Early Detection & Monitoring of Atrial Fibrillation Cases

Data Science 2A Project (Time Series & Data Mining)

Group 123:

Michael & Pedro da Silva

April 2023

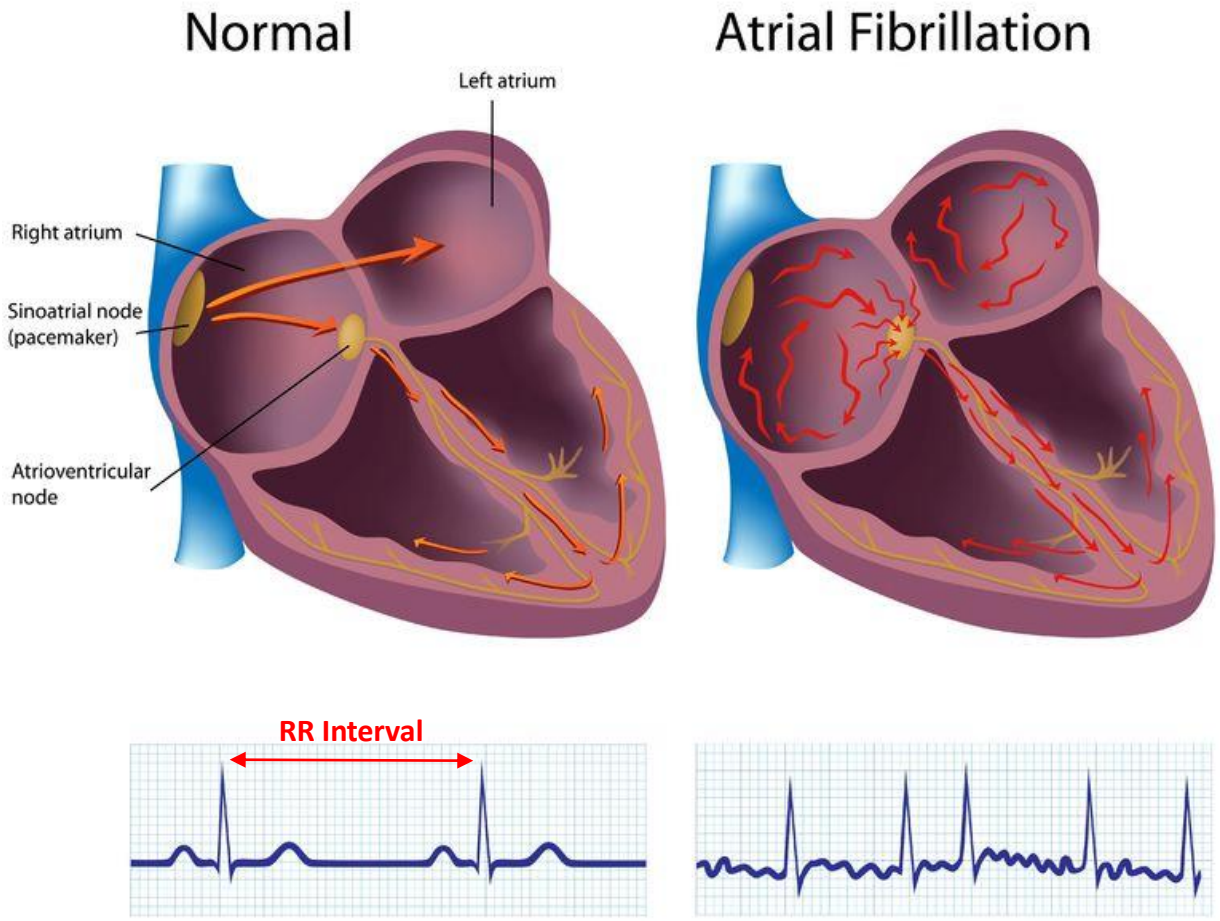
UNIVERSITY OF TWENTE.

Agenda

1. Introduction & Motivation
2. Research Questions
3. Dataset Background
4. Methodologies
5. Results & Discussion
6. Limitations & Future Studies

Introduction & Motivation

Atrial Fibrillation (AF) is a type of irregular **heartbeat**, or arrhythmia, that originates in the upper chambers of the heart (the atria). In AF, the atria beat **irregularly** and **rapidly**, which can cause poor blood flow to the body and increase the risk of blood clots, stroke, heart failure, and other cardiovascular complications.



