.

The maximum value of the nECGs was 10 or 9 in patients without or with RBBB pattern respectively, and the established threshold was 4. If the nECGs value was $\geqslant 4$, acute PE diagnosis was considered; otherwise, the nECGs value suggested a PE negative diagnosis.

After looking at the results achieved by this new score in table 3.1 and testing its superior sensitivity, negative predictive value, and test accuracy, it was concluded it estimated the pre-test probability of acute PE better than the Daniel-ECG score and all the other prediction rules.

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

Table 3.1: The sensitivity, specificity, Test Accuracy (TA) and predictive values of the different investigated methods when they were applied together in a similar test example [4].

[1] It is a diagnostic test used to assess the presence of blood clotting and fibrinolysis (the breakdown of blood clots) in the body. It has high sensitivity but low specificity

[2] Condition characterized by the insufficient blood supply to the right ventricular myocardium that affects the total thickness of the ventricular wall.

[3] Enlargement or expansion of the right ventricle of the heart.

[4] Sudden and severe increase in blood pressure within the pulmonary arteries.

[5] Abnormalities in the electrical conduction system of the heart that specifically affect the right side of the ventricles.

[6] Cardiac conduction abnormality that affects the electrical signals traveling through the right bundle branch, which is a pathway responsible for conducting electrical impulses in the heart's conduction system.

3.1.4 YEARS and PEGeD algorithms

The YEARS and the Pulmonary Embolism Graduated d-Dimer (PEGeD) algorithms are two approaches to PE through the DD test. Specifically, they both establish rules to choose the best suited threshold for DD test on a patient.

The YEARS criteria [23] determines the risk of PE derived from three items in the Wells score that are most predictive of PE:

- Does the patient have clinical signs or symptoms of DVT?
- Does the patient have hemoptysis?
- Is PE the most likely diagnosis?

Based on the previous criteria, if a patient has 0 YEARS items, the DD threshold should be 1000 ng/mL. However, if a patient has at least 1 item, the DD threshold should be 500ng/mL. Finally, after applying the test, if the results are under the defined threshold, PE is considered excluded, otherwise a CT Scan needs to be performed. By that criteria, there is an absolute reduction in CT Scans performed across all ages by 14% when compared to the Wells criteria.

The PEGeD study [24] is another attempt to show the safety of using an adjusted DD threshold. For that, it evaluates the risk of the patient having PE by the application of the Wells score. Based on the patient's risk, there are 4 possible outcomes:

- DD threshold set to 1000 ng/ml for low risk;
- DD threshold set to 500 ng/ml for moderate risk;
- High risk patients are sent directly to chest imaging without DD testing;

Note that, once again, patients who are below the DD threshold are diagnosed as not having PE. The performance of this criteria is similar to the YEARS one.

3.2 Deep Learning Approaches

Applying deep learning to ECGs to interpret them and diagnose cardiovascular diseases is gaining much interest since many deep learning techniques have revolutionized how machines interpret signals. However, models developed prior to the deep learning era were based on handcrafted feature classification, and its diagnosis accuracy was, in most cases, equivalent to random chance.

Focusing only on PE diagnosis, at the time of this research, there was no work found relating the diagnosis of this condition using only ECGs as input (most of the models found use CT Scan images

[25]). The most relevant work found was the one developed by Sulaiman S. Somani et al. [26], where a 1-dimensional (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 ECGs features but also by CT Scan findings and clinical data. Thanks to this clever approach, the results are promising. Nevertheless, they use much more than just ECGs. Because of that, this thesis had to search for inspiration in models relating to other diseases diagnosis using only or almost only ECGs findings in order to accomplish its goals.

In 2020, Al-Zaiti et al. [27] presented various results obtained by applying multiple machine learning-based methods for the prediction of underlying acute myocardial ischemia[7] in patients with chest pain. For this goal, the approach uses 12-lead ECGs, the patient's sex and age as inputs to the models. As a result, its best classifier, a fusion model based on vote count that combines a ANN with logistic regression and a gradient boosting machine, outperformed experts in ECG reading.

Many studies showed that ECG diagnosis accuracy using DNNs improves when residual blocks are added to the network's architecture when compared to the use of CNNs alone. Because of that, based on the ResNet architecture [20], Wenxiao Jia et al. built a 1D-34-layer residual network to address the Cardiology Challenge 2020 [28]. This challenge aimed to identify multiple clinical diagnoses (mainly multiple types of arrhythmia) from 12-lead ECG recordings. This model achieved some promising results, outperforming various methods such as CNN based ones and achieving high and stable performance scores measured by challenge metric. Although it cannot replace doctors in these diagnosing tasks, it can easily be used as an assisting tool for cardiologists.

In 2021, Chao Che et al. [29] proposed an end-to-end deep learning framework based on the combination of 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 main idea behind the combination of these two networks (CNN and transformer) is to acquire complex and special features of the ECGs through the CNN and to capture its temporal features while focusing on context vectors through the use of the embedded transformer. 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 using a transformer and/or attention mechanisms to capture temporal continuity relations of the input data.

Multiple promising approaches could be applied to the PE case scenario. However, this project focused on using mainly two approaches described in the following sections.

[7] Sudden decrease or interruption of blood flow to the heart muscle, resulting in inadequate oxygen supply to the affected area.

3.3 Small 1D-DNN with residual blocks

The first approach applied to the PE diagnosis case scenario was based on the work developed by Ribeiro et al. in 2020 [30]. Here, 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%, proving that neural networks can learn important useful information if fed with a sufficiently large and rich dataset. The DNN architecture is exposed in fig.3.1.

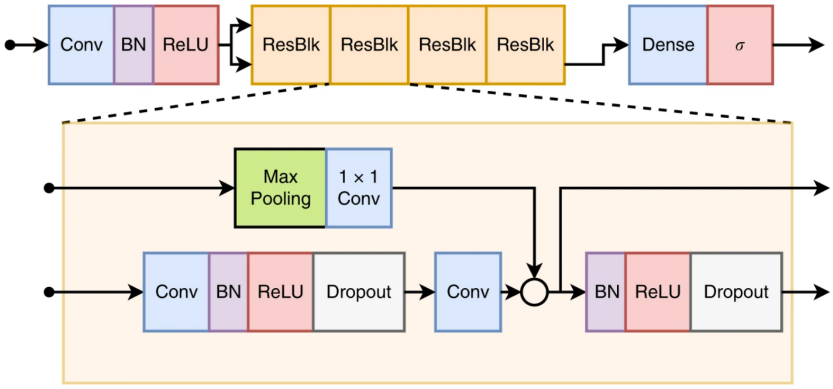


Figure 3.1: The small 1D-DNN residual neural network architecture [30].

