Group 22

Response to critique:

Critique of Group 22 presentation - ECG Diagnosis using DL. Critique by Group 17

The team had a great presentation. It was well-paced and discussed all important features of the project. The proposition to try a different test set to check if the model generalises well seems like an interesting idea and we would like to see the results of the same. The code walk through was smooth and well-planned.
Thank you for your comments!

These are our questions regarding the implementation:

1. The dataset contains ECG recordings of different durations, which would translate to different input shapes. Can you please explain how you dealt with this?
   The DataLoader class that we use loads 30 seconds of data for each ECG used. For 500 Hz, this is 15000 data points. If a record is shorter than 30 seconds, the remainder is filled with 0s. If a record is longer than 30 seconds, only the last 30 seconds is used.
2. Can the model detect the presence of multiple arrhythmias in a single sample? If not, is there a priority for classification based on risk?
   Yes, the model can detect multiple arrhythmias in a single recording. The CPSC dataset contains examples that have multiple labels for the same recording.
3. What is an attention layer? What is its purpose?
   LSTMs, when used in an encoder-decoder architecture, suffer from some drawbacks. All input sequences are encoded into a fixed-length vector, which is then decoded. This hinders performance when very long sequences are given as input. This is because we consider just the hidden state at the last step of the encoder for recovering the entire sequence. Attention mechanism assigns importance to each hidden state that comprises a ‘context vector’ to overcome this issue.
4. Did you perform any hyper-parameter tuning? Can you please discuss implementation details (learning rate, loss function, etc.)?
   Zhang et al. perform hyperparameter tuning through limited trial and error. They use Adam optimizer and cross-entropy loss, with a learning rate of 0.0001 and batch size of 32. We use the same setting and did not perform hyperparameter tuning separately.

Critique of group 22 presentation - ECG Diagnosis using Deep Learning. Critiques by group 11.

Group 22 explains the motivation of the project clearly by carefully discussing the gap in the medical field for its DL aided diagnosis of ECG data. Many of the slides contained

diagrams allowing the audience to present with material not only through text and audio but also through visual aid. All of the presenters talked slowly and clearly and were easy to understand. Information of the datasets were clearly laid out, and so was the baseline architecture used. The presentation had a smooth flow of starting with the background, it's purpose, as well as explaining relevant literature that were used for reference. Their choice of model was clearly defended by backing up with works cited that tackled similar problems. One suggestion was to add more explanation on each of the classes this model was trying to assign the ECG data.
Thank you for your comments!

Question:

- From the diagram on slide 9, there are 9 different classes, normal sinus rhythm (NSR) for negative results, and the remaining 8 were arrhythmia of cardiovascular diseases. However, in your model, you have narrowed down to 4 different classes of cardiac arrhythmia. Did you combine the categories into a broader classification, or did you drop a few of the categories to simplify the datasets?
  We actually trained and tested on all 9 of the classes, we showed only four of them in the presentation due to space limitations (the tables would become too big otherwise). We should have clarified this in the presentation, thank you for pointing it out.
- When recreating the model mentioned on slide 12, what accounts for the differences in the performance reported by Zhang et al. 2020 versus the one replicated?
  The main difference comes from the fact that Zhang et al. report results after 10-fold cross validation, while we did not implement that. Therefore, depending on the randomness in the pre-processing and in the data augmentation (scale+shift) in the data loader, the results would vary slightly.

Critique of group 22 presentation – ECG Diagnosis. Using Deep Learning Critiques by group 5

Overview:

This work aims to detect cardiovascular diseases (CVDs) from ECG signals using a CNN and RNN model. A deep neural network with bidirectional GRUs is experimented with for improvements over existing work.

Good:

The results from literature are successfully reproduced, which defines a good research starting point. Even though BiGRU did not improve upon the baseline results, the reasons for using BiGRU are well given from the literature. The motivation of creating a more efficient classifier is well defined. Further, the results of low number leads and downsampled ECGs are interesting!
Thank you for your comments!

