Identifying Atrial Fibrillation with Stepping Windows

Ari Singh[1], Taran Puvvala[2], Shravan Selvavel[3], and Venkata Vadlamudi[4]

[1]Early College at Guilford

[2]Ardrey Kell High School

[3]Providence Day School

[4]Green Level High School

Dr. C. Chen, Dr. Y. Wang

January 20, 2023

Abstract

An atrial fibrillation (AFib) patient not actively undergoing an arrhythmic episode is hard to diagnose accurately. This hurdle necessitates real-time detection methods that can be used to constantly monitor someone who may be at risk. This project focused on detecting AFib in near real-time using electrocardiogram (ECG) readings and a novel stepping window method. First, the distance between R-waves, or the RR-intervals, in the MIT-BIH AFib dataset's 23 ECG recordings was calculated. Multiple features explored in prior research were chosen for the classification of the RR-Intervals. Features were calculated for each "stepping window"—a subset of 4 RR intervals—while factoring in the features of previous windows. The motivation behind this approach was to allow the model to make more accurate predictions given a very small subset of RR intervals. We attained a maximum accuracy of approximately 94% using a CatBoost classifier and a selected set of features. Thus, we conclude that the stepping window method of feature calculation can be used to classify small sets of RR intervals as AFib. In the future, we hope to explore a wider range of classifiers (including neural networks) and features—including different methods of calculating features in our stepping window system.

Identifying Atrial Fibrillation with Stepping Windows

Introduction

Atrial fibrillation (AFib) is the most common type of arrhythmia, in which the heart contracts irregularly. This irregular rhythm could be slower than normal, faster than normal, or, in some cases, a combination of the two. Normally, a cell known as the sinoatrial (SA) node starts and regulates the heartbeat. The signal emitted by the SA node causes the heart to contract, allowing blood to be sent to the body. During AFib however, the heart will receive signals from areas other than just the SA node, causing the muscle fibers in the atria to contract irregularly, or fibrillate, hence the name "atrial fibrillation"(Cedars Sinai). Common symptoms of AFib include lightheadedness, chest pain, shortness of breath, and extreme fatigue. Without proper treatment, AFib can lead to life-threatening health complications such as an increased chance of stroke and heart failure (NHLBI). According to the Centers for Disease Control and Prevention (CDC), it is estimated that by the year 2030, around 12.1 million people will have atrial fibrillation. As an indirect result of AFib, upwards of 454,000 hospitalizations occur annually, leading to over 150,000 deaths each year (CDC).

Currently, AFib detection procedures use an electrocardiogram (ECG) reading, which measures the heart's electrical activity using two leads placed on the chest, and graphs the voltage over time (Yeo). After taking this measurement, medical professionals examine the ECG to look for signs of AFib. These tests look at heart rate variability (HRV) to analyze the fluctuations in the heart rate of a particular subject. ECG recordings of AFib have a higher HRV, while recordings of normal sinus rhythm have a lower HRV (Thuraisingham).

Recognizing AFib early on in the patient is crucial to prevent any severe, adverse effects. In order to accomplish this, we explored the usage of a machine learning model to recognize AFib through data supplied in near real-time. With the ability to monitor their heart rhythm conveniently, users of this technology will be notified if arrhythmias are detected within seconds of their occurrence. Currently, there are wearable technology options in the market, such as the Apple Watch, that will check the heart rhythm of the wearer about every two hours (Yeo). AFib is extremely difficult to detect if the subject is not undergoing AFib at the time of the ECG test, increasing the chances of an undetected potential diagnosis. Therefore, it is critical for AFib to be detected in near real-time allowing a patient to be notified if they are experiencing atrial fibrillation at any time and can seek further treatment.

Data Overview

We procured the MIT-BIH Atrial Fibrillation database (Goldberger et al.), which resulted from a collaboration between the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (previously known as Beth Israel Hospital) in Boston. This dataset contained ECG recordings for 25 AFib patients, each 10 hours long. ECG signals were included in the data of 23 subjects, so we used these subjects exclusively in our research (Goldberger et al.).