The model architecture is based on a small standard residual network [20], yet it adopts a slight modification to the residual blocks (it uses 4 identical modified residual blocks). The authors tested several different values and configurations over several runs to decide the number of residual blocks, the kernel size, and to tune other hyperparameters. All the input ECGs were resampled to a 400 Hz sampling rate and would vary its length from 7s to 10s. Because of this, these samples are zero-padded, resulting in signals with 4096 samples per lead (raw data). The average cross-entropy is minimized using the Adam optimizer with default parameters and a learning rate of 0.001. The learning rate is also reduced by a factor of 10 whenever the validation loss does not present any improvement for seven consecutive epochs [30].

3.4 The Heartnet architecture

Beyond the capability of extracting complex relevant features from ECG's findings, it is also essential to focus on temporal relations between segments of the input. Since an ECG is composed by extensive sequences of voltage measured values through 12-leads, a 1D-ResNet architecture is not able to capture long-range context by itself. For that, the network needs, 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 [31] 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 as illustrated in fig.3.2 to capture long-range dependencies and temporal information between all the embeddings of the input samples.

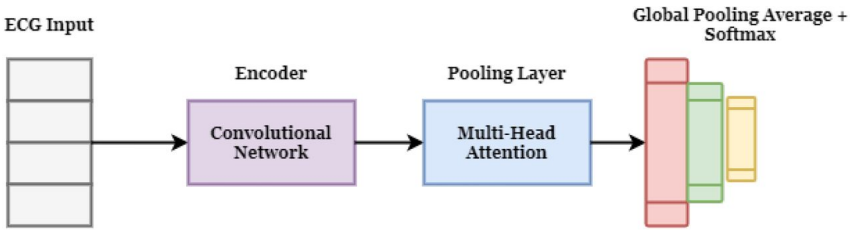


Figure 3.2: Proposed HeartNet architecture [31].

Regarding the specifics of this architecture, starting with the CNN block, it is composed of 4 convolutional blocks: the first two, in addition to the convolutional layer, also apply batch normalization and use the ReLU as the activation function, the third one instead of ReLU uses a softmax layer and, finally, the last one only contains a convolutional layer. This CNN behavior can be interpreted as an encoder that will feed its output embeddings to the multi-head attention layer which follows. This layer applies scaled-dot product attention (equation 2.9) to the embeddings and uses multiple heads (4 or 8 heads). Finally, the output of this layer is fed into the global pooling average layer (instead of a linear layer), followed by a softmax operation.

Regarding the model's performance, it achieved an accuracy of $\approx$ 85% and an F1 score of $\approx$ 86% in diagnosing Atrial Fibrillation (type of arrhythmia) when applied to a dataset composed of 8526 ECG recordings, being 771 of them correspondent to the Atrial Fibrillation class.

4

Data Management and Pre-processing

Contents

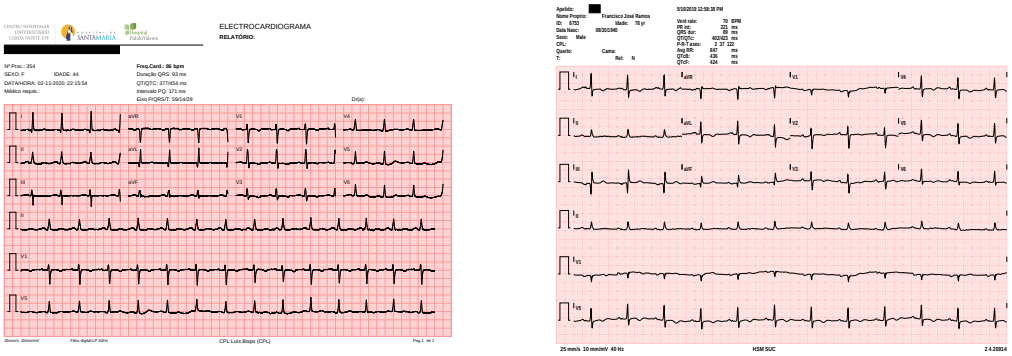
This chapter provides important background information on the dataset used, its management, and pre-processing. It starts with an overview of signal acquisition and initial processing. Secondly, it compares the use of images against the use of the raw data and the pros and cons of each. Then, more detail is given about data augmentation and techniques used to fight the drawbacks of an imbalanced, small dataset. Finally, a background of spectrograms is given since they can be used on a different approach described in further chapters.

4.1 Dataset Acquisition and Selection

The dataset provided by HSM was extremely imbalanced and contained 1014 ECG examples in total: 293 belonging to the positive class (with PE) and the remaining (721) belonging to the negative class. These examples were acquired from patients admitted to the emergency department and in whom CT Scan was performed for suspected PE.

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 before saving the results in a pdf file.

Despite most of the ECG having a standard output format, some provided examples arranged the information differently, while others contained hard-to-use information such as lead attaching noise. Because of that, some of the ECGs were disposed of. Two acquired examples are presented in fig.4.1.



(a) Desired ECG format **(b)** Different ECG format

Figure 4.1: Acquired ECG examples

Consequently, the final useful dataset comprised 929 examples, 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;

- A **test set**, containing 103 examples: 38 of the positive class and 65 of the negative one.

4.2 Image ECG vs. Raw Data

Initially, a dataset of ECG images extracted straight from the selected examples was used to train the initial models. However, the images were excessively large at 2161 x 1121 pixels and lost a significant amount of information when downsampling or size reduction techniques were employed. Additionally, multiple leads overlap in most images, making it challenging for the network to interpret each individually. As a result, the first models trained on this dataset struggled to yield meaningful results or achieve significant learning.

Another path was chosen from this point: using the ECG's raw data. 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 device at a given sampling rate. Since extracting this information from the pdf files was tough, a tool from the company Dotlogic (responsible for the ECG devices and software) was provided, which could extract this information and return it in a standard format XML file. Using this software, it was possible to obtain the same information that was on the previous dataset but in a raw data format, as shown in fig.4.2.

Figure 4.2: 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.

