Automated Diagnosis of Coronary Artery Disease using Electrocardiograms

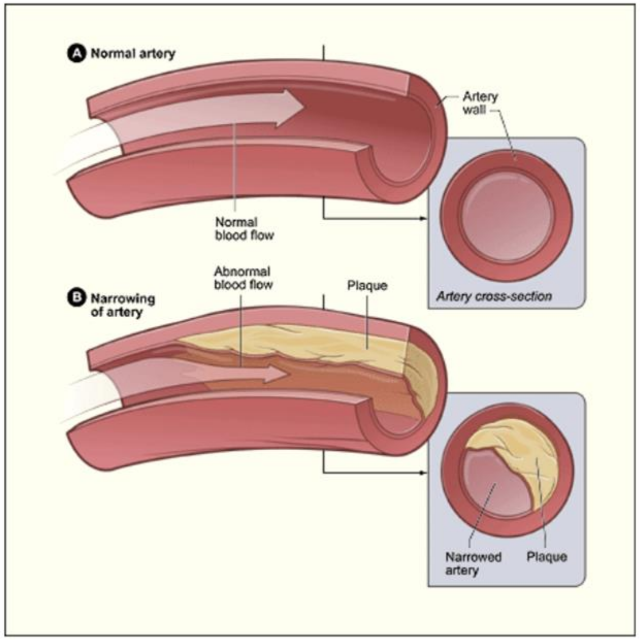
By: Jiachen Guo and Ashwin Vazhayil

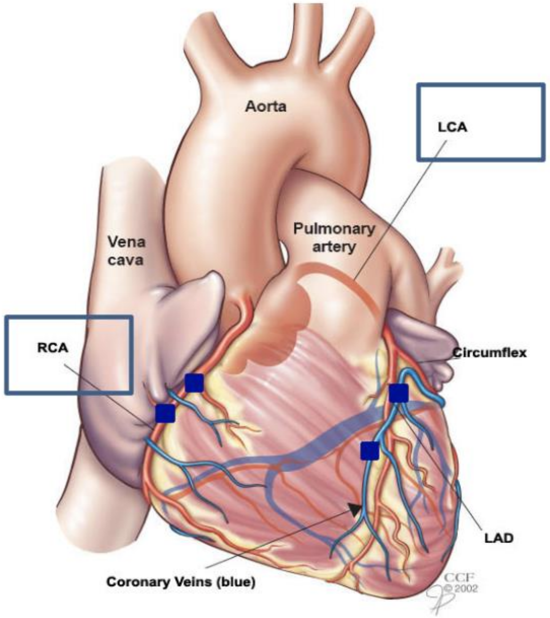
Mentors: Mahsa Tajdari, Xiaoyu Xie, Prof. Marc Laforest, Prof. Wing Kam Liu

2021.08.10

Coronary Artery Disease (CAD)

- CAD is the most common type of heart disease in the world
  - Approximately **400 000** fatalities were caused by CAD in US (2017)
- CAD is usually caused by the buildup of **cholesterol and fatty deposits** (called plaques) inside the arteries, which will **limit or stop blood flow** to the heart

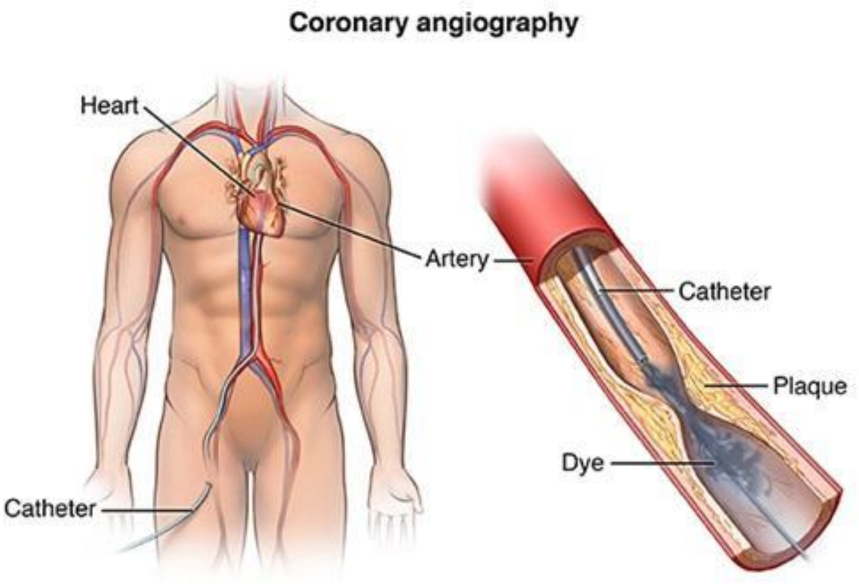




[1] Coronary Artery Disease. (n.d.). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease
[2] https://www.healio.com/cardiology/learn-the-heart/cardiology-review/patients/what-is-coronary-artery-disease-information-for-patients

