Chagas disease classification from electrocardiograms in the time-frequency domain using deep learning

Dylan J.G. Nowee

Layout: typeset by the author using LaTeX.
Cover illustration: Unknown artist

Chagas disease classification from electrocardiograms in the time-frequency domain using deep learning

Dylan J.G. Nowee
14487020

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
Asst. prof. dr. N. Awasthi

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

Semester 2, 2024-2025

Abstract

Chagas disease (CD), caused by *Trypanosoma cruzi*, affects more than 7 million people worldwide, resulting in 10,000 deaths annually. To detect CD related cardiomyopathies, widely available electrocardiograms are the main diagnostic tool. However, the confirmatory serological testing capacity remains limited, calling for a prioritisation of patients. Deep learning (DL) methods combined with a time-frequency analysis of ECG signals have shown success in classification of heart disease, but with little research focused on CD and none using the time-frequency domain.

Therefore, this Thesis examined whether convolutional neural networks (CNNs) can classify CD from 12-lead ECGs represented as the 2D time-frequency distributions obtained through a continuous wavelet transform. To do this, two CNN architectures were trained and evaluated on separate parts of the CODE-15% dataset. Finally, a calibrated model was also assessed.

The results show that the smaller 2D CNN: v1, achieved the highest accuracy (0.75), AUROC (0.83), and PhysioNet challenge score (0.36). Further calibration increased recall (0.96) but significantly reduced accuracy, demonstrating the relevant sensitivity-specificity trade-off for screening in healthcare. All CNNs outperformed a previously proposed one-dimensional model, indicating that the time-frequency method can improve CD detection rates.

The findings suggest that CNNs using time-frequency ECG representations have potential as automated tools for CD screening, but further external validation and serologically confirmed datasets are needed prior to clinical applications.

Contents

Chapter 1

Introduction

Chagas disease (CD), caused by the protozoan parasite *Trypanosoma cruzi*, affects an estimated 7 million people and causes 10,000 deaths annually worldwide (de Fuentes-Vicente et al., 2018). However, the disease is predominantly prevalent in Central and South America, meaning that around 100 million people are at risk of contracting it (World Health Organization, 2025). Despite its impact, CD is considered a neglected disease by the World Health Organization (WHO, 2025), reflected by the low funding and the limited attention it receives. Parasitic transmission of CD can occur orally, congenitally or most often via the triatomine bug (World Health Organization, 2025). The triatomine bug, also known as the kissing bug, transmits the *T. cruzi* parasite through its feces (de Fuentes-Vicente et al., 2018).

Currently, no effective vaccine exists, and the only other preventive measures suggested by the WHO, such as vector-control – reducing the human-vector bug contact – are already adhered to by the affected populations (World Health Organization, 2025). Although vector-control has been the most effective way to prevent CD in Latin America, it is also essential to treat the disease timely and adequately, as its symptoms can be fatal (Nunes et al., 2024). According to research on the progression of CD by Nunes et al. 2024, these symptoms can range from none to mild in the first two months after initial infection — called the acute phase - to severe cardiac disorders in later stages, called the chronic phase. The cardiac symptoms are also occasionally accompanied by digestive or neurological alterations. The cardiomyopathies associated with CD are left ventricular dysfunction, commonly known as a weak heart pump, cardiac arrhythmia (irregular heartbeat), and finally sudden death (Nunes et al., 2024). These and other cardiac abnormalities are examined non-invasively using the standard diagnostic tool for diagnosing diseases of the heart: a 12-lead electrocardiogram (ECG). An ECG is a

diagnostic test that shows the activity of the heart on 12 different axes by measuring the changes in voltage caused by depolarisation of the heart at each contraction (Ashley and Niebauer, 2004). These changes in voltage are visualised as a signal with one beat of the heart forming a wave complex for every axis measured.

However, this signal is not only interpretable for physicians, but also for various machine learning (ML) models in the field of artificial intelligence (AI). Numerous studies have shown improvements to reach physician-level accuracy when automating the classification of various heart diseases using ECG signals (Ribeiro et al., 2020). A systematic review by Luz et al. (2016) has shown that high accuracy in the domain of ECG-based classification is achieved by varying ML models, including models which regularly perform well with temporal data, such as residual networks (ResNets) and long-short term memory (LSTM) models. Another review by Liu et al. (2021) on ECG classification has emphasised a contemporary branch of ML: deep learning (DL), which learns the features that have to be extracted manually in other ML approaches. Liu et al. (2021) have also presented the high accuracy of various DL models, including convolutional neural networks (CNNs), which are becoming increasingly prevalent in the field of ECG-based classification and have high accuracy on multi-dimensional data classification (Mohammed et al., 2024).

While most research in DL ECG classification is focused on arrhythmia and other cardiomyopathy classification, little research has been performed to classify CD specifically. While research by Jidling et al. (2023) focused on the detection of CD using a one-dimensional ResNet model, which considers only the time domain, no research has been conducted on classifying CD using the 2D time-frequency representation (TFR) yet. The TFR of a signal is the representation of that signal over both time and frequency simultaneously, allowing for the analysis in the respective domain. The TFR of ECG data is a complex-valued field where the modulus of the field represents the amplitude of the signal (Hussein et al., 2018). Furthermore, this research will focus on testing the performance of different two-dimensional CNN models on classification of CD from the TFR of ECG signals. Of which one CNN model with few parameters for ECG classification proposed by Jun et al. (2018). Another CNN model with a scaled architecture of the first with more parameters will also be evaluated.

Because the WHO 2025 declared a need for innovative diagnostic tools for chagas disease, DL models, more specifically CNNs, will be considered in this thesis, leading to the following research question:

1. **Can deep learning models, such as convolutional neural networks,**

accurately classify chagas disease from 12-lead electrocardiograms with the use of its time-frequency representation?

To answer this research questions this thesis outlines an experimental approach following the scientific work related to this topic. This literature and other information required to contextualise this thesis will be considered in the background section, and the experimental approach in the methods section. Subsequently, the data and the processes concerning it will be discussed in the data section. The experiments performed will be emphasized in the respective section, as will the results of these experiments. These results and other factors during the study will be discussed in the discussion section, which will lead to a conclusion and definitive answers to the research question. Finally future works will be proposed to reinforce the academic impacts of this study.

Chapter 2

Background

To situate this research into existing work an overview of relevant concepts and related research will be discussed. Research in the field of automated heart disease classification aims to aid professionals in diagnosing patients. But first, key concepts and background information of this field and the research presented will be covered.

2.1 2025 George B. Moody PhysioNet challenge

A goal which lies in parallel to this research goal is the one proposed by Moody et al. (2025), which states there is a need for prioritisation in CD confirmatory testing. The tests used to confirm a CD diagnosis are various serological tests, which from a blood sample, measure the levels of antibodies produced by the body in response to the *Trypanosoma Cruzi* infection (Centers for Disease Control and Prevention (CDC), 2024). According to Moody et al. there are a limited number of serological tests for which a prioritisation queue based on ECG findings is needed to confirm diagnoses and inform individuals on the impacts and treatments for CD. Hence, they invited teams in their annual challenge of 2025 to develop open-source algorithms to identify potential CD cases from 12-lead ECGs. This challenge emphasises the importance and forms a guideline for this study to develop a successful open-source algorithm to detect CD. k

2.2 The electrocardiogram

A normal ECG depicts the heart rhythm called a sinus rhythm, to form this signal, the electric pulses emitted by the heart during depolarisation are displayed. For analysis, a pulse is split into multiple segments like shown in Figure 2.1 (Ashley

and Niebauer, 2004). This includes the P wave, PR interval, QRS wave complex, ST segment, and T wave.
There exist multiple versions of the ECG exam, but the one studied in this research is the 12-lead ECG, which uses a total of 10 electrodes to obtain 12 different views of the heart which are visualised as leads. These 10 electrodes are attached to both arms, both legs and 6 to the chest to surround the heart. The leads these electrodes form are as follows:

- Lead I: vertical plane view from the right arm to the left arm
- Lead II: vertical plane view from the right arm to the left leg
- Lead III: vertical plane view from the left leg to the left arm
- aVR lead: vertical plane view from the right arm
- aVL lead: vertical plane view from the left arm
- aVF lead: vertical plane view from left leg
- V1-V6 leads: horizontal plane views from chest, where V1 is on the right center of the chest and V6 is on the far-left side.

