Analysis of Cerebromicrovascular Disease in Elderly Patients With Diabetes

Matthew Lam
Toronto Metropolitan University
matthew1.lam@ryerson.ca

Aemen Mohsin
Toronto Metropolitan University
aemen.mohsin@ryerson.ca

Lazar Zivkovic
Toronto Metropolitan University
lazar.zivkovic@ryerson.ca

I. Abstract

***Abstract*— Diabetes is one of the most widespread diseases, affecting hundreds of millions of people worldwide. Potentially fatal and often difficult to live with, diabetes has a wide range of effects on the human body and can have devastating impact on lives. Among many of the symptoms diabetes causes, one of the more life threatening is the effect it has on the human heart. Diabetes can cause damage to the heart muscle, blood vessels, and can lead to heart failure. Due to these effects, diabetes can lead to the development of Cerebral Microvascular/Small Vascular Disease (CSVD). The ability to predict the presence of diabetes based on the effect that CSVD has on ECG could provide valuable information in regards to the impact that diabetes has on vascular health and could provide an alternate method of disease prediction for diabetes. The method proposed in this report uses the "Cerebromicrovascular Disease in Elderly with Diabetes" dataset of ECG signals found on the Physionet database. The process follows the Pan-Tompkins method of ECG pre-processing and notch filtration for power line interference removal. Following this, ECG feature extraction was performed and these features were used in three different Machine Learning algorithms. The Logistic Regression model achieved a validation accuracy of 77.8% and testing accuracy of 66.7%, the highest among the three models that were trained on the Physionet dataset. The results are sub-optimal and future testing of this correlation should seek to incorporate a larger dataset and extract a wider variety of features for more thorough classification.**

***Keywords*— Cerebromicrovascular, diabetes, electrocardiogram, machine learning, pre-processing, ECG features**

II. Introduction and Background:

Electrocardiogram (ECG) signals can be used for a variety of applications. Many of these applications include using ECG signals for monitoring the heart and diagnosing arrhythmias, heart disease, heart attack, and cardiomyopia [1]. Diabetes Mellitus (DM), which is a disease that affects the body's response to glucose, is prevalent in around 200 000 Canadians [2]. Although ECG signals cannot be used to directly identify whether an individual has diabetes or not, individuals with diabetes often experience ECG abnormalities due to the strong correlation with heart disease [3]. More specifically, cerebromicrovascular disease, which is a small vessel disease caused by changes in capillaries, arteries and veins can occur, especially in the elderly population [3]. DM is known to alter blood barriers, which affects the metabolism and microcirculatory regulation [4]. Subsequently, this affects the displayed ECG signals of the patients.

For the detection of abnormalities in ECG signals, they need to be extracted and processed initially. This processing can include many techniques to remove noise and artifacts from the signal, including the use of notch filters and the Pan-Tompkins algorithm. These preprocessing stages are necessary to prepare the ECG signal for further analysis through feature extraction.

The main features that can be extracted and analyzed for both normal ECG signals and signals from diabetic individuals include heart rate, which is also the R-R interval, the standard deviation of this R-R interval, the mean and standard deviation of the S-T interval, and lastly the mean and standard deviation of the Q-T interval. The R-R interval is the time between two R waves in the QRS complex [5]. This measures the heart rate [5]. The S-T interval measures the ventricular repolarization and is used to detect myocardial ischaemia and infarction [6]. Lastly, the Q-T interval is used to measure the time between ventricular depolarization and repolarization [7]. Prolonged Q-T waves represent arrhythmias,

while short Q-T waves can indicate an increased risk of atrial and ventricular fibrillation [7].

After the extraction of the features, machine learning algorithms such as the decision tree model, linear discriminant model, and logistic regression model can be used to determine the validity and accuracy of the dataset for both control and diabetic individuals.

III. Methods:

A. *Signal Pre-processing*

Dataset:

The dataset used in the project was the CVES database titled "Cerebromicrovascular Disease in Elderly with Diabetes." This dataset was published on physionet on August 5th, 2022, by Novak and Quispe [4]. The dataset included a total of 120 participants aged 55 to 75, where 60 individuals were diabetic and 60 were control [4]. Their ECG signals were collected over a 24 hour period and sampled at 1000 Hz [4]. 15 control signals were extracted along with 18 diabetic signals to create a total of 33 signals for analysis.

Method 1:

The first method used in processing the signals included notch filtering. Power-line Interference (PLI) occurs due to electromagnetic interference in the power supply [8]. It is an interference occurring at around 50-60 Hz [8]. To counteract the noise caused by powerline interference, a notch filter was used. A notch filter is created by placing zeros on the unit circle at the spot corresponding to the point of noise being removed; in this case, a zero was placed at 60 Hz. By adding this zero, the filter is able to remove noise occurring at the 60 Hz power-line interference. The notch filter was calculated using the following equations (Equation 1, 2, 3):

$$\Theta = 2\pi\frac{f0}{fs} \quad (1)$$

$$H(z) = (1 - z1 * z^{-1})(1 - z2 * z^{-1}) \quad (2)$$

The final transfer function of the notch filter was determined to be:

$$H(z) = 1 - 1.86\, z^{-1} + z^{-2} \quad (3)$$

Method 2:

The second method used for signal processing and filtering included using multiple signals in series in a process called the Pan-Tompkins algorithm. ECG signals are riddled with noise and artifacts including both high-frequency and low-frequency noise, muscle noise, powerline interference, and baseline wander noise [9]. The Pan-Tompkins algorithm accounts for this by including a bandpass filter to remove noise. The bandpass filter has an approximate passband of 5-15 Hz [10]. To create this bandpass filter, a low pass filter is used to remove the high frequency noise including EMG noise and T wave noise, while a high-pass filter is used to remove low frequency noise such as baseline wander [10]. The equation for the low pass and high pass filter is defined below (Equation 4, 5):

$$HLP(z) = \left(\frac{1-z^{-6}}{1-z^{-1}}\right)^2 \quad (4)$$

$$HHP(z) = z^{-16} - \frac{1}{32}\left(\frac{1-z^{-32}}{1-z^{-1}}\right) \quad (5)$$

