ResNets, Transformers or Both for ECG Arrythmia Detection

Adam Vert
Georgia Institute of Technology
avert3@gatech.edu

Daniel McCallum
Georgia Institute of Technology
dmccallum3@gatech.edu

Abstract

This paper investigates the comparative efficacy of Residual Networks (ResNets), Transformers, and a combined ResNet + Transformer model for electrocardiogram (ECG) arrhythmia detection, leveraging the modified MIT-BIH dataset with over 100,000 individual beats across five different classifications. Employing class-balanced focal loss to counteract the data imbalance, the performance of the combined model was evaluated against a Transformer (Encoder), a Transformer (Encoder) without positional encoding, and a ResNet. Our experimental findings indicated that the ResNet had a notably better performance than either of the transformer models and the ResNet + Transformer model outperformed the individual models, with a macro F1 score of 0.9345 on the test data. Moreover, class-balanced focal loss exhibited superior performance for the least common class. The insights garnered from this research highlight the synergistic potential of ResNet and Transformer architectures for ECG arrhythmia detection, while also underscoring the merits of class-balanced focal loss for low-sample classes.

1. Introduction/Background/Motivation

1.1. Introduction

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) account for 32% of all global deaths making them the leading cause of death globally [1]. One of the most popular early screenings for CVDs is the use of an electrocardiogram (ECG) to detect irregular heartbeats, known as arrythmias.

Traditionally, arrythmia detection is done by medical experts with specialized training. However, manual ECG detection is a time-consuming task and is still error prone. In recent times, there has been a rapid increase in the amount of ECG data being collected by wearable devices [2]. This includes popular commercial devices such as the Apple Watch and FitBit which have both introduced features to detect specific arrythmias in the last few years [3][4]. As the popularity and complexity of these devices continues to increase, the ability to create reliable and automated methods to detect arrythmias could provide significant health benefits to society at large.

Due to the importance of the problem, there has been a wide range of research that has investigated using machine learning methods to automate the detection of arrythmias [5][6][7]. Most notably, in 2017 Hannun et al. published a paper utilizing a convolutional neural network (CNN) that yielded a higher F1 score than expert cardiologists [6]. More recently, researchers have found success using the transformer architecture which is notable for its effectiveness on sequential data such as time-series data. Recent implementations of these models have been constructed by inputting the outputs of a CNN with residual connections (ResNets) alongside a positional encoding into the transformer encoder which is then passed through a series of fully connected layers leading to a final prediction [8][9][10].

In this paper, we utilize a modified MIT-BIH dataset available at [11] which has over 100,000 individual beats labelled to 5 different classifications. We use this data to investigate the performance of a combined ResNet + transformer model with its component models. More specifically, we are measuring how the combined model performs versus; 1. Transformer (Encoder) 2. Transformer (Encoder) with no positional encoding 3. ResNet.

Each of these models will be trained using class-balanced focal loss [12] to mitigate the wide class-imbalance found in the data. We also provide a comparison between class-balanced focal loss and cross-entropy loss for the combined ResNet + Transformer model. Further, each of these models will be individually tuned using the Python package Optuna [13].

This research aims to provide insights into the value of each component model by providing performance metrics for researchers to examine the comparative advantages and drawbacks of various approaches to the arrhythmia detection problem. Furthermore, we strive to shed light on the reasons behind the superior performance of certain models not only to illustrate which models outperformed others but