- Electrocardiogram (**ECG**) is capable of capturing the random heartbeat patterns associated with Atrial Fibrillation (AF) episodes. The randomness is reflected by the **RR intervals**.
- Although AF is a common cause of many dangerous cardiovascular complications, it can be **very difficult to detect**, requiring tedious manual works. AFs are very short in the ECG signals and there are a lot of noises in the signals themselves.
- Given the inefficiency and high error rates of manual AF detection, there is a **strong demand** for an **automatic, early AF detection** for effective treatment.

Research Questions

The challenge of this project is to develop a **predictive model** using **machine learning algorithms** that is capable of **automatically detecting** episodes of **Atrial Fibrillation** from ECG data.

Main Research Question:

To what extent can we accurately detect atrial fibrillation using the RR intervals from ECG signals and machine learning techniques?

Sub-Research Question:

1. What features of ECG signals are most informative for detecting atrial fibrillation?
2. Which machine learning algorithms are most effective for detecting atrial fibrillation using ECG data?

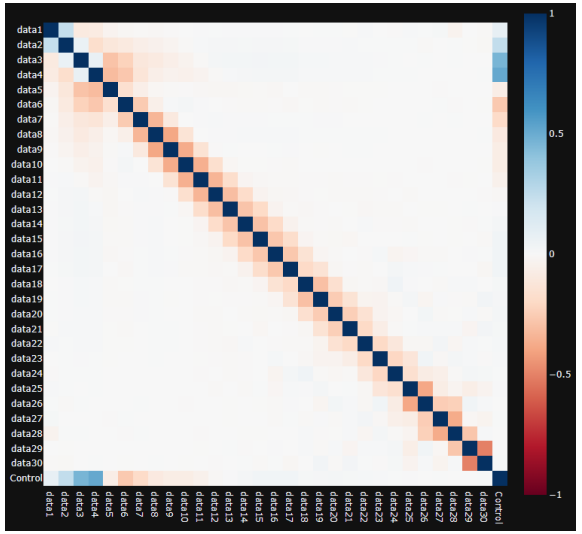
Dataset Background

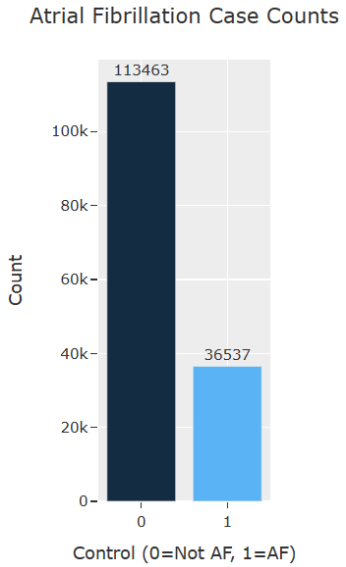
Data Source:
ECG data from the Erasmus Medical Centre in Rotterdam of the department electrophysiology. Data was obtained within 10 days post-operatively of CABG surgery.

Preprocessed data

- 150,000 observations
- 30 dependent variables (data1, data2, …, data30)
- 1 independent variable (Control)

data1	data2	data3	data28	data29	data30	Control
0	0	0	0	0	0	0
0	0.083333	0.916667	0	0	0	1
0	0	0.013333	0	0	0	1
0	0	0	0	0.032258	-0.03226	0
0	0	0	0	0	0	1
0	0	0.015873	0	0	0	1
0	0	0	0	0	0	0
0	0.363636	0.636364	0	0	0	0
0	0	0	0	0	0	0