4.3 Raw Data 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 caused by the attachment or detachment of the pads on the patient. This value was discussed with HSM's specialists and set to 4mV. By that, all values surpassing this threshold were assigned to this value.

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 from 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 [5].

$$Zscore = \frac{x - mean(x)}{std(x)} \tag{4.1}$$

$$MinMax = \frac{x - min(x)}{max(x) - min(x)} \tag{4.2}$$

4.4 Data augmentation

Deep Learning models usually require large amounts of good-quality data to perform at their best. As in this case the dataset was extremely imbalanced and small, it was imperative to make the most use of the data available. One way to address this challenge is by using data augmentation techniques to artificially increase the size and diversity of the dataset [32].

Data augmentation involves applying various transformations to the original data, introducing variations, and improving the developed models' robustness, generalization, and accuracy. Multiple transformations can be applied to ECGs [33], but the main ones used in this work are the following:

- **Shift**: Performs a signal shift in time. In this work, this is made by changing the starting point of all the leads of an ECG from the second 0 to the second 1,808, having each lead the length of 8,192 seconds;
- **Flip**: Performs a flip of all the leads over the x-axis by multiplying each one by -1;
- **Scale**: Performs a scaling operation overall leads by multiplying all of them by a constant;

- **Random drop**: Randomly sets some of the ECG lead's values to 0 with a probability p;
- **Section drop**: The signal is broken into small sections, and each of the sections can be set to 0 with a probability p;
- **Lead drop**: Randomly drops 0 to a certain number of leads by setting each one to 0;
- **Sin sum**: A sine wave is added to the entire sample;
- **Square pulse sum**: A square pulse is added to the entire sample.

4.5 Spectrograms

In some cases, one approach that may improve ECG's analysis is to use spectrograms [34, 35]. A spectrogram is a visual representation of the frequency content of a signal over time. Essentially, 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.

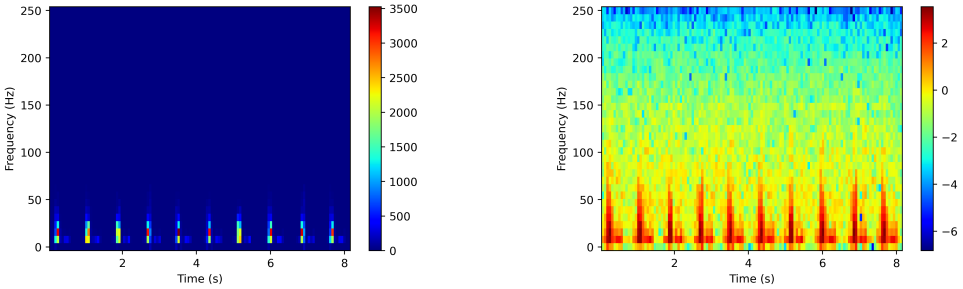
This technique has been widely applied to audio signal analysis [36], but it also demonstrated relevant results on ECG's research [34]. To create a spectrogram from an ECG, first, the signal is broken down into small, overlapping time windows. Then, to which one of these, a Discrete Fourier Transform (DFT) is obtained using the Short-time Fourier Transform (STFT) method, and, therefore, the frequency component is obtained for each segment in time [37]. 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 4.3.

$$S(t, f) = \left| \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f} \right|^2 dt \tag{4.3}$$

where $S(t, f)$ is the time-frequency representation, $x(\tau)$ is the input signal and $w(t)$ is the observation window [37]. The choice of $w(t)$ can significantly 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 our context, but the most common ones are the Hamming window, the Blackman window, and the Tukey window (this last one is widely used on 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 [34].



(a) ECG Spectrogram. **(b)** ECG Log-Spectrogram.

Figure 4.3: Spectrogram and log-spectrogram example for the same ECG.

5

Proposed Model Development

Contents

This chapter will give an insight into how the related work influenced the final model's development and the creation of the chosen approach to the problem.

5.1 The Baseline (Small 1D-DNN)

After careful consideration, the first approach chosen to be applied to the PE diagnosis case scenario was based on the work developed by Ribeiro et al. in 2020 [30] described in section 3.3.

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 on its context, the simple network architecture, and the details given by the authors relative to the training and learning process, made this idea a valuable approach.

Regarding its implementation on the PE dataset, we kept the architecture the same as described on fig.3.1, since it was ready to receive 12-lead ECGs with length of 4096 samples per lead. Furthermore, after testing the use of different numbers of residual blocks, the number defined by this research performed best.

Finally, as no PE based models were found using only ECGs, the results of this network applied to our PE dataset were used as a 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 [28] of adapting standard residual networks [20] with multiple depths to 1D format to get the network to learn more information from deeper features while having in mind that usually networks for ECG diagnosis benefit from residual blocks. After testing several standard ResNet formats with 18, 34, 50, and 101 layers, the ResNet-18 architecture was chosen for being able to retrieve complex features from the ECGs without making it too tricky for its weights to be tuned by such a small dataset as the one presented. When deeper ResNets were used (with more than 50 layers), it was tough for them to achieve relevant learning derived from its large size - bigger models require larger datasets.

The ECG raw data on the dataset is composed of an extensive sequence of 4096 samples per lead. Thus, the 1D-ResNet is not able to capture long-range context by itself. As noted by the work developed by Hasan and Young in 2021 [31], a network applied to ECG's analysis needs also a mechanism capable of providing a global interpretation of the input sequence such as a multi-head attention layer.

The HeartNet performance [31] is relevant not only because of the scores it achieves but also because they were obtained with a network trained on a relatively small dataset. Consequently, it proves

the worthiness of using a DNN output as an input to a multi-head attention layer. By gathering all the information presented until this point, a new model was developed within the scope of this thesis: an attention-enhanced residual network.

This new model is composed of 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 crucial 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.5.1.