These leads are each informative signals, but when combined form the best non-invasive way of inspecting the electrical activity of the heart. Full ECG exams, as seen in Figure 2.2a can be indicative for serological tests, which can confirm a CD diagnosis.

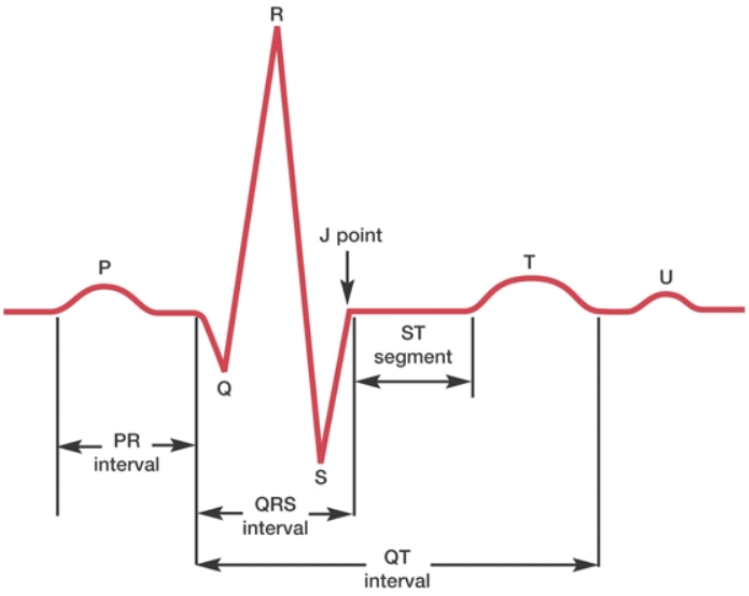


Figure 2.1: The basic pattern of electrical activity across the heart.

2.2.1 Normal and abnormal ECGs caused by chagas disease

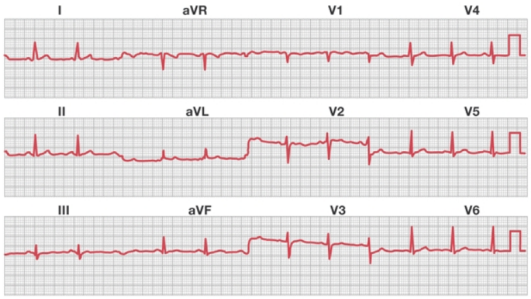
An abnormal ECG - such as one showing characteristics of chagas disease - regularly show various cardiomyopathies. According to a systematic review by Rojas et al. (2018) the most common cardiac problems caused by CD include: complete right bundle branch block (RBBB) - as seen in Figure 2.2b, left anterior fascicular block (LAFB) as seen in Figure 2.2c, a combination of complete RBBB/LAFB, first-degree atrioventricular block (AVB), atrial fibrillation (AF) or flutter - as seen in Figure 2.2d, and ventricular extrasystoles (VE).

In normal heart conduction, electrical signals travel evenly through both the left and right bundle branches, causing the septum to activate from left to right and creating small Q waves in the lateral leads.
In RBBB the left ventricle is activated normally, but the right ventricle is activated later because the signal has to cross from the left side (Burns and Buttner, 2024b). As a result, the ECG shows a second R wave (R') in the chest leads visible in the V1 lead in Figure 2.2b and a broad, slurred S wave in the lateral leads, such as lead I shown in the lower half of Figure 2.2b.

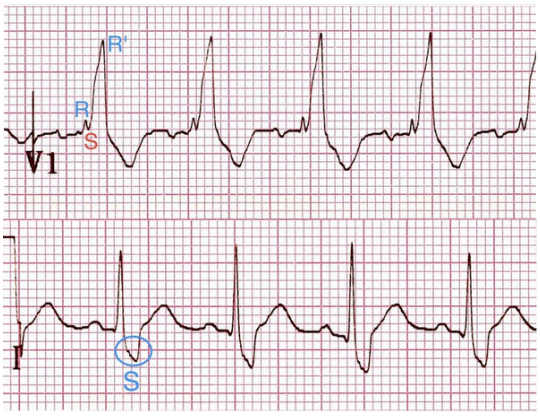
A left anterior fascicular block occurs when the electrical signal in the heart is

delayed or blocked as it travels through the left anterior fascicle of the left bundle branch. This causes the left ventricle to contract after the right ventricle and not simultaneously (Cleveland Clinic, 2025).
In Figure 2.2c different characteristics can be seen. As the rS complexes in leads II, III, aVF, contain small R waves and deep S waves. Further qR complexes in leads I, aVL, contain small Q waves and tall R waves. And finally, the left axis deviation (LAD) in leads II, III and aVF are negative and leads I and aVL are positive (Cleveland Clinic, 2025).

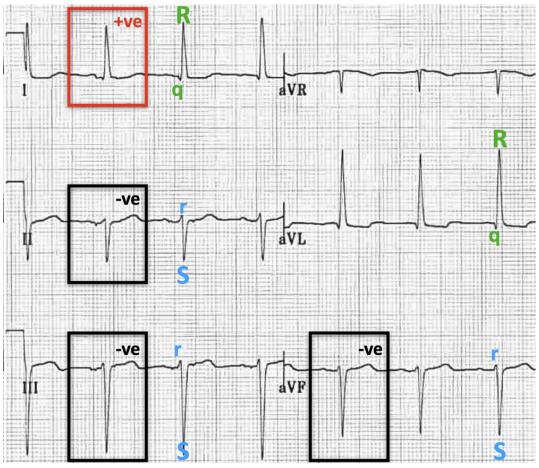
On the other hand, atrial fibrillation on the ECG is characterised by an irregularly irregular rhythm with no distinct P waves. The ventricular rate is variable, and the QRS complexes are generally narrow (<120 ms). Instead of P waves, fibrillatory waves may be seen, which can appear fine (amplitude <0.5 mm) or coarse (amplitude >0.5 mm) like in Figure 2.2d. These fibrillatory waves can sometimes resemble P waves, which may lead to misdiagnosis (Burns and Buttner, 2024a).



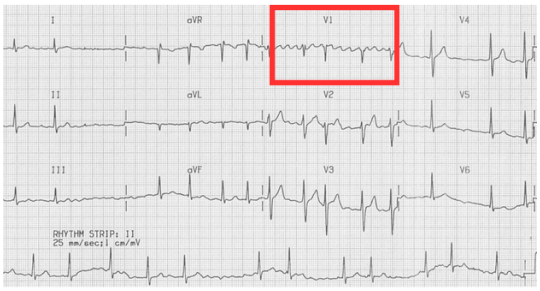
(a) Example of a normal 12-lead ECG exam



(b) Abnormal ECG V1 and I leads with signs of RBBB



(c) Abnormal ECG limb leads with signs of a LAFB



(d) Abnormal 12l-ECG with V1 fibrillatory waves, distinct to AF

Figure 2.2: Normal and abnormal ECG exams commonly caused by chagas disease.

2.3 Time-frequency analysis

Time-frequency analysis is a signal processing approach used to study how the distribution of strength over frequency evolves over time. To do this, time-frequency methods like the continuous wavelet transform (CWT) capture both temporal and frequential characteristics simultaneously to form a time-frequency representation (TFR). Techniques like the short-time Fourier transform and wavelet transform are commonly applied to biomedical signal analysis of ECGs and electroencephalograms alike (Pradhan et al., 2023), (Morales and Bowers, 2022). In these signals relevant features may only appear within specific time intervals or frequency bands.

And by representing the signal in the joint time-frequency domain, these methods provide a means for improved characterisation of transient patterns and non-stationary behaviour. This has proved useful for tasks like arrhythmia detection and disease classification (Pradhan et al., 2023)

2.4 Related work

Previous research in the field of automated CD detection using ECG signals is scarce. However, one study published by Jidling et al. in 2023 investigated whether deep neural networks can detect CD from 12-lead ECGs and be used as an automated screening tool. The authors trained a one dimensional convolutional residual network using two large Brazilian datasets: CODE - with over two million ECG recordings using self-reported CD-labels - and Sami-Trop - with serologically confirmed CD patients. The performance was evaluated on two external test cohorts: REDS-II - comprising 631 serologically confirmed patients - and ELSA-Brasil - comprising 13,739 civilians. The ELSA-Brasil and CODE have a CD prevalence of 2.0%, which is most representative of the Latin American population.

