Project Report : CS 7643
Detecting Chagas Disease from Electrocardiograms with Transformers

Marc Lafargue
Georgia Institute of Technology
225 North Avenue NW, Atlanta, GA
mlafargue3@gatech.edu

Matheus Rama Amorim
Georgia Institute of Technology
225 North Avenue NW, Atlanta, GA
matamorim@gatech.edu

Trevor Gratz
Georgia Institute of Technology
225 North Avenue NW, Atlanta, GA
tgratz6@gatech.edu

Abstract

Chagas disease, often undiagnosed until its chronic, cardiac-damaging stages, poses a significant global health challenge due to limitations in current diagnostic methods. Electrocardiograms (ECGs) offer a potentially accessible screening tool. This project addresses the problem of detecting Chagas disease from 12-lead ECGs using deep learning. We propose and evaluate three architectures: a baseline Convolutional Neural Network (CNN), a pure Transformer, and a hybrid CNN-Transformer, trained on balanced data from the CODE-15 and SaMi-Trop datasets. Our results show that Transformer-based models outperform the baseline CNN (AUROC 0.85 vs. 0.82). The pure Transformer achieved an AUROC of 0.8566 with data augmentation, while the hybrid model reached 0.8437. While further augmentation and pre-training on the hybrid model yielded a higher AUROC (0.8736), it significantly reduced recall to a clinically unacceptable 66.1%, highlighting a critical trade-off. We conclude that the hybrid CNN-Transformer without extensive augmentation offers the best balance for potential diagnostic use, demonstrating the promise and challenges of using deep learning for ECG-based Chagas detection.

1. Introduction/Background/Motivation

Chagas disease is a parasitic illness transmitted by insects, affecting approximately 8 million people globally, [6] with over 4,750 deaths each year. [10] The vast majority of people living with Chagas disease develop no signs or symptoms upon initial infection. [6] However, chronic infection can cause heart disease, including heart failure. [12]

Because the acute phase of Chagas disease is often asymptomatic, it goes undiagnosed in up to 95% of vector-borne cases i.e. those via insect transmission. [6] If Chagas disease is suspected, then multiple serological tests may be needed to correctly diagnose it. Furthermore, access to serological testing remains limited. [6] There is a critical need for new diagnostic approaches to detect Chagas disease, particularly for the large number of undiagnosed cases.

Chronic cases of Chagas disease can cause changes in the Electrocardiogram (ECG) readings of positive patients, specifically changes associated with a right bundle branch block. Recent work has shown that low-cost widely available diagnostic tools, such as ECGs paired with deep learning algorithms, may be utilized to diagnose Chagas disease. [7] Jidling *et al*. (2023) developed a convolutional neural network with residual connections achieving a 0.8 Area Under the Receiver Operating Characteristic curve (AUROC) in a binary classification task of Chagas disease.

The lack of empirical studies using deep learning for Chagas disease detection is surprising given the well-developed field of deep learning based methods for diagnosing other disease using ECGs. For instance, state of the art models for ECG classification have been developed by combining convolution layers with layers that handle the sequential nature of ECGs i.e. RNN or LSTM layers. [8] Other work has explored using convolution layers in combination with transformer blocks for arrhythmia classification. [4]

Our goal is to create a deep learning model that can accurately diagnose Chagas disease from ECG readings, offering a transformative tool that could save millions of lives by enabling timely diagnosis and treatment. In this paper

we will cover the three models we built to diagnose Chagas disease from ECG data. The first model is a convolutional neural network with residual connections, the second is a transformer model, and the third is a hybrid model that combines the two. We will also discuss the challenges we faced in building these models, including class imbalance and computational costs. Finally, we will present our results and discuss their implications for future research in this area.

2. Approach

We will capitalize on the existing 2025 George B. Moody PhysioNet Challenge public challenge that also aims to address this issue and provides training datasets and helper functions to process it: CODE-15 [13] and SaMi-Trop. [3] CODE-15 and SaMi-Trop are comprised of 12-lead ECG readings with associated meta-data indicating the age and gender of the patient, as well as labels for the positive or negative presence of Chagas. Both data sets were collected from Brazil between 2010 and 2016 (CODE-15) or 2010 through 2011 (SaMi-Trop), have a sampling frequency of 400 Hertz and are between 7.3 and 10.2 seconds long. In the challenge data there are over 343,424 samples in CODE-15 and 815 in SaMi-Trop. Unlike CODE-15, all SaMi-Trop records are positive and have been validated by serological tests.

