Capstone - Sprint 0

Calvin Chan

February 2024

1 Area of Interest

The area that I am interested in is biomedical signal processing. I am interested in seeing how I can utilize data science to increase the accuracy of reading medical data, specifically an electrocardiogram (ECG). A systematic review and meta-analysis published in 2020 looked at 78 studies that assessed the accuracy of ECG interpretations by physicians with different levels of training and specialization, including medical students, residents, physicians in non-cardiology practice, and cardiologists [1]. It was found that accuracy scores varied significantly between the different physician levels, ranging from 4% to 95%. Even amongst cardiologists the accuracy still fluctuates, averaging at 75%. This poses a problem for patients and specialists as different ECG results can lead to vastly different decision and outcomes. In this project, we will attempt to utilize data science techniques to read ECG data and provide a result, giving physicians and specialist an additional support for interpretation.

2 User Benefits

As mentioned earlier, ECG interpretations are typically done by physicians and possibly nurses. With a system or model that can read ECG data and provide a result, it can aid these specialist when they are working in fast and high pressure environments, such as hospitals. During high pressure environments, making a quick decision can be easy, but the results and implication can be significant. Knowing that the accuracy of ECG interpretations can be low depending on the physician's level of training and experience, having a model that can give a second opinion could drastically improve physician decisions and patient treatment care.

3 The Big Idea

Machine learning (ML) can play a significant role in classifying whether or not a patient has a certain heart condition by looking at their ECGs. It is possible to train a ML model using previous data with known outcomes to look for certain

patterns and waveforms in ECGs that do have a heart condition versus ones that do not. Although it is highly likely that even the best ML model would not give a result with 100% accuracy. However, with that in mind professionals can still draw insights into these results from ML models which can enlighten decisions made by physicians. There have been previous approaches of using ML models to classify ECG data, one of which utilized deep learning methods. In this previous approach by Parsa Kamali [2], a BrainStation graduate in 2023, he utilized a Binary and Multi-class classification approach to sort ECG signals. As next steps, he had mentioned the implementation of a 1-D convolutional neural network instead, which could possibly be the scope of this project. This would result in a categorical prediction using ML models.

4 Impact

In terms of societal impact, I believe this project can be significant if it was scaled to a larger setting. Being able to have access to this tool in a hospital setting should definitely help physicians of different levels to have better and more accurate diagnosis. It is possible to help medical students and residents in training when they are still learning about how to interpret ECG signals. In terms of business value, I believe this could also be implemented in current technologies. Imagine having a smart watch that would take your ECG on a regular basis and have feedback about your heart condition. This would be significant for people who may have underlying unknown heart conditions that have not seen a doctor yet. To have a device tell you that you may have an underlying condition advising you to see a physician can be lifesaving.

5 The Data

Several datasets were found as potential data used for this project. The main one is called PTB-XL [3], which contains 21799 entries collected from a total of 18869 patients using a 12-lead ECG. Each of these entries is an ECG time series with a length of 10 seconds. What makes this data set different from the others is the extensive use of annotations by up to two cardiologists and the vast amounts of heart conditions as outcomes. It is also paired with extensive metadata on demographics, signal characteristics and signal properties that can be extremely useful when we are looking for details in classifying the signals. Another dataset found is the MIT-BIH Arrhythmia Database which contains 48 half hour ECG readings taken from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979 [4]. With significantly less data, this dataset is less probable in terms of being used. Although it has also been used as a standard for arrhythmia detecting research, allowing us to potentially train our model for detecting arrhythmia. For a third dataset, we have the PTB Diagnostic ECG Database [5], which contains 549 ECG records from 290 subjects between the ages 17-87 years old. This dataset also utilizes a 12-lead ECG with a wide

range of diagnostic classes. This can be useful for us as the more range of type of diagnosis we have to train our model, the more the model can recognize after training.

6 The Alternative

An alternative area that also interests me would be in the field of astronomy. As we are moving towards a Big Data era, there are more and more data being collected with not enough people to process it. This applies to astronomy as well, where telescopes have been collecting more data with a limited amount of people to analyze them. A current problem in astronomy is how we can analyze these data quicker, and the solution has been gearing towards using machine learning. People have been attempting to utilize machine learning techniques to classify galaxies in images in order to process the amount of data we have quicker. For an alternative field of interest, astronomy interests me due to my background in astrophysics.

References

[1] Cook, D. A., Oh, S. Y., & Pusic, M. V. (2020). Accuracy of Physicians' Electrocardiogram Interpretations: A Systematic Review and Meta-analysis. JAMA internal medicine, 180(11), 1461–1471. https://doi.org/10.1001/jamainternmed.2020.3989

[2] https://dvrfp7vt6y4co.cloudfront.net/77c73063-bbb7-406b-b274-ad5a51797277/Parsa%20Kamali%20Resume.pdf

[3] Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2022). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet. https://doi.org/10.13026/kfzx-aw45

[4] https://www.physionet.org/content/mitdb/1.0.0/

[5] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. https://www.physionet.org/content/ptbdb/1.0.0/