Their findings for internal validation include an AUROC of 0.80 and an F1 score of 0.64. On external evaluation, performance dropped: AUROC 0.68 and F1 score 0.61 on REDS-II; AUROC 0.59 and F1 score 0.06 on ELSA-Brasil. When strictly classifying patients with chronic CD, which has more distinct characteristics, their results changed: AUROC 0.82 and F1 score 0.59 on REDS-II; AUROC 0.77 and F1 score 0.02 on ELSA-Brasil. In external evaluation, sensitivity ranged from 0.36 to 0.52, while specificity from 0.76-0.77.

Grad-CAM interpretability analysis indicated that the network focused on QRS regions often associated with conduction disturbances like RBBB, typical of chronic CD.

Finally, the authors conclude that while their model can reliably detect chronic CD, it is less effective at classifying early-stage CD.

2.4.1 Deep learning in electrocardiogram diagnosis

A 2021 review by Liu et al. provided a comprehensive overview of the DL techniques that had been applied in general ECG diagnosis. This highlighted the transition from traditional rule-based and feature-engineered methods toward data-driven DL approaches. This change resulted from the advances in CNNs, recurrent neural networks (RNNs) and hybrid models.

The architectures most effective for ECG diagnosis were discussed and included: CNNs, which proved effective for local feature extraction; RNNs (especially LSTMs and GRUs), which captured temporal dependencies well; and hybrid CNN-RNN models, whose usage had increased over time.

These DL models have been applied to ECG signals for arrhythmia detection, myocardial infarction diagnosis, atrial fibrillation detection, heart failure, sleep apnea, and even cardiovascular risk stratification.

The studies included in this review have used a variety of datasets, including the MIT-BIH arrhythmia database, PTB diagnostic ECG database, and larger population datasets like CODE and UK Biobank.

Resulting DL model performance was consistently higher than traditional machine learning methods, often matching or surpassing expert cardiologists in arrhythmia classification tasks specifically.

Limitations found included critical barriers for deployment such as dataset bias. Dataset bias arises when the training data fails to adequately reflect the real-world conditions in which the model is intended to be deployed (IBM, 2024). This can lead to biased or skewed outputs, and when used can predict incorrectly more often. This could obviously lead to detrimental outcomes in healthcare, which is why it must be avoided. Another limitation found was the lack of generalisability and interpretability of DL models.

This is why the authors suggested that combining ECGs with epidemiological and multimodal data and a focus on explainability would be key for clinical adoption in the future.

2.4.2 Time-frequency methods for electrocardiogram analysis

Another review: Pradhan et al. (2023), explored time-frequency analysis techniques applied to ECG signals. This analysis represents ECG signals simultaneously in both time and frequency domains. unlike standard time-series data representation of ECGs, the TFR offers richer insight into the heart's electrical activity.

The review provided a detailed overview of principal methods including: short-time Fourier transforms, continuous and discrete wavelet transforms (CWT and DWT), empirical mode and wavelet packet decompositions, Wigner-Ville distributions, and Hilbert Huang transforms. These techniques are utilised across various tasks in ECG analysis, including: ECG denoising, arrhythmia or sleep apnea detection, and biometric identification.

While the DWT is the method most used, especially for denoising, techniques that convert ECGs into 2D representations are increasingly integrated with DL models for classification tasks like arrhythmia detection.

The benefits of TFR also included limitations, as features extracted via time-frequency techniques often lacked intuitive interpretability as humans are generally not used to a representation of this form. Also, the higher dimensionality of feature sets resulting from TFR posed a challenge for traditional machine learning, which DL could mitigate by inherently engineering features. Finally, the authors called for more deployment of advanced TFR methods beyond arrhythmia detection, emphasising the need for research on emerging domains.

2.4.3 Literature gap

The limitations discussed and the future implications called for are the reason this research aims to include epidemiological data from the CODE dataset, representative of the endemic populations. And as only one other documented research is available on CD classification, this study will propose a novel approach to classifying CD from 12-lead electrocardiograms. As the continuous wavelet transform 2D time-frequency distribution of ECGs is used as an information enriched input for two-dimensional CNNs. This concludes the gap in literature this research aims to fill.

Chapter 3

Method

This study utilises a supervised learning approach to classify ECGs as either CD-positive or CD-negative using different CNNs. The classification will be based on a two-dimensional time-frequency distribution of the ECG signal, as this representation leverages the temporal and frequential patterns in the of ECGs which are indicative of various cardiac abnormalities (Pradhan et al., 2023). The supervised learning approach was adopted as this was best suited given the labelled data and the goal of binary classification. This research aims to gain insight into the performance of CNNs in classifying CD from ECG signals transformed to the time-frequency domain.

3.1 Data description

The data used in this research is the CODE-15% dataset, which was obtained through stratified sampling 15% of the CODE dataset, and contains 345,779 exams from 233,770 different patients. The data was used in a previous study by Ribeiro et al. and was collected by Telehealth Network of Minas Gerais (TNMG), a public telehealth system in Minas Gerias, Brazil Ribeiro et al. (2020). The dataset contains ECG exams, with a recording duration from 7 to 10 seconds and sampling frequencies ranging from 300 to 600 Hz, and comprises 18 batches. The dataset has a heavily imbalanced CD to non-CD ratio of 1:49, which is representative of reality in Brazil, but forms a challenge in the handling of DL models, see section 3.4 to see how this was circumvented. Along with a unique exam ID and an ECG recording, one exam also contains labels for gender, age, predicted age by an ANN, patient ID, death, time between exams, automated systems' binary flag for a normal ECG, the exam its batch number, and finally multiple labels for strictly validated cardiac abnormalities. As mentioned in Ribeiro et al., 2020 in the process of data collection for the CODE dataset done by TNMG exams were

sent from medical facilities to TNMG for analysis. When analyzed at TNMG by an experienced cardiologist, a diagnosis of CD is not necessarily validated and is therefore not included in CODE-15%.

The ECG recordings are stored in 18 batches as HDF5 files, which each contain two datasets called 'exam_id.' and 'tracings.' The 'exam_id' dataset is a tensor of dimension $(N,)$, where N is the number of exams in the batch and 'tracings' is a tensor of dimension $(N, 4096, 12)$ where N corresponds to the different exams, 4096 to the signal samples, and 12 to the different leads in the ECG recording. The 12 leads are stored in the following order: I, II, III, AVR, AVL, AVF, V1, V2, V3, V4, V5, V6. For an explanation on how each lead measures and represents cardiac activity see section 2.2. The signals, sampled at 400 Hz, have a duration of 7 seconds, comprising $7 \times 400 = 2800$ samples, to 10 seconds, comprising $10 \times 400 = 4000$ samples. Finally, Ribeiro et al. (2020) has padded all recordings in the CODE-15% dataset with zeros on both sides to get a length of 4096 samples for all recordings, which is also the length of the signals used in this study.

3.2 Data preparation and preprocessing

This study has refined and enhanced the original CODE-15% data to allow the application of CD classification through annotation. The annotations used in this study are self-registered CD diagnoses by patients of the CODE study and may or may not be validated Moody et al. (2025). These CD-labels were provided by PhysioNet who published this for the 2025 George B. Moody PhysioNet challenge mentioned in section 2.1. To use the CODE-15% data stored in HDF5 files and the CD labels given by PhysioNet, this study utilises the following data preparation pipeline.

Before data preparation had commenced all available 'exams_part{i}' HDF5 files - with i corresponding to the 18 batches - and the 'exams.csv' file were downloaded from the CODE-15% online publication. These files were then uploaded to the Snellius-SURF server for which access was provided by the *Universiteit van Amsterdam* to perform computationally expensive tasks performed in this study, such as training and evaluating machine learning models.

3.2.1 Preparation stage

In the first stage of the data pipeline, called the preparation stage, the relevant recordings from the HDF5 files were extracted and converted into WFDB-

compatible files, which is a format frequently used for representing digitized signals. In this WFDB-format data is stored as '.dat' and '.hea' file pairs. Subsequently, for each record the corresponding exam ID was matched with demographic data and a binary CD label. The physical ECG signals were then scaled to digital units using a fixed gain factor and pinned to the 16-bit integer range to maintain compatibility with standard ECG formats (Moody et al., 2025). To conclude this stage, the data were saved in 12-lead WFDB-format with respective metadata and corrected checksums often used to verify data integrity (Meylan et al., 2020).