A differentiator is then applied to the signal to differentiate between the high peaks of the QRS complex from the low peaks of the P and T waves [10]. The equation of the differentiator is defined below (Equation 6):

$$H(z) = \frac{1}{8}(2 + z^{-1} - z^{-3} - z^{-4}) \quad (6)$$

This signal is then squared to only display the positive values of the signal and to amplify the peaks of the QRS complex, while de-amplifying those of the P and T waves [10]. The equation for the squaring operator is defined below (Equation 7):

$$y(n) = [x(n)]^2 \quad (7)$$

Lastly, the signal is passed through a moving window integrator with a window length of 150 samples with a sampling frequency of 1000 Hz [10]. Each sample is approximately 150 ms in width. The equation for the moving window integrator is listed below (Equation 8):

$$H(z) = \frac{1}{150}(1 + z^{-1} + z^{-2} + \ldots + z^{-149}) \quad (8)$$

This output can now be used for peak detection and feature analysis due to the visible QRS peaks.

B. Feature Analysis

Features extraction
One method to detection: Find Peaks
Two methods feature extraction: 1) Find QRS point
2) Find ST points

Equations Used:

Heart Rate:

$$\text{HR (bpm)} = \frac{60}{RR\ interval} * 1000 \quad (9)$$

Mean:

$$\overline{X} = \frac{\sum X}{N} \quad (10)$$

Standard deviation:

$$\text{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}} \quad (11)$$

The next step is feature extraction where important portions of the dataset are extracted to be classified. ECG has very distinct features in the time domain shown at point PQRST in the image above. ECG toolbox was used to find all five points' locations on the ECG. This is followed by finding the mean and standard deviation intervals between these five points. This is a lot of features that may hinder machine learning classification. Therefore, only the 6 important features were sent to the classification which is 1) heart rate, 2) mean interval from S-T, 3) std interval from S-T, 4) mean interval Q-T, 5) std interval Q-T and 6) std from R-R. These features are chosen due to the fact that these features are most affected in diabetic patients in ECG. [11]

Method 1: Find the PQRST points

Using the moving window average plot, peak detection was applied to find all points of R in the ECG. Since the S point is located right after the R peak, the ECG toolbox searching for the minimum point is located between the R point and 0.1*sampling frequency [12]. Since the Q point is before the R peak, the program searches for the minimum point between the R point and 0.08* sampling frequency. P and T peaks were detected by searching half of the R-R distance before the R peak but after the S peak. The Heart rate, mean R-R, and standard deviation of R-R were calculated using the equation above.

Method 2: Finding the PQRST intervals

Since the locations of PQRST peaks were found in approach 1, the intervals were found also using the ECG toolbox. This ECG toolbox created by Rohan Sanghavi uses and finds the closest R points and pairs them up with the closest P, Q, S, and T points [12]. These points need to be close to R peaks, if it is further away than the period length, it is voided. The intervals were calculated by finding the difference between the points. The mean and standard deviation was calculated using the equation above.

C. Machine Learning

The machine learning algorithms that were selected for classification were selected for their variety and diversity of method. Three models were used in order to try and observe the general behavior of the dataset across multiple different classification methods and to attempt to find a model among many that is reliable and robust for the classification of diabetes on the basis of ECG.

Method 1:

The decision tree model works by assessing each feature sequentially and splitting the “tree” down binary paths along each feature node. The predicted classification is dependent on the overall predisposition for a certain class to fall along each binary grouping along the tree pathway until it reaches the final feature [13]. All of the previous nodal splits indicate what the classification should be predicted as and the model predicts accordingly.

Method 2:

The second model that was used was the linear discriminant model. This model functions by seeking to establish a linear combination of features that can quantify a prediction as belonging on either side of a linear divisor. The model does this by assigning the discriminant line along the path that sections off the most labels into their appropriate class and weighing the value of the features in accordance with their ability to discriminate accurately [14]. The sum of the squares of all weighing factors is equal to 1 and the model uses these factors to determine an overall score which determines the side of the classification line a subject should reside in. Features with higher weighting

factors will sway a subject's score to a higher degree and thus have a stronger impact on the model's prediction.

Method 3:

The final model that was tested was the logistic regression model. The logistic regression model functions similarly to the linear discriminant model except that instead of dividing the classifications along a linear discriminating boundary, the logistic regression models attempts to fit the most accurate 'S' shaped predictor line to the data. The class labels are placed on one axis while the feature at hand is on the other. A logit line is then fitted on various points along the scale with the intent of getting as many labels to fall correctly in place along the line as possible. The shifting of the line is done in order to map the most labels correctly as possible, that line is then used as the linear regression line for its respective feature. The degree to which a feature's regression line may accurately map the labels also determines that feature's weight with respect to the other features in the dataset [15]. The higher weighted features exhibit a stronger influence on the prediction of the model.

IV. Results:

A. Signal Pre-processing

a. Control signal:

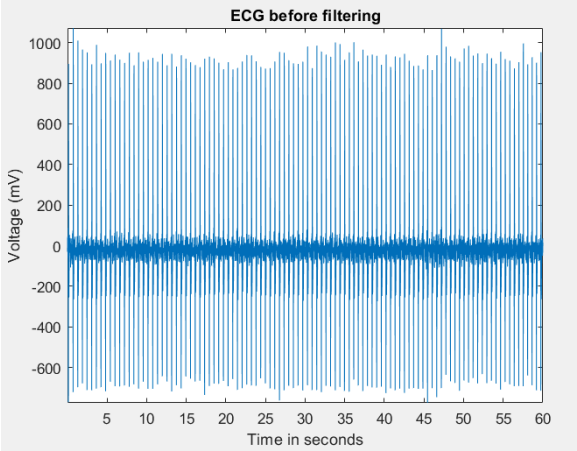

Fig. 1: ECG signal without filtering for subject 4, control.