As shown in Figure 1, there is a distinct difference in the ECG recording of a normal sinus rhythm and that of an AFib arrhythmia. The consistency of the R-Peaks of the normal sinus rhythm graph eludes to a low HRV. However, the irregular pattern of the AFib graph entails a higher HRV.

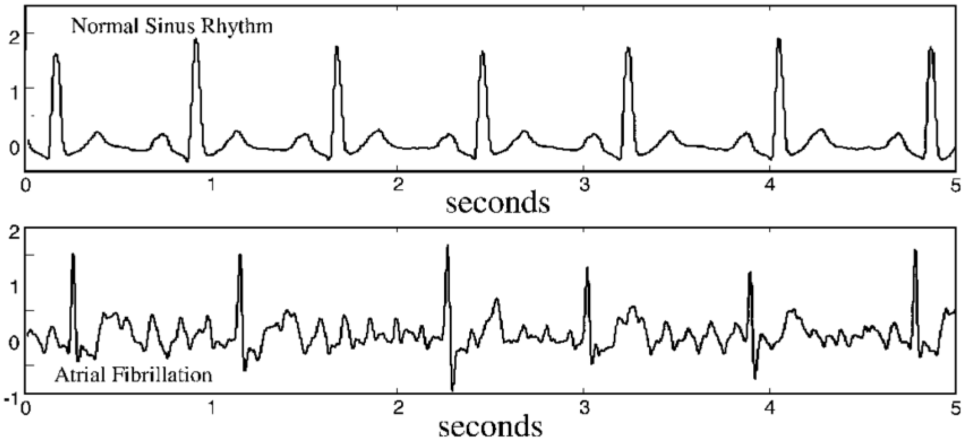
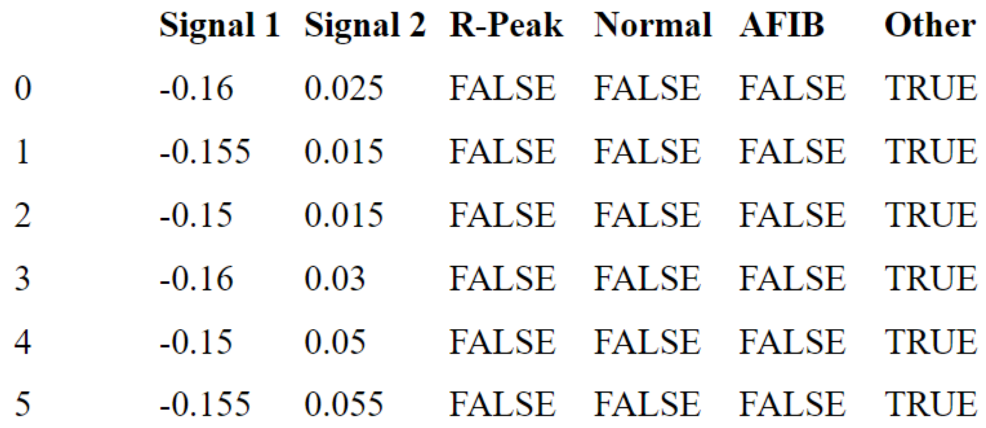


Figure 1: Normal sinus rhythm vs. an AFib rhythm on an ECG(Ramon et. a).

The MIT-BIH team analyzed the signals for each subject in the dataset to classify them as AFib or not at a given moment, packaging those labels with the dataset as rhythm annotations. The database includes six variables: “Signal 1”, “Signal 2”, “R-peak”, “Normal”, “AFib”, and “Other”. “Signal 1” and “Signal 2” are not as relevant to us as they simply measure the electric signal of the heart. The remaining four variables are instrumental, however. “R-peak” indicates when the heart contracts and pushes out blood in the body. “Normal, “AFib”, and “Other” represent categorical variables that classify the type of heart rhythm taking place.