Improvements:

- A definition of what “lead” means in this context would have been beneficial.
  Conventional ECG is captured by sticking many electrodes on the human body and measuring the voltage differential between various pairs. Pairs of electrodes are specifically chosen to measure various views of the electrical impulses in the body. Each pair is called a “lead”, and there are 12 such leads in a conventional ECG. We’ll include some text defining this in the report.
- The literature review mentions traditional methods, but their results are not shared. It would be nice to compare your results with those as well.
  Our baseline DL model from Zhang et al. 2020 performs much better than traditional methods. They compare their model to these methods and show that the best traditional model achieves an average F1 score of 0.619 (Section IV. C in the paper), in comparison to the DL model’s average F1 of 0.813. We chose to omit this comparison since it’s already covered in the paper.

Questions/Improvements:

- It seems like 3-lead is the best tradeoff between accuracy and number of leads. What is the accuracy of 3-lead when downsampling 5x down?

  Yes, it does seem like 3-lead offers a “best of both worlds” scenario. We have performed this experiment and reported the results in Table 2. It turns out that the 3-Lead ECG with 5x down performs the same as the original 12-Lead model on all metrics. We therefore get 20x improvement at no cost!
- What is the lowest amount of parameters that can be achieved when downsampling? Or what is the percentage of reduction achieved between the numbers that are presented (e.g. number of 12-lead parameters downsampled / number of 3-lead parameters downsampled)?

  Since we’re downsampling the data only, while keeping the model hyperparameters the same, the reduction is directly proportional to the amount of downsampling. This is due to the fact that convolutions have a complexity proportional to (kernel size x signal length). Reducing signal length by 5 reduces the overall activation memory and computations by 5 (20% of original). Reducing the number of leads has a linear relationship with the computations. An example:
  12 Lead 500 Hz: X (X is the number of operations, or activation size)
  12 Lead 100 Hz: X/5
  3 Lead 100 Hz: X/20
  1 Lead 100 Hz: X/60
  We have included this explanation in the text.
- Even though RNNs may be faster to train, do you think it is possible that LSTMs might produce better results?

  Both LSTMs as well as Bi-GRUs are better than vanilla RNNs and are able to handle the gradient vanishing/exploding problem. We tried using LSTM as well, and got very similar results. Given that we didn’t see a considerable difference in performance between LSTM and Bi-GRU, we decided to stick to Bi-GRU since it is faster to train.
- Why did you choose a 1996 dataset? (A better comparison might be to a more current dataset.)

The CPSC dataset is from 2018, while the PTB-XL dataset is from 1996 (although it was made publicly available only in 2020). We mainly use the CPSC dataset, since Zhang et al. 2020 uses it too. PTB-XL is relevant only to understand model generalization.

GROUP 22: ECG ANALYSIS USING DEEP LEARNING

Raghav Subbaraman, Savitha Srinivasan, and Shivani Athavale

University of California San Diego, La Jolla, CA 92093-0238,

ABSTRACT

Deep Learning methods for electrocardiogram (ECG) diagnosis have seen a lot of interest in recent years. The ease of obtaining ECGs with the dearth of expert physicians has created the necessity for automatic diagnosis. Our goal is to use this ECG data to build a deep learning model for automatic diagnosis of various heart arrhythmia. Using [1] as inspiration, we create a new model with a Bi-GRU layer that performs at par. Then, we simplify the model by lowering the input sample rate and data dimension and observe the performance. We show that a 3-Lead ECG at 100 Hz rate provides 20x reduction in inference resources at no performance cost. Finally, we test model generalizability across unseen datasets from a different demographic. Our project is available at: `https://github.com/savitha0602/ecg-diagnosis`

***Index Terms*—** Electrocardiogram, ECG, Deep Learning, Gated Recurrent Unit, GRU, Medical, Diagnosis, Arrhythmia

1. INTRODUCTION

Cardiovascular diseases (CVDs) are a major cause of deaths, and put a lot of burden on the economic and health systems in the US and in other countries [2]. The electrocardiogram (ECG) is a non invasive method to measure the electrical activity of the heart. The 12-lead ECG is the traditional method which uses leads/electrodes at different locations in the body to obtain time-series signals. Our goal is to use this data to build a deep learning based model for automatic diagnosis of various heart arrhythmia. The main obstacle in conventional ECG diagnosis is the requirement of expert physician input. Traditional computer aided tools can automate part of the process, but require expert knowledge and feature extraction.