Method 1:

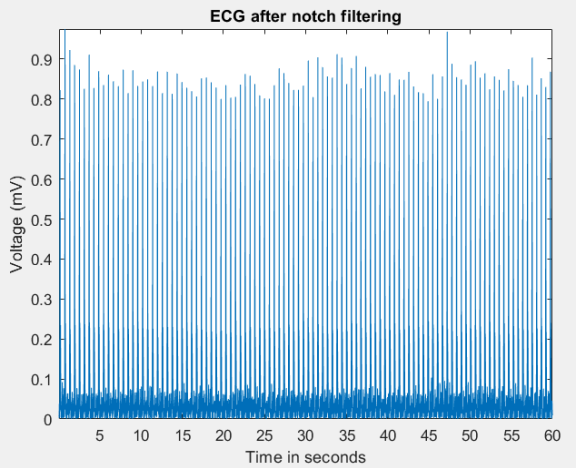

Fig. 2: ECG signal after notch filtering for subject 4, control.

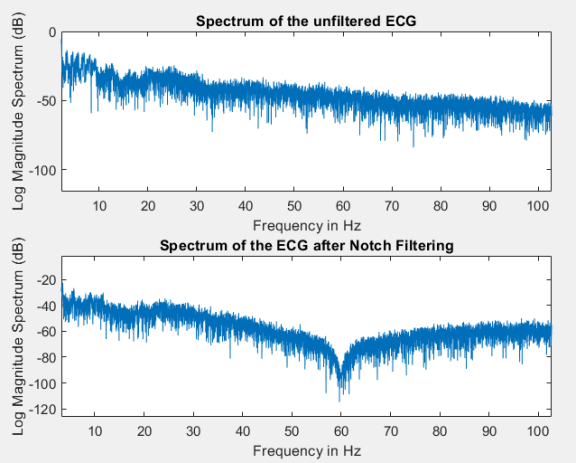

Fig. 3: Frequency domain of ECG signal before and after notch filtering for subject 4, control.

Method 2:

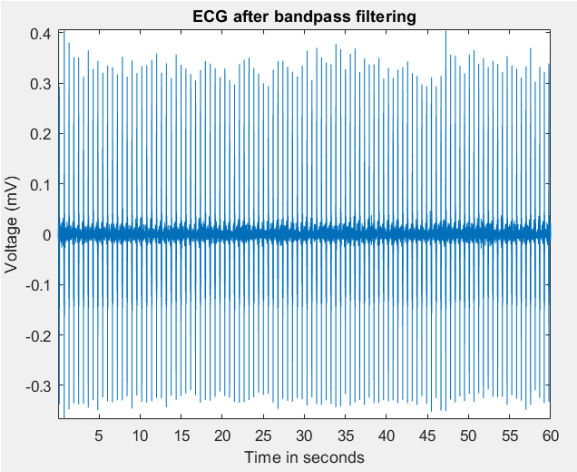

Fig. 4: ECG signal after bandpass filtering for subject 4, control.

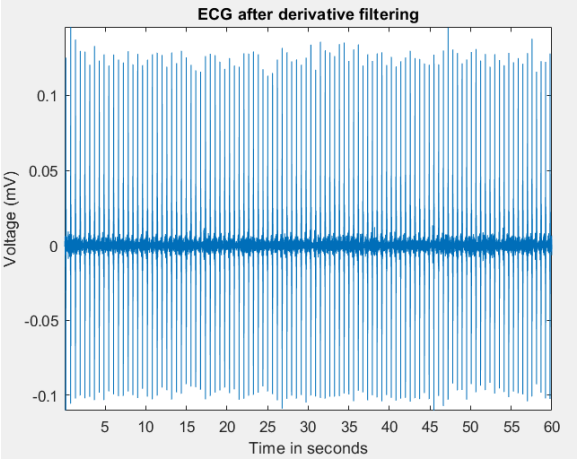

Fig. 5: ECG signal after derivative filtering for subject 4, control.

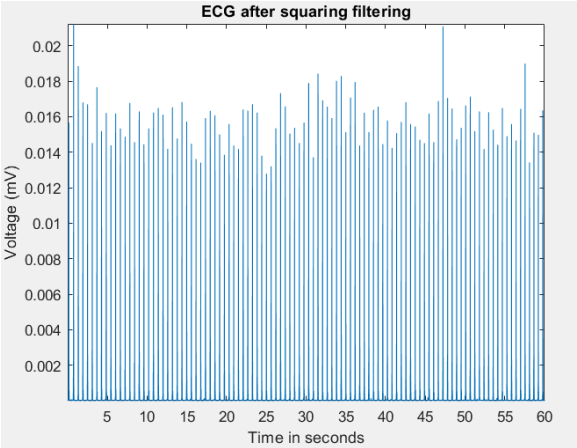

Fig. 6: ECG signal after squaring operator for subject 4, control.

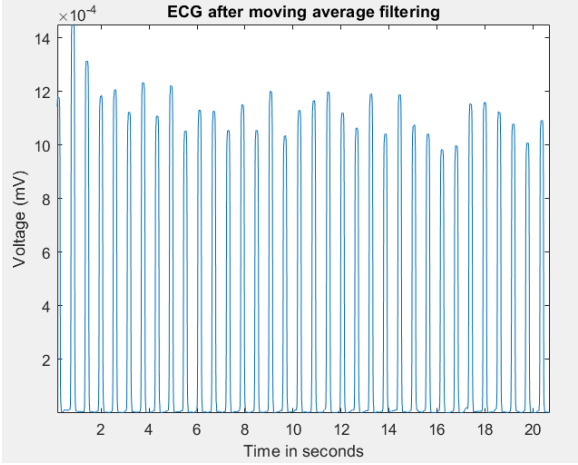

Fig. 7: ECG signal after moving window integrator filtering for subject 4, control.