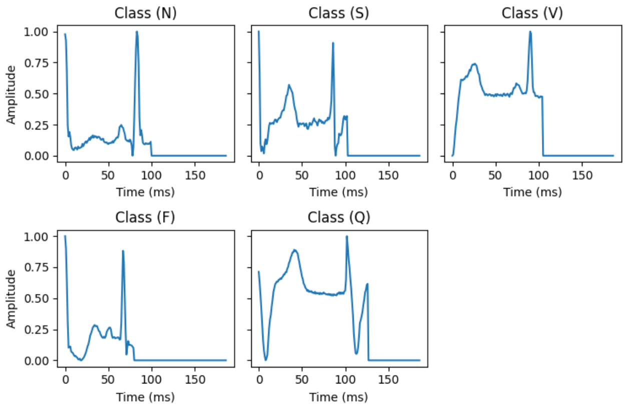

Figure 1. Example ECG sample for each class

Class	Annotations
N	• Normal • Left/Right bundle branch block • Atrial escape • Nodal escape
S	• Atrial premature • Aberrant atrial premature • Nodal premature • Supra-ventricular premature
V	• Premature ventricular contraction • Ventricular escape
F	• Fusion of ventricular and normal
Q	• Paced • Fusion of paced and normal • Unclassifiable

Table 1. Definitions of classes

also to unravel the underlying mechanism that contributed to their efficacy.

1.2. Datasets

In this paper, we use the pre-processed dataset created in [5] and publicly accessible in [11]. This dataset utilizes labelled electrocardiogram (ECG) records from PhysioNet MIT-BIH Arrhythmia. The MIT-BIH dataset contains ECG recordings from 47 different subjects at the sampling rate of 360Hz. This dataset uses ECG lead II re-sampled to 125Hz with individual beats already extracted. The final dataset includes 109,446 total beats with 21,892 of those designated for a testing set. We then split up the remaining samples using an 80/20 split resulting in training and validation sets of 70,004 and 17,510 beats respectively. Each beat is annotated by at least two cardiologists. Annotations were used to create five different categories in accordance with Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard. Typical examples for each category can be found in figure 1 and category descriptions can be found in table 1. Correctly diagnosing arrhythmia is a challenging task that is usually reserved for medical professionals with specialized training. As shown in table 1, class N and F appear to have similar patterns to the casual observer.

2. Methods/Approach

2.1. Models

We propose a model architecture that combines the best inductive priors from convolutional neural networks and transformers. Given that the problem of classifying arrhythmia in time series ECG signals is inherently a visual task for cardiologists, we utilize a series of convolutional layers that act as feature detectors for learning shapes and patterns in the ECG signal. This is implemented with a ResNet like style architecture with 1d convolutions and a series of residual blocks with skip connections. The residual blocks with skip connections make it easier for the model to learn the identity function. This helps ensure that adding additional layers improves performance and enables better gradient flow for the network. [14] The learned embedding from the residual blocks is then passed into a standard Transformer encoder as described in [15] with a linear layer at the end for classification. The Transformer encoder acts as a more generalized MLP that uses self-attention to learn a richer representation of the convolved sequence and what part of the sequence is most important for classification.

To validate our proposed architecture, we isolate its key components and train them as 4 separate models. We train (a) a standard Transformer encoder as defined in [15] with a linear layer at the end for classification, (b) the same Transformer encoder but without the positional encoding, (c) a ResNet only architecture as defined in [5] and (d) our proposed architecture as seen in figure 2 that combines the ResNet with the Transformer encoder.

2.2. Model Training

To evaluate our approach, models were implemented in PyTorch and trained on an A100 GPU using the Adam optimizer with a Noam learning rate scheduler as defined in [15]. The Noam learning rate scheduler increases the learning rate linearly for a number of warm-up steps and then decreases the learning rate proportionally to the inverse square of the step number to ensure good convergence. Due to class imbalance, class-balanced focal loss was used as is explained in section 2.5. Macro F1 Score (defined in section 3.1) was used as a performance metric and learning curves were plotted to evaluate the performance of each model. Early stopping was implemented to select the model with the best validation macro F1-Score and stop training after 30 consecutive epochs with no new best score. Loss

