Automated Classification of Cardiac Abnormalities with a Single ECG Lead: Final Report

Keith Willard, Claire Chen, Brandon Herrera, and Michelle Galarneau

Abstract

We developed and tested a series of deep neural networks to classify cardiac rhythms using electrocardiogram (ECG) recordings. Our goals were to 1) determine the performance degradation associated with reducing model inputs from 12 ECG leads to 1 ECG lead and 2) identify the 1-lead model with superior performance. The first 1-lead model was based on a standard resnet50 model. The second 1-lead model was adapted from a top-performing neural network from the Physionet Challenge 2020, a deep learning competition that challenged participants to classify cardiac rhythms using the standard 12 ECG leads. The latter 1-lead model had the best classification performance, with an F1 score of 0.60 and area under the Precision-Recall curve (AUC-PR) of 0.44. Compared to the 12-lead version of that model, this represents an 11% reduction in F1 score and a 17% reduction in AUC-PR.

1 Background

Electrocardiography is a tool commonly used in clinical practice for diagnosing cardiac abnormalities. The current gold standard cardiac diagnostic method, 12-lead electrocardiography involves placing electrodes externally on the patient's limbs and chest to form 12 vectors, or leads, from which to visualize the electrical activity of the heart. The standard leads are six limb leads (I, II, III, aVR, aVL, aVF) and six precordial (chest) leads (V1, V2, V3, V4, V5, V6). The standard electrocardiogram (ECG) thus consists of 12 printed cardiac electrical recordings that, combined, reveal rich information about the heart's rate, rhythm, and conduction patterns to aid in differential diagnosis.

While these standard 12 leads are still widely used in clinical settings today, placing the electrodes and setting up the rest of the ECG system can be tedious. In addition, reviewing a high volume of recordings and making diagnoses contributes to a significant burden on the part of clinicians. Moreover, this method is intended for an acute look at the patient's cardiac electrical activity; the ECG recordings are typically only ten seconds in duration. While some abnormalities may be detected within a matter of seconds, other arrhythmias may be paroxysmal in nature, requiring longer term, ambulatory monitoring for detection. Historically, a Holter Monitor with two or three leads has been used to provide up to 48 hours of ambulatory monitoring. However, the Holter system is considered bulky and not user-friendly. As a result, smaller and more self-contained devices have been developed for cardiac monitoring via a single ECG lead. Some of these devices adhere externally to the patient's chest and can last for several weeks (e.g., Zio Patch), while others can be implanted for up to five years of monitoring (e.g., Medtronic's LINQ device, Medtronic (2022)).

These and similar devices often have automatic arrhythmia detection features. While algorithm enhancements have been made in recent years, most of the algorithms are relatively basic in the types of arrhythmias that are detected and the modes by which they are detected. For example, while LINQ's new atrial fibrillation (AF) detection feature utilizes a cloud-based neural network for classification of the ECG episode, LINQ's brady and tachy detection features are based solely on rate and duration. The device is not able to subclassify different types of conduction blocks, such as first-degree atrioventricular (AV) block or left or right bundle branch blocks. It would be clinically useful to detect these and other types of arrhythmias with high sensitivity and specificity.

2 Problem Statement and Clinical Significance

The aim of the present work is to develop a deep neural network to classify a large database of 10-second ECG recordings using a single ECG vector. The long-term goal of developing this model would be to use it to automatically classify rhythms in an ambulatory single-lead ECG device. We are also interested in the classification performance degradation associated with reducing model inputs from 12 ECG leads down to 1 ECG lead.

The neural network will classify the ECG recordings into the following six classes: Bradycardia, First-degree AV Block, Bundle Branch Block, Atrial Fibrillation/Flutter, Normal Sinus Rhythm, and Other. The clinical significance for detecting each rhythm type is as follows:

- **Bradycardia**: Depending on the rate and frequency, a slow heart rate, i.e., bradycardia, can indicate the need for a pacemaker. Note: Sinus bradycardia, AV block, and sinus node dysfunction are subclasses of brady, but unfortunately not all of these labels exist in the database of interest.
- **First-degree AV Block**: AV block is typically regarded as a progressive disease of the conduction system. While first-degree block does not involve an abnormally slow ventricular rate, it may be an early indication that the patient may go on to develop higher-grade AV block, which warrants permanent pacemaker implantation. Thus, it may be useful to proactively monitor the frequency of first-degree AV block over time. Correlation with patient symptoms may also be useful in generating an indication for a pacemaker.
- **Bundle Branch Block**: Right or left bundle branch block occurs when the depolarizations and subsequent repolarizations of the right and left ventricles do not occur simultaneously and so result in elongated QRS complexes. Over time, this lack of synchrony can have lead to pathophysiologic changes in the heart's structure and function (i.e., heart failure). Depending on the severity, bundle branch block can be treated using cardiac resynchronization therapy (CRT) or, more recently, conduction system pacing (CSP). Ambulatory detection of this conduction pattern could prove useful in enabling early treatment to prevent adverse cardiac remodeling.
- **Atrial Fibrillation/Flutter**: While atrial fibrillation and atrial flutter alone are not life threatening arrhythmias, they put the patient at risk of stroke, adverse cardiac remodeling, heart failure, and more severe cardiac arrhythmias. Depending on the Afib/Aflutter burden, treatment via cardiac ablation may be warranted.
- **Normal Sinus Rhythm**: Because the rhythm types above are not inclusive of all cardiac abnormalities, it is useful to know when a recording is normal vs. when it is abnormal and falls into one of the above classes vs. when it is abnormal and does not fall into one of the above classes.
- **Other**: This class is used to indicate recordings that have one or more cardiac abnormalities but do not fall into any of the above classes.

The ECG data used in this analysis was aggregated from the Physionet/Computation in Cardiology Challenge 2020, in which the aim was to develop an automated interpretation algorithm for identifying 27 clinical diagnoses from 12-lead ECG recordings. This database consists of 43027 clinical 12-lead ECG records of approximately 10-second duration (5000 samples per record at 500 Hz sampling) from four different sources:

- Southeast University, China, including the data from the China Physiological Signal Challenge 2018 (two datasets from this source)
- St. Petersburg Institute of Cardiological Technics, St. Petersburg, Russia
- The Physikalisch Technische Bundesanstalt, Brunswick, Germany (two datasets from this source)
- Georgia 12-Lead ECG Challenge Database, Emory University, Atlanta, Georgia

All diagnoses are encoded with SNOMED-CT codes. The frequency distribution of diagnoses in the database can be found in Figure 1. Note that nearly half of all datasets are considered to be normal sinus rhythm, i.e., healthy. Regarding demographic make-up of the data, fifty-two percent of

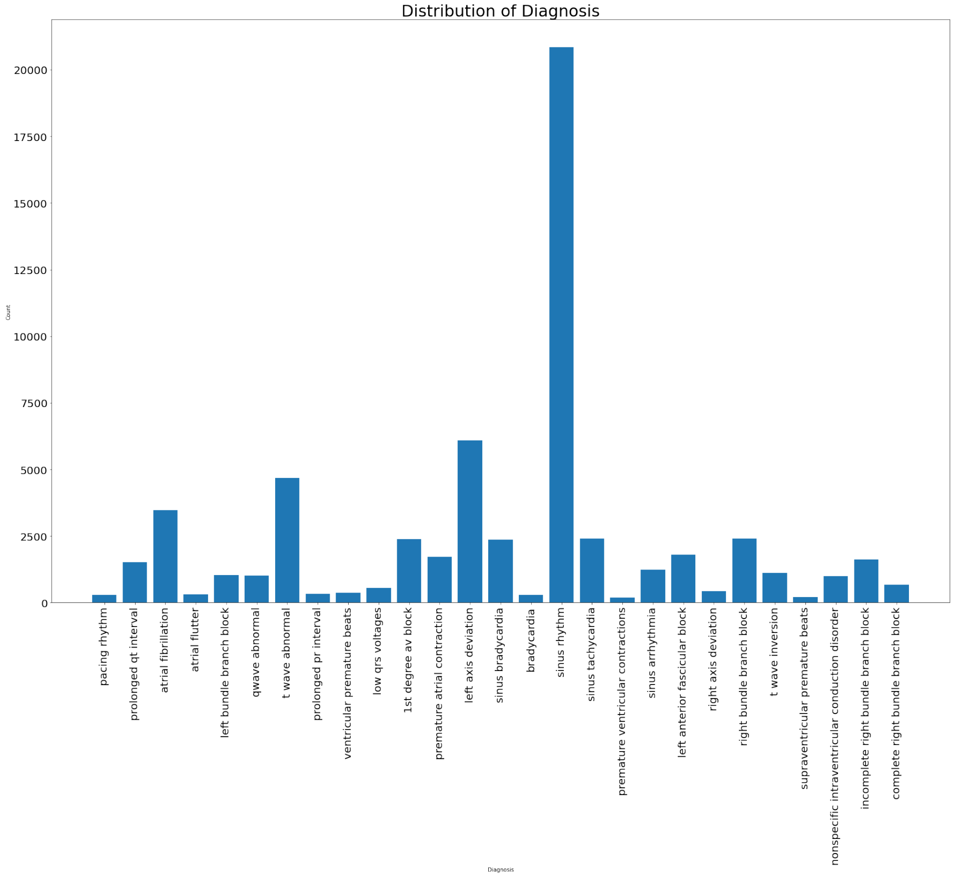

Figure 1: Distribution of Diagnoses in Physionet/Computation in Cardiology Challenge 2020 ECG Database

the recordings are from male patients while forty-eight percent are from female patients, and patient ages range from 0 to 95 years.