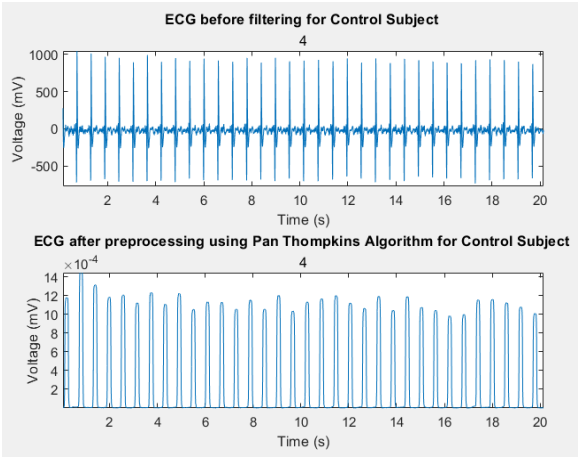

Fig. 8: Comparison of ECG signal before and after notch and Pan Tompkins filtering for subject 4, control.

b. Diabetic Signal

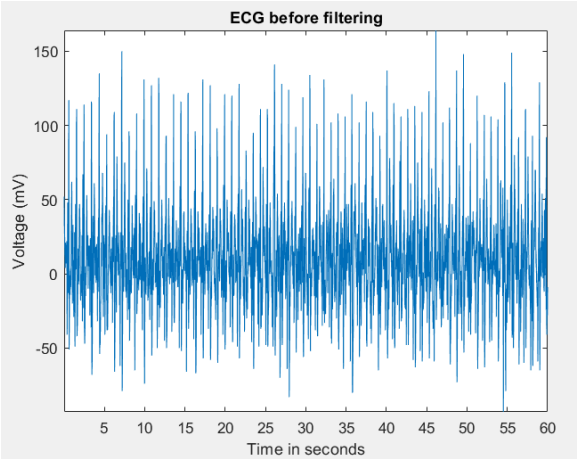

Fig. 9: ECG signal without filtering for subject 20, diabetic.

Method 1:

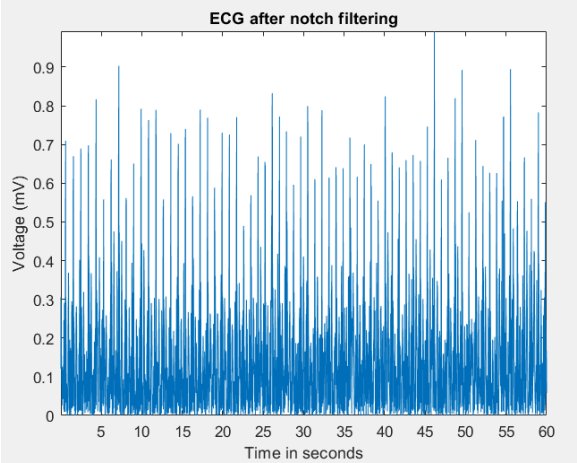

Fig. 10: ECG signal after notch filtering for subject 20, diabetic.

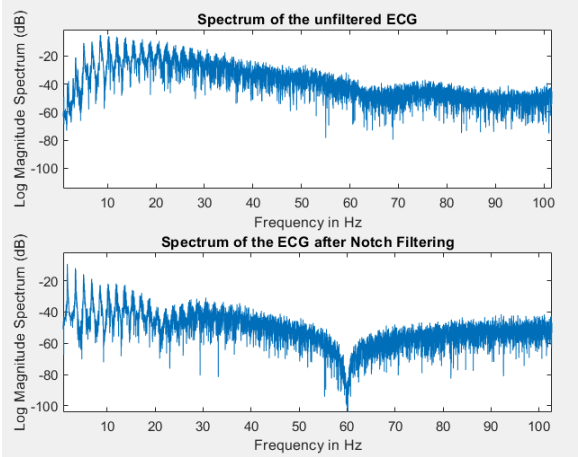

Fig. 11: Frequency domain of ECG signal before and after notch filtering for subject 20, diabetic.

Method 2:

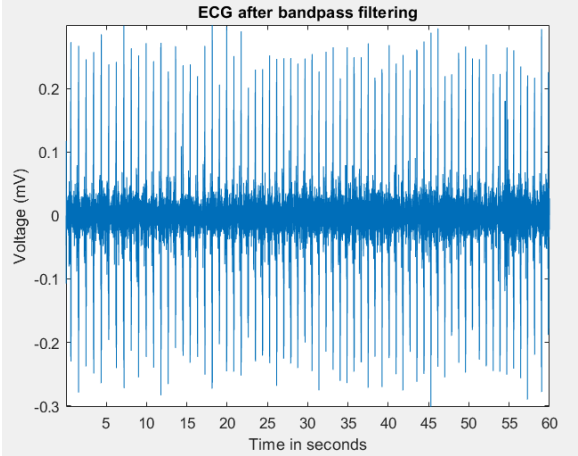

Fig. 12: ECG signal after bandpass filtering for subject 20, diabetic.

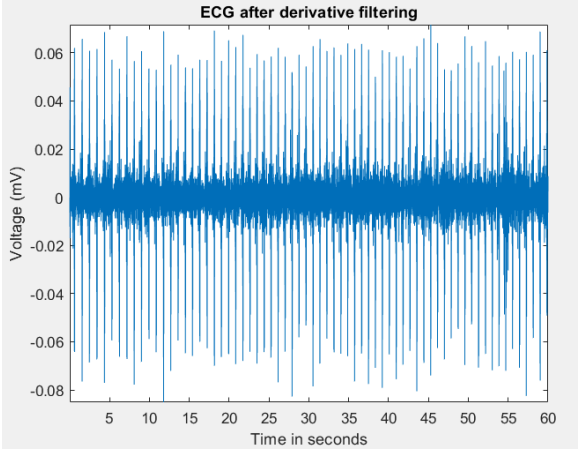

Fig. 13: ECG signal after derivative filtering for subject 20, diabetic.

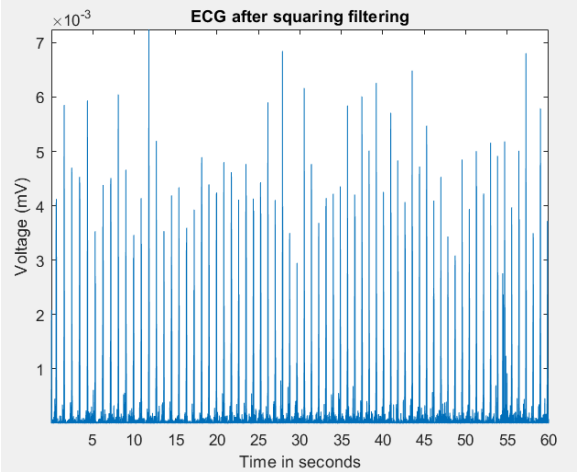

Fig. 14: ECG signal after squaring operator for subject 20, diabetic.