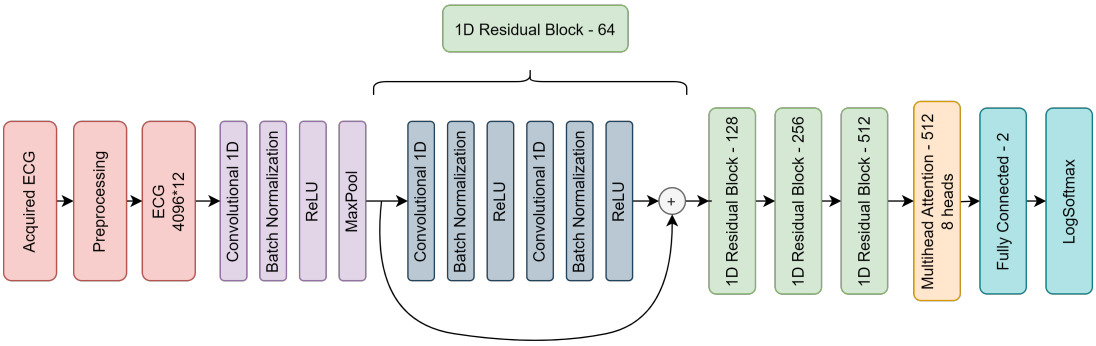


Figure 5.1: Developed model architecture.

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 chapter 4 and the Z-score normalization (equation 4.1). 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 4096 with 12 different channels. After that, the input goes through the entry block (purple zone on fig.5.1) 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.5.1) 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 patients without PE, and class 1 represents patients with it.

5.3 Spectrogram based model

Another approach that could improve performance in solving the PE diagnosis problem is the use of ECG's log-spectrogram representation as input data to the DNN. Notice that spectrograms have two dimensions, which makes it possible for pre-trained 2-dimensional (2D) networks to be used in this case. Nevertheless, although this was possible, pre-trained models such as standard ResNet demonstrated poor results when applied to this dataset, probably because they had not seen many spectrograms during their training.

In the research developed by Martin Zihlmann et al. [34], a special network called Convolutional Recurrent Neural Network (CRNN) is applied to an ECG's spectrogram dataset. Their architecture follows the same idea presented by the model developed in section 5.2, being composed of a CNN for deep, complex feature extraction and a Long short-term memory (LSTM) block [38] for capturing long term temporal dependencies. Since the architecture presented in fig.5.1 has all the benefits brought by this CRNN, adding even others, such as the skip connections, and in order to understand if there was a benefit in using spectrograms over raw data ECGs, the Attention-enhanced residual network presented in fig.5.1 was modified to be able to be fed with 2D samples. As a consequence, 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. Moreover, before the self-attention operation, a flatten operation is applied to reduce the number of dimensions of the embeddings. This new model adapted to spectrogram's input data is shown in fig.5.2.

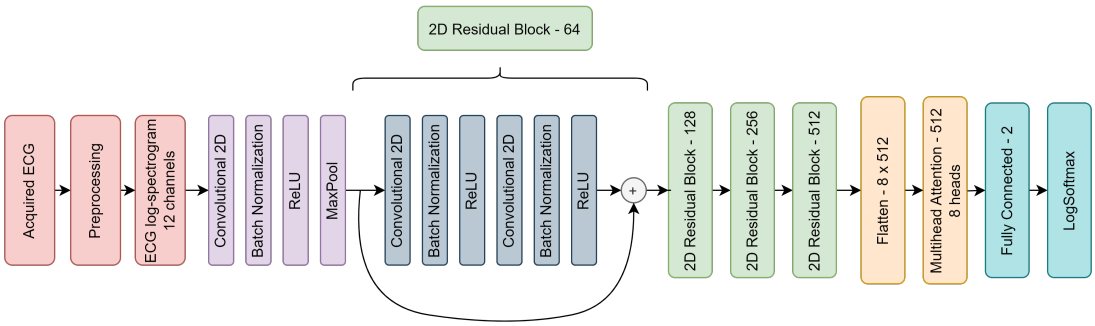


Figure 5.2: Developed model architecture adapted to spectrogram's input.

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 4.2) is applied following the recommendations of the work developed by Hongzu Li et al. [35]. Afterward, the spectrogram is built following the steps presented in section 4.5. 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

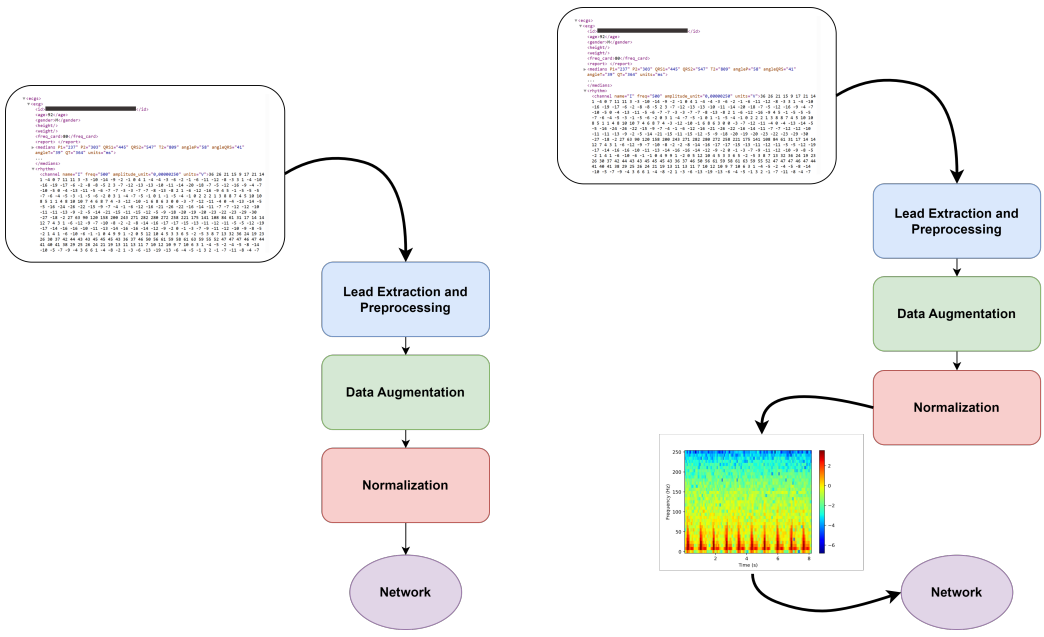
Experimental Results

Contents

6.1 Setup of the experiments

The experiments were focused on three different models explained in chapter 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 [12] and trained using an NVIDIA 32GB V100S installed in a DELL PowerEdge C41402 server at Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa (INESC-ID).

All these models were trained on the PE training set presented in section 4.1. 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 was 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.6.1.



(a) Raw data case scenario **(b)** Log-spectrogram case scenario

Figure 6.1: Data 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 4.4 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.