3.2.2 Preprocessing stage

In the second stage of the data pipeline, called the preprocessing stage, each ECG signal underwent a series of signal processing steps designed to standardise input and enhance feature representation. In this stage 4th-order Butterworth high-pass filter with a 0.5 Hz cut-off was applied to each lead separately. This is highly effective in removing common artifacts in ECG recordings such as baseline wander and low-frequency drift. The 4th-order Butterworth high-pass filter is described by the following formula:

$$y[n] = \sum_{k=1}^{M} b_k \cdot x[n-k] - \sum_{k=1}^{N} a_k \cdot y[n-k] \tag{3.1}$$

Where:

- $x[n]$ is the input signal at time index n,
- $y[n]$ is the filtered output,
- b_k and a_k are the filter coefficients (determined via a Butterworth design),
- M and N denote the filter order (in this case, both are 4).

Furthermore, each lead was standardised independently using z-score normalisation which centres the signal and scales it to unit variance. This step standardises the amplitude distributions of the ECG waves and ensures stable training (Fei et al., 2021). for each lead the z-score normalisation is given by the following formula:

$$\tilde{x}_t = \frac{x_t - \mu}{\sigma} \tag{3.2}$$

where:

- x_t is the amplitude value of the signal at time index t,
- $\mu = \frac{1}{N}\sum_{i=1}^{N} x_t$ is the sample mean of the signal,
- $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_t - \mu)^2}$ is the sample standard deviation,
- $\tilde{x_i}$ is the normalised signal value,
- N is the number of samples.

To avoid division by zero, a small constant or floor value was substituted in case of $\sigma = 0$.

And finally to capture both time and frequency characteristics of the ECG signals in the form of a time-frequency representation (TFR), a continuous wavelet transform (CWT) was applied to the signal, using complex Morlet wavelets. Unlike other methods like the Fourier transform, which assumes stationarity of a signal, the CWT utilises localised time-frequency analysis with the use of wavelets like the Morlet wavelet, see **??**. The localised time-frequency analysis makes it well suited for physiological signals like ECGs that exhibit non-stationary behaviour in the mean and variance of the signal.

In this study the CWT of an ECG signal was computed using neurodsp, where the CWT is implemented as a convolution between the ECG signal and a set of complex Morlet wavelets tuned to a specific range of frequencies. In this convolution each wavelet passes frequencies within a certain time range of the signal and rejects frequencies outside of that time range, enabling the extraction of frequency-specific signal strength over time.

The implementation of the complex Morlet wavelet $\psi(t;\omega)$ for time range τ and the number of cycles ω is mathematically formulated as:

$$\psi(\tau;\omega) = \pi^{-1/4}(e^{i\omega\tau} - e^{-\omega^2/2})e^{-\tau^2/2} \tag{3.3}$$

Where:

- $\tau \in [-2\pi s, 2\pi s]$ in this implementation and is the range of time in which the wavelet resides,
- in which s scales the time window, not the wavelet itself - set to 1,
- and ω is the number of cycles or wave-oscillations in the wavelet - set to 5,

- $e^{-\omega^2/2}$ is the term subtracted to ensure a zero-mean needed to qualify as a mother wavelet Ψ.

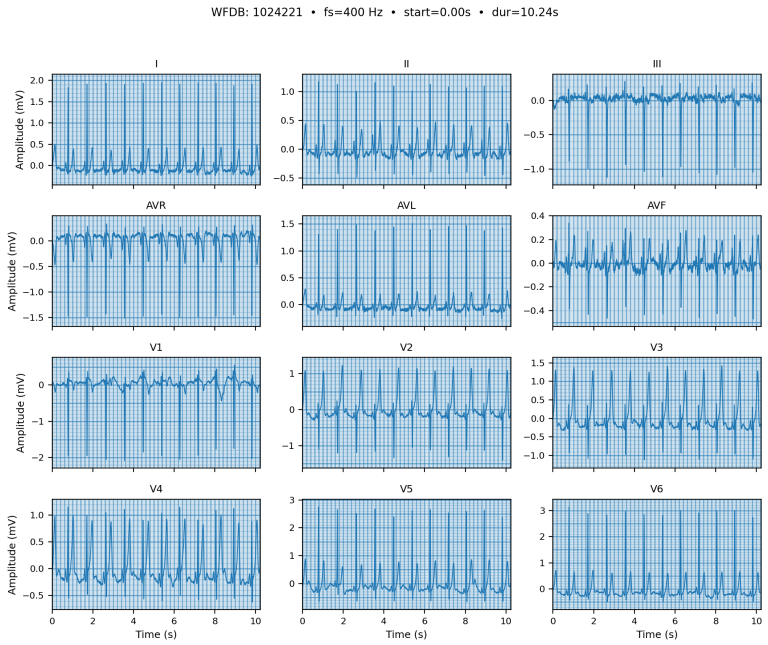
Subsequently, the Morlet wavelet is used as the mother wavelet to compute the CWT. Mathematically, the CWT is represented using

$$CWT_{\tau,s,\Psi}(x) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \Psi^*_{\tau,s}(t) dt \tag{3.4}$$

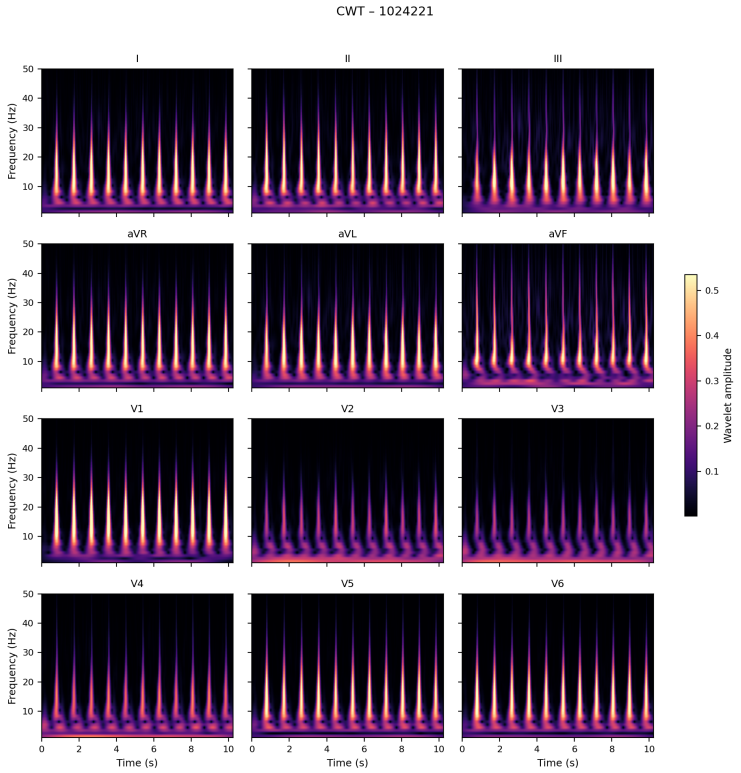
Where:

- $\Psi_{\tau,s}(t)$ is the mother wavelet and * represents the conjugate function.
- $\frac{1}{\sqrt{s}}$ is used to normalise the mother wavelet Ψ

The TFR resulting the CWT is computed for the frequency range of 1 to 50 Hz for all 12 leads of the ECG recordings and is the final input to all DL models in training and evaluation. A comparison of a regular ECG and its TFR is shown in Figure 3.1.



(a) Standard 12-lead ECG exam 1024221



(b) Time-frequency distribution of ECG exam 1024221

Figure 3.1: Comparison of regular and TFR ECGs

3.3 Model architectures

In this study multiple model architectures were compared, including models based on research conducted by the authors Tan and Le on more parameter efficient versions of their original EfficientNet models proposed 2 years prior (Tan and Le, 2019). The EfficientNetV2 model series, consisting of the small, medium and large variants, outperformed earlier iterations of the EfficientNet architecture in ImageNets top-1 accuracy scores, CIFAR, Flowers-, and Cars datasets. This and the fact that EfficientNetV2 models have slightly better parameter efficiency, yet train and infer faster than the originals, was pivotal when considering CNNs for this study. As the models suited for this CD classification task should not necessarily have many parameters.