With reducing the class size from 27 down to 6 and additional data pruning, we ended up using 40719 of the total 43027 datasets in our model development and testing. We used 80% of these for training (30575 datasets) and 20% for testing (10144 datasets). When splitting the data between training and testing, the class distribution was preserved.

3 Previous Work

So far, applications of deep learning to the automated classification of cardiac rhythms on recorded electrocardiograms have yielded significant progress (Karthik et al. (2022), Rawi et al. (2022)). As mentioned above, conferences have been held that focused on using deep neural networks to classify many rhythm types using the standard 12-lead ECG datasets (Perez Alday EA;Robichaux C;Ian Wong AK;Liu C;Liu F;Bahrami Rad A;Elola A;Seyedi S;Li Q;Sharma A;Clifford GD;Reyna MA; et al. (2020), Perez Alday et al. (2022)), with more recent work tackling the more technically challenging 2-lead ECG datasets (moo (2022)). In addition, Medtronic is one medical device company that developed and implemented a deep neural network to improve the detection of atrial fibrillation and atrial flutter (AF) using a single-lead ECG in the LINQ insertable cardiac monitor.

Despite this progress, there has been relatively little published work on multi-rhythm classification using 1-lead ECG data due to inherent data limitations. However, for a future in which edge computing on consumer devices is becoming increasingly feasible, 1-lead ECG analysis of multiple rhythm types represents an attractive technical challenge.

In this analysis, we investigated whether one of the standard 12 ECG leads yields acceptable performance for detecting our diagnostic classes of interest. Results from this analysis may provide insights into the design of an ECG-based cardiac monitoring device for detecting an individual or subset of rhythm abnormalities. Performance will be assessed by evaluating the area under the curve of the receiver operating characteristic curve (AUC-ROC), area under the curve of the Precision-Recall curve (AUC-PR), F1 score, and accuracy. F1 score will be especially useful for this analysis, given the class imbalance in the data sample.

4 Methods and Results

The general approach for classifying cardiac abnormalities using single-lead ECG data was to use a convolutional neural network. Convolutional neural networks are well known for analyzing image data, as they take into account the spatial structure of the data. Because ECG waveforms are time-series data and have strong 1-dimensional locality, this information can also be extracted by convolutions.

Our first attempt was to take a pretrained resnet50 model and adapt it to our classification problem and see if we could create a model comparable to more specialized contest models from the Physionet Challenge 2020 (Perez Alday EA;Robichaux C;Ian Wong AK;Liu C;Liu F;Bahrami Rad A;Elola A;Seyedi S;Li Q;Sharma A;Clifford GD;Reyna MA; et al. (2020), Perez Alday et al. (2022)) using 12-lead, and then reduce the data source down to only 1-lead.

We replaced the top fully connected layer with our own two-layer classification tier: one layer to match the output tensor shapes, a batch normalization layer, and the final reducing layer to the match the number of six class groupings we desired to predict driven by a cross entropy loss function since our grouping approach eliminated the multi-label aspect of this multi-class problem. Our training consisted of using an 80/20 or 80/10/10 stratified (by class) split, depending on whether we were using a validation and test, or just a test split.

We attempted several preliminary trials at variations on this model, including allowing the resnet50 model parameters to train or be frozen, as well as removing the top convolutional tier. We trained these attempts for at least 100 epochs and the loss function change was almost zero. Other than the usual resnet preprocessing, no ecg specific preprocessing was performed. No approach improved over simply replacing the top classification tier and freezing the resnet50 model parameters.

The final training run of the 12-lead resnet50 model consisted of 100 epochs and a batch size of 500. After training the input and classification layers of the model, performance on the test dataset was determined: F1 score was 0.34, AUC-ROC was 0.66, AUC-PR was 0.25, and accuracy was 0.40. The confusion matrix for the six diagnostic classes of interest is shown in Figure 2. Note that it has been normalized to account for the class imbalance.

Even though these results were a disappointment we completed our analysis of this approach by training a 1-lead version of this model. The final training run of the 1-lead resnet50 model consisted of 100 epochs and a batch size of 100. After retraining the input and classification tiers of the model, performance on the test dataset was determined: F1 score was 0.20, AUC-ROC was 0.55, AUC-PR was 0.18, and accuracy was 0.28. The normalized confusion matrix for the six diagnostic classes of interest is shown in Figure 3.