The CODE-15 and SaMi-Trop datasets combined contain more than 100 gigabytes of data. In addition, only 2.1% of the data have positive chagas labels. To address the computational costs and class imbalance, we limited our training dataset to 14,000 ECG recordings made up of equal and number of positive and negative chagas labels.

Literature shows that few deep learning models have been formally studied such as Jidling *et al*. (convolutional neural network with residual connections) [7]. Furthermore, transformers are increasingly being used for time-series analysis, including ECG classification in general [1].

We have implemented three models to classify Chagas disease from ECG data:

1. A convolutional neural network with residual connections
2. A transformer encoder stack with signal processing and data augmentation
3. A convolutional neural network with residual connections signal processing combined with a transformer encoder classifier

For the first model we recreated the network used by Jidling *et al*. (2023). [7]. This network will serve as a baseline for the other two. Figure 1 showcases the overall architecture, containing four residual blocks with convolution layers that are then summed and normalized and activated using a ReLU function. We also used similar parameters from [13], which can be found in the project's repository, where this architecture was first implemented to predict heart conditions using ECG data. In that way, we kept the model consistent with previous papers, and could evaluate whether our other approaches with transformers improved the model predictive power. We did made a few small adjustments, that we believe will not impact the final conclusion of our project: we concatenated small embeddings for age group and sex with the output of the residual blocks, so the difference in performance can be assessed using the same variables; we trained for 10 epochs instead of 70, so the results are more comparable to our other models that were also trained for fewer epochs; finally, we reduced batch size from 64 to 32, for computational performance improvement.

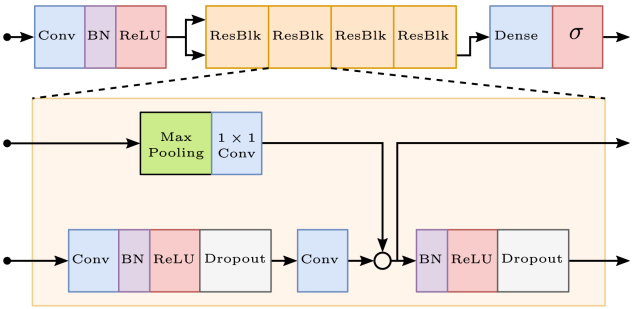

Figure 1. Convolution with Residual Connections: Architecture

The second model makes use of a BERT style transformer encoder stack, similar to that found in the paper by Choi *et al*. [5]. It was selected because it is better suited than a fully CNN based network to factor the sequential nature of an ECG signal, as well as being able to pay attention to any part of its signal whether forward or backward. Figure 2 shows the architecture of the transformer model, along with the final hyperparameter values that were selected. To remove artefacts and reduce the memory and computational burden, the inital ECG signal was processed "classically" with a 0.5 hz highpass filter followed by an 160hz resampling and an 80hz low pass filter. The model was trained on 32 epochs.

Given that transformer based models scale well with data, we also tested the impact on classification performance of including synthetically augmented training data: after the model was trained for an initial 16 epochs on unmodified data, it was trained for and additional 16 epochs (total 32) on augmented data. The original dataset was transformed by applying 3 random tranformations to each ECG sequence (e.g. random phase and amplitude shifts, time dilation, etc.). Figure 3 details an example of an augmented ECG signal. Finally, the classification was performed with a multilayer perceptron on the learned "CLS"

token, which is the first token of the sequence and is used to aggregate information from all other tokens.