3.3.1 2D CNNs

The models discussed in the section are based on the EfficientNetV2 models, where minor modifications allowed for passage of the 2D, 12-lead data used in this research (Tan and Le, 2021). The small and large variations of the model have both been imported via their PyTorch implementations and altered to accept a 12 channel input. This was done to make the models capable of accepting all 12 leads of an ECG recording, which is more optimal for model-reasoning as it avoids unnecessarily truncating the data channels. Although the 50×4096 input was dynamically supported by both models, both required an additional linear layer to output a binary classification for CD. Both models extensively use the MBConv and Fused-MBConv operations described in Tan and Le (2021). EfficientNetV2 employs Fused-MBConv operations in stage 1 to 3, followed by MBConv operations in stage 5 to 6, each with multiple layers of the respective operation in which it uses a 3×3 kernel. Stage 7, which leads into the output layer, comprising a regular convolution operation, average pooling and a fully connected layer. A visual representation of the smaller, 2D CNN-v1 model implemented is depicted in Figure 3.2.

In Tan and Le, 2021 EfficientNetV2-s was scaled to obtain EfficientNetV2-M/L using compound scaling proposed in 2019 by Tan and Le. Where in 2021 the authors describe how they restricted inference image size to 480×480 and gradually added more layers in later stages to increase network capacity without increasing the runtime. The same structure and additional layers were used in this study to form the second, larger 2D CNN based on EfficientNetV2-L which will be referred to as 2D CNN-v2.

The threshold parameters in the output layers used for these 2D CNNs were 0.5 for the base v1 and v2 and 0.11985953479190482 for v1 after calibration using Platt scaling, which will be covered in subsection 4.1.2.

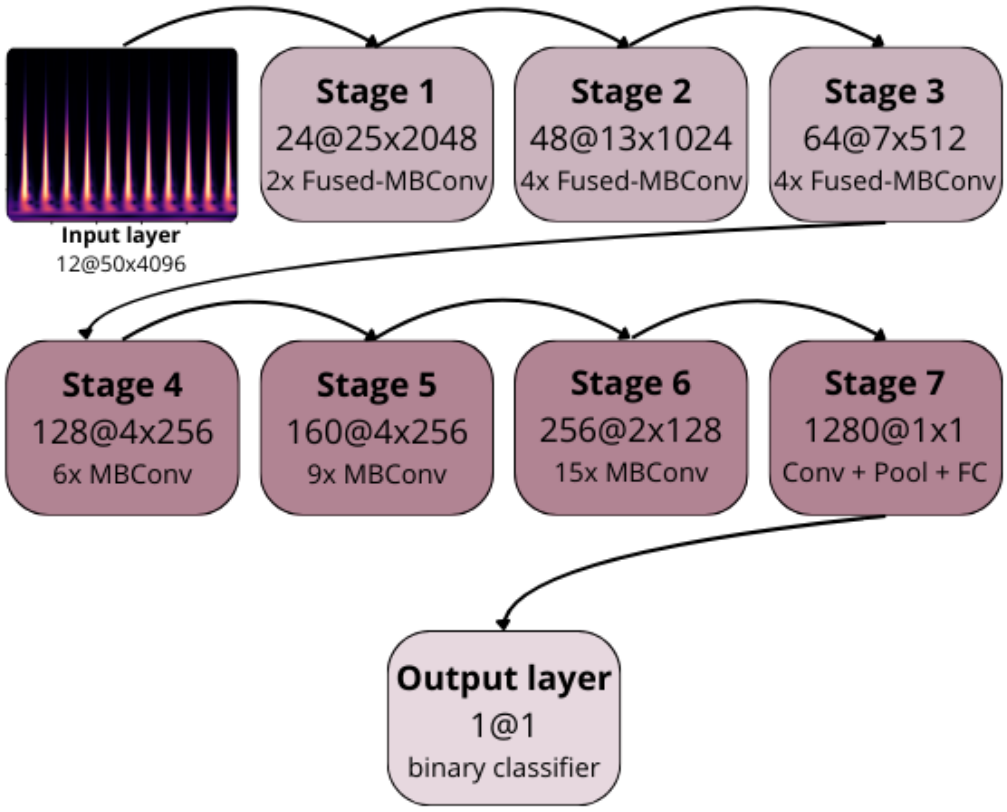


Figure 3.2: Architecture of 2D CNN-v1 used to classify chagas disease. Number of channels, dimensionality, the amount of layers and type of operation are also denoted

3.4 Training procedure

For all implementations in this study the 18 CODE-15% batches were split to form the following train-test-validation split: batch 0 to 2 - comprising 59,841 records - is considered the test set, batch 3 - comprising 19,946 records - is the validation set, and finally batch 4 to 17 - comprising 277,639 records - is the training set. This train set was restructured to a balanced set, as the original dataset was heavily imbalanced, see section 3.1 and a balanced dataset, which represents all categories equally, improves the ability for DL models to distinguish different classes in classification. Practically, this meant that during training the models

would be exposed to equal numbers of CD-positive and negative cases, granting them the opportunity to learn from the patterns in the ECG data more optimally. To accomplish this, firstly, 20% of 5,047 CD positive cases in the train set were randomly up-sampled to get 6,056 positive records. Secondly, this was then balanced with random negative cases to create a 1:1 ratio of the labels, which, after faulty exam recordings, resulted in a training set of 11,872 recordings. No data augmentation was performed on the test and validation sets. However, they were also processed to clear faulty ECG recordings, resulting in a test and validation set comprising 59,700 and 19,904 exams respectively. The test set was the set used to evaluate the models and the validation set to calibrate v1 using Platt scaling.

All models were trained using the binary cross-entropy with logits loss function (BCEL) which, in contrast to the binary cross-entropy (BCE) loss function, possesses the desirable property of numerical stability. Numerical stability is the absence of numbers that are either too large or too small to store in memory, leading to overflow and underflow, resulting in unusable logits (raw model output). The BCEL loss function is more numerically stable as it applies a sigmoid activation to the logits of the models, in contrast to the BCE loss function, where the sigmoid activation must be applied separately from BCE, possibly leading to over- or underflow. The BCEL loss function: $BCEL_{x,y}$ is formulated as:

$$BCEL_{x,y} = -(y \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))) \tag{3.5}$$

Where:

- $x \in \mathbb{R}$ is the model's raw output (logit) before activation,
- $y \in 0, 1$ is the true label, either not CD or CD,
- $\sigma(x) = \frac{1}{1+\exp -x}$ is the sigmoid activation of a logit x

When the training pipeline is executed for 5 epochs it splits the training data into batches of 16 recordings. For each batch the gradients are cleared, and a forward pass is executed. Subsequently, the BCEL loss is calculated, and a backward pass computes the gradients. Finally the parameters are adjusted according to the Adam optimiser algorithm.

3.5 Metrics

To evaluate the performance of the proposed models extensively, several metrics were employed. The choice of metrics was based on the characteristics of the

dataset, the clinical priority of classifying CD, and the evaluation standards established in the related PhysioNet challenge. The metrics used are: accuracy, recall, the PhysioNet challenge score, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), average precision (AP), and the receiver operator characteristic- and precision-recall curves.

Accuracy was included as a general, intuitive measure of proportion of correctly classified cases. However, it can be misleading in imbalanced datasets, since it is disproportionately influenced by the majority class: CD negative cases. For this reason, other metrics are reported alongside accuracy.

Recall, equivalent to sensitivity or the true positive rate, measures the proportion of correctly identified CD positive cases. This metric is of high clinical importance, as false negatives could lead to missed diagnoses. It is thus emphasised to reflect the value of minimising undetected positive CD cases in a screening context. Recall was also the premature scoring format for the 2025 George B. Moody PhysioNet challenge before they ultimately changed it to a custom challenge score loosely based on the recall.