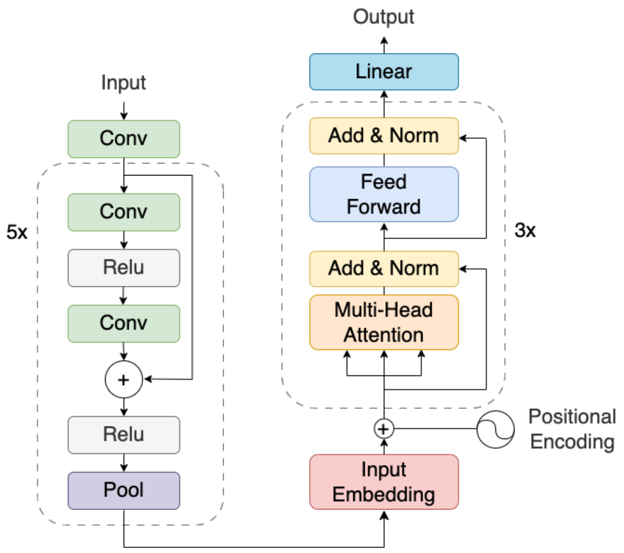

Figure 2. Combined ResNet + Transformer Encoder Architecture

and accuracy curves can be seen in the /plots/ section in the GitHub repository provided.

3. Class Imbalance & Focal Loss

As discussed in Section 2.1, there are 5 unique classifications within the MIT-BIH dataset. However, within the data there is a wide imbalance between the number of samples representing each class.

	N	S	V	F	Q
% of Samples (N samples)	82.8% (90589)	2.5% (2779)	6.6% (7236)	0.7% (803)	7.3% (8039)

Table 2. Percentage and number of samples for each class for the entire dataset

As can be seen in Table 2, the dataset is dominated by the “N” class which represents a normal beat and is the label for 82.8% of all samples. This is in stark contrast to classes “S” and “F” which represent 2.5% and 0.7% of the dataset respectively. In order to quantify how this class imbalance would affect the final model, we trained the final model using cross entropy as well as focal loss.

Focal loss is first proposed in [16] and later the class-balanced version described in [12]. Class balance focal loss was created as a method that aims to mitigate class imbalance by weighting the cross-entropy loss based on the frequency of the class labels. Further, focal loss also helps adjust learning rates based on how “easy” they were to predict using a modulating factor. For instance, in our example the loss function for a label of “N” would be weighted less than one of sample “F”. In practice, this allows the models update steps to not be overwhelmed by prevalent examples and still allows for tangible contributions from less common classes.

Class-balanced focal loss is defined as:

$$\text{CBfocal}(z, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_{ti})^{\gamma} \log(p_{ti}) \quad (1)$$

where β is a hyperparameter that represents how much to reweight based on the class imbalance with $\beta = 0$ meaning no change, γ is a hyperparameter that exponentially scales the modulating factor when $\gamma = 0$ focal loss is equivalent to cross entropy, p_t is the models outputted probability for the true class, and n_y is the number of samples for label y. For our implementation we used $\beta = 0.9999$ and $\gamma = 1$.

3.1. Hyperparameter Tuning

To optimize our hyper-parameters, we used the Tree-structured Parzen Estimator algorithm described in [17] through the Python Package Optuna [13]. Each of the 4 models’ hyper-parameters described in Section 2.3 was optimized separately by training for 10 epochs for 50 trials. Table 3 summarizes the selected hyper-parameters for each model.

3.2. Code Availability

The code and models used in this paper are available at https://github.com/VertAdam/Arrhythmia-Detection-Transformer.