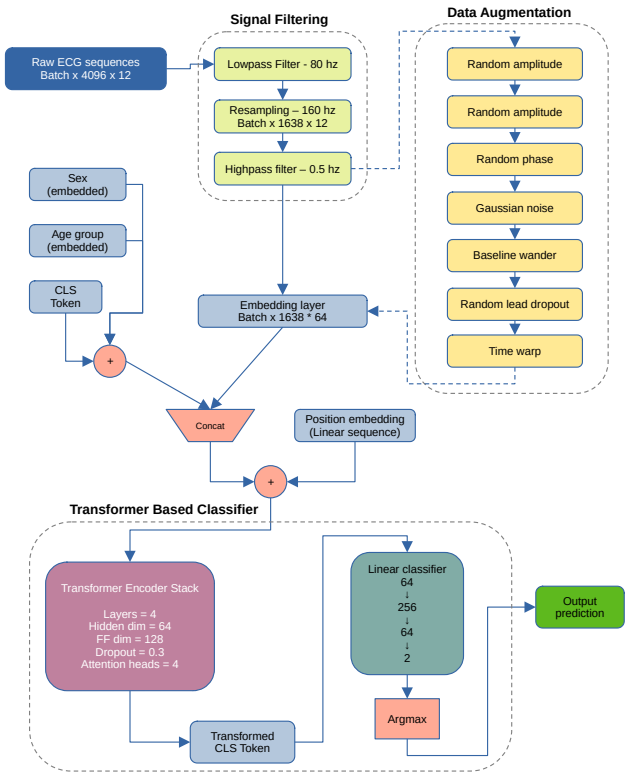

Figure 2. Transformer Model: Architecture

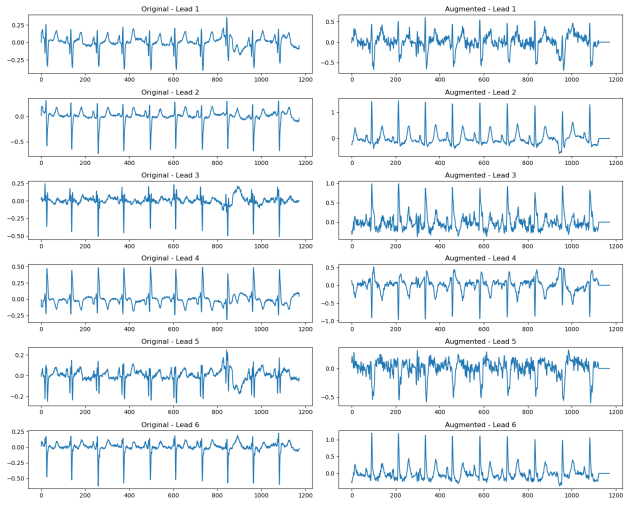

Figure 3. ECG Data Augmentation: Example of Augmentation

Lastly, we combine these two models into one, so as to incorporate the advantages of each into a single model e.g. the feature extractive capabilities of convolution and the sequential advantages of transformers. The first section of the model consists of three parallel tracks of one-dimensional convolution blocks. The three tracks vary the kernel size to capture features constructed from different widths. [14] A single convolution block is a convolution layer with 32 filter, batch normalization, a Rectified Linear Unit, drop out, and a residual connection. Within each track the convolution blocks are stacked in three sequential layers. The output of these layers is projected into a dimension compatible with our transformer, down sampled by a factor of 2 to reduce computational complexity, and then fed into the bidirectional transformer encoder. As discussed above, we extract the "CLS" token for classification. Next we created age and sex embeddings and concatenated them with our classification token. This concatenated vector was fed through a fully connected linear layer with layer normalization, a rectified linear unit layer, and dropout. This structure allows ECG data to interact in non-linear ways with demographic data on patients. This is important as there are known differences in ECG data by age and sex. [9] The intent of this architecture was to use a 1-dimensional convolutional neural network to perform feature extraction as a form a signal processing and then to use a transformer to model the sequential nature of ECG data. The full architecture can be seen in Figure 4. The loss function used was cross-entropy, data were zero-padded to sequence length of 4,096 (10.2 seconds at a sampling rate of 400 per second), and fed through a bandpass filter (0.5 and 100 for high and low pass filters respectively). Dropout was applied with a hyper parameter value of 0.1.

Given the noted challenges with data set size and the sparsity of positive labeled data, we attempted to augment the positive labeled data. Specifically, a random amount of time between 0 and 1 second from the front of the positive sequence was cropped, provided that the crop did not reduce the total length of the sequence to less than 7.3 seconds. Similarly, we cropped the positive-labeled sequences at the end of the sequence. This left three versions of a single positive sequence. We then sampled new negative labeled data to create a balanced data set. Moreover, we performed pre-training on negative labeled data from the PTB-XL dataset and performed full fine-tuning. For pre-training, we masked a continuous random half-second of the data and made the model predict the mask. For reasons discussed in the next section, we prefer the model trained on the unaugemented dataset with no pre-training.

