

Arrhythmia prediction using QRS complex detection with genetic data

Data Science Industry Project 1 - Group 28

The University of Melbourne

Janya Kavita Pandya

1291944

pandyaj@student.unimelb.edu.au

Soham Dighe

1219439

sdighe@student.unimelb.edu.au

Hieu Nguyen

940627

mnguyen9@student.unimelb.edu.au

Wesley Zhang

1240942

zhaoyangz1@student.unimelb.edu.au

Ryan Chow

1003648

lrchow@student.unimelb.edu.au

May 27, 2022

1 Introduction

Reports generated by the World Health Organization (WHO) on deaths suggest that cardiovascular diseases (CVD) are the leading cause of death globally, accounting for more than 30 percent of deaths yearly. Many of these CVDs called arrhythmias are caused when the heart beats with an irregular or abnormal rhythm. These arrhythmias can be detected by real-time observation of the ECG signals generated by the heart. Algorithm design for ECG signal analysis has been an area of research for more than 40 years now and a natural question arises do these arrhythmias occur in a certain pattern? Do factors such as demographics and genetics affect the occurrence of these arrhythmias? In this project, we are considering ECG and genetic data to detect the occurrence of 2 specific arrhythmias to understand the role of the genetics behind the occurrence of those arrhythmias.

The client we are currently working with is Applied Precision Medicine Pty Ltd which brings together consultancy, implementation, and training to deliver genetic tests into commercial clinical use as quoted from their website. We are directly in communication with Richard Rendell the Managing Director, Christopher Pendlebury the Chief Science Officer, and Yuhong Qin who does RD for medical devices. After being in contact with them we determined that they wanted to create a device that would be able to detect ECG signals and be able to determine if the user is having heart conditions in real-time. Thus we aim to use the data science techniques we've learned to produce machine learning algorithms and to support them in this endeavor.

One of the challenges of machine learning algorithms in medical applications and on our project specifically is the accuracy of the algorithm must be at least 99 percent accurate as any errors or inaccuracies may potentially lead to fatal accidents. Due to this, we need to verify our algorithms and run many tests to ensure the safety of the product. Another problem we encountered was the lack of computing resources available to us to compute the machine learning algorithms which would require long processing times. We have solved this by reaching out to Applied Precision Medicine for them to share their resources with us by giving us access to their servers. Another option we also considered was the Melbourne Research Cloud (MRC) Spartan provided to us by the university. The data available to us for this project was also an issue as currently, we are using Physionets open-source long-term atrial fibrillation (AF) data which is lacking as there are only data of 84 patients. Furthermore, the dataset does not include any genetic data which would make it even more difficult for us to achieve our goal of understanding the role of the genetics behind the occurrence of arrhythmias. Currently, we have decided to create some dummy genetics data as a temporary solution to this.

2 Related Work

Brief literature of more than 6 papers has been conducted. The review has been done keeping in mind the diversity and challenges in ECG signal generation, signal pre-processing, feature extraction, and QRS peak complex detection.

2.1 Signal Generation

Continuous and accurate generation of ECG signals is an integral part of QRS complex detection process, these are few of the methods that are used for signal generation are listed below

- Holter systems: these were used earlier but they give issues in prolonged monitoring of signals[2]
- The electrocardiogram (ECG) is an important tool in assessing cardiac health. Regular ECG measurements are desirable for long-term monitoring.[2]

2.2 Signal Pre-processing

QRS detection is not a simple peak-finding problem. Difficulties arise in QRS detection due to factors such as muscular noise, electrode artifacts, baseline drift, pathological signals, low signal-to-noise ratios, QRS-like artifacts, and tall P-and T-waves. Some of the prominent signal pre-processing techniques are listed below

- Band Pass Filtering: reduces high/low frequency noises, improves the signal-to-noise ratio and permits the use of lower thresholds.[1]
- Wavelet Decomposition: Dyadic, Discrete (DWT), Quadratic Spline (QSWT): Used along with filtering[3]
- Differentiation, integration, squaring: These techniques are based on the well-known Pan and Tompkins algorithm[1]

Pre-Processing	Performance	Complexity
Wavelet Transform [3]	High	High Time Consuming
Hilbert Transform [3]	Low, Several commonly presented artifacts could be mistaken as R peaks	High Normally used together with another technique
Differentiation, Integration, Squaring	Low Require complex detection technique	Low
Empirical Mode Decomposition[3]	Medium	High Used together with other techniques to improve performance

Table 1: Pre-processing Methods and their complexity

2.3 Signal Analysis:QRS complex Detection

The QRS complex is the most distinguishable pattern in the ECG signal, and therefore, its detection usually serves as the reference for locating other waves and segments. Features including R-R interval, QRS interval, P-R interval, heartbeats, and the amount of P Wave per QRS Complex are extracted using this method.[4]

- Pan Thompkins: In the PT algorithm, the ECG is band-pass filtered, differentiated, and squared, then moving window integration is performed, and two sets of adaptive thresholds are applied to both the filtered signal and the integrated signal for QRS detection.
- GR Algorithm: It is a recently-published QRS detector using a finite-state machine to update the detection threshold which is applied to the pre-processed ECG (derivative, then a moving average, then squaring)

2.4 Feature Extraction

Feature Extraction Once the output from the peak detection algorithm is obtained it is important to analyze the peaks and extract relevant features to detect arrhythmias

Five-step Analysis that could be applied to detect arrhythmias:

- Step 1: Is the rhythm regular or irregular? This can be determined by comparing the maximum duration of R-R interval and the minimum duration R-R interval.