Data Extraction

To begin our initial data extraction and analysis, we extracted the data into 23 data frames, each with around nine million rows that reflect the individual ECG samples. The first six different subjects are shown in Figure 2.



	Signal 1	Signal 2	R-Peak	Normal	AFIB	Other
0	-0.16	0.025	FALSE	FALSE	FALSE	TRUE
1	-0.155	0.015	FALSE	FALSE	FALSE	TRUE
2	-0.15	0.015	FALSE	FALSE	FALSE	TRUE
3	-0.16	0.03	FALSE	FALSE	FALSE	TRUE
4	-0.15	0.05	FALSE	FALSE	FALSE	TRUE
5	-0.155	0.055	FALSE	FALSE	FALSE	TRUE

Figure 2: First six samples of data from the first patient.

Our base feature extracted from the data was the length in samples between each R-peak as reported in each data frame, known as RR-Intervals. RR-Intervals have been explored extensively for their use in arrhythmia classification, and we decided to use them as our primary data source in this project. After saving these intervals, we removed outliers that were over 500 samples in length. 500 samples or longer between R-peaks would signal an improbably low heartbeat, which would most likely be attributed to a missed R-peak in the creation of the database. After outlier removal, we correlated the intervals with the type of rhythm reported at the end of each interval. This could include normal rhythm, AFib, or another arrhythmia.

Developing a Stepping Window Classification System

Literature Review

To guide our research, we analyzed the approaches used by various institutions conducting similar research. We particularly paid attention to the various models used, allowing us to see which ones foster the highest success rate. We explored the usage of a sliding window format for RR-Intervals, as well as other strategies for detecting AFib in near real-time.

In their paper on "An arrhythmia classification system based on RR-Interval signal", Tsipouras et al. suggest the usage of a moving window of 3 RR-Intervals for classification (Tsipouras). This moving window technique allows for arrhythmia detection very close to real-time, as the subject data refreshes routinely within the length of only 3 RR-Intervals. Despite this, the research does not involve machine learning techniques. Instead, it relies on a manually determined set of rules that come together to form an advanced classification algorithm for ventricular flutter, fibrillation, premature ventricular contractions, and 2nd-degree heart block. Each arrhythmia had its own set of conditions. Once paired with their sliding window, they were able to put the 3 RR-Interval subsets through the conditions to classify the type of rhythms present in the window (Tsipouras).

A major limitation of this research was that an algorithm like this could not be easily generalized as a consistent set of conditions and cannot fully apply to every dataset. This was shown when the algorithm was applied to different datasets, and varying results were shown throughout the positive predictive value percentage, sensitivity, and specificity (Tsipouras).