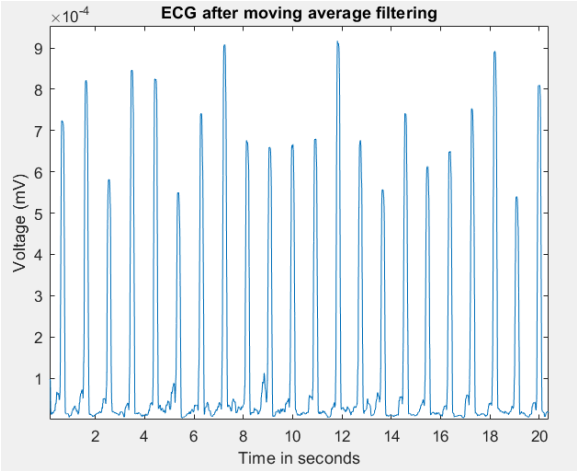

Fig. 15: ECG signal after moving window integrator filtering for subject 20, diabetic.

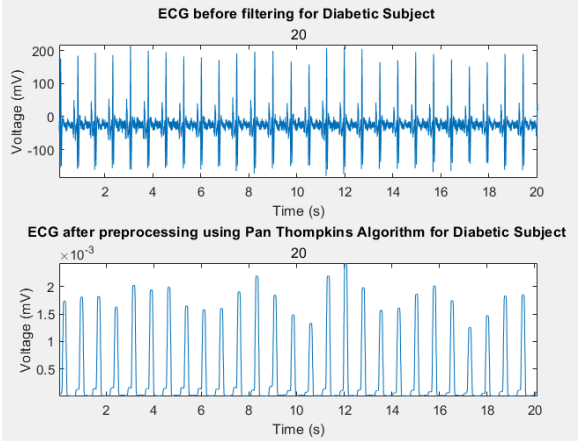

Fig. 16: Comparison of ECG signal before and after notch and Pan Tompkins filtering for subject 20, diabetic.

B. Feature Analysis

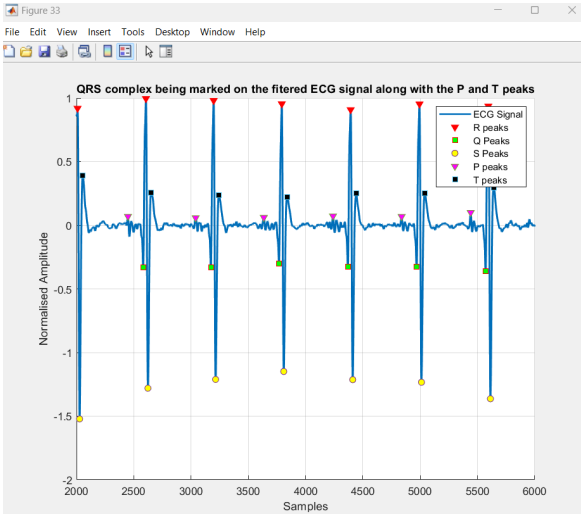

Fig. 17: Shows the all points PQRST in the ECG signal using peak detection

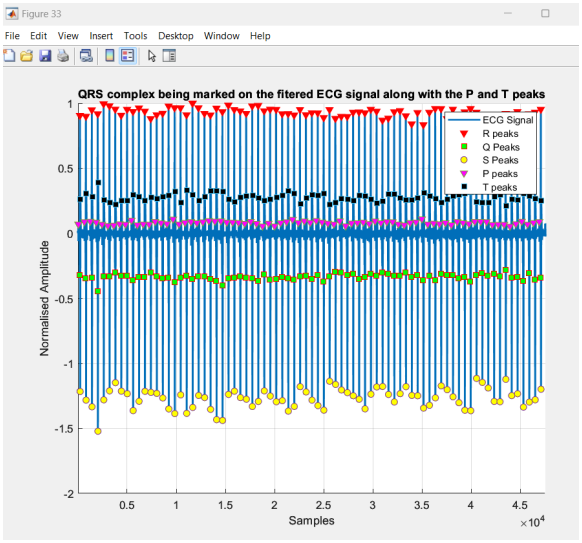

Fig. 18: Zoomed out all points PQRST in the ECG signal for a whole minute

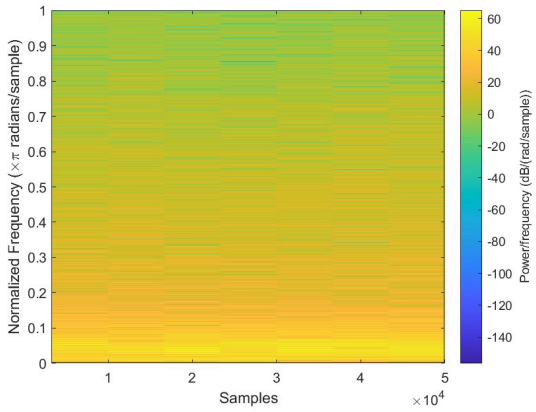

Fig. 19: A Spectrogram of a sample ECG signal

Table 1: Shows the sample features of 4 trials: where 0 is control and 1 is diabetes

#	Label	HR (bpm)	meanST (ms)	std_ST (ms)	RR_mean (ms)	mean_QT (ms)	RR_std (ms)	std_QT (ms)
1	0	87.2596617	34.9130435	3.8184503	687.6029412	72.8115942	17.2030366	10.2645526
2	0	106.972548	59.4166667	7.93225584	560.8915663	105.6190476	8.99595037	8.04341347
3	1	84.812919	168	82.9954362	707.4393939	234.5522388	4.93068504	82.1429433
4	1	101.795378	108.3125	77.2240852	589.4177215	161.3625	17.1206934	76.3289423

C. Machine Learning

There were 3 machine learning algorithms that were used for the accuracy testing of the extracted features. These algorithms were the Decision Tree model, the Linear Discriminant analysis, and Logistic Regression. A 5 fold cross validation was used instead of a 10 fold standard due to the limited data pool. The testing group was 20% of the data and was used to test the models after validation. The validation and testing accuracy results are visible in Table 1.

Table 1: Validation and test accuracy for Tree, Linear Discriminant, and Logistic Regression Models.