Given the relatively poor ability of both our 12-lead and 1-lead resnet50 based models discriminate between cardiac abnormality classes, we explored modifying one of the top-performing models from the Physionet challenge in an effort to improve classification performance. After several failed attempts at reproducing reported model performance of several other groups we were able to reproduce in our hands, the contest performance of a model developed by LaussenLabs that was the third-overall best contest performer (and was only narrowly behind the top two scores) (Goodfellow et al. (2020)) from their publicly available github repo. `https://github.com/Seb-Good/physionet-challenge-2020`. Figure 4 shows the overall structure of the LaussenLabs deep neural network.

As seen in Figure 4, the LaussenLabs model resembles a standard resnet model but has several interesting features. First, there is fairly extensive feature extraction to infer the P, R, and T waveforms, which correspond to atrial depolarization, ventricular depolarization, and ventricular repolarization,

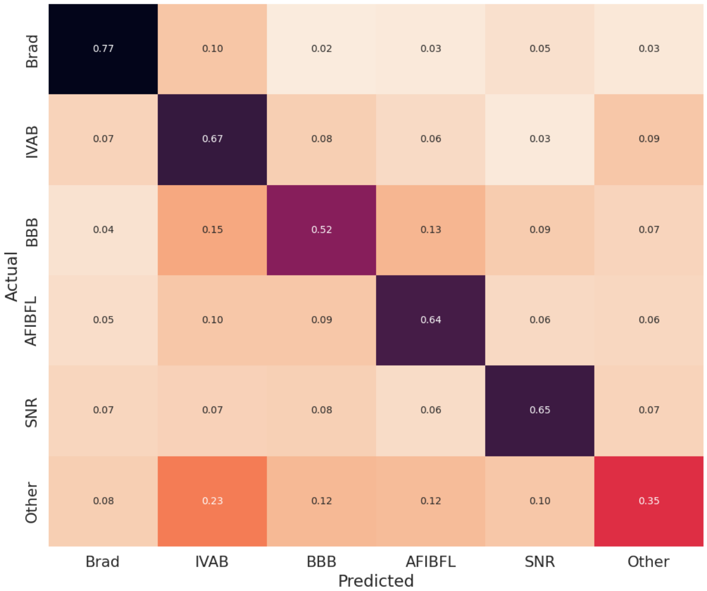

Figure 2: Normalized confusion matrix for 12-lead resnet50 model. Brad = bradycardia, IVAB = first-degree atrioventricular block, BBB = bundle branch block, AFIBFL = atrial fibrillation/flutter, SNR = normal sinus rhythm, Other = any rhythms that do not fall into the preceding categories.

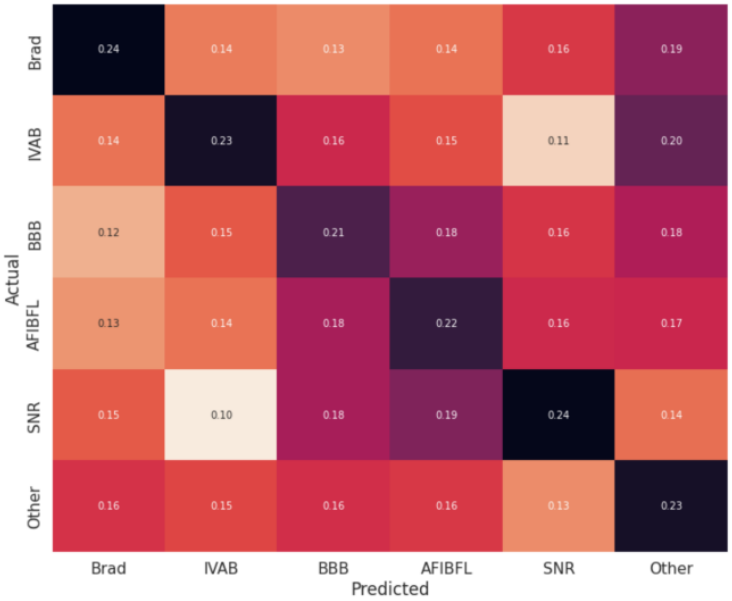

Figure 3: Normalized confusion matrix for single-lead resnet50 model

respectively. Analysis of these waveforms provides critical insight into the cardiac rhythm, hence why these three extracted waveforms are added as additional channels of data in the model. Each input waveform is upsampled from 5000 points to 19000 points.