- Step 2: Are QRS complexes narrow? The duration of the QRS complex should not exceed 0.1 seconds. But if QRS complex is too narrow, it will be considered as an abnormal event as well.
- Step 3: Are All P Waves Similar and Are PR Intervals Normal?

This needs to identify whether P Wave is absent first. Then, if a P Wave appears, the P-R interval will be measured. The P-R interval is normally 0.12-0.2 seconds.

- Step 4: Is the Rate Normal?

The normal range of heartbeats is 60-100 beats per minute. If the heart rate is less than 60 minutes, it is called bradycardia. If the heart rate is more than 100 beats per minute, it is tachycardia.

- Step 5: Do Waves and Complexes Proceed in Normal Sequence? Each P Wave should be followed by QRS Complex and T Wave. The normal sequence of they proceed indicates the normal sequence for each cardiac cycle

2.4.1 Q and S Waves Detection

After the position of R Wave detecting, Q Wave and S Wave can be identified with the morphological characteristics. Q and S peaks occur around R peak within 0.1 seconds. Therefore, the turning point that connects the baseline to the falling edge is the beginning of Q Wave; while the turning point that connects the rising edge to the baseline is the end of S Wave. Between the two points, there should be two minimum voltages which are Q peak and S peak respectively.

2.4.2 P and T Wave Detection

[5] For the normal ECG signal, P Wave, QRS Complex and T Wave appear alternately. The gap between P Wave and QRS Complex is usually no more than 0.2 seconds, therefore, the maximum voltage point within 0.2 seconds before Q Wave could be suggested as P peak. Once the P Wave is detected, the maximum voltage point between S peak and the next P peak shall be the T peak.[5]

2.5 Models

Recently, a new novel approach to detect arrhythmia has been to use an object detection deep learning model on ECG images. It relies on the increasingly fast and accurate object detection model from the field of computer vision. Most notable is the You-Only-Look-Once model or YOLO. And recent research paper so far suggests that the result is quite significant. The advantages are that it is fast and can achieve real-time detection. As well as input simply being raw ECG images, very processing

lite technique is needed, different from the traditional model used in ECG. Another potential model that can be explored is the ArrhythmiaNet which input is processed from ECG graphs, yet it is one of the state of the art model in Arrhythmia detection.

2.6 Performance Metrics

The QRS detection performance was evaluated as the QRS sensitivity ($Se = TP/(TP + FN)$) and positive predictivity ($P+ = TP/(TP+FP)$), where TP is the number of true positive detections, FP is the number of false-positive detections, and FN is the number of false-negative detections (missed QRS complexes). Some have also used Error Rate that is defined as $Er = FP+FN/TB*100$. [5]

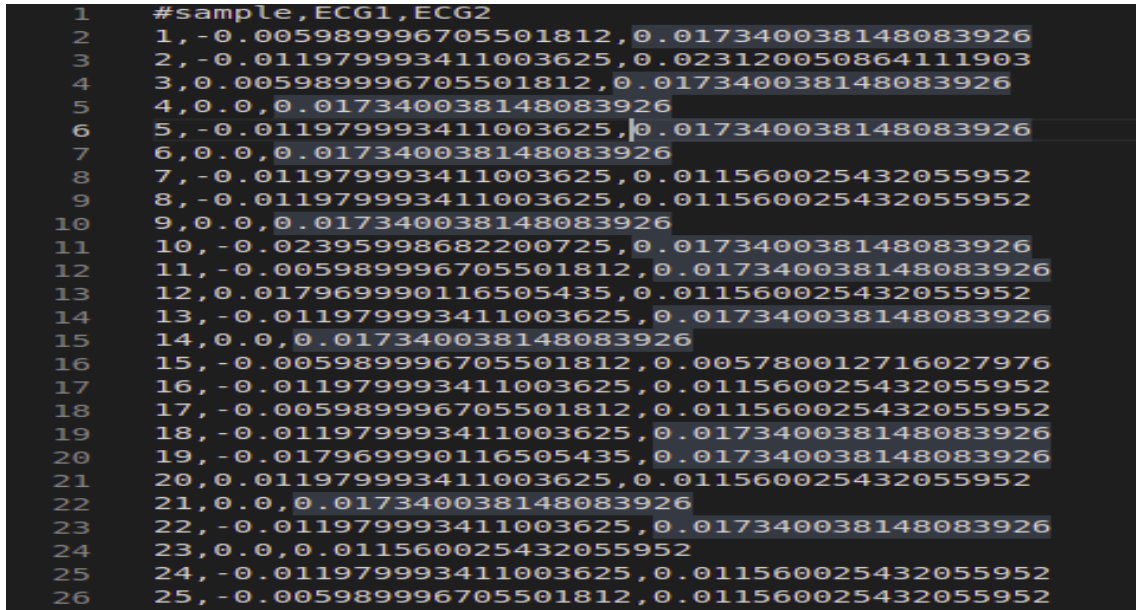
3 Data Analysis

For this project on our client's recommendation, we are using the PhysioNet dataset. This is an open-source data repository for biomedical research data, managed by the MIT Laboratory for Computational Physiology.

3.1 Data Description

We had information for 84 long-term ECG recordings of subjects with paroxysmal or sustained atrial fibrillation (AF). Each record contains two simultaneously recorded ECG signals digitized at 128 Hz with 12-bit resolution over a 20 mV range; record durations vary but are typically 24 to 25 hours. The 2 ECG values helped in stabilizing the ECG output.