4. Experiments and Discussion

4.1. Experimental Setup

In order to quantify the differences between component models and the ResNet + Transformer model, as well as the efficacy of using class-balanced focal loss, we decided to use the F1-score as our primary metric. The F1-score is the harmonic mean of precision and recall. For an individual class, the F1-Score is defined as

$$F1_c = 2 \cdot \frac{TP_c}{TP_c + FP_c + FN_c} \quad (2)$$

Where TP, FP, and FN represent true positive, false positive and false negative respectively. However, in order to create more summarized values, the values that will be reported in the results will be the macro F1 score, which takes the mean F1 score over each class as well as the weighted average, which gives the total F1 score over all samples. The macro F1-score can be interpreted as better representing the performance of all classes, even the rare ones. The weighted F1-score can be interpreted as a metric that gives the performance treating all samples individually.

$$F1_{\text{Macro}} = \frac{1}{N} \sum_{c} F1_c \quad (3)$$

	ResNet + Transformer	ResNet	Transformer w/ PE	Transformer w/o PE
In channels	480	512	-	-
Out channels	480	512	-	-
Kernel size	7	7	-	-
Stride	1	1	-	-
Padding	same	same	-	-
Pool kernel size	5	5	-	-
Pool stride	2	2	-	-
Num res blocks	5	5	-	-
PE dropout rate	0	-	0	-
Number of heads	8	-	40	10
Feed forward dimensions	4096	-	2048	2048
Transformer dropout rate	0.1	-	0	0.1
Number of Transformer layers	3	-	6	9

Table 3. The optimal hyper-parameters selected for each model after hyper-parameter tuning

$$F1_{\text{Weighted}} = F1_{\text{all}} \tag{4}$$

Along with reporting the F1 score of each model, we will also show the confusion matrices with focal loss and with cross-entropy loss. All models were trained, validated, and tested on the same data.

4.2. Results

The results of the F1-Score performance can be seen in Table 4.

The confusion matrix for cross-entropy loss and class-balanced focal loss can be seen in figure 3.

4.3. Discussion

4.3.1 Positional Encoding

Our experiments surfaced some intriguing insights, shedding light on the performance of various machine learning models. The Transformer encoder, with and without positional encoding, exhibited comparable results, yielding a macro F1 score of 0.9020 and 0.9024 respectively on the test data. Surprisingly, the Transformer encoder without a positional encoding seemed to gain no advantage from the addition of a positional encoding.

Our hypothesis suggest that this may be, in part, due to the fact that this dataset consists of only individual beats, the sequential nature of the data becomes less important. The value the transformer in this single-beat scenario might not stem from its ability to understand the sequence order, but from its ability to contextualize absolute beat values. It is possible that for a dataset with continuous data, such as the unaltered MIT-BIH dataset, positional encoding may be more valuable, however more research would be needed to confirm this.

These findings do provoke an interesting question though. When cardiologists diagnose arrythmias, It is standard for them to do so largely through sequential means. They even have standardized names for each segment of the waveform. It appears, however, that our results indicate that transformer models could identify these arrythmias using a different logic than that employed by cardiologists. Even in the absence of positional encoding, self-attention still managed to learn a representation of the data with reasonable accuracy.

4.3.2 ResNet vs Transformer

In our experiments, an interesting finding is the superior performance of the ResNet model when compared to both of the Transformer encoder models. The ResNet yielded a macro F1 score of 0.9237 on the test set making it over 2% better than either of the transformer models.

The ResNets better performance likely is a result of the visual nature of the original MIT-BIH dataset, where labels were assigned through visual inspection by experts. ResNets, which utilize convolutional neural networks, excel in the area of visual tasks as their underlying architecture enables them to analyze complex localized spatial features.

With that being said, recently a subset of Transformers known as Vision Transformers have been shown to be outperforming ResNets even on visual tasks when there is a sufficiently large amount of data available [18] [19]. If this task is as much of a visual task as this experiment implies, it may be a promising avenue of research to apply vision transformers to this same problem.