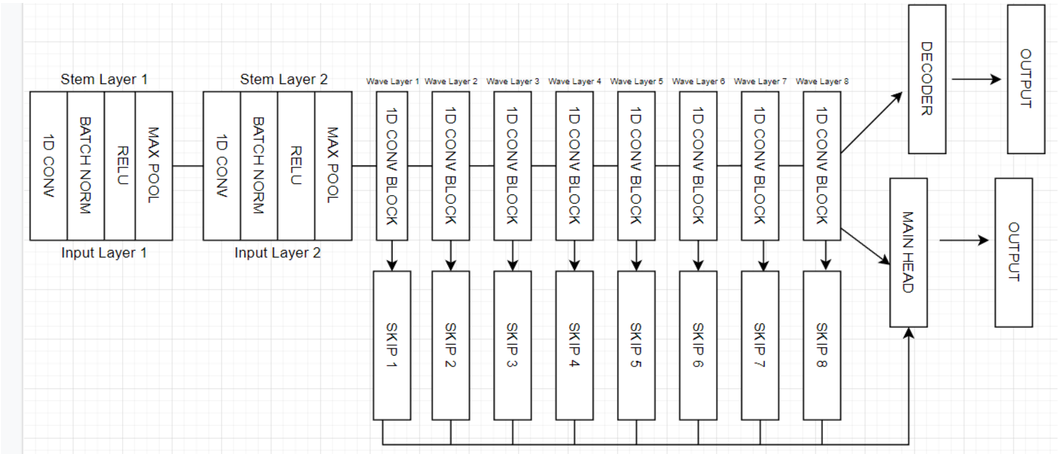

Figure 4: Structure of LaussenLabs model for classifying 12-lead ECG data from the Physionet Challenge 2020.

The LaussenLabs convolutional model itself is a multilayer convolutional model with skip connections but is one dimensional rather than two. The model has two heads, one is the the normal classification head used to predict the class labels driving a binary cross entropy loss function. The other is an encoder head driving a mean squared loss function (output layer with input layer). The overall training loss used is the sum of loss function from each head. In addition, since the original contest required predicting 27 class labels (multilabel allowing for more than one label per waveform group) as well, their prediction step was not as simple as determining the probability exceeding 0.5 threshold. Instead, they constructed a prediction loss function and used it to learn a more optimal set of thresholds during training.

The LaussenLabs group also adopted a more sophisticated training approach by splitting their training data into six cross folds (keeping train/test class stratification). Each fold was used independently to train distinct model instance, resulting in six trained models. At classification time the average of the ensemble of their predictions was used.

We first applied their 12 lead model by grouping their results from 27 classes predictions to our desired 6 classes. Figure 5 shows the normalized confusion matrix for the six diagnostic classes of interest for this model that is clearly a step up from our resnet50 model and good starting point for a reduced lead model.

We tried two different ways of adapting their 12-lead model to the l-lead case. Because we could not use their serialized trained model if we structurally changed the model to accept only one lead, we first tried to side step that issue by synthesizing the other 11 leads from one lead and then continuing to use their serialized trained 12-lead model.

We trialed this idea by training a series of autoencoders. We took a single ECG lead (e.g., lead II) signal and attempted to produce realistic signals to represent the other 11 leads. These 11 autoencoders would be added on to the front-end of the model to effectively use a single input without changing the input structure of the LaussenLabs serialized models we had available. This approach ended in failure and Figure 6 demonstrates why. Note that the example input waveform used was sinus rhythm, but the synthetic lead V6 output waveform more closely resembled ventricular fibrillation (VF) than it did sinus rhythm.

With this failure under our belt we then shifted to our second but more more time consuming strategy of modifying their model to be intentionally focused on a 1-lead model. But this meant that we needed to retrain a large parameter model from scratch. Fortunately the access we were provided to the MSI cluster gave us just barely enough resources in the time we had left to accomplish a reasonable training cycle. Training on the MSI cluster with 8 GPU's we took five calendar days of and completed just in time to give us a few hours of analysis before our deadline.