The challenge score is a complex metric introduced by the authors for the 2025 George B. Moody PhysioNet challenge, designed to give a benchmark for the performance of a model in a scenario where only 5% of cases seen can be classified as CD positive. It calls for adequate prioritisation by the model evaluated, as the 5% represents the estimated capacity of serological tests used to confirm a CD diagnosis.
on the webpage for the challenge it is mathematically defined that X is a dataset comprising n ECG recordings. Also let p_x be the model's estimated probability of a positive case of CD for any ECG recording $x \in X$. Now let $x_1 = \text{argmax}_{x \in X} p_x$ be the record with the highest probability, x_2 the record with the second highest probability, etc. They define $X_\alpha = \{x_k \in X : k \leq \alpha n\} \subseteq X$ as the collection of recordings containing the $k = \lfloor \alpha n \rfloor$ highest probabilities. Then, the challenge score is the true positive rate for X_α, in other words, the fraction of CD positive cases in X_α out of the total number of CD positive cases in X, for $\alpha = 0.05$ (Moody et al., 2025).

The AUROC metric summarises a model's ability to discriminate between positive and negative cases across all possible thresholds. It is threshold-independent and a higher AUROC indicates that a model is better at ranking positive cases above negative cases, regardless of the threshold chosen. It does give a biased view of performance of models evaluated on highly imbalanced datasets, as the denom-

inator in the x axis: false positive rate contains the number of true negatives $FPR = \frac{FP}{FP+TN}$. In an imbalanced dataset this number is much larger than the nominator containing the number of false positives, minimising the impact false positives have on this scale and increasing the area under the receiver operator curve.

This is why the AUPRC is included, as it is more informative than AUROC when class imbalance is present. Precision-recall analysis emphasises the trade-off between sensitivity and the proportion of false positives, which is particularly relevant for screening methods. The average precision score is also reported as a single summarising statistic for the PR curves.

In addition to scalar metrics, ROC and PR curves are plotted to provide a visual presentation of model behaviour across thresholds. The curves display the trade-offs between TPR, FPR and precision, and enable a more nuanced interpretation of model performance than numeric metrics. As the ROC curve possesses a left-skewed bias due to class imbalance, the PRC elaborates on the sensitivity-precision trade-off to complete the comprehensive evaluation.

Chapter 4

Experiments

4.1 2D CNN-v1

In this section the experiments performed with a 2D CNN mentioned in subsection 3.3.1, which has a relatively low number of parameters, will be highlighted. Two main experiments were performed evaluating the 2D CNN-v1 on the test set: the baseline model and a calibrated version for which Platt scaling was used.

4.1.1 Baseline

The first experimental setup trained the baseline 2D CNN-v1 model on a balanced subset of the CODE-15% dataset defined as the training set in section 3.4. Each ECG recording was preprocessed using a continuous Morlet wavelet transformation, followed by normalisation and zero padding. In line with section 3.5, model performance was assessed on binary classification of CD using accuracy, recall, AUROC, AUPRC, AP, ROC and PRC, and the challenge score designed by PhysioNet (Moody et al., 2025). More metrics, including: specificity, precision and F1 score were calculated to allow indirect comparison to the previous automated CD screening results in Jidling et al. (2023).

4.1.2 Platt scaling calibration

After assessment of 2D CNN-v1, the baseline model was further subjected to calibration to improve recall and adjust the decision threshold accordingly. Calibration was performed using Platt scaling on the validation set, and this version was then evaluated on the same held-out test set as the baseline. Calibration resulted in a threshold of 0.1198595347919048, which was substantially lower than baseline. Upon evaluation, a major improvement to the model its recall was evident, in-

dicating improved detection of CD. But this came at the cost of a reduction in accuracy, precision and F1 score, in turn indicating a less balanced model. Confusion matrices for baseline and calibrated v1 were also made to further assess and compare the different v1 models.

4.2 2D CNN-v2

The second architecture, 2D CNN-v2, was a model with an increased number of convolutional filters and deeper layers to test the impact of capacity on classification performance. Training and evaluation were performed under the same conditions as 2D CNN-v1, ensuring the comparability of results.

Chapter 5

Results

The results from the experiments mentioned in chapter 4 will be outlined in this chapter, as will further metrics of performance regarding the various models presented in chapter 3. When evaluating the different models on the test set, the recall (or true positive rate), accuracy, ROC, AUROC, PRC, AUPRC, AP and challenge score were considered as substantiated in section 3.5. They will provide the overview of the models' performance on this binary classification task or were metrics of interest to the 2025 George B. Moody PhysioNet challenge Moody et al. (2025).

Table 5.1 presents the evaluation metrics for the baseline and calibrated 2D CNN models. Recall, Accuracy, AUROC and challenge score are shown in Table 5.1, while the AUPRC, AP, ROC and PRC are shown in Figure 5.1. 2D CNN-v1 outperforms v2 on all of these metrics except the recall. In addition, 2D CNN-v1 achieves the highest overall accuracy (0.745), AUROC (0.826), and challenge score (0.356). Calibrating 2D CNN-v1 increased its recall to 0.96 while reducing its accuracy significantly (0.404). 2D CNN-v2 achieved slightly higher recall than the base v1 model (0.756), but lower values for accuracy (0.698), AUROC (0.806), and the challenge score (0.331).

Table 5.1: Performance metrics for baseline and calibrated 2D CNN models

Model type	**Recall**	**Accuracy**	**AUROC**	**Challenge score**
2D CNN-v1	0.741	**0.745**	**0.826**	**0.356**
Calibrated 2D CNN-v1	**0.960**	0.404	**0.826**	**0.356**
2D CNN-v2	0.756	0.698	0.806	0.331

When inspecting the confusion matrices of the small 2D CNN before and after calibration in Table 5.2 it becomes clear that the base model shows a higher number of correctly classified negative cases with a larger proportion of false negatives. After calibration, the number of correctly identified CD positive cases increased, with only 48 false negatives compared to 315 before calibration. This increase in sensitivity came at the cost of a considerable increase in false positives. These changes are consistent with the observed trade-off between recall and accuracy depicted in Table 5.1 and further addressed in chapter 6.

Table 5.2: Comparison of confusion matrices before and after model calibration. Green = correct predictions; red = errors.

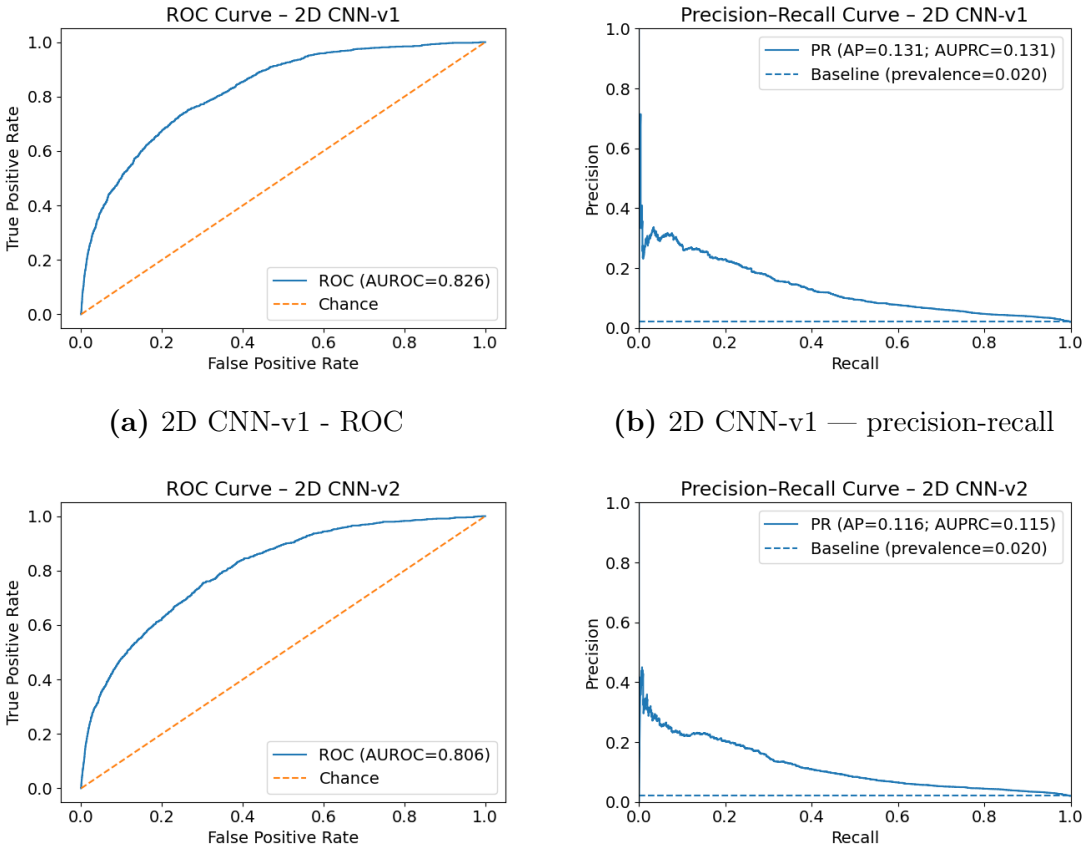
Baseline 2D CNN-v1 model confusion matrix