The image below shows us what every data file looks like. Here as we can see the sample column has the timestream divided into tiny samples at a frequency of 128Hz which means every sample lasts for 1/128 seconds which is approximately 0.46 secs per sample. This takes the total number of samples for each patient to about 900,000 each.



```
1 #sample, ECG1, ECG2
2 1, -0.005989996705501812, 0.017340038148083926
3 2, -0.011979993411003625, 0.023120050864111903
4 3, 0.005989996705501812, 0.017340038148083926
5 4, 0.0, 0.017340038148083926
6 5, -0.011979993411003625, 0.017340038148083926
7 6, 0.0, 0.017340038148083926
8 7, -0.011979993411003625, 0.011560025432055952
9 8, -0.011979993411003625, 0.011560025432055952
10 9, 0.0, 0.017340038148083926
11 10, -0.02395998682200725, 0.017340038148083926
12 11, -0.005989996705501812, 0.017340038148083926
13 12, 0.017969990116505435, 0.011560025432055952
14 13, -0.011979993411003625, 0.017340038148083926
15 14, 0.0, 0.017340038148083926
16 15, -0.005989996705501812, 0.005780012716027976
17 16, -0.011979993411003625, 0.011560025432055952
18 17, -0.005989996705501812, 0.011560025432055952
19 18, -0.011979993411003625, 0.017340038148083926
20 19, -0.017969990116505435, 0.017340038148083926
21 20, 0.011979993411003625, 0.011560025432055952
22 21, 0.0, 0.017340038148083926
23 22, -0.011979993411003625, 0.017340038148083926
24 23, 0.0, 0.011560025432055952
25 24, -0.011979993411003625, 0.011560025432055952
26 25, -0.005989996705501812, 0.011560025432055952
```

Figure 1: ECG Mve vs time data snapshot

These records were divided into 4 different files for each patient. The files were data, header, annotated and qrs file. The data file contained data of the 2 ECG streams for the patient with their values given in mV. The Header file had background information on the conditions in which data was collected like number of channels, base time, signal length, etc. In the remaining 2 files, we had atr annotations were obtained by manual review of the output of an automated ECG analysis system and qrs annotations

were produced by an automated QRS detector.

As we can see the annotated file has the fields corresponding to the sample number, symbols denoting what type of beat it is, which channel is working(in this case we have just 1 channel), its subtype, and auxiliary note which is the actual annotation. The image below shows what each symbol represents in the given dataset in addition to these symbols the auxiliary notes have descriptive symbols as well like AFIB, VT, etc. In these annotation files, all detected beats (including occasional ventricular ectopic beats) are labeled N, detected artifacts are labeled '—', and AF terminations are labeled T.

```

1  sample,symbol,subtype,chan,num,aux_note
2  83607,+,0,0,0,(N
3  83628,N,0,0,0,
4  83736,N,0,0,0,
5  83844,N,0,0,0,
6  83953,N,0,0,0,
7  84062,N,0,0,0,
8  84172,N,0,0,0,
9  84282,N,0,0,0,
10 84392,N,0,0,0,
11 84503,N,0,0,0,
12 84611,N,0,0,0,
13 84721,N,0,0,0,
14 84832,N,0,0,0,
15 84942,N,0,0,0,
16 85054,N,0,0,0,
17 85164,N,0,0,0,
18 85274,N,0,0,0,

```

Figure 2: Annotated sample

label_store	symbol	description
0	0	Not an actual annotation
1	1	N Normal beat
2	2	L Left bundle branch block beat
3	3	R Right bundle branch block beat
4	4	a Aberrated atrial premature beat
5	5	V Premature ventricular contraction
6	6	F Fusion of ventricular and normal beat
7	7	J Nodal (junctional) premature beat
8	8	A Atrial premature contraction
9	9	S Premature or ectopic supraventricular beat
10	10	E Ventricular escape beat
11	11	j Nodal (junctional) escape beat
12	12	/ Paced beat
13	13	Q Unclassifiable beat
14	14	~ Signal quality change
16	16	Isolated QRS-like artifact
18	18	s ST change
19	19	T T-wave change
20	20	* Systole
21	21	D Diastole
22	22	" Comment annotation
23	23	= Measurement annotation
24	24	p P-wave peak
25	25	B Left or right bundle branch block
26	26	^ Non-conducted pacer spike
27	27	t T-wave peak
28	28	+ Rhythm change
29	29	u U-wave peak
30	30	? Learning
31	31	! Ventricular flutter wave
32	32	[Start of ventricular flutter/fibrillation
33	33] End of ventricular flutter/fibrillation
34	34	e Atrial escape beat
35	35	n Supraventricular escape beat
36	36	@ Link to external data (aux_note contains URL)
37	37	x Non-conducted P-wave (blocked APB)
38	38	f Fusion of paced and normal beat
39	39	(Waveform onset
40	40) Waveform end
41	41	r R-on-T premature ventricular contraction

Figure 3: Annotation list

Now, let's take a closer look at what an AFIB rhythm wave would look like. Here the waves around

the 600 sample point follow the typical QRS peaks with R waves rising so far above the Q and S wave and as will be discussed in the following sections the distance between consecutive R peaks is said to be a good indicator of Atrial fibrillation.