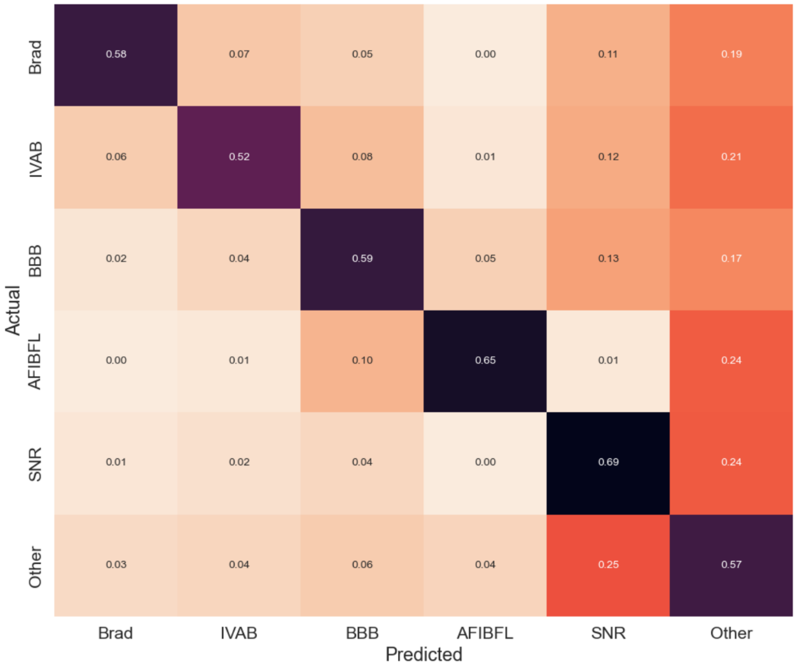

Figure 5: Normalized confusion matrix for 12-lead LaussenLabs model

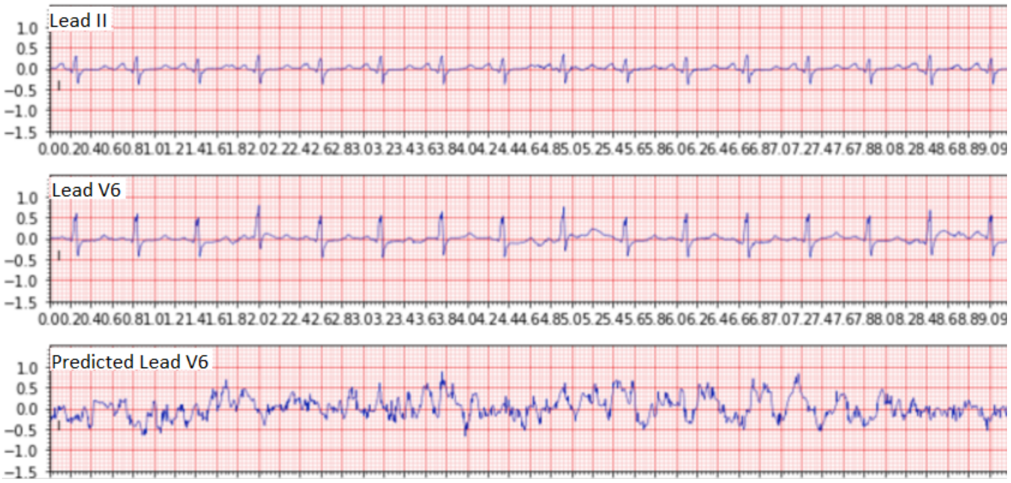

Figure 6: Results from an autoencoder intended to take a lead II waveform as input and produce a synthetic lead V6 waveform.

The classification evaluation was run on the identical test dataset as used with all the other models and as seen in Table 1 the F1 score was 0.60, AUC-ROC 0.83, AUC-PR 0.44, accuracy was 0.50, and the normalized confusion matrix seen in (Figure 7).

Table 2 includes the numbers of parameters used in all four models included in this analysis. Note that the numbers of parameters in the LaussenLabs models are multipled by six, since an ensemble of six models was used, and unlike for the resnet models, all the parameters had to be trained.

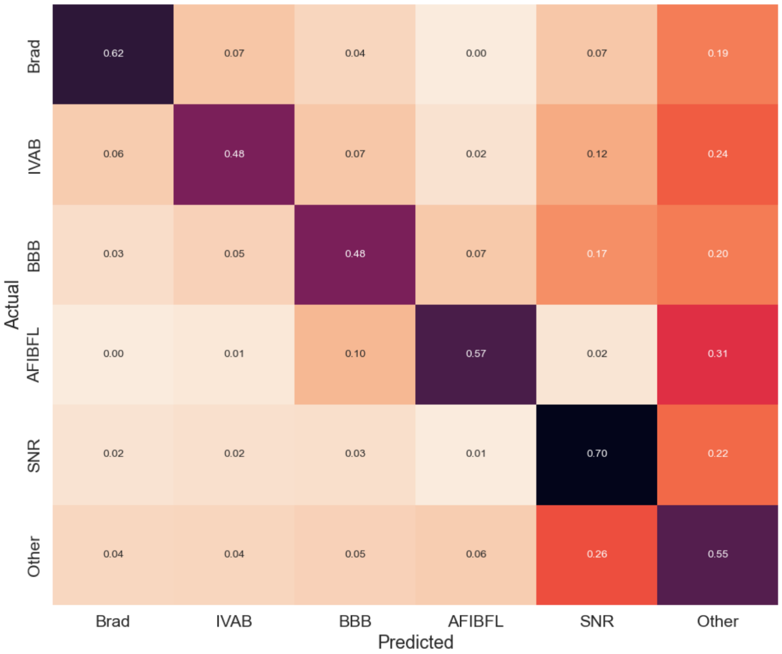

Figure 7: Normalized confusion matrix for single-lead version of LaussenLabs model