Model	Validation %	Testing %
Decision Tree	70.4	66.7
Linear Discriminant	63.0	50
Logistic Regression	77.8	66.7

The confusion matrices for each of the three models are visible in Figures 20, 21, and 22 below. The 0 label designates the control subjects and the 1 label designates the diabetic subjects.

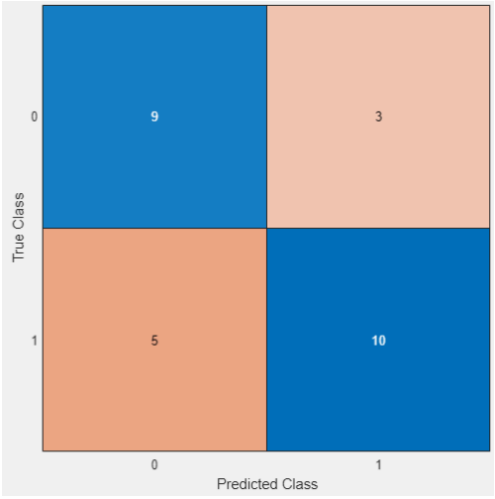

Fig. 20: Decision Tree confusion matrix.

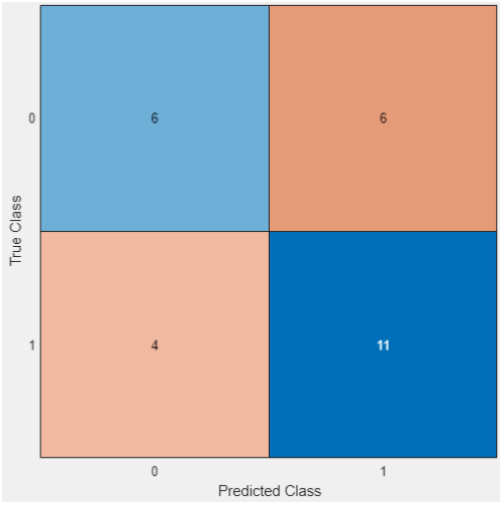

Fig. 21: Linear Discriminant confusion matrix.

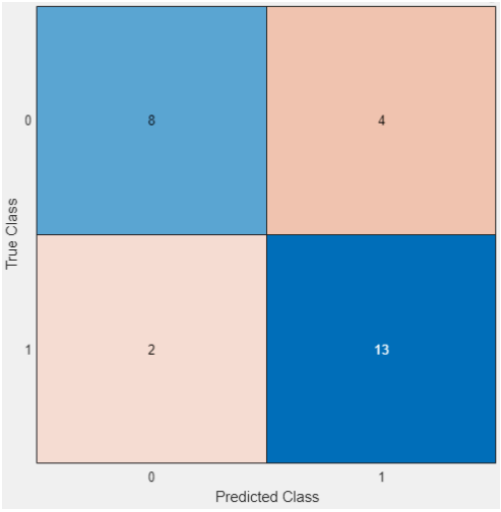

Fig. 22: Logistic Regression confusion matrix

IV. Critical Analysis and Discussion:

A. Signal Pre-processing

The notch filtering at 60 Hz was effective in removing the power-line interference. Although a significant difference was not observed in the time domain, when plotted in the frequency domain using a fast fourier transform, the signal displayed a significant noise reduction at the 60 Hz location. For the frequency spectrum response of the control signal (Fig. 3) before and after notch filtering, a clear difference was observed at the 60 Hz location, since visible powerline noise was detected in the original signal. However, when analyzing the frequency response of the diabetic signal (Fig. 11), the original signal did not experience visible power-line interference due to the overall noise of the signal. The notch filter; however, was effective in removing the noise at the 60 Hz location. When comparing the results to other sources, ECG signals typically experienced a power-line interference at 50 or 60 Hz [16]. The dataset used in this report did not experience as much noise at 50 Hz, so the notch filter was applied at 60 Hz.

After the application of the notch filter, the Pan-Tompkins algorithm was applied. The algorithm applied three filters including bandpass, differentiator and moving window integrator, along with the squaring operation. The bandpass filter, which was a cascade of the high-pass and low-pass filters, displayed the signal after the filtration of both high-frequency and low-frequency noise present in both the control (Fig. 4) and diabetic signals (Fig. 12). The control signal experienced a greater suppression of low frequency noise compared to the diabetic signal. Afterwards, the derivative filter and squaring operator amplified the peaks and analyzed the positive signals. The control signal (Fig. 5, 6) for both operators displayed a signal with definitive peaks and no low-frequency noise. The diabetic signal (Fig. 13, 14); however, still experienced low frequency noise in both these operators. Lastly, the moving average filter was applied, which used a window size of 150 samples. The number of samples were increased compared to applications, since many used a sample size of 30 windows for a sampling frequency of 200 Hz [10]. This would yield in a sample size of approximately 150 ms. The dataset used in this report included a sampling frequency of 1000 Hz so more samples were used to achieve the same results. 150 samples were used which also yielded a window size of approximately 150 ms. The larger sampling frequency was not downsampled so the sample size was decided accordingly.

Each step of the algorithm displayed results that were expected when compared with those of accredited sources.

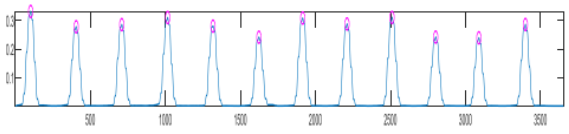

Fig. 23: ECG signal after application of Pan Tompkins algorithm from Journal of Physics [10].

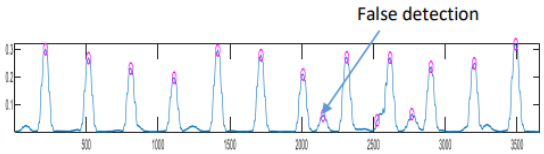

Fig. 24: ECG signal after application of Pan Tompkins algorithm with false peak detection from Journal of Physics [10].