Implementation of a Stepping Window Classification System

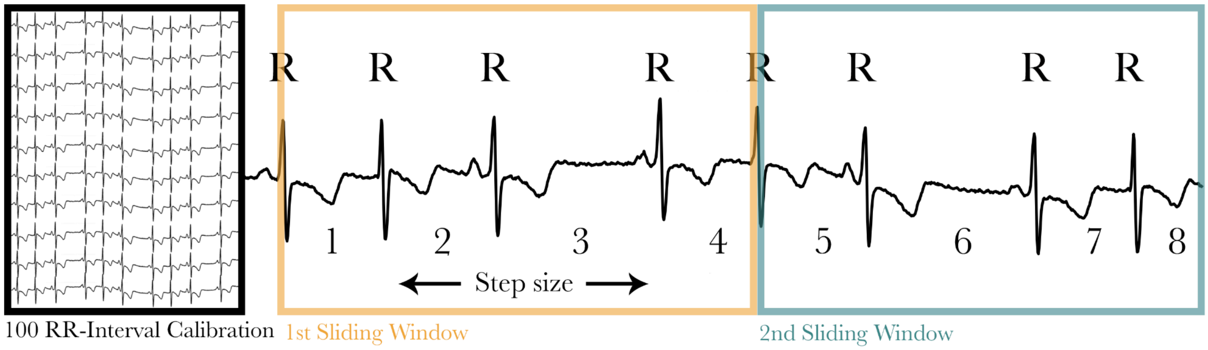


Figure 3: Example of Stepping Window

We developed our own feature calculation system that enabled machine learning classification with short subsets of data. First, the RR-intervals for each subject were broken into subsets of 4 intervals each. Then, when the features for one window are calculated, the features from the previous window are given a weight and included in the calculation. This technique essentially creates a trend of the features over time–enabling more accurate classification of a shorter window. Without a compensation method like this one, machine learning classification of such little data would yield inaccurate results.

Before the model can perform in real-time, there is a calibration period of 100 RR intervals. This "burn-in" period is necessary for the features to stabilize slightly before being calculated for each small subset. The first stepping window includes the calibration period's features, and the cumulative calculation continues from there.

Through this stepping window technique, the model does not have to rely on features calculated purely from the very short window. Instead, it can form more accurate predictions fuelled by all of the features that came before the current window. This method capitalizes on the sequential nature of heart rhythm data to power stronger classification.

Classification Models

We used the extracted features with several models to classify the subsets of RR intervals as AFib or normal. Through testing, we can understand which model would work best and be the most robust given our extracted features on the MIT-BIH database. The models we tested are Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM, and Catboost. We compared the performance of each model by analyzing the various metrics: average accuracy, the standard deviation of accuracy, sensitivity, specificity, precision, runtime, and F1-Score

Overall Results

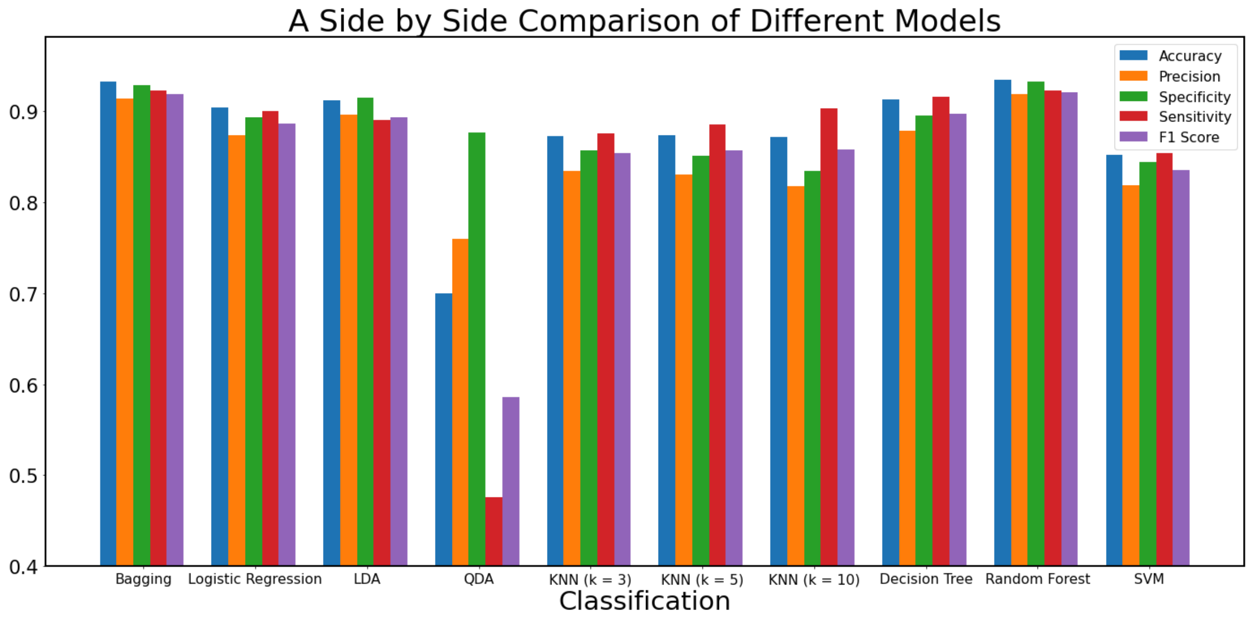
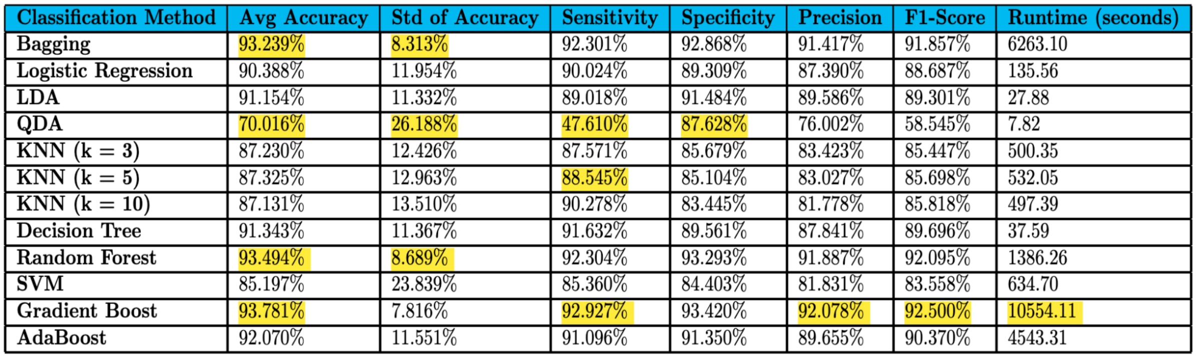