Table 1: Performance metrics, 1-lead and 12-lead versions of the resnet50 and LaussenLabs models

Model	**F1 Score**	**AUC-ROC**	**AUC-PR**	**Accuracy**
12-lead resnet50	0.34	0.66	0.25	0.40
1-lead resnet50	0.20	0.55	0.18	0.28
12-lead LaussenLabs	0.68	0.89	0.53	0.60
1-lead LaussenLabs	0.60	0.83	0.44	0.50

Table 2: Number of parameters used in each model

Model	**Number of Trainable Parameters** $\times 10^6$	**Total Number of Parameters**$\times 10^6$
12-lead resnet50	4.20	27.72
1-lead resnet50	4.20	27.72
12-lead LaussenLabs	2.14 ×6	2.14 ×6
1-lead LaussenLabs	2.14 ×6	2.14 ×6

5 Discussion

It became clear as we were preparing our progress report that we had significantly underestimated the amount of effort it would take to produce a reasonably performing 1-lead model, which was to be the basis of the key questions we sought to answer at the outset of the project. Thus, the project scope narrowed to focus on producing a reasonably performing cardiac abnormality classification model. In the end, we feel we accomplished that basic goal, with certain caveats.

The 12-lead LaussenLabs model, which can be considered representative of the state-of-the-art of 12-lead classification models, had the best performance (F1 score = 0.68) of our four models tested in this analysis. It is noteworthy that reducing the input of the LaussenLabs model from 12 leads down to 1 lead resulted in an only 12% reduction in the F1 score (F1 score = 0.60) and 17% reduction in AUC-PR (AUC-PR = 0.44). Categorically, predictions were correct more often than not for both the 12-lead and 1-lead LaussenLabs models. In assessing the confusion matrices (Figures 5 and 7), it is apparent that the biggest issue occurs when an episode that falls into one of the specified cardiac abnormality categories is classified as ”Other.” Depending how this information is used, this may not be a significant problem clinically, as ”Other” still indicates some type of cardiac abnormality as opposed to a healthy rhythm. Conversely, ”Other” cardiac abnormalities are mistaken as healthy rhythms approximately one-quarter of the time. This could manifest as a patient safety issue, since dangerous cardiac abnormalities could go undiagnosed and untreated. It is important to note that, while there is currently no definitive performance bar set for cardiac rhythm classification performance for clinical use, it is likely that these models are not yet suitable as a substitute for ECG interpretation by a clinician.

While the 12-lead LaussenLabs model was not our original work, we did perform original and significant work in adapting it to accept a single ECG lead input and were successful in adapting and subsequently training the model. We believe that a model that takes only 1 lead instead of 12 as input, a 1-lead model may be more clinically useful. Given that 1-lead ECG monitors can be much more portable and less cumbersome compared to standard 12-lead ECG monitors, 1-lead ECG monitors can be worn for extended periods of time. They can even be implanted in the body and used to continuously monitor cardiac rhythms for up to five years (Medtronic (2022)). The power of being able to monitor and detect cardiac abnormalities while the patient is ambulatory and over an extended period of time may trump the relatively small performance enhancement gained by adding 11 more ECG leads.

Moreover, we were able to compare the work previously published by LaussenLabs to a fairly naïve but potentially powerful application of transfer learning using the resnet50 model adapted to ECG waveform input for both the 12-lead and 1-lead data inputs. The former represents a more complex model with many domain-specific features, while the latter represents a more simplistic model (albeit, deep neural network) without many specific adaptations to the task at hand. While the F1 scores of the resnet50 models were poor compared to that of the LaussenLabs models (12-lead: 0.34 vs. 0.68, 1-lead: 0.20 vs 0.60, for resnet50 vs. LaussenLabs, respectively), it is important to note that the resnet50 models were significantly smaller in trainable size compared to the LaussenLabs models (Table 2). If we had had double the amount of MSI time available, we would have looked at training a resnet model from scratch on our ECG datasets. In addition, we would have benefited from more time to systematically explore which of the domain-specific features in the LaussenLabs models had the largest performance impact.

We suspect that their data pre-processing that was tuned to picked up certain classification issues (the inferred R-, P-, and T- waves) as well as the learnable prediction thresholds had the largest impact on performance. Both of these ideas could be adapted to our resnet50 models. The additional autoencoder head contributing half of the driving loss was also an intriguing component of the LaussenLabs model, as the basis for it was not readily apparent to us; perhaps, it is smoothing out, without too much distortion, the overall loss function, which is still primarily driven by the classification loss. This is suggested by the difference in training loss functions for the two 1-lead models (Figure 8).

