**AstraVision/National Institute of Cardiology- México City Project Notes**

**Heart Project Overview**

The National Institute of Cardiology- Mexico, working with AstraVision leaders, Instituto Tecnologico Autonomo de Mexico researchers, and Indiana University Data Science Masters students, is attempting to find a more accurate way to classify arterial blockages of the heart. The team hopes to develop a machine learning model that will classify heart conditions based upon ECG/EKG readings.

1. ***Team Introductions***

**AstraVision Team**

AstraVision is a Latin American, primarily Mexican, non-profit technology company working to provide technological support to the healthcare sector. Led by Pilar Espinosa and Karlo Valentín Rodríguez Rangel, AstraVision develops models that interpret medical device readings to help medical practitioners more accurately identify illnesses and injuries. To date, AstraVision has provided model interpretations of X-Ray images of the spine and lungs, helping to identify spinal injuries and a variety of respiratory illnesses, including COVID.

**Instituto Tecnologico Autonomo de Mexico Research Team**

Led by Dr. Edgar (Paco) Roman-Rangel, the ITAM researchers have provided oversight of and advice to the student teams working to develop some of the AstraVision models. Dr. Roman has significant data science experience and knowledge of machine learning.

**Indiana University Data Science Masters Team**

The Indiana University Data Science Masters program is one of the top 10 data science graduate instructional programs in the United States. The students are a mix of graduate students, both online and residential, with a range of work experience and skills. Almost all of them are working while getting their degrees. A new team is assigned to work with the AstraVision organization each term and that team can be somewhat to very experienced. The strongest IU Data Science Masters teams working with AstraVision have been those with one or more highly capable data science project managers and one or more highly experienced data science machine learning engineers. Due to the non-disclosure nature of the work, very few will have any experience working in healthcare or medicine.

**Heart Project Progress To Date**

The Heart Project was introduced to the full project team in mid-February 2022. The IU Data Science team quickly found an existing machine learning model, coded in the Python language, that identified Atrial Fibrillation from ECG/EKGs. This model, called the Conrad/Weimann model, was based on 1-lead heart monitor readings augmented by some 12-lead readings. Unfortunately, the full model could not be copied for National Institute use as the 1-lead readings could only be used for academic research, not replicable commercial use. The IU Data Science team adapted the model to identify arterial blockages using the main lead reading from a commercially reusable 12-lead ECG reading dataset, called the PTB-XL dataset, containing over 21,837 12-lead readings. However, they found too few arterial blockage readings of a key type in that dataset, so the adapted model was not as strong as it should have been. The PTB-XL 21,837 ECG observations include 241 ISCI and 700+ ISCA labelled readings.

1. ***Project Challenges***

**Not Enough ECG Observations**

There are three available options for increasing the number of arterial blockage observations/ readings used in training the model. The first approach is to find more publicly available labelled ECG observations of arterial blockages of several types. David Ross is working on that and has found another dataset of 10,344 12-lead ECG readings, the Georgia dataset. He is also investigating a third set of 10,530 of 12-lead ECG readings, the Chinese datasets. The second approach is to create simulated ECG readings from the existing samples from the PTB-XL 21,837 12-lead ECG readings dataset, and potentially other datasets. This method is called Generative Adversarial Networks. Nisha Abdul-Jabbar is looking into this. A potential third approach is to use more than just the main lead ECG reading to create more observations. This third way may create secondary problems, due to the model adaptation.

**Model Adaptation**

The current model was adapted from 1-lead heart monitor ECG observations. Because the National Institute of Cardiology observations, the PTB-XL dataset, and other commercially reusable datasets are based upon 12-lead ECG readings, the adapted model may include some incomplete interpretations of the 12-lead ECG readings. Over time, the model will likely need further adaptation to use all the elements of the 12-lead ECG readings. This work will most likely happen in the fall of 2022 with the incoming IU Data Science team. David is currently searching to determine if a 12-lead ECG interpretation model was published from a 2020-2021 Physionet/CinC challenge that could assist in the further adaptation of the 1-lead model. Please understand that it is much simpler to adapt an existing working model used to identify/classify atrial fibrillation than to completely create a new model.

**Model Training and Testing Resource Requirements**

Another challenge for the IU Data Science teams is potential computer (GPU) constraints for training and testing the model. Each time the model is changed, it may require retraining/retesting that can take up to 12 hours of dedicated cloud computing time, which can be quite expensive. However, so long as the model is only being tested on publicly available datasets, such as the PTB-XL, Georgia, and Chinese datasets or on simulated ECG readings, it can be pretrained/ tested on mostly free Indiana University cloud servers. The pre-trained model can then be refined using the National Institute of Cardiology ECG data on the more expensive AWS cloud servers on which AstraVision has a subscription.

**Model Balancing**

A final challenge is balancing the model. An optimal model from a data science perspective is built on near equal representations of each heart condition that the model will work to predict. However, that optimal situation does not reflect medical reality. While the patient population is much more likely to include persons with heart illnesses, it is still not likely to include similar numbers of patients with each selected heart condition. Even if the initial model is trained on ECG readings with the exact number of patients with each of the selected heart conditions, as the model is retrained and/or refined periodically with National Institute of Cardiology ECG readings, that refining of the model will be subject to differing numbers of patients with each condition, potentially leading the model to better recognize one condition compared to another.

There are multiple ways of handling this conceptual challenge. The Fall 2022 IU Data Science team may also explore multiple model approaches due to this conceptual challenge.