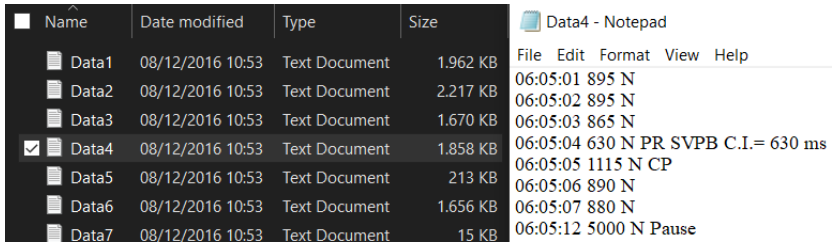




Raw data

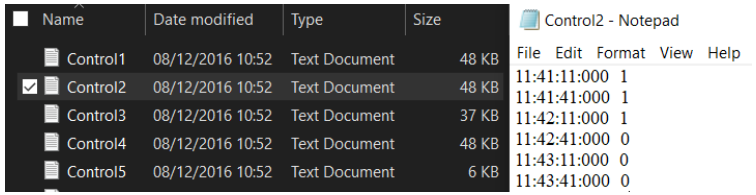
ECG Data

- 804 text files (Data1 to Data804)
- A semi-automatic program (Synescope) is used to analyze the ECGs for R peak annotation, followed by manual auditing by a physician to label AF cases.



Class Data

- 804 text files (Control1 to Control804)
- The AF and no-AF episodes were transformed into 1 and 0 in 30-second intervals by considering if more than 75% of the period was AF.



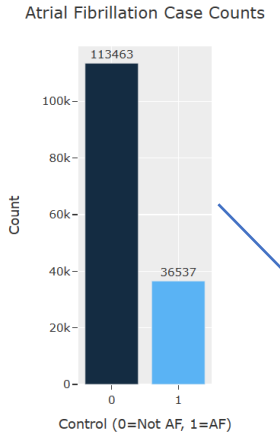
Methodologies

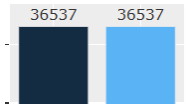
Preprocessed Data → **EDA**

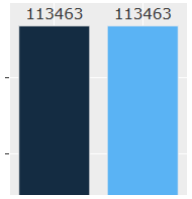
Handle Imbalance
- Original

- Undersampling

- Oversampling (SMOTE)







Model Building
- Train-test split
- Feature scaling
- Model training & testing**
- Result analysis

Raw Data → **EDA**

Data Preprocessing
- Text file transformation
- Merge both raw files, sliding windows of 30 seconds & 50% overlap
- Data cleaning

Feature Engineering
- Feature generation*
- Handle missing data
- Feature selection

Model Building
- Handle class imbalance
- Train-test split
- Feature scaling
- Model fitting & predicting**
- Result analysis

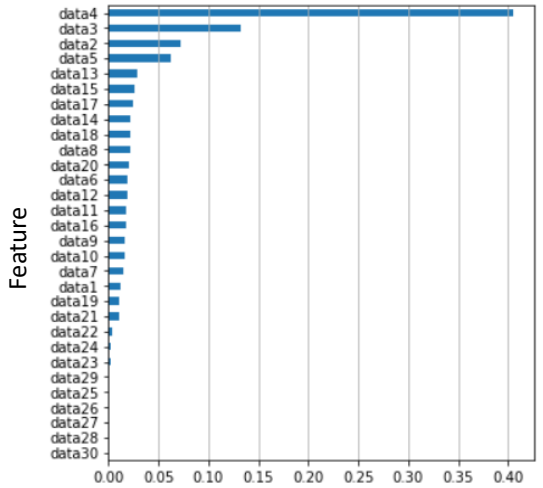
*features: time domain, frequency domain, geometrical, CSI&CVI, poincare plot (refer to the appendix for in-detail feature list)
**models: logistic regression, elastic net, KNN-DTW, naïve bayes, decision tree, random forest, XGboost, ANN, K-means clustering

Results & Discussion

Preprocessed Data

Machine Learning Model*	Original Data			
	Recall		Accuracy	
	Train	Test	Train	Test
Logistic Regression	0.81	0.81	0.93	0.93
LogReg + Elastic Net	0.82	0.80	0.94	0.93
KNN-DTW	0.72	0.70	0.91	0.89
Naïve Bayes	0.91	0.92	0.69	0.67
Decision Tree	0.85	0.86	0.93	0.93
Random Forest	0.88	0.89	0.95	0.95
XGBoost	**0.92**	**0.92**	**0.96**	**0.96**
Artificial Neural Network	0.91	0.92	0.95	0.96