4.3.3 ResNet + Transformer

Combining the ResNet and the Transformer architectures by feeding the output of the ResNet as an embedding to the Transformer achieved the best performance with a macro F1 score of 0.9345 on the test data. Based on our other results, we believe the ResNet backbone of this architecture is the key component for learning a rich set of visual features

Model	Loss	Validation Macro F1	Validation Weighted F1	Test Macro F1	Test Weighted F1
ResNet + Transformer	Class-balanced Focal Loss	0.9533	0.9912	**0.9345**	0.9888
ResNet + Transformer	Cross-Entropy	**0.9561**	**0.9926**	0.9312	**0.9892**
ResNet	Class-balanced Focal Loss	0.9490	0.9904	0.9237	0.9877
Transformer w/ Positional Encoding	Class-balanced Focal Loss	0.9182	0.9822	0.9020	0.9807
Transformer w/o Positional Encoding	Class-balanced Focal Loss	0.9148	0.9828	0.9024	0.9805

Table 4. Macro and Weighted F1 performance for all model types

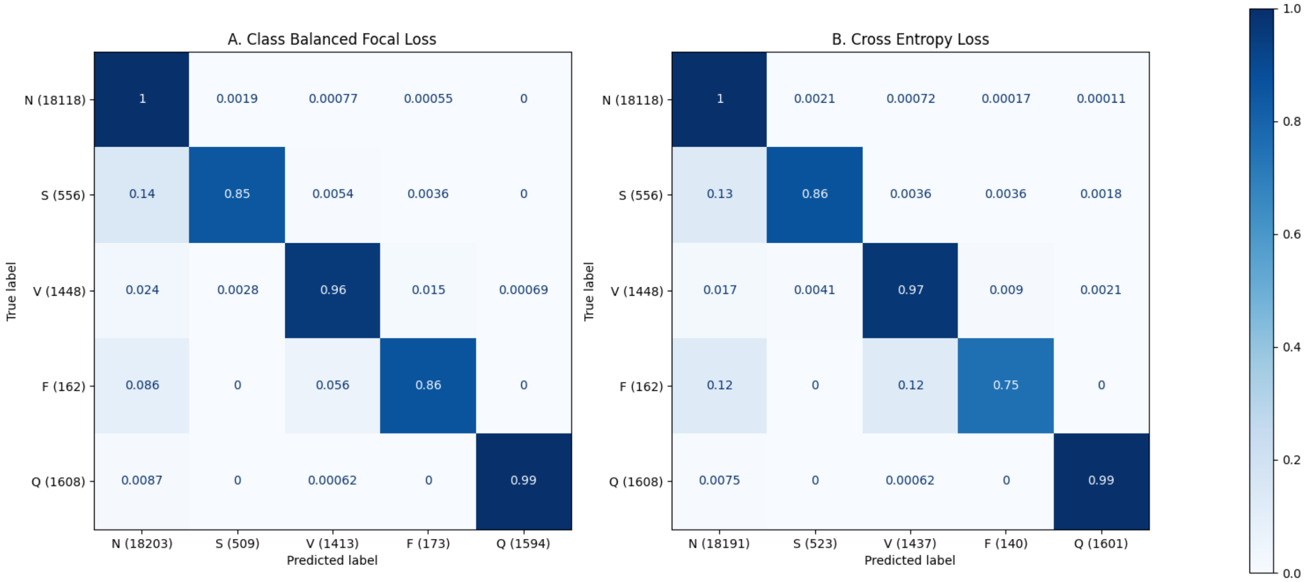

Figure 3. Confusion matrix for the Resnet + Transformer normalized over true values (i.e., recall) for A. class-balanced focal loss and B. Cross-Entropy loss

from the data and it seems that the addition of the Transformer encoder helps push up the F1 score another percentage point. In this model the ResNet provides the ability to learn localized spatial features as discussed in section 3.3.2, these features are then fed into a transformer that utilizes self-attention to learn a richer representation of the convolved sequence and the most important features for classifying each arrhythmia.

4.3.4 Cross Entropy Loss vs Class-balanced Focal Loss

When looking at the F1 performance metrics for the ResNet + Transformer with cross entropy loss, vs the same model with class-balanced focal loss, there is minimal difference, and it would be understandable to assume that the two loss functions are making little difference. However, in figure 3 we display the confusion matrices for the two loss functions.