The ECG signal of the control after the Pan Tompkins algorithm (Fig. 8) resembled the signal processed in the reference with normal peak detection (Fig. 23). This was due to minimal noise present in the control signal. On the contrary, due to a larger amount of noise present in the diabetic signal (Fig. 16), the signal experienced false peak detection as shown in the reference (Fig. 24). Even after the false QRS peak detection, the QRS complex for the signals were still very visible and could be used for feature analysis. Due to this, both the notch filtering and the Pan-Tompkins algorithm was successful in preprocessing the signal for feature extraction.

B. Feature Analysis

Peak detections on the moving average were very affected in finding the R waves. This was because the moving average filter made the peaks obvious. The peak detection was very effective in finding the QS points in an ECG. This is because minimum peaks were clear and the R peaks were used as a reference.

The peak detection to find T S was not as affected. Some T waves were hidden by the high fluctuations of the signal while some of the P waves had very large amplitudes and could have been misinterpreted as R waves. Some of the P or T points were unable to be detected and as a result, could not be used. To counteract this, a large amount of ECG pulses was calculated and the mean was taken from each of the signals to account for some of the points not detected.

Six important features were sent to the classification:

1) Heart Rate / RR-mean and standard deviation R-R

Research shows that a normal elder aged 45 to 70 has an average heart rate between 60 to 100 bpm [17]. This is similar to the data table extracted where the maximum and minimum heart rate gathered is 106.9 bpm to 58.47 bpm. Patients with diabetes have an increased heart rate reaching more than 86 bpm. [18] For every 12-bpm increase in heart rate, the incidence of type 2 diabetes rises by 13.3% [18]. This is similar to the results found in the table where the range increases from 126 bpm to 65 bpm. Since the heart rate is calculated from the RR-mean, the treads are the same.

A reduction in heart rate variability is linked to type 2 diabetes (HRV) [19]. Therefore, it is assumed that the standard deviation should be lowered in diabetes then the control. However, this is not the case, there was not much of a trend between diabetes and control. This could be due to the small dataset. Only 33 patients were measured which makes it difficult to determine a relationship.

2) mean interval from S-T / std interval from S-T

According to Shlomo Stern, patients with diabetes have high changes in S-T intervals [11]. This conclusion is supported by our results where the standard deviation of S-T intervals is way higher than the standard deviation of the S-T intervals on the constant patient. Furthermore, the results show that the mean S-T is higher in patients with diabetes.

A normal patient has an ST mean internal is 80-120 ms which is higher than our results [20]. This is because our results calculated the difference between the peaks of S and T. This differs from the interval calculations in research where the intervals include the point T ends. Therefore, our results will be smaller than the intervals found in the research.

4) mean interval Q-T/ std interval Q-T

According to Shlomo Stern, patients with diabetes have long Q-T intervals [20]. The results reflect this where mean Q-T is higher in comparison to normal patients. Furthermore, our results show that the diabetic QT intervals have a high standard deviation when compared to the control patients. Therefore, there is a large change in variation in diabetic patients.

As mentioned before, the interval found in the research will be higher than the results in the table. This is because our findings estimated the separation between the peaks of S and T. This is different from interval computations used in research, which take into account the ends of point T.

C. Machine Learning

The decision tree method garnered an acceptable validation accuracy of 70.4%, the second best validation accuracy among the three models. It

also had a testing accuracy of 66.7%, similar to its validation accuracy albeit slightly worse. Although the accuracy is high enough to observe a general trend, ~70% accuracy is generally considered to be a poor result and it leaves too much potential for misclassification, especially in a clinical setting. Due to the simplistic nature of the binary nodal classification, the decision tree model is prone to overfitting, a problem wherein the predictability of a model is too closely fitted to the particular data used to train it, rendering it incapable of accurately predicting for a set whose features fall too far disconnected from the boundaries of the trained model [13]. The rigidity of the decision tree model causes it to be heavily influenced by outliers and skews, particularly when dealing with a smaller dataset as is the case here.

The linear discriminant model was the second model that was tested and it was notably the worst performing model with a validation accuracy of 63.0% and a testing accuracy of only 50%. Linear discriminant analysis works heavily on the assumption that there will be some features that have a clear gaussian distribution which allows for the class division along a linear path. This assumption struggles to account for large populations of outliers, especially in very sporadic datasets. These limitations are likely the reason why both the validation and testing results of the linear discriminant model were the worst performing of the three models. Smaller datasets especially are particularly prone to outliers which is precisely what appears to have affected the linear discriminant accuracy in this case.

The logistic regression model yielded the highest validation accuracy of the three and had the same testing accuracy as the decision tree model. Logistic regression suffers from similar overfitting issues that the decision tree and linear discriminant models suffer from in that they cannot deal well with the higher outlier content present in smaller datasets. Although the validation accuracy was impressively high at 77.8%, the testing accuracy reflects a similar weakness to the decision tree, an accuracy of only 66.7%. As mentioned previously, the testing group was 20% of the overall dataset however, there were only 33 subjects, 15 control and 18 diabetic. Such a small sample size is very problematic when attempting to predict binary labels as even 1 overly skewed outlier would shift the accuracy downwards by ~17%. As such it is still hopeful that the logistic regression validation accuracy was as high as it was, potential for better testing accuracy may be realized with a larger dataset and more patients to train and test on.

V. Conclusion:

Overall, we were able to classify diabetes patients with an accuracy of 77.8%. First, the ECG signal was preprocessed using a notch filter and the Pan-Tompkins Algorithm to remove any abnormalities. Secondly, feature extraction was performed where peak detection was implemented to find the R peaks. An ECG tool box was used to find the PQRST point in which the heart rate, which is also the R-R interval, its standard deviation, the mean and standard deviation of the S-T interval, and finally its mean and standard deviation of the Q-T interval were extracted. Finally, the features were imputed into a machine learning classification where three models were used. These were the Decision Tree model, Linear Discriminant model, and Logistic Regression Model.

VI. References:

[1] “Electrocardiogram (ECG),” NHS choices, 07-Sep-2021. [Online]. Available: https://www.nhs.uk/conditions/electrocardiogram/#:~:text=An%20ECG%20is%20often%20used,dizziness%20and%20shortness%20of%20breath. [Accessed: 28-Nov-2022].

[2] P. H. A. of Canada, “Government of Canada,” *Canada.ca*, 17-Aug-2020. [Online]. Available: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/diabetes-canada-highlights-chronic-disease-surveillance-system.html. [Accessed: 28-Nov-2022].