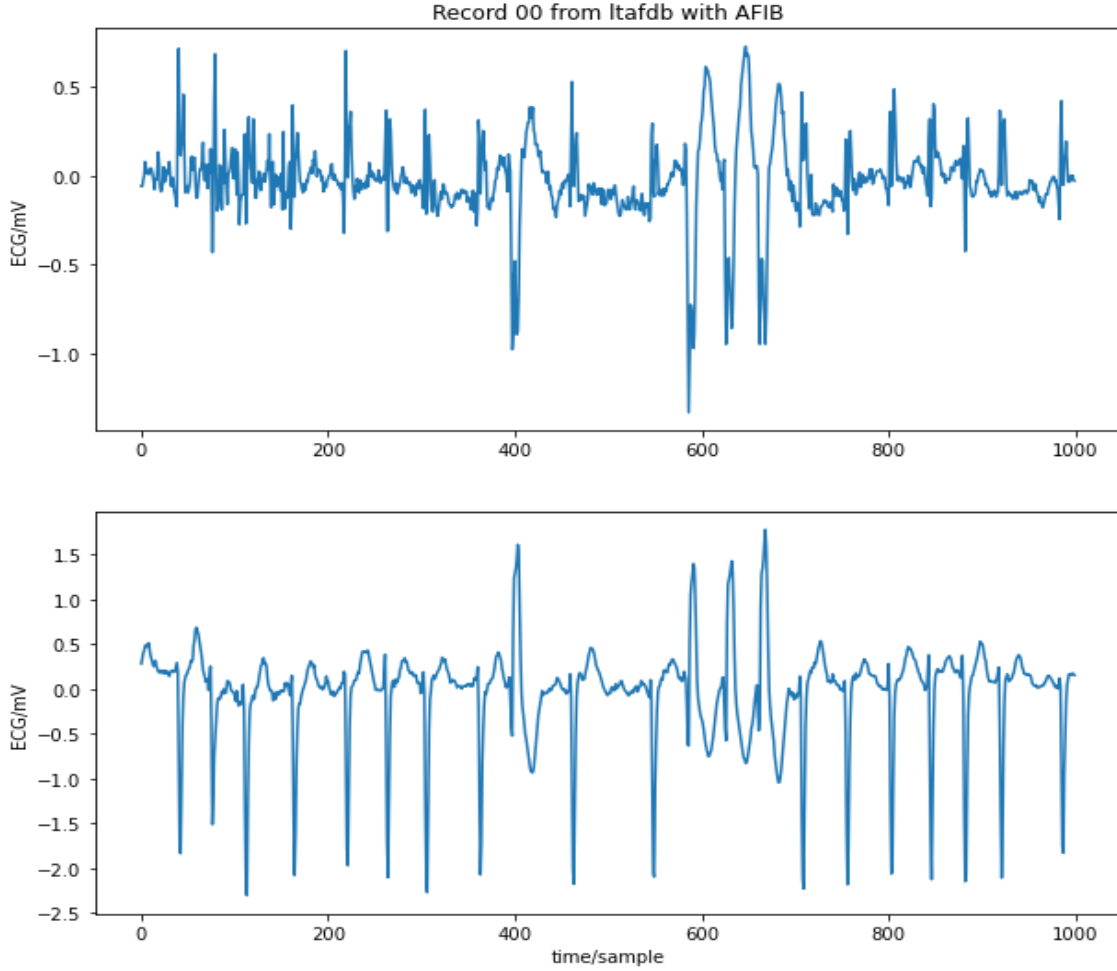


Figure 4: ECG

From the above graph we can see that the two different readings give similar but not the same graph. The baseline value (dn) and root mean square (rms) value of the cubed ECG signal, $m[n]$, can be used to calculate the threshold value for detecting upward and downward R-peaks. The threshold voltage was calculated using the formula :

- For upward R-peaks region: $dn + rms / 2$
- For downward R-peaks region: $dn3 \times rms$
- Frequency density (Fd) = frequency / width of interval
- $\sigma = \max(Fd1, Fd2, \dots, Fdn1, Fdn)$
- mean = interval width of sigma

- $\text{time-limit} = ((\text{upper limit of mean}) + (\text{lower limit of mean})) / 2$, $\text{dn} = \text{time-limit}/4$

3.2 Data Pre-Processing

Since the original data was huge, we have taken 1 record as a sample and used it to implement Pan thomkins method for qrs peak detection, The diagram below shows the pre-processing flow of the algorithm: The raw ECG signal is firstly band-pass filtered, then it is differentiated, and squared,

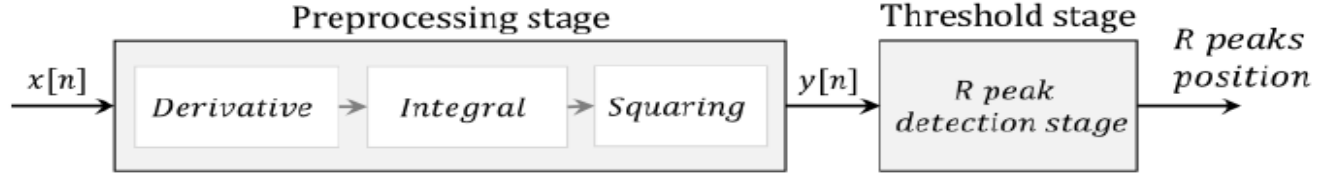


Figure 5: Pan-Thomkins Algorithm

following that moving window integration is performed, and two sets of adaptive thresholds are applied to both the filtered signal and the integrated signal for QRS detection.