Figure 4: Performance of Different Classification Methods



Classification Method	Avg Accuracy	Std of Accuracy	Sensitivity	Specificity	Precision	F1-Score	Runtime (seconds)
Bagging	93.239%	8.313%	92.301%	92.868%	91.417%	91.857%	6263.10
Logistic Regression	90.388%	11.954%	90.024%	89.309%	87.390%	88.687%	135.56
LDA	91.154%	11.332%	89.018%	91.484%	89.586%	89.301%	27.88
QDA	70.016%	26.188%	47.610%	87.628%	76.002%	58.545%	7.82
KNN (k = 3)	87.230%	12.426%	87.571%	85.679%	83.423%	85.447%	500.35
KNN (k = 5)	87.325%	12.963%	88.545%	85.104%	83.027%	85.698%	532.05
KNN (k = 10)	87.131%	13.510%	90.278%	83.445%	81.778%	85.818%	497.39
Decision Tree	91.343%	11.367%	91.632%	89.561%	87.841%	89.696%	37.59
Random Forest	93.494%	8.689%	92.304%	93.293%	91.887%	92.095%	1386.26
SVM	85.197%	23.839%	85.360%	84.403%	81.831%	83.558%	634.70
Gradient Boost	93.781%	7.816%	92.927%	93.420%	92.078%	92.500%	10554.11
AdaBoost	92.070%	11.551%	91.096%	91.350%	89.655%	90.370%	4543.31

Table 1: Performance of classification models using all features

As shown in Figure 4 and Table 1, some models had more consistent and accurate results, whereas other models had less accurate results. Bagging and Gradient Boost both had high average accuracies of 93.239% and 93.781%, respectively. In contrast, classification methods such as QDA had lower accuracy, sensitivity, and specificity. The inconsistency in the QDA classifier can also be seen in the model's standard deviation of 26.188%. This high spread of accuracy shows the model's inability to perform in a typical fashion. We determined that the quadratic decision function created by the QDA model was not a good fit for our data, resulting in the poor performance that we observed. Along with the accuracy and standard deviation, other key performance measures are sensitivity and specificity. These measures are crucial as they demonstrate whether the model is unable to detect AFib or normal rhythm accurately. For instance, the QDA classification method has a low sensitivity of 47.610% but a relatively high specificity of 87.628%. This exemplifies how a particular method can be efficient in classifying a particular rhythm (in this case, normal rhythm) but still not be able to properly classify another type of rhythm (in this case, AFib).