Other than the source data and data processing helper functions from the Pysionet challenge repository (repository), everything else in the codebase was developed by our team: the preprocessing code, the transformer code, the convolution plus transformer code, and the code to evaluate the performance of these models.

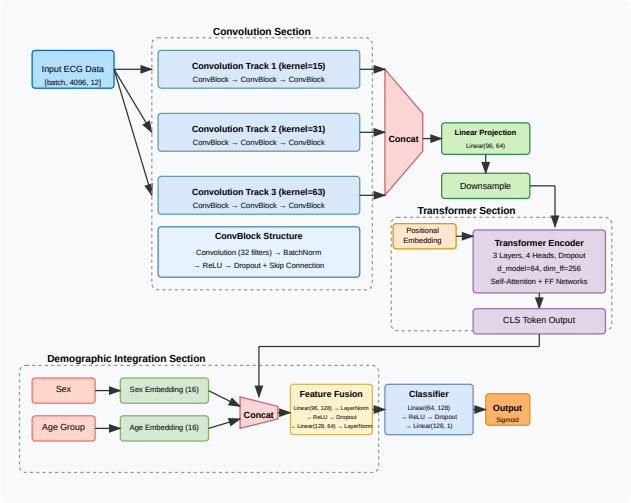

Figure 4. Convolution + Transformer Model: Architecture

Figure 5. Convolution with Residual Connections: Training and Validation Loss

3. Experiments and Results

3.1. Model 1: Convolution with Residual Connections

In our first experiment, in which we replicated the model architecture built in [7] and [13], and tested it in a validation set representing 20% of the observations, we achieved an AUROC of 0.8180, accuracy of 71.85%, precision of 67.35% and a recall of 86.40%. Table 1 below presents the confusion matrix, and Figure 5 showcases the loss curves for this model. The AUROC that we obtained is very similar to the one shown in [7] of 0.80, with ours being slightly better most likely because of the fact that we are using a balanced subset of the data, while Jidling *et al.* (2023) were using all the observations in the CODE and SaMi-Trop datasets. Nevertheless, the proximity between our results with this model and the results shown in [7] indicates that this is a proper baseline for us to evaluate our experiments.

Actual \ Predicted	Positive	Negative
Positive	1285	204
Negative	623	826

Table 1. Convolution with Residual Connections: Validation Confusion Matrix

3.2. Model 2: Transformer

In our second experiment, we trained a transformer model on a dataset with a training/validation/test ratio of 70/20/10. The transformer model was trained for 32 epochs total. The hyperparameter tuning was done by iteratively changing a parameter and running a simulation. The final hyperparameters are described in Figure 2. Below are the performance indicators of the classifier trained on non-augmented data:

- AUROC: 0.8494
- Accuracy: 0.7679
- Precision: 0.7650
- Recall: 0.7764
- F1-Score: 0.7707

Actual \ Predicted	Positive	Negative
Positive	573	165
Negative	176	555

Table 2. Transformer Model without Data Augmentation: Validation Confusion Matrix

Below are the performance indicators of the classifier trained on augmented data, where the augmented dataset was used for the last 16 epochs of training:

- AUROC: 0.8566
- Accuracy: 0.7720
- Precision: 0.7560
- Recall: 0.8062
- F1-Score: 0.7803

Actual \ Predicted	Positive	Negative
Positive	595	143
Negative	192	539

Table 3. Transformer Model with Data Augmentation: Validation Confusion Matrix

Figure 6 shows the evolution of training and validation loss over the 32 epochs of training. Though ideally multiple runs with different seeds should be performed to extract