Since the goal of the project was to achieve the greatest confidence when predicting a positive PE patient (class 1 on the dataset), the most important metric taken into account 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 the F1-score, specificity, and overall accuracy.

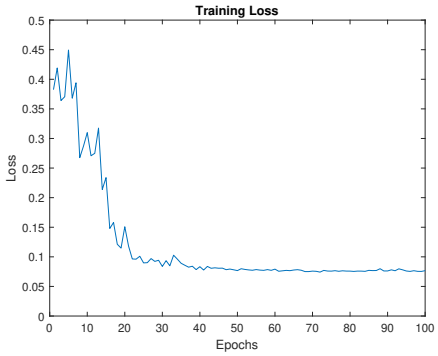
Regarding hyperparameters and training routines, all models were trained several times during 50 epochs to perform hyperparameter tuning. Those were chosen among the following options: kernel size of the entry layer - odd numbers from 3 to 31, batch size - [4,8,16,32,64,128], initial learning rate - [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

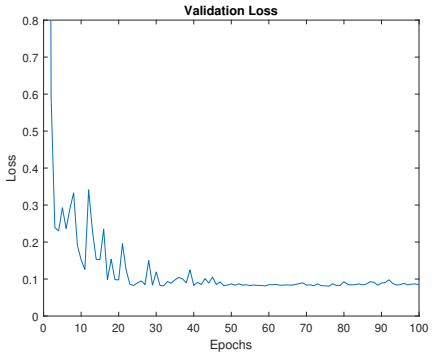
After tuning the hyperparameters and determining the final training conditions for each model, the selected values were used to train each network for 100 epochs. All weights were initialized following the Xavier uniform distribution [39], and biases were assigned 0 value. Each model's state was saved every time both PPV and recall were at least 0.6 and 0.3, respectively, measured on the validation set. Considering all the metrics mentioned before, the best state for each model was selected from the filtered ones. This state was achieved using a kernel size in the first layer of 17, batch size of 32, the AdamW optimizer with default parameters [40], 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 losses are exposed in fig.6.3. Finally, the best models' performance was evaluated on the test set, whose results are revealed in the table 6.1.

Table 6.1: 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

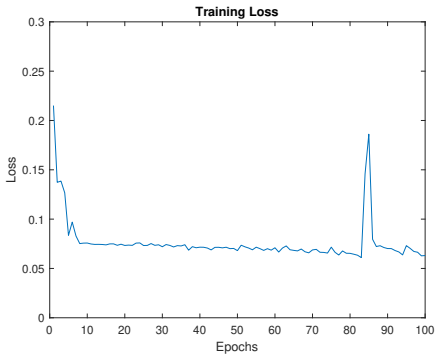


(a) Training loss

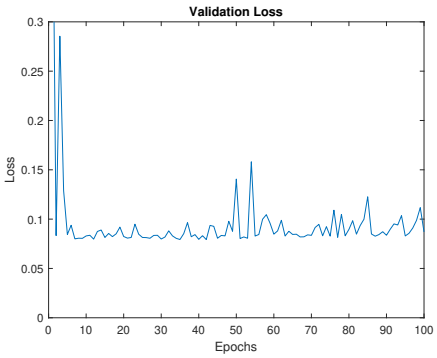


(b) Validation loss

Figure 6.2: Baseline loss charts.

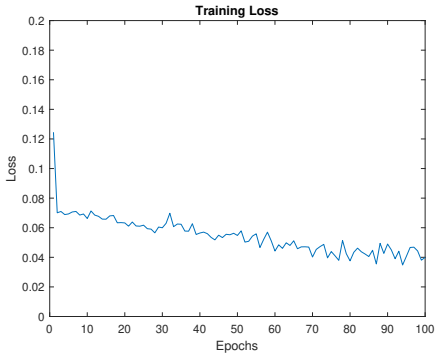


(a) Training loss

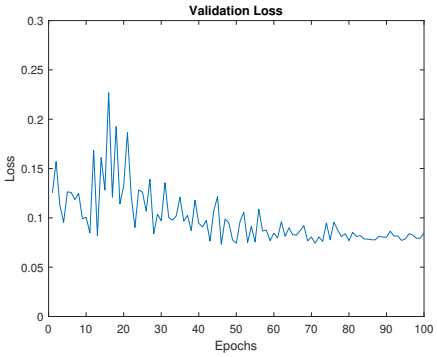


(b) Validation Loss

Figure 6.3: 2D-Attention Resnet (spectrograms) loss charts.



(a) Training loss



(b) Validation loss

Figure 6.4: 1D-Attention Resnet loss charts.

The 1D attention-enhanced residual network performed the best among these three models, surpassing the baseline. 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 when diagnosing PE, temporal-related features are more meaningful 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 the best-developed model and clinical approaches

Once the final developed model was trained and set (the 1D-Attention ResNet), all it remained was comparing its scores against the ones achieved by current common clinical approaches. Unfortunately, no results were provided by HSM on this test set using the nECGs. However, several results on the test set were obtained using several clinical guideline-recommended approaches. Those results are compiled in the table 6.2.

Table 6.2: Diagnostic accuracy of Wells and Geneva's scores combined with a fixed and 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]

All the clinical results were associated with a confidence interval. To ensure a fair comparison, we calculated also the confidence intervals for the machine learning model predictions using the same method of normal approximation.[1]

Moreover, the ROC was drawn, and the AUC was calculated. Notice that the AUC value is around 0.75, which is superior to the value of all the clinical approaches' scores acquired on this metric. The ROC is represented in fig.6.5.