[3] C.-Ł. J. D. E. Z. A. P. B; “Cerebral Small Vessel Disease: A Review,” Advances in clinical and experimental medicine : official organ Wroclaw Medical University. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33768739/#:~:text=Cerebral%20small%20vessel%20disease%20(CSVD,over%2080%20years%20of%20age. [Accessed: 28-Nov-2022].

[4] V. Novak and R. Quispe, “Cerebromicrovascular disease in elderly with diabetes,” Cerebromicrovascular Disease in Elderly with Diabetes v1.0.1, 05-Aug-2022. [Online]. Available: https://doi.org/10.13026/00bm-0x81. [Accessed: 28-Nov-2022].

[5] “RR interval,” *RR Interval - an overview | ScienceDirect Topics*. [Online]. Available: https://www.sciencedirect.com/topics/nursing-and-health-professions/rr-interval#:~:text=The%20RR%20i

nterval%2C%20the%20time,as%20well%20as%20autonomic%20influences. [Accessed: 28-Nov-2022].

[6] E. Burns, R. Buttner, and E. B. and R. Buttner, “The St Segment,” *Life in the Fast Lane • LITFL*, 16-Mar-2022. [Online]. Available: https://litfl.com/st-segment-ecg-library/. [Accessed: 28-Nov-2022].

[7] E. Burns, R. Buttner, and E. B. and R. Buttner, “QT interval,” Life in the Fast Lane • LITFL, 14-May-2022. [Online]. Available: https://litfl.com/qt-interval-ecg-library/. [Accessed: 28-Nov-2022].

[8] S. Mian Qaisar, “Baseline wander and power-line interference elimination of ECG signals using efficient signal-piloted filtering,” Healthcare technology letters, 03-Sep-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7494370/#:~:text=Power%2Dline%20interference%20(PLI),0.7%5D%20Hz%20%5B2%5D. [Accessed: 28-Nov-2022].

[9] M. Elgendi, “Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases,” PloS one, 16-Sep-2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3774726/. [Accessed: 28-Nov-2022].

[10] M. A. Z. Fariha1, R. Ikeura2, S. Hayakawa2, and S. Tsutsumi2, “IOPscience,” Journal of Physics: Conference Series, 01-Jun-2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1532/1/012022#:~:text=The%20Pan%2DTompkins%20Algorithm%20is,quality%20clinical%20ECG%20signal%20data. [Accessed: 28-Nov-2022].

[11] S. Stern, Shlomo Stern From The Hebrew University of Jerusalem, S. Sclarowsky, Samuel Sclarowsky From The Hebrew University of Jerusalem, and C. to S. Stern, “The ECG in diabetes mellitus,” Circulation, 20-Oct-2009. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/circulationaha.109.897496. [Accessed: 28-Nov-2022].

[12] “ECG Signal pqrst peak detection toolbox,” *MathWorks*. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/73850-ecg-signal-pqrst-peak-detection-toolbox?s_tid=srchtitle_ROHAN+SANGHAVI+ecg+_3. [Accessed: 28-Nov-2022].

[13] Y.-Y. Song and Y. Lu, “Decision tree methods: Applications for classification and prediction,” Shanghai archives of psychiatry, 25-Apr-2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/. [Accessed: 28-Nov-2022].

[14] A. Araveeporn, “Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity,” International Journal of Mathematics and Mathematical Sciences, 06-Sep-2022. [Online]. Available: https://www.hindawi.com/journals/ijmms/2022/7829795/. [Accessed: 28-Nov-2022].

[15] S. Sperandei, “Understanding logistic regression analysis,” Biochemia medica, 15-Feb-2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/#:~:text=Logistic%20regression%20is%20used%20to,the%20observed%20event%20of%20interest. [Accessed: 28-Nov-2022].

[16] A. links open overlay panelLeifSörnmoPabloLaguna, LeifSörnmo, and PabloLaguna, “EEG signal processing,” *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, 02-Sep-2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780120375529500039. [Accessed: 28-Nov-2022].

[17] A. J. Rodriguez, “Normal heart rate for elderly adults,” Griswold Home Care, 07-Sep-2021. [Online]. Available: https://www.griswoldhomecare.com/blog/2021/september/normal-heart-rate-for-elderly-adults/#:~:text=A%20normal%20heart%20rate%20for,heart%20rate%20with%20regular%20exercise. [Accessed: 28-Nov-2022].

[18] Organization, Healio, and ImageObject, “Higher heart rates related to type 2 diabetes risk in normal-weight adults,” Healio, 18-Feb-2019. [Online]. Available: https://www.healio.com/news/endocrinology/20190215/higher-heart-rates-related-to-type-2-diabetes-risk-in-normalweight-adults#:~:text=The%20researchers%20found%20that%20heart,%3D%201.24%3B%2095%25%20CI%2C. [Accessed: 28-Nov-2022].

[19] H. Media, “Heart rate variability in type 2 diabetes,” *Endocrinology Advisor*, 23-Sep-2022. [Online]. Available: https://www.endocrinologyadvisor.com/home/topics/

diabetes/type-2-diabetes/heart-rate-variability-in-type-2-diabetes/. [Accessed: 28-Nov-2022].

[20] M. D. Lawrence Rosenthal, “Normal electrocardiography (ECG) intervals,” *Normal Electrocardiography Intervals*, 16-Jul-2021. [Online]. Available: https://emedicine.medscape.com/article/2172196-overview. [Accessed: 28-Nov-2022].

[21] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, “A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods,” *Computational Intelligence and Neuroscience*, 15-Apr-2022. [Online]. Available: https://www.hindawi.com/journals/cin/2022/3820360/. [Accessed: 28-Nov-2022].

V. Appendix

Table 1: Contribution by each student.

Student	**Contribution**
A. Aemen (33%) Signature: AM	- Preprocessing - Introduction - Methods - Results - Discussion
B. Matthew (33%) Signature: ML	- Feature Analysis - Methods - Results - Discussion - Conclusion
C. Lazar (33%) Signature: LZ	- Machine Learning - Abstract - Methods - Results - Discussion