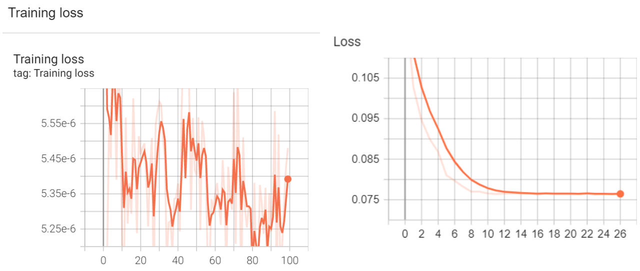

Figure 8: 1-lead resnet50 training loss curve (left), 1-lead LaussenLab training loss curve (right)

This work suggests that domain specific preprocessing can still be valuable even in the day of deep multilayer pretrained models. In addition to adding the inferred R, P, T waveform information to the input, we think it would be more feasible to apply Fast Fourier Transform (FFT) to the raw ECG signal. FFT can convert the ECG signal from time domain to the frequency domain Prasad & Parthasarathy (2018). This makes the network gain insight into peak values and isolate useful extractions while filtering out the noise by learning. By doing so, the input is not merely the time series data but a combination of time and frequency data. Next, while still adopting the resnet50 model we could tune the structure using Bayesian optimization Snoek et al. (2012). This step is expected to take time and effort to make things compatible, but once we implement it into our model, Bayesian optimization can guide us tuning structural hyperparameters (e.g., number of hidden units, number of layers) and optimizer hyperparameters (e.g., batch size, learning rate, optimizer). It is possible this resnet50-based model we envision will be competitive with the LaussenLab model.

Depending on system constraints, it may be required to implement the model directly on the cardiac monitoring device hardware rather than on an edge AI device or cloud-based system. Because GPU/CPU power for these monitoring devices remains a limitation, the goal of creating a 1-lead model that could be implemented onboard the device while retaining something close to our best 1-lead model represents another future challenge.

6 Conclusions

We learned through this work to appreciate deeply that rhythm classification even with 12-lead ECG data is challenging, and that further reducing down to a single lead is even more so. However, while including more ECG leads as inputs improves classification performance, it may not be feasible given the constraints of designing a cardiac monitoring device. For example, size is a critical parameter for an implantable monitor. Beyond model complexity, there are other avenues that could be explored to improve the performance of a portable cardiac monitoring device. Episodes of greater than 10 seconds in duration may be useful in model training by providing more context for a given cardiac abnormality. And, as mentioned earlier, feature preprocessing could still play a useful role in these limited devices. As challenging as this area is, the value in creating a reliable ambulatory 1-lead ECG monitoring device will drive further improvements.

References

Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021, 2022. URL `https://moody-challenge.physionet.org/2021/`.

Sebastian Goodfellow, Dmitrii Shubin, and Bobby Greer. Seb-good/physionet-challenge-2020: Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020, 2020. URL `https://github.com/Seb-Good/physionet-challenge-2020`.

S. Karthik, M. Santhosh, M.S. Kavitha, and A. Christopher Paul. Automated deep learning based cardiovascular disease diagnosis using ecg signals. *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, 42(1):183–199, 2022. ISSN 0267-6192.

Medtronic. Linq ii - cardiac monitors, 2022. URL `https://www.medtronic.com/us-en/healthcare-professionals/products/cardiac-rhythm/cardiac-monitors/linq-ii.html`.

Erick Andres Perez Alday, Annie Gu, Amit Shah, Chengyu Liu, Ashish Sharma, Salman Seyedi, Ali Bahrami Rad, Matthew Reyna, and Gari Clifford. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020, Jul 2022. URL `https://physionet.org/content/challenge-2020/1.0.2/`.

EA Perez Alday EA;Robichaux C;Ian Wong AK;Liu C;Liu F;Bahrami Rad A;Elola A;Seyedi S;Li Q;Sharma A;Clifford GD;Reyna MA;, A Gu, AJ Shah, C Robichaux, AI Wong, C Liu, F Liu, AB Rad, A Elola, S Seyedi, and et al. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020, 2020. URL `https://pubmed.ncbi.nlm.nih.gov/33176294/`.

B. V. P Prasad and Velusamy Parthasarathy. Detection and classification of cardiovascular abnormalities using fft based multi-objective genetic algorithm. *Biotechnology, biotechnological equipment*, 32(1):183–193, 2018. ISSN 1310-2818.

Atiaf Ayal Rawi, Murtada Kalafalla Albashir, and Awadallah Mohammed Ahmed. Classification and detection of ecg arrhythmia and myocardial infarction using deep learning: A review. *Webology*, 19(1):1151–1170, 2022. ISSN 1735-188X.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms, 2012. URL `https://arxiv.org/abs/1206.2944`.