1. ***Research Findings***

**Physionet/CinC 2020-2021 ECG Interpretation Challenge**

While not known to the Spring 2022 IU Data Science team, Physionet/CinC hosted a data science challenge from 2020 through 2021 for teams to create models interpreting 2-lead, 3-lead, and 12-lead ECG readings. The 12-lead challenge is documented by the Physionet/CinC team here (<https://physionet.org/content/challenge-2020/1.0.1/>). The results were announced on 26-Jan-2022 and can be found here (<https://moody-challenge.physionet.org/2020/>). A history of past Physionet/CinC challenges can be found here (<https://cinc.org/physionet-cinc-challenge-awards/>).

**Challenge Datasets**

The challenge used the PTB-XL dataset, the Georgia dataset, the two Chinese datasets, and some other unpublished datasets for model training and testing. The published datasets can be found here (<https://www.kaggle.com/general/199422>). While the PTB-XL and Georgia datasets can be used in commercial models, the use terms of some of the other datasets must still be determined. As noted previously, the PTB-XL dataset contains 12,837 labelled 12-lead ECG readings, each 10 seconds long at 500hz. The Georgia dataset contains 10,344 labelled 12-lead ECG readings, each also 10 seconds long at 500hz. The two Chinese datasets contain 10,530 labelled ECG readings, each of lengths ranging from 6 to 60 seconds long at 500hz. There was also a much smaller dataset of 74 ECG readings. All of the datasets contained SNOMED-CT labels, documented in the Appendix below.

**Challenge Outcome**

217 teams competed and submitted 707 working models to the challenge. A list of the 12-lead challenge winners can be found here (<https://cinc.org/physionet-cinc-challenge-awards/>). As the challenge was extended into 2021, some of the information and models from that challenge may not yet have been published. The challenge stated that all models submitted for judging needed to be published as open source permitting commercial reuse. The code used to evaluate the 12-lead ECG interpetation models can be found here (<https://github.com/physionetchallenges/evaluation-2020>). While this evaluation process may be very useful to the project team, the team must test its accessibility as it may require access to certain datasets beyond the PTB-XL, Georgia, and Chinese datasets.

**Challenge Model(s)**

To date, David has found some models with Python language code from the challenge. Many of these models have associated research papers.

* The 1st Rank Philips Research North America team’s model description can be found here (<https://www.physionet.org/files/challenge-2020/1.0.1/papers/CinC2020-107.pdf>). *The model code has not yet been found.*
* The 2nd Rank Xian Jiaotong University team’s model and code can be found here (<https://github.com/ZhaoZhibin/Physionet2020model>). The model is a PyTorch implementation. The model allows commercial adaptation and re-use.
* The 20th Rank Singstad/Tronstad model and code can be found here (<https://www.kaggle.com/code/bjoernjostein/physionet-challenge-2020/notebook>). The associated research paper can be found here (<https://paperswithcode.com/paper/convolutional-neural-network-and-rule-based>). This model is well documented. This model allows commercial adaptation and re-use, with attribution.
* The 36th Rank CVC-University of Alberta team’s model and code can be found here (<https://github.com/awwong1/physionet-challenge-2020>. This model allows commercial adaptation and re-use.
* The unknown-rank Laussen Labs model and code can be found here (<https://github.com/Seb-Good/physionet-challenge-2020>). This model allows commercial adaptation and re-use.
* The unknown rank Universidad Politecnica de Madrid model and code can be found here (<https://github.com/pabloi09/physionet-challenge-2020>). This model allows commercial adaptation and re-use.
* The unknown rank UFMG Brazil model and code can be found here (<https://github.com/antonior92/physionet-12ecg-classification>). This model allows commercial adaptation and re-use.

These models may provide a better foundation for the final project model than the Conrad/Weimann model. We will continue to search for other published models with code.

**Challenge Model Review**

The Physionet/CinC 2020/2021 challenge 41 top ranked models were further analyzed by a team of Chinese and Singaporean researchers. The results of their analysis can be found here (<https://www.frontiersin.org/articles/10.3389/fphys.2021.811661/full>). This analysis provides significant guidance on the benefits and weaknesses of different teams approaches to solving the challenge. It should be required reading for all of the data scientists on the project team.

**Appendix 1: Outstanding Summer 2022 Questions:**

**Data – PTB-XL**

Can we split the PTB-XL 10 second result into multiple parts if each three second interval resembles each other?

How much deviation is normal between contiguous 3 second ECG readings?

Could we use other leads that resemble the main lead for additional observations? How much deviation would be permitted – can that be mathematically defined?

**Data – Astravision/Mexican Hospital**

When can we expect the first 350 observations to be labelled?

When can we expect the second set of 350 observations and related labelling?

What is the estimated percentage of each year’s observations that are likely to be in the two key classifications of ISCA and ISCI?

**Data – Georgia and other Kaggle 2020/2021 Atrial Fibrillation datasets**

The Georgia dataset contains 10,344 12-lead ECGs of 10 second length with a sampling frequency of 500 Hz. <https://moody-challenge.physionet.org/2020/>.

There are also two Chinese datasets containing a total of 10,000+ ECG readings. However, these readings are from a mix of 3, 6, and 12-lead ECG readings.

In the Georgia dataset and other datasets used in the Kaggle 2020/2021 challenge, which of the sub-classifications listed below are similar to the ISCI and ISCA subclasses from the PTB-XL dataset?

**(Continued on following page)**

**Kaggle 2020/2021 Diagnostic Codes Lists**

Please note that the total column includes observations from the two additional Chinese datasets that may or may not be usable in the project’s model training. Both the PTB-XL and Georgia dataset observations will be used.