	Predicted True	Predicted False
True label	900	315
False label	14927	43558

Calibrated 2D CNN-v1 model confusion matrix

	Predicted True	Predicted False
True label	1167	48
False label	35504	22981

While the ROCs in Figure 5.1a and Figure 5.1c show models which appear to distinguish the positive and negative cases well, due to the imbalanced nature of the data it gives a slightly biased view of their performance. This is because the x-axis in a ROC curve: the false positive rate, contains the number of true negatives in its denominator, thus skewing the curve more left on the plot, consequently increasing the AUROC. This is why the precision-recall curves (PRC) are also shown, as they provide another view of the models' performance on a heavily imbalanced dataset. In a PRC the baseline AUPRC score for random classification is equal to the prevalence of the true label, which in this case is 0.02 or 2%. The small and large 2D CNNs AUPRCs of 0.131 and 0.115 respectively indicate that the models perform better than baseline. But when inspecting the PRCs in Figure 5.1b and Figure 5.1d it is evident that both models are unlike the sloping shape of a PRC expected from DL methods and are far from a perfect PRC with AUPRC=1 and a curve approaching (1,1).



(a) 2D CNN-v1 - ROC

(b) 2D CNN-v1 — precision-recall

(c) 2D CNN-v2 — ROC

(d) 2D CNN-v2 — precision-recall

Figure 5.1: Comparison of ROC and PR curves for v1 and v2 models.

An indirect comparison between the 2D CNNs and the model proposed by Jidling et al. in 2023 is presented in Table 5.3. Here the evaluation of the 1D CNN by Jidling et al. was done externally, on a dataset that consisted of 13,739 patients with 280 serologically validated CD patients. The dataset is called ELSA-Brasil and has the same prevalence as the CODE-15% dataset used for evaluation in this study of 2.0%, which makes this a good evaluation for comparison. The metrics included in this table are all the metrics used in Jidling et al. (2023). It is clear that the 2D CNNs score higher than the 1D CNN on almost all metrics. It also stands out that the base 2D CNN-v1 model scores higher than the 1D CNN across all metrics.

Table 5.3: Performance metrics for 1D CNN evaluation on ELSA-Brasil and 2D CNN base and calibrated models on CODE-15%.

Metrics	**1D CNN**	**2D CNN-v1**	**calibrated 2D CNN-v1**	**2D CNN-v2**
Recall	0.36	0.74	**0.96**	0.76
Specificity	0.76	0.99	**1.00**	0.70
Precision	0.03	**0.06**	0.03	0.05
F1 score	0.06	**0.11**	0.06	0.09
AUROC	0.59	**0.83**	**0.83**	0.81
Average precision	0.04	**0.13**	**0.13**	0.12

Chapter 6

Discussion

The base 2D CNN models for CD detection showed reasonable accuracy and performance on distinguishing CD positive cases from negative cases. Furthermore, the calibrated v1 model showed a great increase in recall at the cost of a major reduction in accuracy and an increased number of false positives. Further AUC analysis shows the ability of models v1 and v2 to distinguish CD from non-CD cases. Considering a left-skewed bias which occurs in highly imbalanced datasets, the models still show good ability to distinguish CD from non-CD . Moreover, PRC analysis shows model v1 outperforming v2 once more where both v1 and v2 models classify CD five to six times better than random baseline. Finally, when indirectly compared to previous work evaluated on a similar evaluation set all scores previously achieved by a 1D CNN implemented in Jidling et al. (2023) are bested.

6.1 Interpretation of results

When screening for any disease on a large scale, vital considerations must be made. For this, criteria have been constructed over 50 years ago by Wilson and Jungner which are still regarded as the golden standard today (Andermann et al., 2008). One such criteria is the acceptability of the test, stating the test must be accurate, reliable, easy to perform and acceptable to the target intervention. With no clear definition of accuracy given by Wilson and Junger, one could argue an accuracy score of 0.745 by v1 is reasonable, but still insufficient to be viable for healthcare implementation. However, with a challenge score of 0.356 by v1, resembling a real-world scenario where only 5% of cases can be classified as CD it could be argued that v1 and v2 would perform well above average in real life screening implementation too, as this is much higher than baseline.

Furthermore, the high number of false positives caused by the reasonable accu-

racy combined with the low prevalence of CD conflicts with another criteria: the cost-effectiveness of screening. When too many patients come for confirmatory testing after the automated CD-ECG screening done by one of these CNNs, the costs of all false positive patients unnecessarily taking serological testing would be excessive and will possibly not outweigh the benefits.

Looking at these criteria, it becomes evident that in the accuracy-recall trade-off between base v1 and its calibrated version, the base model is preferred as it has a higher accuracy and predicts less false positives, which are cost-inefficient.

Additionally, when comparing v1 and v2, the higher accuracy of v1 possibly outweighs the slightly higher recall v2 achieves as this aligns better with the acceptability of a test. When comparing AUROC and AUPRC it also becomes clear that v1 is the preferred model for classifying CD from ECGs, as it achieves higher scores than v2 for these metrics.

Furthermore, from ROC analysis it can be concluded that both models' curves are a biased view of their ability to distinguish CD from non-CD cases as the data is highly imbalanced. But their ability to distinguish CD and non-CD remains a strength of these models.

Finally, the PRCs, with a more representative view of performance, show v1 and v2 struggling to generalise. The steep curve at low recall (bottom left of Figure 5.1b and Figure 5.1d) indicates that for thresholds with lower recall the models recognise characteristic features indicative of CD, but produce more false alarms at higher recall thresholds. Meaning they cannot generalise to more complex features of CD cases very well. This means the v1 and v2 models identify a small subset of high-confidence CD cases, but struggle to generalise to a more heterogeneous patient population. This makes sense, as CD can be hard to distinguish at an early stage, as it can present as numerous cardiomyopathies but also be asymptomatic.

6.2 Contributions

Comparison between the evaluations done in this study and the validation of the 1D CNN model for CD screening proposed by Jidling et al. (2023) shows v1 outperforming the 1D CNN in AUROC, recall and specificity. However, the CNN introduced in Jidling et al. (2023) outperforms 2D CNN-v1 on all other metrics, mostly due to the low precision of v1 (0.06) which in turn affects its F1 score and average precision. This can be attributed to the different evaluation sets used. As Jidling et al. (2023) used a mix of CODE-15% and Sami-Trop data to manually

increase CD prevalence in the (validation) data to create a balanced validation set. Nevertheless, external validation conducted in that study used datasets of lower CD prevalence, including ELSA-Brasil, which has the same prevalence as the test set used for this study (0.02). When comparing resulting metrics of this evaluation to base and calibrated 2D CNN-v1, and base 2D CNN-v2 in Table 5.3, the 1D CNN is outperformed on all but three metrics. 2D CNN-v1 even outperforms it on all metrics, therefore indirectly indicating it possesses better chagas disease classification capability and could thus receive higher implication to incorporate in chagas disease screening programs.

6.2.1 Novelty and practical contribution

The use of the time-frequency distribution of ECG signals as input is a novelty within the field of CD classification. This is backed by a strong methodological foundation with a wavelet preprocessing pipeline that allows for use of the TPR, and multiple metrics to evaluate the models' performance. Possibly leading to practical contributions to the medical field in the form of DL models with the ability to classify and screen for CD with reasonable accuracy.

6.3 Limitations and possible confounders

In contrast to its strengths, a comparison of a model, excluding its input dimensions, could have been used to directly compare performance between using the TPR and using the 1D ECG signal as input. Now, only TPR input was used and compared to a previous study.