Deep Learning (DL) is a natural fit for this problem, since ECG data is easier to obtain than expert diagnosis. Added to this, DL has shown success in various medical applications such as medical image recognition and computer vision for diagnosis and surgery. Since ECG signals can be represented at 1-D time series data, convolutional filters and recurrent neural networks would presumably work well in characterizing the temporal and spatial relationships. These DL models find use cases when experts are not available and in wearables for real-time computing. In this paper, we use [1] as our primary reference, where a DL model is used on raw 12-lead ECG data to classify various types of arrhythmia. We build on the model built by [1] using a Bi-GRU layer. Then, we explore data downsampling and lead reduction to make the model more efficient. Finally, we explore model generalizability outside of the dataset used in [1].

2. RELATED WORK

There is a large body of prior work on ECG diagnosis using traditional methods as well as deep learning. Traditional methods use signal processing and feature engineering based ML methods, requiring expert knowledge or intervention. [3] uses a computer program to detect STEMI (ST elevation myocardial infarction) using 12-lead ECG signals. It uses the age and sex based criteria by the Glasgow program [4] and the ESC/ACC criteria [5], achieving a sensitivity of 89%. However, this method required cardiologists to verifying the final diagnosis.

Recently, there has been great interest in DL based methods for ECG diagnosis, with CNNs and LSTMs are observed to be the most common architectures [6]. In [7], the authors use segmented ECG signals and DL to classify NSR, SVEB, VEB, and fusion of Ventricular and NSR. In specific, they use multi layer perceptron network with ReLU activation, and obtained an accuracy 99.68 percent. The authors in [8] use a wavelet transform and multiple LSTM recurrent neural networks for binary classification of VEB and SVEB. While these works demonstrate high performance on a limited set of arrhythmia, we explore DL based diagnosis for a much wider set.

We primarily build on the DL model developed in [1] for 12-Lead ECG records that predicts between 10 different classes according to the guidelines in [9]. This work uses a CNN along with residual connections to obtain an average F1-score of 0.813 on their dataset. The authors in [10] propose a different architecture based on RNN on the same dataset and obtain better performance. This analysis provided the inspiration for us to try and combine these models.

3. DATASET AND FEATURES

CPSC2018: [1] uses the China Physiological Signal Challenge 2018 (CPSC2018) database [9]. This database comprises of 6877 12-lead ECGs lasting between 6 seconds and

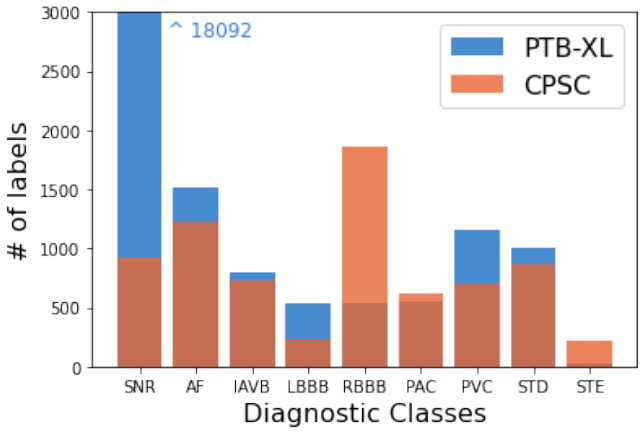

Fig. 1. Class distribution within the CPSC 2018 [9] and PTB-XL [12] Datasets. There is imbalance in both datasets, but they favor different labels. RBBB and AF are most frequent in CPSC2018, while SNR and AF are most frequent in PTB-XL.

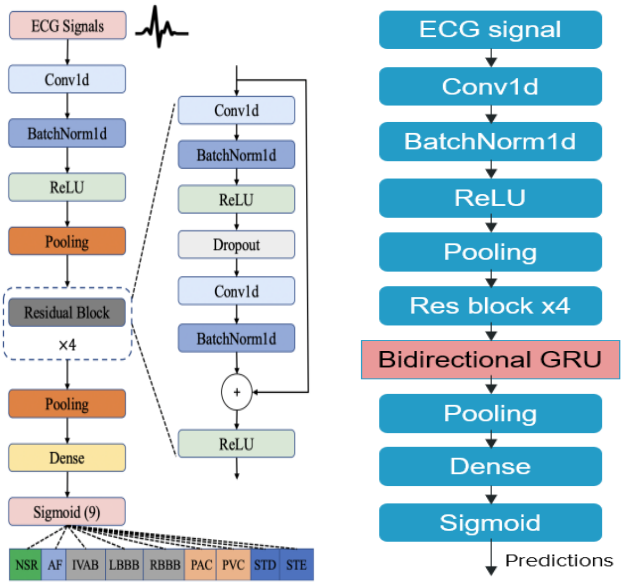

Fig. 2. (a) Network architecture proposed in [1] used as our baseline (b) Modified architecture