The most notable difference is looking at the performance of class F. Class F is by far the least common class with only 162 values in the test set. The cross-entropy loss performed significantly worse on this class resulting in a recall of only 0.75, whereas the class-balanced focal loss yielded a recall of 0.86. This is what is expected as the class-balanced focal loss puts a higher emphasis on class labels with lower samples, such as F. Even with the implementation of class-balanced focal loss, we do not see a notable decrease in the weighted F1 score with the two scores only having a difference of 0.0002 on the test set. This shows that despite the fact that the class-balanced focal loss is performing better on the low-sample classes, it is still performing as well as the cross-entropy loss on the more popular classes.

5. Additional Thoughts

5.1. Challenges and Unreleased Analysis

Over the course of this project we ran into a number of challenges. The one that is addressed directly in this paper, was class imbalance. Initially, we started training our models using cross-entropy but after noticing the discrepancy in the class sizes, we decided to add in an additional analysis on the usefulness of class-balanced focal loss. Another challenge that we ran into was computational limitations. Specifically, when tuning the hyperparameters we had to limit the number of epochs to 10 which wasn't ideal.

Along with these challenges, we also had some code and analysis that was never seen in this paper. At first, we were attempting to recreate and improve on the Wide and Deep model presented in [8]. This model was full coded and ready for implementation however this model was designed for continuous data and we came to the decision that we would be looking at individual beats. We also downloaded and extracted individual beats from the Icentia11k dataset [20]. This is an extremely large dataset that comes with continuous data, we created a methodology to extract individual beats from this data and implemented it extracting 100,000 individual beats. After this however, we determined that due to the page limit of this report, it would be infeasible for us to adaquetly add an additional dataset to analyze.

5.2. Project Success

Our goal for this project was to provide insight into the different methodologies that have been utilized in state of the art arrythmia detection algorithms. In this, we believe we succeeded. Most notably, we think that our results showing the difference in performance between ResNets and Transformers shows that this problem may be one of a much more spatial/visual nature than originally thought. Further, we think that our work provides interesting questions as to why the value of the positional encoding seems so marginal. All in all, our work not only illuminated key aspects of state-of-the-art arrhythmia detection methodologies but also, as is often the case in research, paved the way for a multitude of intriguing questions that await further exploration.

5.3. Limitations and Future Work

Our experiments highlighted the strong inductive priors of ResNet for learning visual representations, particularly in scenarios with limited training data such as in this study with approximately 87,000 samples. In such contexts, we anticipate Convolutional Neural Network (CNN) based approaches to continue demonstrating robust performance.

However, it's important to consider that as the volume of training data increases, the necessity for the inductive biases of CNNs, as embedded in the model architecture, may diminish. In these instances, transformers, renowned for their capability to capture complex dependencies in data, might offer a distinct advantage in environments with abundant data, as they could potentially discern intricate patterns without relying on a specific feature-extraction backbone such as ResNet.

Supporting this premise, recent studies have shown that a subset of Transformers known as vision transformers have been surpassing CNNs on image classification benchmarks when trained with large datasets[18][19]. Consequently, a potential avenue for future research in this field could be to replicate our experiments using a Vision Transformer and a larger datasets, such as the Icentia11k dataset[20], which encompasses ECG signals from 11,000 patients and 2 billion labelled beats. Undertaking such experiments may help evaluate the effectiveness of Transformers for time series classification tasks when there is ample training data available.

Moreover, we should emphasize that our results, specifically concerning positional encoding, may not be transferrable to datasets containing multiple beats. Future research should investigate if the observed lack of dependency on positional encoding persists in scenarios involving multiple beats.