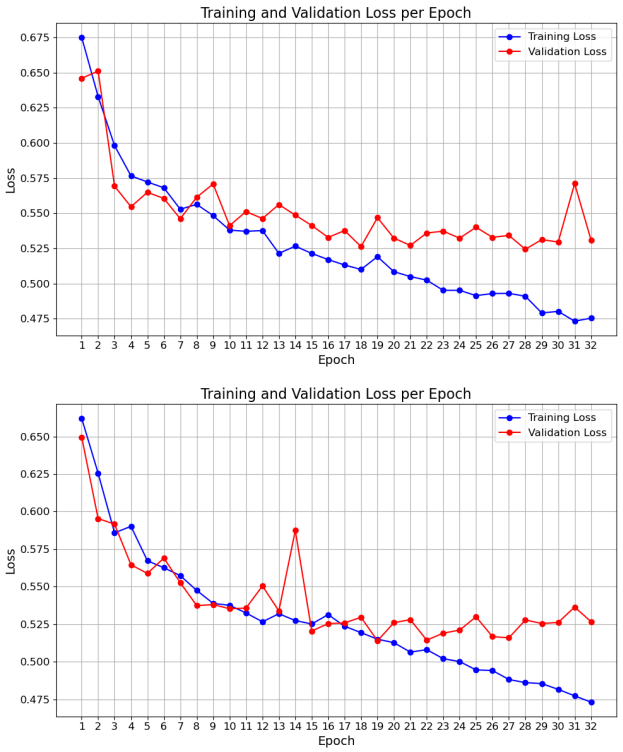

Figure 6. Transformer Model: Loss curves without (top) and with (bottom) data augmentation

meaningful statistical conclusions, we can already see that the model without augmentation has lower training loss but higher validation loss from when the augmentation is applied. This indicates that the model without augmentation is overfitting to the training data, while the model with augmentation is able to generalize better to unseen data. The performance metrics of the model with augmentation are slightly better than those of the model without augmentation, but the difference is not significant. We also not that the model with augmentation has a higher true positive rate (recall) and lower false positive rate than the model without augmentation, indicating that it is better at distinguishing between positive and negative samples. Table 2 and Table 3 show the confusion matrices for the models without and with augmentation, respectively.

This model outperformed the baseline model in most aspects, but it should be noted that the number of epochs it was trained on was higher (32 vs 10). Also, despite augmenting the dataset to 4 times the original size, we still see a strong tendency for the model to overfit. This is likely due to the fact that there is still room for hyperparameter optimization, and that the augmentation was not very aggressive, so the model was still able to memorize the training data.

3.3. Model 3: Convolution + Transformer

Prior to settling on the model described in Figure 4 for the convolution plus transformer model, a number of models designed to reduce the computational complexity of the model were created. Our first model replaced the full convolution layers with depthwise separable convolutions, each convolution track was only two deep, and we omitted the integration of the age and sex characteristics. For all models discussed below, the performance metrics were calculated on a validation set representing 20% of the instances.

The first model achieved an AUROC of 0.8334. While adding the fully-connected layers necessary to incorporate age and sex data is computationally intensive, adding these data may significantly improve performance. Our second model iterated on the first by included the demographic integration section depicted in Figure 1. With the inclusion of these variables the AUROC validation performance jumped to 0.8449.

Additional models varied this second architecture slightly. For instance, we increased the depth of the convolution tracks to three and we increased the number of convolution track to four, with the fourth convolution track using a kernel size of 127. The architecture with a wider convolution showed signs of overfitting in the later epochs. Thus, we conducted an additional experiment where we increased the dropout rate from 0.1 to 0.2 and increased the number of training epochs to 20. However, this variation failed to produce continued improvements.

We focused on variations in the convolution architecture because the existing literature achieved relatively good results utilizing convolution only. For instance, Jidling *et al*., (2023) was able to achieve a 0.8 AUROC using only convolutions. However, all variations of architecture, with the exception of the model excluding age and sex, achieved AUROCs between 0.84 and 0.85.

In the end, our preferred model presented in Figure 4, achieved a validation AUROC of 0.8437, an accuracy of 75.5%, precision of 72.6%, and a recall of 83.1%. Table 4 below presents the confusion matrix followed by Figure 7, the loss curves.