60 seconds at a sampling rate of 500 Hz. The ECGs are labeled with 9 diagnostic classes: normal sinus rhythm (NSR), atrial fibrillation (AF), first-degree atrioventricular block (IVAB), left bundle branch (LBBB), right bundle branch block (RBBB), premature ventricular contraction (PVC), ST-segment depression (STD), ST-segment elevation (STE). Fig 1 shows the count of each of the nine labels in the dataset. It is apparent that there exists a class imbalance issue, particularly in the case of STE and LBBB.

PTB-XL: The PTB-XL ECG dataset is a large dataset of 21837 clinical 12-lead ECGs from 18885 patients of 10 second length at a sampling rate of 500 Hz. Only labels common with CPSC2018 are used for our generalizability study. The label frequency is plotted in Fig 1, and shows significant class imbalance, different from that of CPSC2018. **This dataset is used only in the generalizability study**. Both datasets are publicly available in the WFDB format and are identically organized [11].

Feature extraction: Since we use an end-to-end deep learning technique, there is no explicit feature extraction required and the raw ECG data is used. The 1D convolutional layers in the architecture are automatically trained to extract deep features.

4. METHODOLOGY

We use the architecture proposed in the 12-lead interpretable deep learning paper [1] as our baseline. The authors of the paper developed a deep neural network based on 1D convolutional layers for automatic diagnosis of 12-lead ECG recordings. The network architecture utilizes residual blocks with shortcut connections to make the model training tractable. The inputs to the model are raw ECG signals $x \in \mathbb{R}^{nstepsX12}$, where nsteps is the length of each input(optimally chosen to be 15000, which denotes a duration of 30s at a sampling rate of 500 Hz). The network outputs a multi-label classification result $y \in \mathbb{R}^{1x9}$. We also incorporate a bidirectional gated recurrent unit(GRU) to the baseline architecture, as proposed in [13].

4.1. Baseline

Figure 2 (b) shows the network architecture used as our baseline. The network consists of 34 layers, including 4 stacked residual blocks that are used to extract deep features. Each residual block contains the following layers:

- Two 1D convolutional layers: For automatic feature extraction
- Two batch normalization layers: For faster and stable training
- ReLU: For non-linear activation
- Dropout: To reduce overfitting

The features extracted by the residual blocks are pooled using adaptive average pooling and max-pooling. The pooled features are then concatenated and sent to the output layer that uses sigmoid activation to classify into one of nine diagnostic classes.

4.2. Gated Recurrent Unit

The Gated Recurrent Unit (GRU) is a variant of RNN that can effectively retain long-term dependencies in sequential data

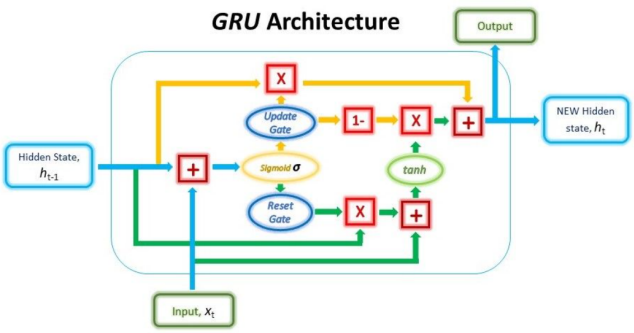

Fig. 3. Inner architecture of a GRU cell [14]

[14]. It uses gating mechanisms to manage information flow between cells in the network. Just like a vanilla RNN, GRU also maintains a "hidden state" that is fed back into the cell together with the next input in the sequence. The hidden state holds both long-term as well as short-term dependencies. The GRU cell contains two gates: update and reset. These gates selectively filter out information that is not useful. The architecture of the GRU cell can be seen in 3

4.3. Modified Architecture

Since the input is a 1D time-series signal, RNNs are popularly used to learn the time correlation between ECG signal points. Chen et al. [13] propose using a birectional RNN in addition to the convolutional layers. The paper also reports better results on the same CPSC2018 dataset. Inspired by this, we use a bidirectional GRU layer after the stacked residual blocks in an effort to improve the performance of the baseline model. We choose a GRU over a vanilla RNN to account for long-range dependencies. Another popular choice is Long Short Term Memory(LSTM) instead of GRU. Since GRUs have fewer weights and parameters compared to LSTM layers, they are faster to train.