Cardiac Catheterization/angiogram

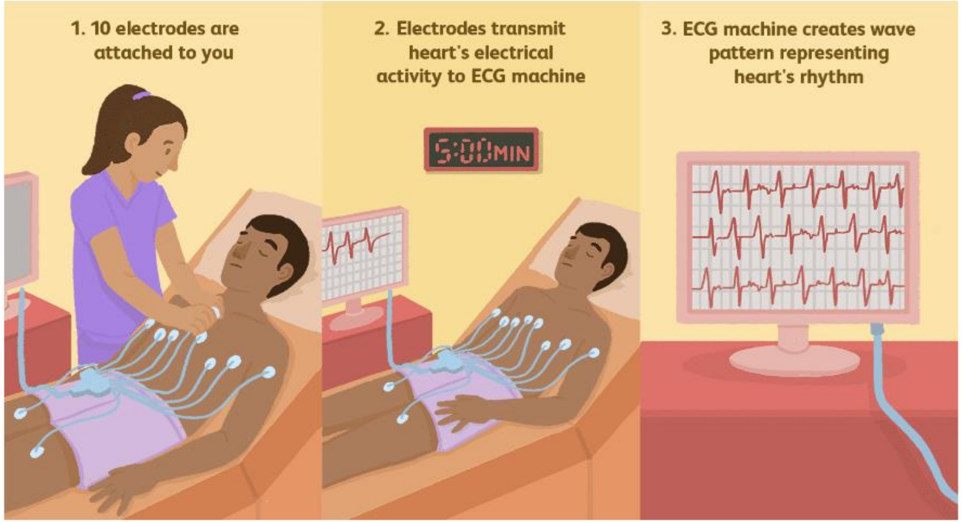
- Catheter is inserted into the **artery/vein**
- Dye is **injected** into catheter
  - Tracking blood flow
  - Stenosis/ buildup
- **INVASIVE**
- Does not apply to all **age** groups
  - Radiation, reactions
- Expensive ($4-6K)

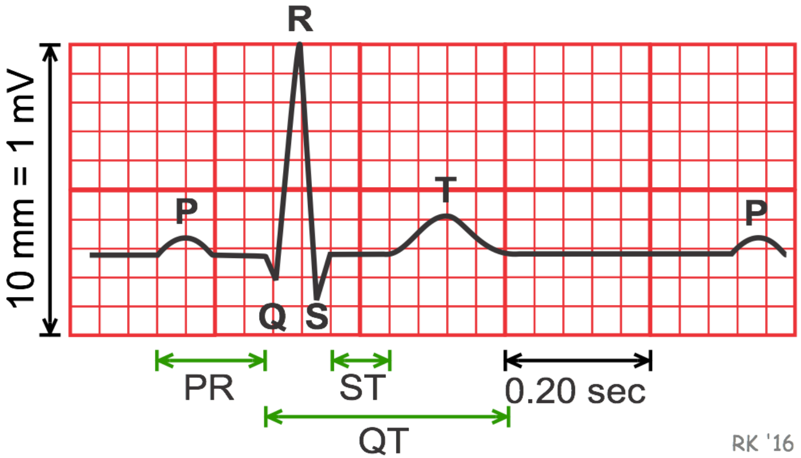


[1] https://www.mayoclinic.org/tests-procedures/cardiac-catheterization/about/pac-20384695

Electrocardiogram (ECG) in CAD diagnosis

- ECG is a simple test that can be used to check the **heart's electrical activity**
- ECG signals are **nonstationary and nonlinear** in nature
- **Noise**: baseline wander (low freq.) and muscle tremors (high freq.)
  - Denoising
- Accuracy of diagnosis is highly dependent on physician's **training and knowledge** (12 million people are misdiagnosed each year)

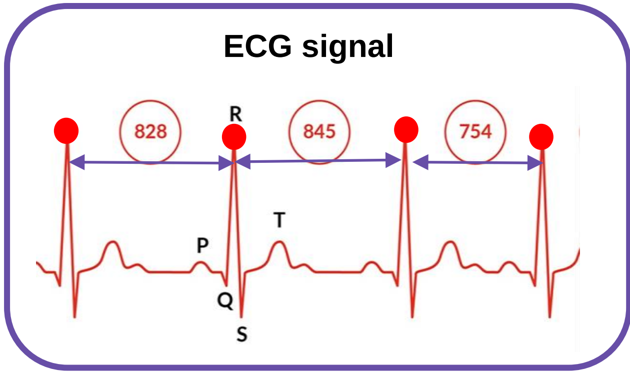




[1] https://www.verywellhealth.com/the-electrocardiogram-ecg-1745304
[2] https://www.cvphysiology.com/Arrhythmias/A009