[1] Visit https://www.degruyter.com/document/doi/10.1515/tjb-2020-0337/html for details. Accessed in May 2023

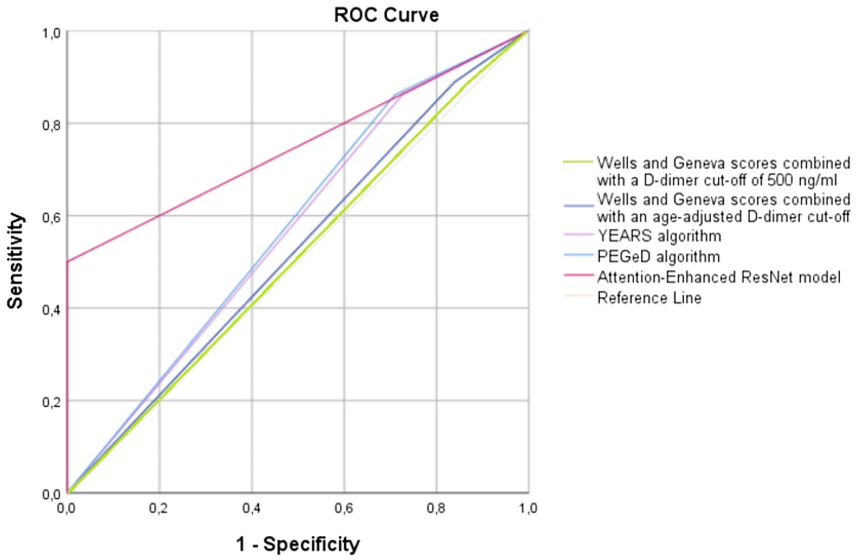


Figure 6.5: ROC curve demonstrating the diagnostic performance of different decision rules to predict pulmonary embolism..

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

Conclusion and Future Work

Contents

7.1 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 quickly and accurately diagnose PE, especially acute PE. In the scope of this thesis, a review over ECG analysis using deep learning was performed with a particular 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 to be 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 DD, have extremely high sensitivity for PE but very low specificity. Because of that, applying this model to a patient subjected to a DD test highly assists medical decisions toward a good diagnosis (we get a joint test with high specificity and high sensitivity). At the time of this work, to the author's knowledge, there was no previously deep learning model 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, it was extremely difficult to extract relevant ECG findings from the samples and, consequently, achieve even better results than the ones accomplished. One explanation for the lack of positive examples is the little importance many hospitals give to labeling. Information is power, and although there are many positive 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 challenging 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.

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

Bibliography

[1] E.-O. Essien, P. Rali, and S. C. Mathai, "Pulmonary Embolism," *The Medical Clinics of North America*, vol. 103, no. 3, pp. 549–564, May 2019.

[2] V. Vyas and A. Goyal, "Acute Pulmonary Embolism," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK560551/

[3] K. R. Daniel, D. M. Courtney, and J. A. Kline, "Assessment of Cardiac Stress From Massive Pulmonary Embolism With 12-Lead ECG," *Chest*, vol. 120, no. 2, pp. 474–481, Aug. 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012369215514554

[4] 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*, vol. 130, pp. 143–151, Sep. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002914920305531

[5] W. Hao and K. Jingsu, "Investigating deep learning benchmarks for electrocardiography signal processing," *arXiv pre-print 2204.04420*, 2022. [Online]. Available: https://arxiv.org/abs/2204.04420

[6] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0870255123001853

[7] M. AlGhatrif and J. Lindsay, "A brief review: history to understand fundamentals of electrocardiography," *Journal of Community Hospital Internal Medicine Perspectives*, vol. 2, no. 1, p. 14383, Jan. 2012. [Online]. Available: https://www.tandfonline.com/doi/full/10.3402/jchimp.v2i1.14383

[8] E. Khan, "Clinical skills: the physiological basis and interpretation of the ECG," *British Journal of Nursing*, vol. 13, no. 8, pp. 440–446, Apr. 2004. [Online]. Available: http://www.magonlinelibrary.com/doi/10.12968/bjon.2004.13.8.12778

[9] H. Tung, C. Zheng, X. Mao, and D. Qian, “Multi-Lead ECG Classification via an Information-Based Attention Convolutional Neural Network,” *Journal of Shanghai Jiaotong University (Science)*, vol. 27, no. 1, pp. 55–69, Feb. 2022. [Online]. Available: https://doi.org/10.1007/s12204-021-2371-8

[10] S. Safari, A. Baratloo, M. Elfil, and A. Negida, “Evidence Based Emergency Medicine Part 2: Positive and negative predictive values of diagnostic tests,” *Emergency*, vol. 3, no. 3, pp. 87–88, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4608333/

[11] D. D. Valle, “Artificial Intelligence-A Modern Approach (3rd Edition).” [Online]. Available: https://www.academia.edu/45007883/Artificial_Intelligence_A_Modern_Approach_3rd_Edition_

[12] “PyTorch documentation — PyTorch 2.0 documentation.” [Online]. Available: https://pytorch.org/docs/stable/index.html

[13] K. Janocha and W. M. Czarnecki, “On Loss Functions for Deep Neural Networks in Classification,” *Schedae Informaticae*, vol. 25, 2017. [Online]. Available: https://doi.org/10.4467/20838476SI.16.004.6185

[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*, vol. 42, no. 2, pp. 318–327, Feb. 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2858826

[15] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, Apr. 1980. [Online]. Available: https://link.springer.com/article/10.1007/BF00344251

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, vol. 25, 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[18] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark,” *Neurocomputing*, vol. 503, pp. 92–108, Jun. 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.06.111

[19] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” in *30th International Conference on Machine Learning, ICML 2013*, vol. 28. [Online]. Available: https://proceedings.mlr.press/v28/pascanu13.pdf

[20] 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)*, 2016, pp. 770–778.

[21] 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*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[22] D. D. Wong, G. Ramaseshan, and R. M. Mendelson, “Comparison of the Wells and Revised Geneva Scores for the diagnosis of pulmonary embolism: an Australian experience,” *Internal Medicine Journal*, vol. 41, no. 3, pp. 258–263, 2011. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20214691/

[23] T. van der Hulle, W. Y. Cheung, S. Kooij, L. F. M. Beenen, T. van Bemmel, J. van Es, L. M. Faber, G. M. Hazelaar, C. Heringhaus, H. Hofstee, M. M. C. Hovens, K. A. H. Kaasjager, R. C. J. van Klink, M. J. H. A. Kruip, R. F. Loeffen, A. T. A. Mairuhu, S. Middeldorp, M. Nijkeuter, L. M. van der Pol, S. Schol-Gelok, M. ten Wolde, F. A. Klok, M. V. Huisman, A. J. Fogteloo, L. J. M. Kroft, M. P. Brekelmans, R. M. J. Vermaire, H. Bastiaansen-Bergsma, J. S. Biedermann, A. Klijn, S. van der Voort, A. W. E. Lieveld, P. Y. de Jong, C. G. Schaar, and A. I. del Sol, “Simplified diagnostic management of suspected pulmonary embolism (the years study): a prospective, multicentre, cohort study,” *The Lancet*, vol. 390, no. 10091, pp. 289–297, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673617308851

[24] C. Kearon, K. de Wit, S. Parpia, S. Schulman, M. Afilalo, A. Hirsch, F. A. Spencer, S. Sharma, F. D’Aragon, J.-F. Deshaies, G. Le Gal, A. Lazo-Langner, C. Wu, L. Rudd-Scott, S. M. Bates, and J. A. Julian, “Diagnosis of pulmonary embolism with d-dimer adjusted to clinical probability,” *New England Journal of Medicine*, vol. 381, no. 22, pp. 2125–2134, 2019. [Online]. Available: https://doi.org/10.1056/NEJMoa1909159

[25] 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*, 2022, pp. 695–702.