3.2.1 Snapshot of Sample Data

The snapshot has 2 columns, timestamp and ECG measurement in Mv, and it has a total of 934 time samples

	timestamp	ecg_measurement
0	96813044	2.233627
1	96816964	2.179863
2	96820892	2.150538
3	96824812	2.160313
4	96828732	2.155425
...
929	100456148	2.438905
930	100460108	2.394917
931	100464068	2.375367
932	100468028	2.365591
933	100471988	2.306940

934 rows × 2 columns

Figure 6: Data Snapshot

3.2.2 Pre-processing stages

- Raw ECG Signal

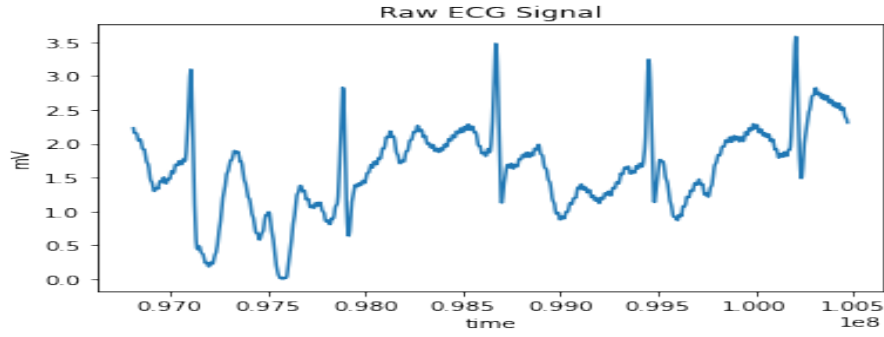


Figure 7: Raw ECG Signal

- Bandpass filtering: The figure below shows the outputs after low pass and high pass filtering of the raw signal

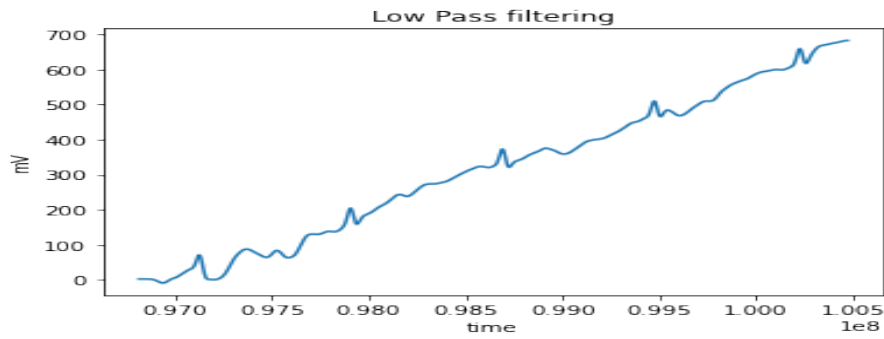


Figure 8: Low pass filtering on ECG Signal

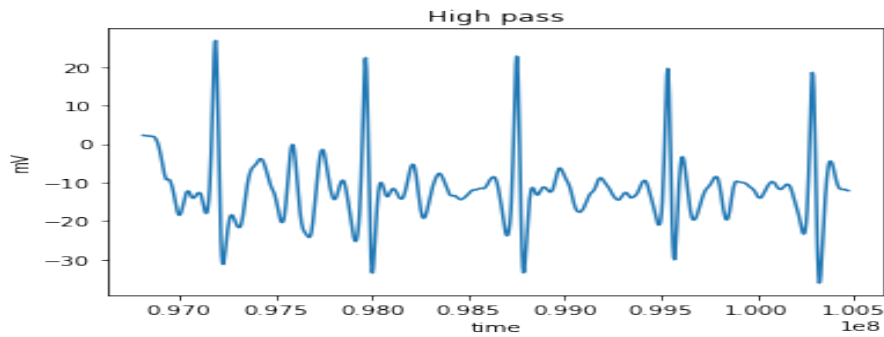


Figure 9: High Pass filtering on ECG Signal

- The Filtered signal was squared and derivative was taken and then integrated to generate the clean pre-processed signal that can be fed to qrs detector