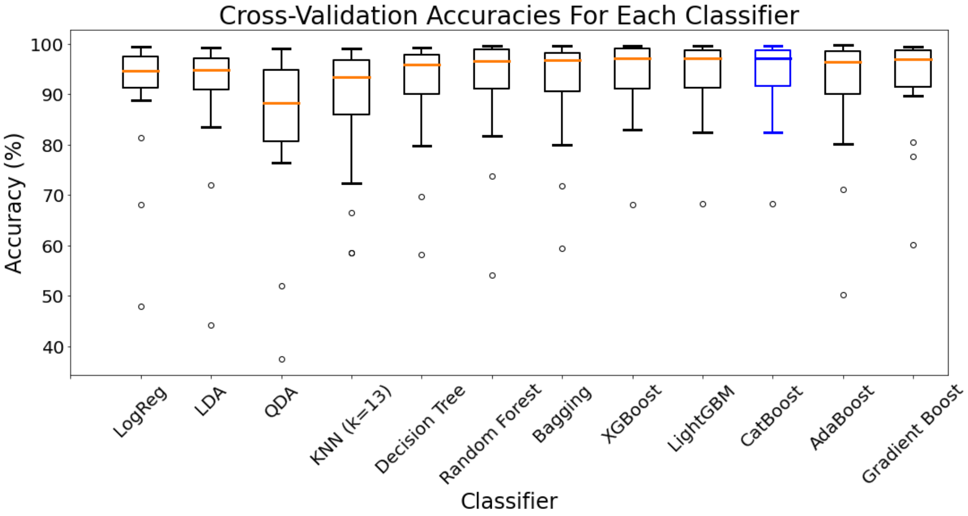


Figure 5: Cross-Validation Accuracies For Each Classifier

As shown in Figure 5, the different models we looked at had other spreads of accuracy. Certain models such as decision trees and logistic regression had lower spreads of accuracy, indicating more consistency throughout the model. In contrast, models such as QDA and KNN (k = 13) had their accuracies vary significantly more. These models had much less consistent outliers that were extremely far from the center of the data.

After creating our models, we used cross-validation to analyze the efficiency and accuracy of the models. Cross-validation is the process of using data to validate the model following training. We primarily utilized the technique for cross-validation to leave one person out (LOPO). LOPO was implemented into our models by leaving out one subject for testing and training the model on the remaining 22 subjects. Each subject was used for testing as there was a rotation throughout all the subjects. We calculated the average accuracy by looking at the percentage of correct detections averaged across the 23 LOPO cross-validation folds. To further analyze the classifiers, we computed the standard deviation of the accuracies for each of the 23 folds to see how inconsistent the model performance was.

Feature Selection

In our feature selection, we identified certain features as necessary. These features had relatively high-performance measures throughout and high consistency in their results.