Feature Importance
(based on XGBoost Model)



*model hyperparameters are tuned using GridSearch. Cross Validation with 5 folds were implemented for model training & testing
**importance = normalized average gain of each feature in all trees of the XGBoost model, representing the improvement in accuracy when it is used in split

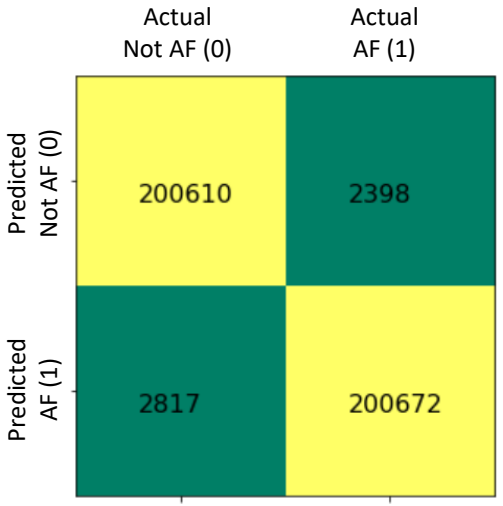
Results & Discussion

Raw Data

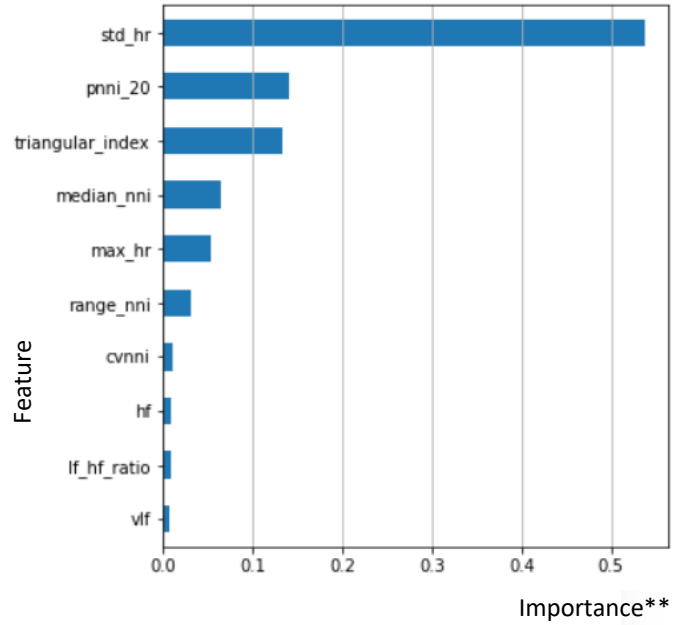
Model Result

Machine Learning Model*	Oversampling (SMOTE)			
	Recall		Accuracy	
	Train	Test	Train	Test
Logistic Regression	0.84	0.84	0.89	0.89
LogReg + Elastic Net	0.84	0.84	0.89	0.89
KNN-DTW	0.88	0.90	0.89	0.90
Naïve Bayes	0.83	0.83	0.89	0.89
Decision Tree	0.89	0.88	0.92	0.92
Random Forest	0.89	0.90	0.93	0.93
XGBoost	**0.99**	**0.99**	**0.99**	**0.99**
Artificial Neural Network	0.87	0.86	0.91	0.91

Confusion Matrix
(XGBoost Model, test dataset)



Feature Importance
(based on XGBoost Model)



- XGBoost Model has the best recall (0.99) and accuracy (0.99) compared to all models in both train and test datasets.
- Time domain features (std_hr, pnni_20, median_nni) have higher feature importance than frequency domain features (hf, lf_hf_ratio, vlf)

- Unsupervised learning: Tried K-means clustering with no success validating the data

*model hyperparameters are tuned using GridSearch. Cross Validation with 5 folds were implemented for model training & testing
**importance = normalized average gain of each feature in all trees of the XGBoost model, representing the improvement in accuracy when it is used in split

Limitations & Future Studies

- **Limited domain knowledge** and information about the data collection process. Expertise in signal filtering is required, and more time should be invested in this step to extract better Heart Rate Variability features.
- **Availability of computational resources**, particularly in computationally expensive algorithms such as KNN-DTW, resulting in the use of only 25% of the raw data (200 out of 804 files) for model training.
- **More input data from patients**, using RR intervals is only one of the methods to detect Atrial Fibrillation
- **Data quality**, validity and standards of the dataset came into check during the development of the testing models
- **Deep learning techniques** exploration for AF detection from ECG signals, which may improve the accuracy and efficiency of the detection process.
- **Shorter window size** in AF labeling, as 30 seconds might be too long to detect AF signals. It might also reduce class imbalance in the pre-processing part.