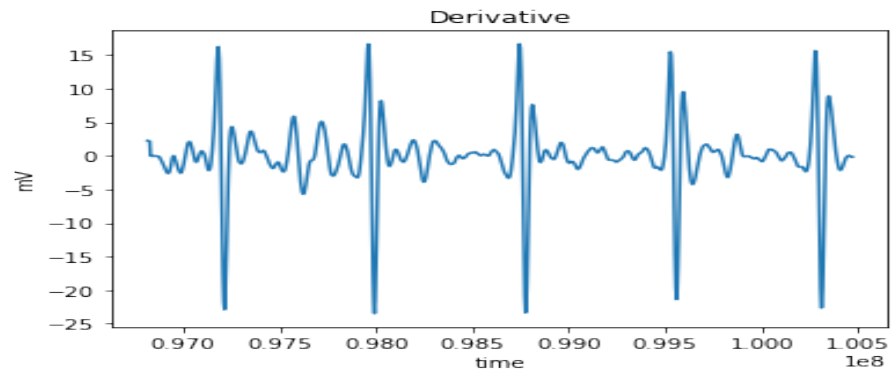


Figure 10: Derivative of ECG Signal

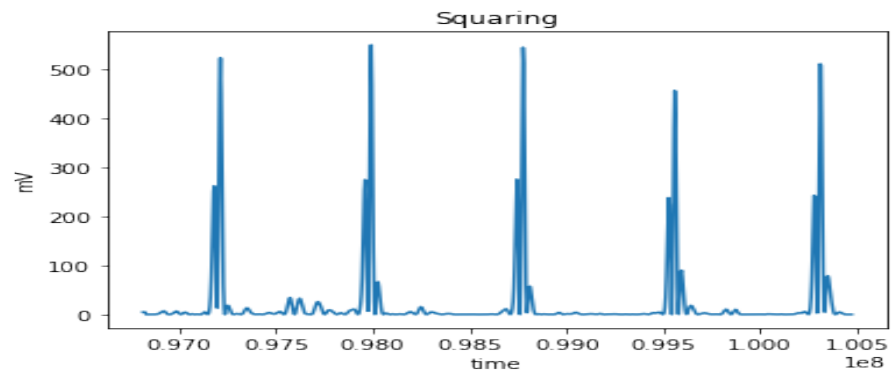


Figure 11: Squaring of ECG Signal

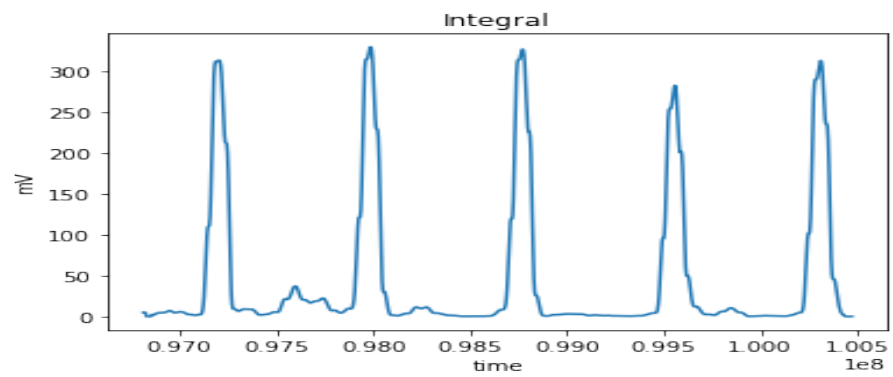


Figure 12: Output after integration of ECG Signal

3.3 QRS detection and correction

The GQRS algorithm attempts to locate QRS complexes in an ECG signal in the specified record. The detector algorithm is new and as yet unpublished. The output of gQRS is an annotation file (with annotator name qrs) in which all detected beats are labeled normal ("N"). The *subtyp*, *chan*, and *num* fields of each annotation respectively indicate the detection pass (0 or 1) during which the QRS complex was detected, the signal number on which it was detected, and the peak amplitude of the detector's matched filter during the QRS complex.