4.4. Metrics

Our baseline [1] reports Precision, Recall, F1 score and area under the receiver operating characteristic curve(AUC) for each of the diagnostic classes. The metrics are calculated as given below:

$$ACC_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (1)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$Precision_i = \frac{TP_i}{TN_i + FP_i} \quad (3)$$

$$F_{1i} = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (4)$$

Model	Prec	Recall	F1	AUC	Acc
Baseline	0.821	0.812	0.813	0.970	0.966
Reproduced	0.84	0.801	0.806	0.975	0.963
Our Bi-GRU	0.809	0.808	0.805	0.966	0.966

Table 1. Comparison of the baseline 12 Lead model with our reproduction of [1] and the Bi-GRU model. The metrics are averaged over all diagnostic classes.

where TP_i, TN_i, FN_i represent the number of true positive samples, number of true negative samples, number of false positive samples and the number of false negative samples for class i respectively. The average of these scores are also reported for to evaluate the performance of the model.

5. RESULTS AND DISCUSSION

5.1. Experimental Setup

Data preprocessing: The CPSC2018 dataset is preprocessed in order to make all input signals the same length. The optimal input length was found to be 15000 (30 seconds at a sampling rate of 500 Hz) by [1]. Inputs that are longer than 30 seconds are cropped, and the ones that are shorter are padded with zeros. When a lower number of leads need to be used, data corresponding to other leads are simply dropped. **Data Augmentation:** Scaling and shifting has been used as data augmentation strategies to help reduce overfitting and enhance model robustness. **Training details:** The code is developed using PyTorch framework. Adam optimizer is used with a learning rate of 0.0001. A batch size of 32 and a kernel size of 15 for the 1D CNNs is used. The model is trained for 30 epochs.

5.2. Bi-GRU based model performance

From Tabe 1, we note that the addition of the BiGRU layer did not affect the results significantly, for this fold. However, it does outperform the baseline when another fold is used. Performing 10-fold cross-validation might give us a better intuition about the effectiveness of the added layer. We can, however, infer that there is a significant contribution of the residual blocks to the model performance. Additionally, as [13] suggests, recurrent units might also require attention for better results.

5.3. Case Study: Reducing complexity and inference time

The 12-Lead method is the clinical standard for ECG. However, it is expensive and requires expert assistance to set-up since 12 leads (leads are electrodes) need to be placed accurately on the human body. In contrast, 6-Lead, 3-Lead and even 1-Lead ECG are commonplace today [15], fuelled by the growing tech market for personal care. Various authors have

Model (Leads, Rate)	Prec	Recall	F1	AUC	Acc
(12, 500 Hz)	0.84	0.801	0.806	0.975	0.963
(6, 500 Hz)	0.825	0.81	0.815	0.966	0.962
(3, 500 Hz)	0.801	0.803	0.796	0.968	0.935
(1, 500 Hz)	0.798	0.736	0.745	0.955	0.949
(12, 100 Hz)	0.805	0.822	0.805	0.974	0.961
(3, 100 Hz)	0.813	0.8	0.805	0.971	0.961
(1, 100 Hz)	0.782	0.72	0.722	0.947	0.946

Table 2. Model performance when number of leads and sample rates are varied

also demonstrated that ECGs with lower leads may be used for diagnosis with minimal loss in performance [16, 17, 11]. In addition to having lower number of leads, these devices also typically have lower processing power for signal processing. For example, the AppleWatch is a 1-Lead ECG that may be required to perform quick inference on-board. To see how well the baseline model scales to these scenarios, we perform experiments with lower number of leads in the data. We also explore data downsampling as a method to reduce model complexity improve inference time.