Furthermore, a limitation of the evaluation was the fact that the models were not evaluated on the external test set on the George B. Moody PhysioNet 2025 challenge website, as the storage needed for inference exceeded the limit the organisation had set for one submission and this form of external evaluation was thus unfeasible. This meant the challenge score in this study had to be computed on the test set and not an external set as provided by the challenge. This caused the inability to directly compare these models with others submitted to the challenge. However, an indirect comparison between a challenge score of 0.356, achieved by v1, and the submitted models on the website would place this model in 25th out of 295 submissions, as of August 20th, 2025.

Also, despite its higher parameter count and results from previous studies (Tan and Le, 2021), 2D CNN-v2 achieved underwhelming results. This can be attributed to a restriction set by the original EfficientNetv2-L implementation that restricts

image size to 480 as input for inference. This smaller input size possibly reduced the part of the input that was at disposal for computation during inference, consequently diminishing the performance of 2D CNN-v2 on evaluation.

Further, a possible confounder for 2D CNN-v1 and v2 outperforming the 1D CNN introduced in Jidling et al. (2023) is that the superior performance was caused by the choice of this study its base model (EfficientNet-v2) which could have higher classification task performance than the 1D CNN, regardless of the TPR preprocessing pipeline used in this study.

Finally, the CODE-15% dataset used in this study utilises self-reported CD labels, which may or may not have been validated (Moody et al., 2025). This means there is a possibility that the models have tried learning some patterns, which are not truly characteristic of CD, but just a result from a faulty self-reported CD label. To combat the impact of this choice, in the evaluation of the models another external evaluation could have been performed.

6.4 Future implications

Future research should focus on the use and construction of larger serologically validated datasets for CD cases, where the Sami-Trop dataset is a good example and was not used in this study because of its smaller size.

Another focus should be comparing different ways of training various DL models (not just CNNs) on imbalanced datasets. In this study training on a balanced dataset was opted for, but another method proven to be effective is the use of focal loss for training, where the weights of false predicted samples are adjusted to reduce the impact of an imbalanced positive class (Mulyanto et al., 2021).

A broader application could be realised, allowing for multi-class classification where CD is included in a list of numerous cardiac diseases.

A final implication is the use of the Mexican hat wavelet transform for ECG-based CD classification instead of the Morlet, as the Mexican hat has shown outstanding results in ECG classification (Wang et al., 2021).

Chapter 7

Conclusion

In conclusion, all 2D CNNs implemented in this study show potential for chagas disease screening, evident through their ability to classify chagas disease from 12-lead electrocardiograms with reasonably high accuracy. Their further chagas disease distinguishing ability and signs of confident classification in precision recall trade-off proven are other strengths of the 2D CNNs. However, they cannot classify chagas disease from electrocardiograms sufficiently accurate and sensitive to replace all screening methods. Nevertheless, indications of the time-frequency methods used to improve overall performance over current methods are noticeable, but undisputed evidence remains absent. These models, specifically 2D CNN-v1, do possess the potential to improve chagas disease detection over other implementations currently available.

Bibliography

Andermann, A., Blancquaert, I., Beauchamp, S., and Déry, V. (2008). Revisiting wilson and jungner in the genomic age: a review of screening criteria over the past 40 years. *Bulletin of the World Health Organization*, 86(4):317–319.

Ashley, E. A. and Niebauer, J. (2004). *Cardiology Explained*. Remedica, London. Accessed: 2025-08-07.

Burns, E. and Buttner, R. (2024a). Atrial fibrillation - ecg library. Published October 8, 2024.

Burns, E. and Buttner, R. (2024b). Right bundle branch block (rbbb) – ecg library. Published October 8, 2024.

Centers for Disease Control and Prevention (CDC) (2024). Clinical testing and diagnosis for chagas disease.

Cleveland Clinic (2025). Left anterior fascicular block. Last reviewed May 24, 2025.

de Fuentes-Vicente, J. A., Gutiérrez-Cabrera, A. E., Flores-Villegas, A. L., Lowenberger, C., Benelli, G., Salazar-Schettino, P. M., and Córdoba-Aguilar, A. (2018). What makes an effective chagas disease vector? factors underlying *Trypanosoma cruzi*-triatomine interactions. *Acta Tropica*, 183:23–31.

Fei, N., Gao, Y., Lu, Z., and Xiang, T. (2021). Z-score normalization, hubness, and few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 142–151.

Hussein, A. F., Hashim, S. J., Abdul Aziz, A. F., Rokhani, F. Z., and Adnan, W. A. W. (2018). Performance evaluation of time-frequency distributions for ecg signal analysis. *Journal of Medical Systems*, 42(1):15.

IBM (2024). What is data bias? Published October 4, 2024.

Jidling, C., Gedon, D., Schön, T. B., Oliveira, C. D. L., Cardoso, C. S., Ferreira, A. M., Giatti, L., Barreto, S. M., Sabino, E. C., Ribeiro, A. L. P., and Ribeiro, A. H. (2023). Screening for chagas disease from the electrocardiogram using a deep neural network. *PLOS Neglected Tropical Diseases*, 17(7).

Jun, T. J., Nguyen, H. M., Kang, D., Kim, D., Kim, D., and Kim, Y.-H. (2018). Ecg arrhythmia classification using a 2-d convolutional neural network. `https://arxiv.org/abs/1804.06812`. arXiv preprint arXiv:1804.06812.

Liu, X., Wang, H., Li, Z., and Qin, L. (2021). Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187.

Luz, E. J. d. S., Schwartz, W. R., Cámara-Chávez, G., and Menotti, D. (2016). Ecg-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127:144–164.

Meylan, A., Cherubini, M., Chapuis, B., Humbert, M., Bilogrevic, I., and Huguenin, K. (2020). A study on the use of checksums for integrity verification of web downloads. *ACM Trans. Priv. Secur.*, 24(1).

Mohammed, F. A., Tune, K. K., Assefa, B. G., Jett, M., and Muhie, S. (2024). Medical image classifications using convolutional neural networks: A survey of current methods and statistical modeling of the literature. *Machine learning and knowledge extraction*, 6(1):699–735.

Moody, G. B., Mark, R. G., Goldberger, A. L., and Clifford, G. D. (2025). Physionet/computing in cardiology challenge 2025. `https://moody-challenge.physionet.org/2025/`. Accessed: 2025-06-14.

Morales, S. and Bowers, M. E. (2022). Time-frequency analysis methods and their application in developmental eeg data. *Developmental Cognitive Neuroscience*, 54:101067.

Mulyanto, M., Faisal, M., Prakosa, S. W., and Leu, J.-S. (2021). Effectiveness of focal loss for minority classification in network intrusion detection systems. *Symmetry*, 13(1):4.

Nunes, M. C. P., Bern, C., Clark, E. H., Teixeira, A. L., and Molina, I. (2024). Clinical features of chagas disease progression and severity. *The Lancet Regional Health – Americas*, 37.

Pradhan, B. K., Neelappu, B. C., Sivaraman, J., Kim, D., and Pal, K. (2023). A review on the applications of time-frequency methods in ecg analysis. *Journal of Healthcare Engineering*, 2023(1):3145483.

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

Rojas, L. Z., Glisic, M., Pletsch-Borba, L., Echeverría, L. E., Bramer, W. M., Bano, A., et al. (2018). Electrocardiographic abnormalities in chagas disease in the general population: A systematic review and meta-analysis. *PLOS Neglected Tropical Diseases*, 12(6):e0006567.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*. PMLR.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR.

Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C., and Ou, C. (2021). Automatic ecg classification using continuous wavelet transform and convolutional neural network. *Entropy*, 23(1):119.

World Health Organization (2025). Chagas disease (also known as american trypanosomiasis). Accessed: 2025-05-15.

Appendix

7.1 Resources used

To time efficiently train the models used in this study I used the Snellius supercomputer services, for which on SLURM jobs the gpu_h100 partition was used. Programming languages, packages and dependencies used in this study include: Python 3.9.18, joblib 1.4.2, numpy 2.0.2, pandas 2.2.2, scikit-learn 1.6.1, wfdb 4.3.0, h5py 3.13.0, torch 2.7.0+cu118, torchvision 0.22.0+cu118, tqdm 4.67.1, neurodsp 2.3.0, matplotlib 3.9.4, and scipy 1.13.1.

7.2 Git

The code for this research can be found here.