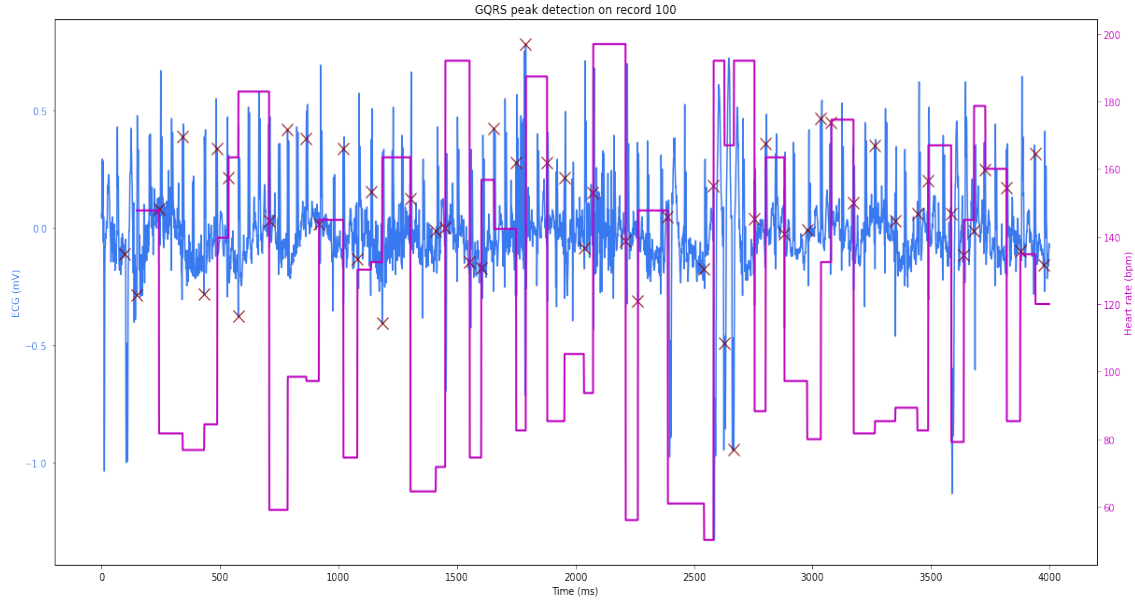


Figure 13: GQRS peak detection for given sample

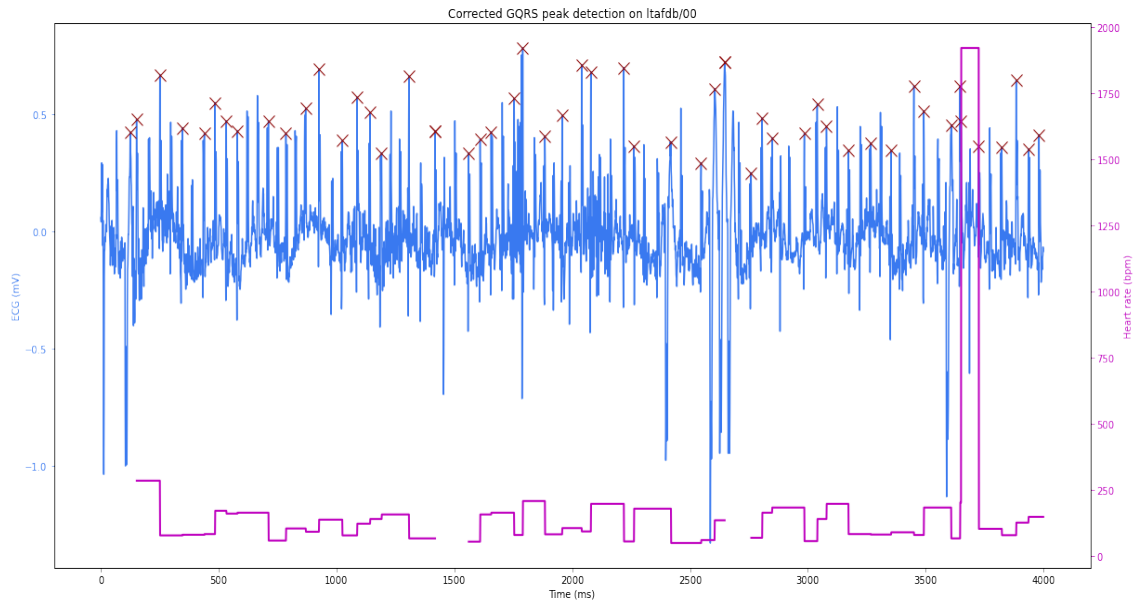


Figure 14: GQRS peak correction

As we see in the figure above after Peak correction we can see that only the part of the wave with maximum irregularity in the graph is marked which gives us a clue as to where the Atrial Fibrillation has taken place.

4 Proposal for the next Semester

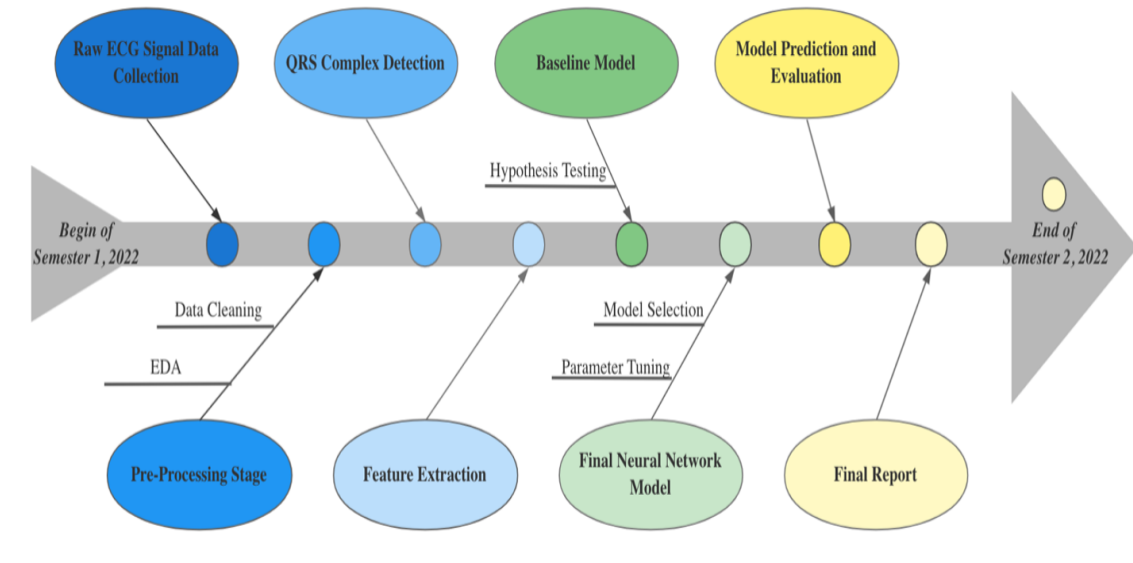


Figure 15: Plan for the next coming semester

4.1 How do you propose to answer the client's question

It's a good idea to start by assessing the client's managerial goals. Agree on what KPIs are most important for the client's goals and how they are currently evolving. Compare and contrast different ways to tackle the objective and implement step by step processes on how the team proceed with the project. The next stage is to think about how the outcome of the project will help the client. What result from the analysis would be considered success and useful. Exploratory data analysis questions are essential for guiding the team through the process and focusing on significant insights. This framework can assist the team delving further into the more detailed insights stakeholders seek. The figure above depicts the major project milestones that the team will be completing throughout the semester.

4.2 How do you plan to use the data (in terms of models, algorithms, and other technique) for analysis and prediction. What techniques from the Related work section will you use.

The team plans to use the neural network model as the final model for prediction. The realization of this idea is divided into the following steps. The first step is data preprocessing. Before the second semester, the team will further compress data, reduce noise, and segment the data. The second

is feature extraction, GR algorithm mentioned in related work can be used as the method of QRS detector and object detection using image recognition by the Yolo model. The third is modeling, where traditional machine learning methods may first be used to verify possible assumptions. But because traditional machine learning must use annotated electrocardiograms, and deep learning methods can automatically learn features, neural networks are used as predictive models. Compared with numerous neural network models, CNN and LSTM are the two most widely used methods for this situation. This is because most of the current studies on ECG focus on ECG classification and disease recognition, and CNN can be used to learn features and classify them. However, since there is no connection between each CNN node and the ECG presentation is time-series, LSTM is also used for ECG diagnosis. Our final plan is to combine CNN and LSTM to make a hybrid neural network structure model.

4.3 How might you adapt them

The team proposes to implement a hybrid neural network model. CNN and LSTM are combined to form a multi-layer network. The structure of the network includes the convolutional layer, max pooling layer, convolution layer, and other alternate methods are passed to the two-way LSTM layer to learn the remote context in the two input directions. The time dimension can be reduced by half by adding max pooling. Finally, specific parameter tuning methods are tried and implemented to improve the prediction results.