[26] 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*, vol. 3, no. 1, pp. 56–66, Mar. 2022. [Online]. Available: https://doi.org/10.1093/ehjdh/ztab101

[27] 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*, vol. 11, no. 1, p. 3966, Aug. 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-17804-2

[28] 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*, 2020, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9344409

[29] 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*, vol. 21, no. 1, p. 184, Dec. 2021. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01546-2

[30] 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*, vol. 11, no. 1, p. 1760, Apr. 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-15432-4

[31] 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*, vol. 10, pp. 100 501–100 512, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9889702

[32] 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*, ser. Proceedings of Machine Learning Research, vol. 174, Apr 2022, pp. 282–310. [Online]. Available: https://proceedings.mlr.press/v174/raghu22a.html

[33] N. Nonaka and J. Seita, "Randecg: Data augmentation for deep neural network based ecg classification," in *Advances in Artificial Intelligence*. Cham: Springer International Publishing, 2022, pp. 178–189. [Online]. Available: https://doi.org/10.1007/978-3-030-96451-1_16

[34] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *2017 Computing in Cardiology (CinC)*, pp. 1–4. [Online]. Available: https://www.cinc.org/archives/2017/pdf/070-060.pdf

[35] H. Li and P. Boulanger, "Structural Anomalies Detection from Electrocardiogram (ECG) with Spectrogram and Handcrafted Features," *Sensors*, vol. 22, no. 7, p. 2467, Mar. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2467

[36] L. Wyse, “Audio spectrogram representations for processing with convolutional neural networks,” *arXiv pre-print 1706.09559*, 2017. [Online]. Available: http://arxiv.org/abs/1706.09559

[37] 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)*, vol. 10(1-16), p. 81–85. [Online]. Available: https://jtec.utem.edu.my/jtec/article/view/4099

[38] 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. [Online]. Available: http://arxiv.org/abs/1909.09586

[39] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 249–256, Jan. 2010. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf

[40] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *arXiv pre-print 1711.05101*, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1711.05101

A

Clinical Scores Sheets

Daniel-ECG-score

Characteristic		Score
Tachycardia (>100 beats/min)?		2
Incomplete right bundle branch block?		2
Complete right bundle branch block?		3
T wave inversion in all leads V_1 through V_4?		4
T wave inversion in lead V_1?	<1 mm	0
	1-2 mm	1
	>2 mm	2
T wave inversion in lead V_2?	<1 mm	1
	1-2 mm	2
	>2 mm	3
T wave inversion in lead V_3?	<1 mm	1
	1-2 mm	2
	>2 mm	3
S wave in lead I?		0
Q wave in lead III?		1
Inverted T wave in lead III?		1
If all of $S_IQ_{III}T_{III}$ is present, add		2

Figure A.1: The Daniel-ECG-score sheet [3].

Table A.1: The Wells score sheet for PE diagnosis.

Clinical variables	Points
Clinical signs and symptoms of DVT	3
An alternate diagnosis is less likely than PE	3
Heart rate $>$ 100	1.5
Immobilization or surgery in the previous 4 weeks	1.5
Previous DVT/PE	1.5
Hemoptysis	1
Malignancy (treatment currently, in the previous 6 months, or palliative)	1

Table A.2: The risk of PE predicted by the Wells score using 3 or 2 categories.

Risk group	Points	Risk of PE
Low risk	0-1	3.6%
Moderate risk	2-6	20.5%
High risk	$>$6	66.7%

Risk group	Points	Risk of PE
Low risk	0-4	5.1%
High risk	$>$4	39.1%

Table A.3: The Revised Geneva score sheet for PE diagnosis.

Clinical variables	Points
Age 65 years or over	1
Previous DVT or PE	3
Surgery or fracture within 1 month	2
Active malignant condition	2
Unilateral lower limb pain	3
Hemoptysis	2
Heart rate 75 to 94 beats per minute	3
Heart rate 95 or more beats per minute	5
Pain on deep palpation of lower limb and unilateral edema	4

Table A.4: The risk of PE predicted by the Revised Geneva score.

Risk group	Points	Risk of PE
Low risk	0-3	8%
Moderate risk	4-10	29%
High risk	11	74%