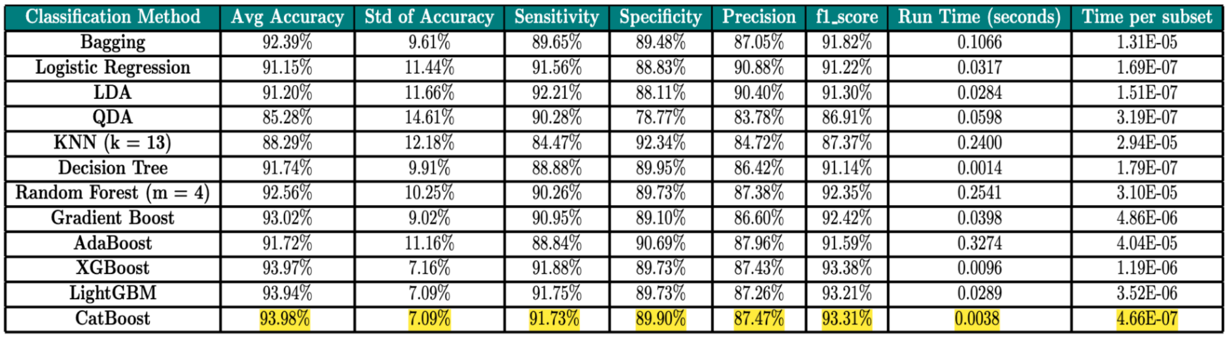
Interquartile range (IQR) measures the distance between the RR-interval lengths of the 1st and 3rd quartiles within a four RR-Interval subset. Due to AFib subjects having more variation in RR-interval measurements, they tend to have more widespread interval lengths. This results in an extensive range between the data quartiles in the subset. On the contrary, subjects with normal sinus rhythm have RR-interval lengths that are much closer together. This causes the data to be less widespread, resulting in a smaller IQR. IQR was quite effective in the classification of AFib and non-AFib. The average accuracy for IQR was greater than 80% for all the different classification methods. The IQR also had a relatively high sensitivity and specificity (80-90%) for most classification models. This demonstrates the feature's effectiveness in detecting both normal rhythm and AFib. IQR also had a lower standard deviation, showing how it had slight variance through all the cross-validation folds. Therefore, we identified IQR as a key feature in detecting AFib and decided to explore it further with our model.

Median Absolute Deviation (MAD) is a measure of the spread of the data. While it functions in a similar fashion to variance and standard deviation, it is more tailored towards more volatile data sets. MAD is calculated by subtracting the mean of the RR-Intervals of a four RR-Interval subset from the current RR-Interval, taking the absolute value, and finding the median of the resulting set of numbers. MAD was an essential feature throughout the classification methods. MAD consistently had an average accuracy of over 80% and maintained a standard deviation below 18%, exemplifying the accuracy of the feature. In addition, the feature had specificity measures greater than 80% and had relatively high sensitivity and precision measures going as high as 94.6% and 95.8% respectively. Overall MAD was an essential feature throughout all of the classification models and played a major role in distinguishing between AFib and normal sinus rhythm.

Root Mean Square of Successive Differences (RMSSD) measures the standard deviation between the differences of each successive RR interval. RMSSD performed quite accurately since it had high sensitivity and specificity reaching up to 87% and 95% respectively. Additionally, it had a high precision of up to 96% when paired with XGBoost, showing the minimal false predictions that occurred. Thus, we could conclude that RMSSD was a crucial feature of our model.

Transition Proportions focus on distinguishing each RR-Interval into one of three categories: short, regular, and long. Using these classifications, the proportions of transitions from category to categories is calculated. For example, using transition proportions, the proportion of short RR-Intervals that change to long or normal RR-Intervals is computed. These proportions can be utilized to find trends and correlations between the proportions and rhythms. Transition proportions was by far the best performing out of all the features, with almost every aspect being close to 90% with every classifier. While some of the run times for models such as Gradient Boost and AdaBoost were very high, the overall high performance proved that transition proportions were the highest-performing feature.

The coefficient of variance measures the spread of data points around the mean, and represents the ratio of the standard deviation to the mean (Hayes). The coefficient of variance also allows for comparison between values or items with different units of measurement. The coefficient of variance had a consistent level of accurate performance. Both the sensitivity and specificity were considerably high, especially when paired with boosting methods such as Gradient boost, AdaBoost, XGBoost, LightGBM, and CatBoost. Therefore we were able to conclude that the coefficient of variance was an important feature to include as it classified between positive and negative results accurately.