6. Conclusion

In this paper, we studied the performances of various machine learning models, including ResNet, Transformer, and their combination, on arrhythmia detection. Our findings suggest that positional encoding provides no significant advantage for Transformer models on individual beat datasets. The success of the ResNet model confirms the importance of visual pattern recognition, demonstrating its superiority in a task where labels are assigned visually.

A hybrid ResNet + Transformer model yielded the best results, combining ResNet's ability to learn complex visual features and Transformer's capacity for creating a richer representation through self-attention.

Lastly, while the cross-entropy loss and class-balanced focal loss had similar overall performance, the latter demonstrated better handling of underrepresented classes. This result encourages the exploration of class-balanced focal loss as a method to manage class imbalance in medical data.

Overall, we believe that we were successful in delivering our goal of providing insight on the performance of various deep learning-based approaches for arrhythmia detection as well as giving a more detailed reasoning behind the comparative performance for the different models and parameters tested.

7. Work Division

A table of the division of work can be seen in Table 5

References

[1] "Cardiovascular diseases." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) 1

[2] C. C. Cheung, A. D. Krahn, and J. G. Andrade, "The emerging role of wearable technologies in detection of arrhythmia," *Canadian Journal of Cardiology*, vol. 34, no. 8, pp. 1083–1087, 2018. 1

[3] "Ecg app and irregular heart rhythm notification available today on apple watch." [Online]. Available: https://nr.apple.com/dE3O8p5G2K 1

[4] "Irregular heart rhythm notifications." [Online]. Available: https://blog.google/products/fitbit/irregular-heart-rhythm-notifications/ 1

[5] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "Ecg heartbeat classification: A deep transferable representation," in *2018 IEEE international conference on healthcare informatics (ICHI)*. IEEE, 2018, pp. 443–444. 1, 2

[6] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, 2019. 1

[7] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015. 1

[8] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, and J. Rubin, "A wide and deep transformer neural network for 12-lead ecg classification," in *2020 Computing in Cardiology*. IEEE, 2020, pp. 1–4. 1, 6, 8

[9] R. Hu, J. Chen, and L. Zhou, "A transformer-based deep neural network for arrhythmia detection using continuous ecg signals," *Computers in Biology and Medicine*, vol. 144, p. 105325, 2022. 1

[10] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–13, 2021. 1

[11] "Kaggle heartbeat dataset." [Online]. Available: https://www.kaggle.com/datasets/shayanfazeli/heartbeat 1, 2

[12] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277. 1, 3

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631. 1, 3

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988. 3

[17] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011. 3

[18] "Papers with code - imagenet benchmark (image classification)." [Online]. Available: paperswithcode.com/sota/image-classification-on-imagenet 4, 6

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 4, 6

[20] S. Tan, G. Androz, A. Chamseddine, P. Fecteau, A. Courville, Y. Bengio, and J. P. Cohen, "Icentia11k: An unsupervised representation learning dataset for arrhythmia subtype discovery," *arXiv preprint arXiv:1910.09570*, 2019. 6

Student Name	Contributed Aspects	Details
Daniel McCallum	Implementation, Experimentation, Analysis	• Implemented and parameterized ResNet, Transformer, Transformer w/o positional encoding, and ResNet + Transformer models. • Implemented training/data pipeline. • Implemented learning curves (f1 score and loss) for training. • Trained models using early stopping with f1 score. • Created figures and tables for the report. • Contributed to various sections of the report and experiment analysis.
Adam Vert	Implementation, Experimentation, Analysis	• Implemented WideAndDeep which seeded our CNN + Transformer approach (Not used in final report) [8] • Implemented hyperparameter tuning pipeline and ran hyperparameter search. • Implemented Noam learning rate scheduler. • Implemented focal Loss. • Implemented and trained models using early stopping. • Created figures and tables for the report. • Contributed to various sections of the report and experiment analysis. • Prepped additional Icentia11k dataset including extracting individual beats (not used in report)

Table 5. Contributions of team members.