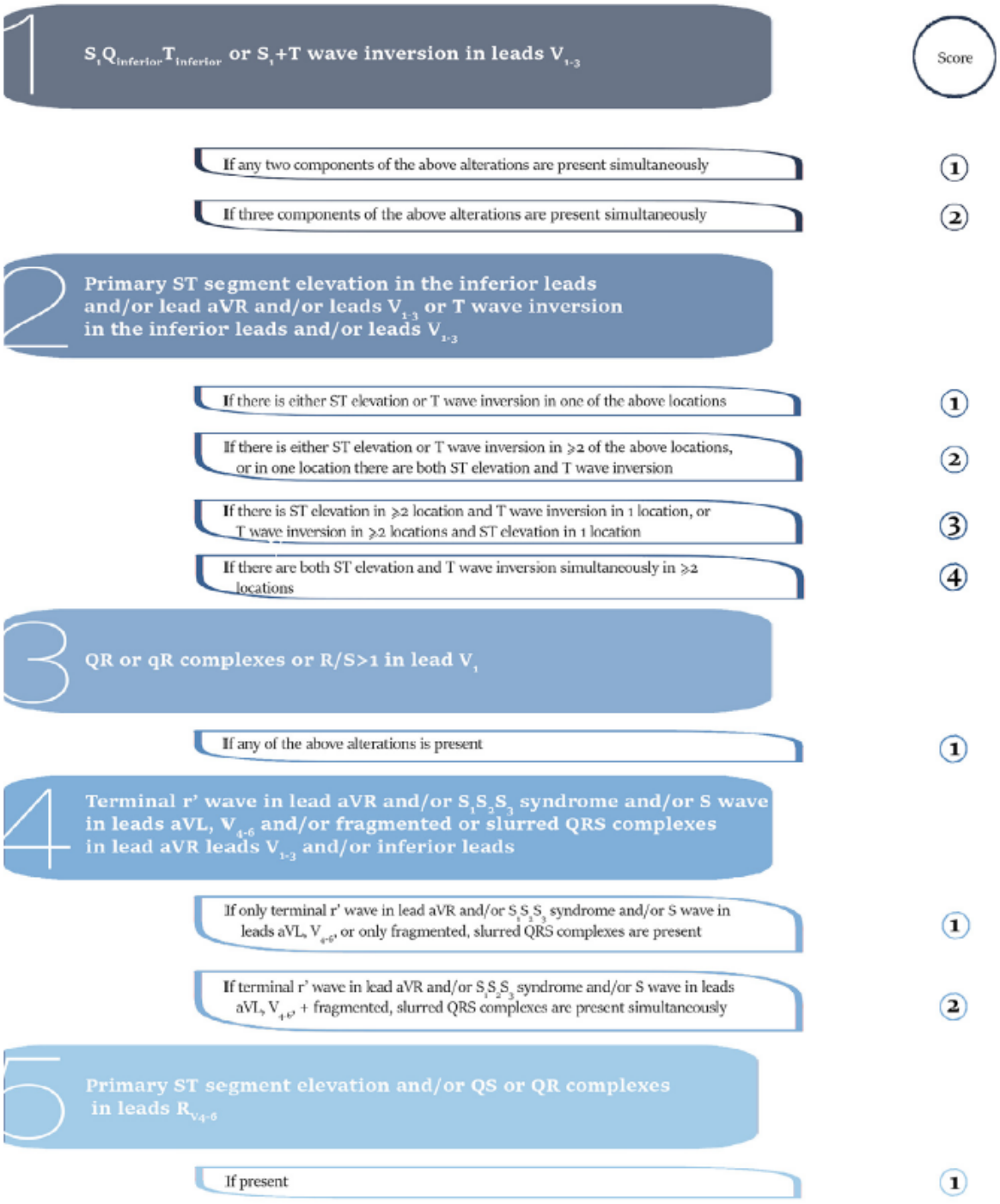


Figure A.2: Novel ECG score sheet for patients without RBBB pattern [4].

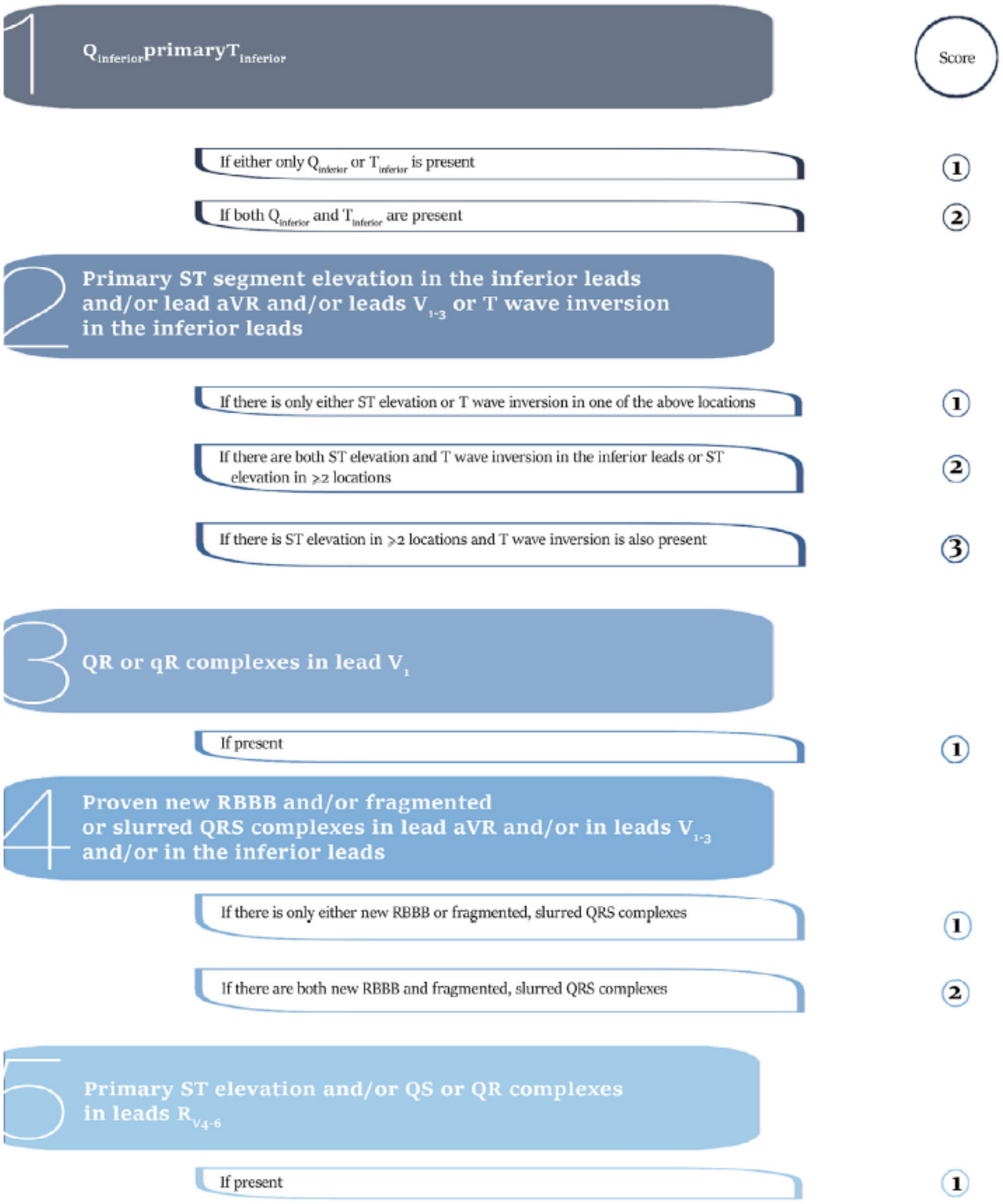


Figure A.3: Novel ECG score sheet for patients with RBBB pattern [4].