Reducing number of leads in data: All recordings in the CPSC2018 dataset contain 12 leads. To observe model performance with varying number of leads, we use the code from [1] to select 6-Lead, 3-Lead and 1-Lead data according to [11]. For each of the Lead configurations, we re-train the model and report test performance averaged across all labels. The results are summarized in Table 2. We see that the average performance does not change significantly even if we use only 3 Leads. When using only 1 Lead (Lead I), the performance significantly degrades on all metrics. This indicates that most of the underlying classes can be identified with just 3 Leads.

Effect of downsampling data: The ECG recordings in the CPSC2018 dataset are sampled at 500 Hz. Typical heartbeat frequencies are close to 1.2 Hz (72 bpm), indicating that 500 Hz is much higher than the required Nyquist rate for ECG time-series. Indeed, [18, 19] argue that 100 Hz is sufficient. Therefore, we downsample the dataset to 100 Hz by picking 1 in 5 samples, and then re-train the model and observe its performance. This downsampling operation results in a 5x improvement in model inference time and a 5x reduction in memory usage (since CNNs scale linearly with input size). The average performance of the models across all labels are shown in Table 2, we see that the performance is almost unchanged. The 12-Lead 100 Hz model's average F1 score is virtually same as that of the 500 Hz model. We have also reported 100 Hz model results for 3-Lead and 1-Lead ECG configurations. In summary, among all the models tested, the 3-Lead ECG at 100 Hz provides the best trade-off between performance and complexity: The average F1 score is the same as that of the 12-Lead 500 Hz model, while being 20 times more efficient in terms of data usage and inference time.

Train	Test	Prec	Recall	F1	AUC	Acc
CPSC	CPSC	0.84	0.801	0.806	0.975	0.963
CPSC	PTB-XL	0.511	0.66	0.505	0.844	0.834
PTB-XL	CPSC	0.567	0.823	0.578	0.879	0.713
PTB-XL	PTB-XL	0.634	0.815	0.661	0.914	0.865

Table 3. Model performance when trained and tested on various dataset combinations.

5.4. Case Study: Model generalizability across datasets

CVDs are prevalent in all parts of the world, therefore, it is of great importance to ensure that the model generalizes across various demographics and recordings. Inspired by the diverse sources of data from [11], we attempt to compare the model performance across the CPSC2018 [9] and the PTB-XL [12] datasets. In specific, we train on each and test on each, resulting in four sets of results summarized in Table 3. We see that the average performance across all metrics is poor whenever PTB-XL is involved. Upon analyzing the label frequency and the distribution of the number of labels per observation, we see a stark difference between both datasets. From Fig 1, we see that the type of class imbalance is different across both the datasets. Additionally, upon analyzing the dataset, we see that the PTB-XL dataset contains up-to 5 labels per observation, compared to only 3 in CPSC; possibly worsening cross-dataset performance.

6. CONCLUSION

Deep Learning based solutions are a good fit for automatic ECG diagnosis. The residual-network based model from [1] is quite capable of classifying various hear arrhythmia based on ECG readings. Adding a Bi-GRU to the network did not improve performance significantly, indicating that the original model itself is quite capable. We observe that the model can be simplified and made more efficient by reducing the number of ECG leads used, as well as by performing downsampling. A 3-Lead ECG with a sample rate of 100 Hz performs the same as a 12-Lead 500 Hz model while being 20x more efficient. We observe that the model does not perform well on the PTB-XL dataset when trained only on the CPSC dataset (and vice-versa), even though the common labels were used. We believe that this is due to non-uniform class imbalance, and differences in diagnostic methods/expertise in ground-truth labelling. While we see that DL models can solve the task at hand, it is still an open challenge to combine datasets and testing methods to make DL based ECG classification generalizable and practical.

7. ACKNOWLEDGEMENTS

We thank the members of Group 17, 11 and 5 for their insightful suggestions and critique. We are grateful to Prof. Gerstoft and the TAs of ECE228 for their guidance and support throughout the course.

8. CONTRIBUTIONS

Raghav: Downsampling experiments, lower-lead experiments, code-review, writing (section 5, 6). Savitha: Bi-GRU model ideation and implementation, results, dataset visualization, writing (section 3, 4). Shivani: Model generalization experiments, lower-lead experiments, dataset visualization writing (section 1, 2). Github link: `https://github.com/savitha0602/ecg-diagnosis`

9. REFERENCES

[1] Dongdong Zhang, Xiaohui Yuan, and Ping Zhang. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *CoRR*, abs/2010.10328, 2020.