Actual \ Predicted	**Positive**	**Negative**
Positive	1240	252
Negative	469	976

Table 4. Convolution + Transformer: Validation Confusion Matrix

Our experiments with different architectures revealed that near the 6th epoch the validation loss and the validation AUROC ceased to improve with more epochs. Combined with our findings on increasing the drop out rate and training epochs, this consistency across minor variations in architectures indicated that further variations of our architecture were unlikely to yield significant gains. As such, we

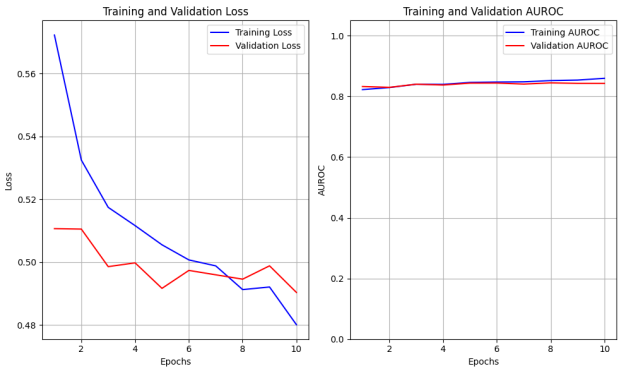

Figure 7. Convolution + Transformer Model: Training and Validation Metrics

attempted to augment our model in two ways: data augmentation and pre-training.

Data augmentation and pre-training were described in the approach section; we augmented the positive labeled classes and sampled an equal number of negative instance and pre-trained on a continuous half second mask of the PTB-XL dataset. Note the PTB-XL data was not used in training other than pre-training. During fine-tuning with the augmented data we created IDs for each unaugmented original file. We sampled these IDs and all versions of a file were placed in either the training or validation sets, but not both. This sampling procedure ensure that there was no data leakage between the training and validations sets.

With a larger dataset we expected the model to learn more and present less chance of overfitting. We increased the complexity of the model by increasing the number of hidden dimensions in the transformer block to 128 and the number of encoder layers to 4. We utilized a fourth parallel convolution track with the additional track having a kernel size of 127.

At first glance, the full-fine tuned pre-trained model trained on augmented data performed significantly better than our preferred model achieving a 0.8736 AUROC on the validation set. However, the model exhibited undesirable behavior. Specifically, the recall drops precipitously to 66.1

Actual \ Predicted	**Positive**	**Negative**
Positive	2888	1481
Negative	7	4394

Table 5. Augmented and pre-trained Convolution + Transformer: Validation Confusion Matrix

The high rate of false negatives is particularly concerning for a medical diagnostic model. Utilization of a model such as this would mean many people with Chagas disease are being told they do not have the disease. A potential remedy would be to lower the classification threshold of 0.5. However, this would likely result in many more false positives. This could induce unnecessary, costly, and invasive serological testing. For these reasons, our preferred model is that depicted in Figure 4.

The pre-trained model with augmented data was intended to address the paucity of positive labeled instances. To that end, pre-training ECG models has received notable recent attention.[2][11] Future work on diagnosing Chagas disease with deep neural networks should invest in leveraging these advancements in pre-training to address the scarcity of labeled data.

4. Discussion

Our primary measure of success was the Area Under the Receiver Operating Characteristic curve (AUROC), supplemented by accuracy, precision, recall, and F1-score analysis using confusion matrices. We conducted experiments comparing three distinct deep learning architectures: a baseline Convolutional Neural Network (CNN) with residual connections, a pure Transformer model, and a hybrid CNN-Transformer model. We also explored the impact of data augmentation and pre-training.

Quantitatively, both the Transformer (AUROC 0.8566 with augmentation) and the hybrid CNN-Transformer (AUROC 0.8437) models outperformed the baseline CNN (AUROC 0.8180), demonstrating the potential of attention mechanisms for ECG-based Chagas disease detection. The Transformer model benefited slightly from data augmentation, showing improved generalization as indicated by loss curves. However, our attempt to further improve the hybrid model using more extensive data augmentation and pre-training resulted in a model with a higher AUROC (0.8736) but a clinically unacceptable recall (66.1%). This highlighted a critical trade-off: optimizing solely for AUROC can lead to models unsuitable for diagnostic purposes due to high false negative rates.

Therefore, while we succeeded in developing models surpassing the baseline, the failure of the augmented/pre-trained model underscores the importance of evaluating models holistically, considering the specific clinical context. The high rate of false negatives in that model would be detrimental in a real-world diagnostic scenario. Our preferred model remains the hybrid CNN-Transformer without extensive augmentation or pre-training, balancing performance across multiple metrics. Future work should focus on advanced pre-training techniques specifically designed to handle the scarcity of labeled positive data in medical imaging without compromising recall.