4.4 How will you evaluate success

The success of the model is measured by evaluation metrics. Four possible evaluation indicators are listed below:

- $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$
- $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$
- $\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$
- $\text{Positive predictive rate} = \text{TP} / (\text{TP} + \text{FP})$

The higher the four indicators are associated with better accuracy of the model. The higher the indicators are, the stronger the predictive classification ability is.

4.5 What hurdles are you likely to face

It's always a good idea to have more data. Instead of depending on assumptions and weak correlations. More useful data leads to more accurate and better models.

In data science projects, often analysts may not have the option of increasing the quantity of the training data. However, it is best to ask for more information if feasible. Working with constrained data sets will be less painful as a result of this.

5 Timeline

The proposed timeline is that the team will continue to perform data pre-processing in the month of July and move on to QRS detection and feature extraction for the entire month of August. For the month of September, the team will finalize the model and tackle parameter tuning for better model accuracy. Then, the team will spend the first half of October to work on the final report and then plans to negotiate and improve the details with the clients in the latter half of that month. By the end of October, the team will be presenting the finalized findings to both the client and the University.

The role of each group member are listed below.

- Soham: data preprocessing with more techniques
- Janya and Minh: QRS detection and feature extraction
- Ryan and Wesley: Build up models and fine tuning
- All: final report and presentation

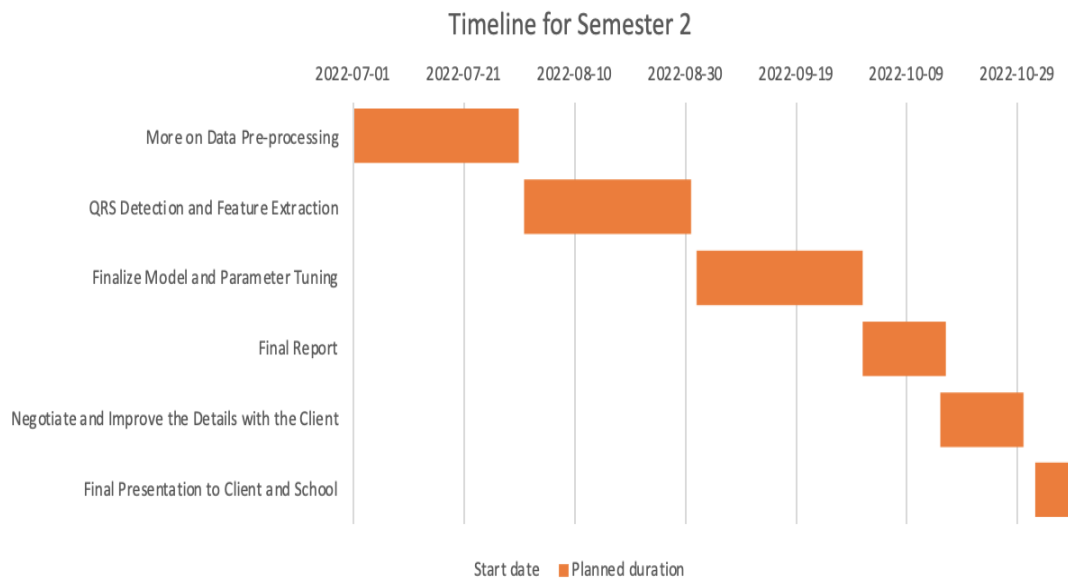


Figure 16: Timeline for the next coming semester

References

- [1] Pan J, Tompkins WJ. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng.* 1985 Mar;32(3):230-6. doi: 10.1109/TBME.1985.325532. PMID: 3997178.
- [2] Saadi, D. B., Tanev, G., Flintrup, M., Osmanagic, A., Egstrup, K., Hoppe, K., Jennum, P., Jeppesen, J. L., Iversen, H. K., Sorensen, H. B. (2015). Automatic Real-Time Embedded QRS Complex Detection for a Novel Patch-Type Electrocardiogram Recorder. *IEEE journal of translational engineering in health and medicine*, 3, 1900112.
- [3] R. Gutiérrez-Rivas, J. J. García, W. P. Marnane and Á. Hernández, "Novel Real-Time Low-Complexity QRS Complex Detector Based on Adaptive Thresholding," in *IEEE Sensors Journal*, vol. 15, no. 10, pp. 6036-6043, Oct. 2015, doi: 10.1109/JSEN.2015.2450773.
- [4] Tanushree Sharma Kamalesh K. Sharma (2016) QRS Complex Detection in ECG Signals Using the Synchrosqueezed Wavelet Transform, *IETE Journal of Research*, 62:6, 885-892, DOI: 10.1080/03772063.2016.1221744
- [5] Khamis H, Weiss R, Xie Y, Chang CW, Lovell NH, Redmond SJ. QRS Detection Algorithm for Telehealth Electrocardiogram Recordings. *IEEE Trans Biomed Eng.* 2016 Jul;63(7):1377-88. doi: 10.1109/TBME.2016.2549060. Epub 2016 Mar 31. PMID: 27046889.
- [6] Hannun, A.Y., Rajpurkar, P., Haghpanahi, M. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 25, 65–69 (2019). <https://doi.org/10.1038/s41591-018-0268-3>
- [7] Rahul, J., Sora, M. Sharma, L.D. Exploratory data analysis based efficient QRS-complex detection technique with minimal computational load. *Phys Eng Sci Med* 43, 1049–1067 (2020). <https://doi.org/10.1007/s13246-020-00906y>