[2] Salim S Virani, Alvaro Alonso, Emelia J Benjamin, Marcio S Bittencourt, Clifton W Callaway, April P Carson, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Francesca N Delling, et al. Heart disease and stroke statistics—2020 update: a report from the american heart association. *Circulation*, 141(9):e139–e596, 2020.

[3] PW Macfarlane, DR Hampton, E Clark, B Devine, and CP Jayne. Evaluation of age and sex dependent criteria for st elevation myocardial infarction. In *2007 Computers in Cardiology*, pages 293–296. IEEE, 2007.

[4] PW Macfarlane, B Devine, and E Clark. The university of glasgow (uni-g) ecg analysis program. In *Computers in Cardiology, 2005*, pages 451–454. IEEE, 2005.

[5] Borja Ibanez, Stefan James, Stefan Agewall, Manuel J Antunes, Chiara Bucciarelli-Ducci, Hector Bueno, Alida LP Caforio, Filippo Crea, John A Goudevenos, Sigrun Halvorsen, et al. 2017 esc guidelines for the management of acute myocardial infarction in patients presenting with st-segment elevation: The task force for the management of acute myocardial infarction in patients presenting with st-segment elevation of the european society of cardiology (esc). *European heart journal*, 39(2):119–177, 2018.

[6] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, page 100033, 2020.

[7] Giovanna Sannino and Giuseppe De Pietro. A deep learning approach for ecg-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems*, 86:446–455, 2018.

[8] Saeed Saadatnejad, Mohammadhosein Oveisi, and Matin Hashemi. Lstm-based ecg classification for continuous monitoring on personal wearable devices. *IEEE journal of biomedical and health informatics*, 24(2):515–523, 2019.

[9] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[10] Tsai-Min Chen, Chih-Han Huang, Edward SC Shih, Yu-Feng Hu, and Ming-Jing Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *Iscience*, 23(3):100886, 2020.

[11] Reyna, Matthew, Sadr, Nadi, Gu, Annie, Perez Alday, Erick Andres, Liu, Chengyu, Seyedi, Salman, Shah, Amit, and Gari Clifford. "Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet - Computing in Cardiology Challenge 2021" (version 1.02). PhysioNet (2021). https://doi.org/10.13026/jz9p-0m02, 2021. [Online; accessed 07-June-2021].

[12] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[13] Tsai-Min Chen, Chih-Han Huang, Edward S.C. Shih, Yu-Feng Hu, Ming-Jing Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *iScience*, 23(3): 100886, 2020.

[14] Gabriel Loye. Gated Recurrent Unit (GRU) With PyTorch. https://blog.floydhub.com/gru-with-pytorch/, 2008. [Online; accessed 07-June-2021].

[15] Kevin Rajakariar, Anoop N Koshy, Jithin K Sajeev, Sachin Nair, Louise Roberts, and Andrew W Teh. Accuracy of a smartwatch based single-lead electrocardiogram device in detection of atrial fibrillation. *Heart*, 106(9):665–670, 2020.

[16] Robert Brockway, Marina Brockway, Brian Brockway, and Robert Hamlin. Comparison of one-and three-lead

ecg to measure cardiac intervals and differentiate drug-induced multi-channel block. *Journal of pharmacological and toxicological methods*, 93:80–89, 2018.

[17] Haakon Tillmann Haverkamp, Stig Ove Fosse, and Peter Schuster. Accuracy and usability of single-lead ecg from smartphones-a clinical study. *Indian pacing and electrophysiology journal*, 19(4):145–149, 2019.

[18] Era Ajdaraga and Marjan Gusev. Analysis of sampling frequency and resolution in ecg signals. In *2017 25th Telecommunication Forum (TELFOR)*, pages 1–4. IEEE, 2017.

[19] Ohhwan Kwon, Jinwoo Jeong, Hyung Bin Kim, In Ho Kwon, Song Yi Park, Ji Eun Kim, and Yuri Choi. Electrocardiogram sampling frequency range acceptable for heart rate variability analysis. *Healthcare informatics research*, 24(3):198, 2018.