5. Work Division

Our three-person team met on a weekly basis to discuss the progress of our project which allowed each member to contribute to all aspects of the project. However, we also divided the work into three main components, each of which was assigned to a different team member. The meetings and discussions allowed each of us to learn from each other and to share our knowledge. The table below summarizes the contributions of each team member.

Name	Contribution	Details
Marc Lafargue	Transformer model, data augmentation	Developed transformer model, implemented data augmentation, ran experiments, wrote results
Matheus Rama Amorim	Data mining, CNN+Resnet model	Prepared datasets, built CNN+Resnet baseline, ran experiments, wrote results
Trevor Gratz	Convolution + Transformer Model	Developed hybrid architecture, ran experiments, wrote results

Table 6. Contributions of team members.

6. Github Repository

The code for this project is available on GitHub at https://github.gatech.edu/mlafargue3/springCS7643project. The repository contains the code for the three models, as well as the data processing scripts and the helper functions used to process the data. The repository also contains the scripts used to run the experiments and to generate the results presented in this paper.

References

[1] Yaqoob Ansari, Omar Mourad, Khalid Qaraqe, and Erchin Serpedin. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017–2023. *Frontiers in Physiology*, Volume 14 - 2023, 2023. 2

[2] Jessica Y Bo, Hen-Wei Huang, Alvin Chan, and Giovanni Traverso. Pretraining ecg data with adversarial masking improves model generalizability for data-scarce tasks. *arXiv preprint arXiv:2211.07889*, 2022. 6

[3] Clareci Silva Cardoso, Ester Cerdeira Sabino, Claudia Di Lorenzo Oliveira, Lea Campos de Oliveira, Ariela Mota Ferreira, Edécio Cunha-Neto, Ana Luiza Bierrenbach, João Eduardo Ferreira, Desirée Sant'Ana Haikal, Arthur L Reingold, et al. Longitudinal study of patients with chronic chagas cardiomyopathy in brazil (sami-trop project): a cohort profile. *BMJ open*, 6(5):e011181, 2016. 2

[4] Chao Che, Peiliang Zhang, Min Zhu, Yue Qu, and Bo Jin. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21(1):184, 2021. 1

[5] Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G. Yhdego, Fatemeh Khadem, and Fatemeh Afghah. Ecgbert: Understanding hidden language of ecgs with self-supervised representation learning, 2023. 2

[6] Zulma M Cucunubá, Sebastián A Gutiérrez-Romero, Juan-David Ramírez, Natalia Velásquez-Ortiz, Soledad Ceccarelli, Gabriel Parra-Henao, Andrés F Henao-Martínez, Jorge Rabinovich, María-Gloria Basáñez, Pierre Nouvellet, et al. The epidemiology of chagas disease in the americas. *The Lancet Regional Health–Americas*, 37, 2024. 1

[7] Carl Jidling, Daniel Gedon, Thomas B Schön, Claudia Di Lorenzo Oliveira, Clareci Silva Cardoso, Ariela Mota Ferreira, Luana Giatti, Sandhi Maria Barreto, Ester C Sabino, Antonio LP Ribeiro, et al. Screening for chagas disease from the electrocardiogram using a deep neural network. *PLoS Neglected Tropical Diseases*, 17(7):e0011118, 2023. 1, 2, 4

[8] Xinwen Liu, Huan Wang, Zongjin Li, and Lang Qin. Deep learning in ecg diagnosis: A review. *Knowledge-Based Systems*, 227:107187, 2021. 1

[9] Peter W Macfarlane. The influence of age and sex on the electrocardiogram. *Sex-Specific Analysis of Cardiovascular Function*, pages 93–106, 2018. 3

[10] Francisco Rogerlândio Martins-Melo, Marcia C Castro, and Guilherme Loureiro Werneck. Levels and trends in chagas disease-related mortality in brazil, 2000–2019. *Acta Tropica*, 220:105948, 2021. 1

[11] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022. 6

[12] Maria Carmo P Nunes, Caryn Bern, Eva H Clark, Antonio L Teixeira, and Israel Molina. Clinical features of chagas disease progression and severity. *The Lancet Regional Health–Americas*, 37, 2024. 1

[13] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of

the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1760, 2020. 2, 4

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3