Classification Method	Avg Accuracy	Std of Accuracy	Sensitivity	Specificity	Precision	f1_score	Run Time (seconds)	Time per subset
Bagging	92.39%	9.61%	89.65%	89.48%	87.05%	91.82%	0.1066	1.31E-05
Logistic Regression	91.15%	11.44%	91.56%	88.83%	90.88%	91.22%	0.0317	1.69E-07
LDA	91.20%	11.66%	92.21%	88.11%	90.40%	91.30%	0.0284	1.51E-07
QDA	85.28%	14.61%	90.28%	78.77%	83.78%	86.91%	0.0598	3.19E-07
KNN (k = 13)	88.29%	12.18%	84.47%	92.34%	84.72%	87.37%	0.2400	2.94E-05
Decision Tree	91.74%	9.91%	88.88%	89.95%	86.42%	91.14%	0.0014	1.79E-07
Random Forest (m = 4)	92.56%	10.25%	90.26%	89.73%	87.38%	92.35%	0.2541	3.10E-05
Gradient Boost	93.02%	9.02%	90.95%	89.10%	86.60%	92.42%	0.0398	4.86E-06
AdaBoost	91.72%	11.16%	88.84%	90.69%	87.96%	91.59%	0.3274	4.04E-05
XGBoost	93.97%	7.16%	91.88%	89.73%	87.43%	93.38%	0.0096	1.19E-06
LightGBM	93.94%	7.09%	91.75%	89.73%	87.26%	93.21%	0.0289	3.52E-06
CatBoost	93.98%	7.09%	91.73%	89.90%	87.47%	93.31%	0.0038	4.66E-07

Table 2: Performance of All Selected Features for Each Model

Conclusions

We were able to see trends that helped us identify the features and classifiers that were the most efficient and effective in distinguishing between AFib and normal sinus rhythm in near real-time. To decide the best model, we looked at the different performance measures and the consistency of the model. All of these factors helped us understand if the model was able to meet the goal of detecting AFib in real-time properly.

In summary, we found XGBoost to be the best model when paired with the subset of features we classified as productive. The stepping window approach we explored proved to classify AFib with shorter subsets than before. Instead of looking at longer stretches of RR-Intervals, the model can classify on only a few seconds of data. This method could move AFib detection to near real-time. We aim to do more research on feature calculation, classifier options, and optimal parameters for the stepping window. We can also explore signal processing to allow us to feed real-world, raw data into our models. With signal processing, we could work towards implementing our findings into an application built for everyday wearable technology. As a result, the burden of the disease will be minimized.

References

Atrial Fibrillation. Cedars. (n.d.). Retrieved July 18, 2023, from https://www.cedars-sinai.org/health-library/diseases-and-conditions/a/atrial-fibrillation.html

Castells, F., Mora, C., Millet, J., Rieta, J. J., Sánchez, C., & Sanchís, J. M. (2004). Multidimensional Ica for separating atrial and ventricular activities from single lead ecgs in Paroxysmal Atrial Fibrillation episodes. *Independent Component Analysis and Blind Signal Separation*, 1229–1236. https://doi.org/10.1007/978-3-540-30110-3_155

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, PhysioToolkit, and PhysioNet. *Circulation*, *101*(23). https://doi.org/10.1161/01.cir.101.23.e215

Hayes, A. (2023, January 18). *Co-efficient of variation meaning and how to use it*. Investopedia.

Retrieved July 13, 2022, from https://www.investopedia.com/terms/c/coefficientofvariation.asp

Thuraisingham, R. A. (2006). Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR Intervals. *Computer Methods and Programs in Biomedicine*, *83*(1), 78–82. https://doi.org/10.1016/j.cmpb.2006.05.002

Tsipouras, M. G., Fotiadis, D. I., & Sideris, D. (2005). An arrhythmia classification system based on the RR-interval signal. *Artificial Intelligence in Medicine*, *33*(3), 237–250. https://doi.org/10.1016/j.artmed.2004.03.007