Thank you!

Questions?

UNIVERSITY
OF TWENTE.

Raw Data Feature List (1/2)

Feature Type	Feature Name	Description
Time domain	mean_nni	The average of all RR-intervals in a given period of time
	sdnn	The standard deviation of all RR-intervals in a given period of time
	sdsd	The standard deviation of differences between adjacent RR-intervals in a given period of time
	nni_50	The number of pairs of adjacent RR-intervals that differ by more than 50 ms in a given period of time
	pnni_50	The percentage of pairs of adjacent RR-intervals that differ by more than 50 ms in a given period of time
	nni_20	The number of pairs of adjacent RR-intervals that differ by more than 20 ms in a given period of time
	pnni_20	The percentage of pairs of adjacent RR-intervals that differ by more than 20 ms in a given period of time
	rmssd	The square root of the mean of the squared differences between adjacent RR-intervals in a given period of time
	median_nni	The median of all RR-intervals in a given period of time
	range_nni	The difference between the maximum and minimum RR-intervals in a given period of time
	cvsd	The coefficient of variation of differences between adjacent RR-intervals in a given period of time
	cvnni	The coefficient of variation of all RR-intervals in a given period of time
	mean_hr	The average heart rate calculated from the RR-intervals in a given period of time
	max_hr	The maximum heart rate calculated from the RR-intervals in a given period of time
	min_hr	The minimum heart rate calculated from the RR-intervals in a given period of time
	std_hr	The standard deviation of the heart rate calculated from the RR-intervals in a given period of time

Raw Data Feature List (2/2)

Feature Type	Feature Name	Description
Frequency domain	vlf	the spectral power in the frequency range of 0.0033 to 0.04 Hz
	lf	the spectral power in the frequency range of 0.04 to 0.15 Hz
	hf	the spectral power in the frequency range of 0.15 to 0.4 Hz
	lf_hf_ratio	the ratio of the power in the LF band to the power in the HF band
	lfnu	the normalized units of the LF component, expressed as a percentage of the total power minus the VLF power
	hfnu	the normalized units of the HF component, expressed as a percentage of the total power minus the VLF power
	total_power	the sum of spectral power across all frequency bands (VLF, LF, and HF)
Geometrical	triangular_index	the width of the distribution of RR intervals, calculated as the number of RR intervals divided by the height of the histogram's mode
	tinn	the estimated width of the distribution of RR intervals by calculating the time between the first and last of three equidistant points along the histogram of RR intervals
CSI & CVI	csi	(Complex Systems Instability): This is a non-linear measure of the complex interactions between different physiological systems, such as the cardiovascular and respiratory systems. It is calculated as the ratio of the power in the high-frequency (HF) range (0.15-0.4 Hz) to the power in the low-frequency (LF) range (0.04-0.15 Hz) of the heart rate variability spectrum.
	cvi	(Composite Variability Index): This is a measure of the overall variability in the heart rate over time. It is calculated as the ratio of the standard deviation of the normal-to-normal intervals (SDNN) to the mean heart rate.
	Modified_csi	An enhanced version of the original CSI that includes a correction for respiratory sinus arrhythmia, which is a natural variation in heart rate that occurs with breathing. It is calculated as the ratio of the power in the high-frequency range (0.15-0.4 Hz) to the power in the low-frequency range (0.04-0.15 Hz) after correcting for respiratory sinus arrhythmia.
Poincare plot	sd1	a measure of the short-term variability of beat-to-beat intervals
	sd2	a measure of the long-term variability of beat-to-beat intervals
	ratio_sd2_sd1	the ratio of sd2 to sd1, which provides an index of the balance between sympathetic and parasympathetic activity

Raw Data Feature Correlation