Heart rate variability (HRV) as an indicator of CAD

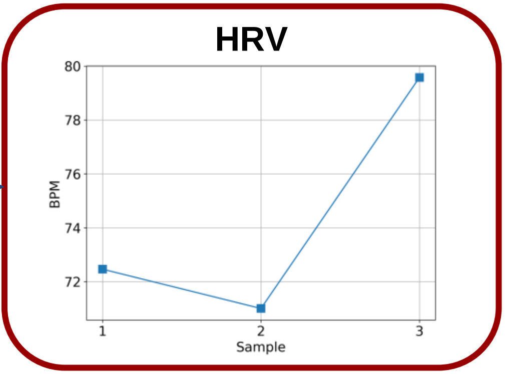
- **HRV** measures the variation of **heart beat** from beat to beat



Instantaneous heart rate

$$HR = \frac{60}{\Delta t}$$

Unit: bpm



https://www.firstbeat.com/en/blog/what-is-heart-rate-variability-hrv/

- **CAD patients** exhibit altered **HRV** in multiple aspects
  - Reduced rhythm of HRV
  - Reduction of low frequency and high frequency power
  - Reduced magnitude of HRV in time-frequency domain

Dua, S., Du, X., Sree, S. V., & VI, T. A. (2012). Novel classification of coronary artery disease using heart rate variability analysis. Journal of Mechanics in Medicine and Biology, 12(04), 1240017.

MDS approach for automated CAD diagnosis

Data collection

- Raw ECG signals from PhysioNet
- Data preparation with data balance
- ECG denoising
- R-peaks detection
- Heart rate variability

Feature extraction

- Time domain
- Frequency domain
- Time-freq domain

Classification

- Support Vector Machine

System & design

- Automated diagnosis mini-app

Data Collection (import raw data)

- Open source data from **PhysioNet** (a physiological database)

Import raw ECG

Data balance

ECG denoising

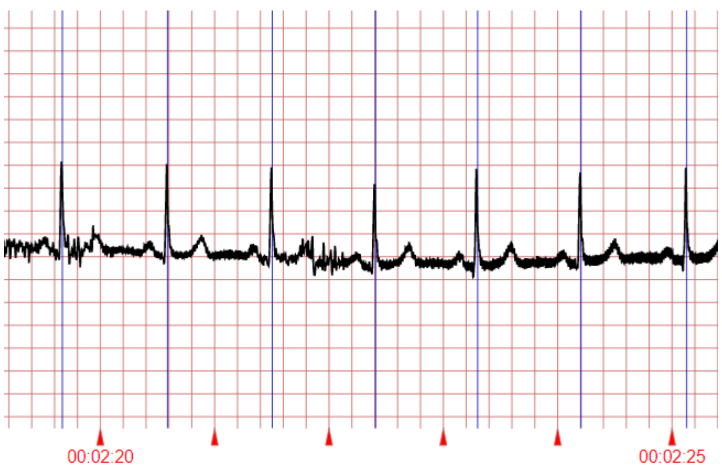
R-peaks detection

Heart rate variability

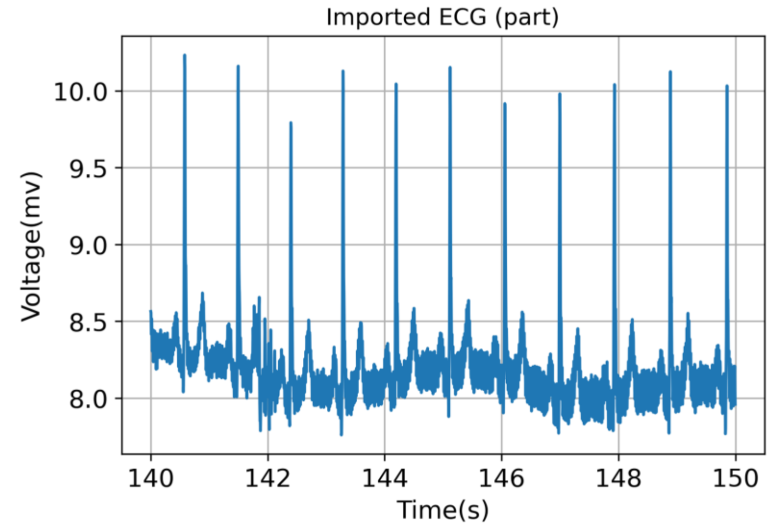
① Fantasia database
- **40 healthy subjects**
- **120 min** ECG signals
- **250 Hz**

② St Petersburg database
- **7 CAD patients**
- **30 min** ECG signals
- **257 Hz**

ECG signal in the database (segment): in special format



PyECG library



Data Collection (data balance)

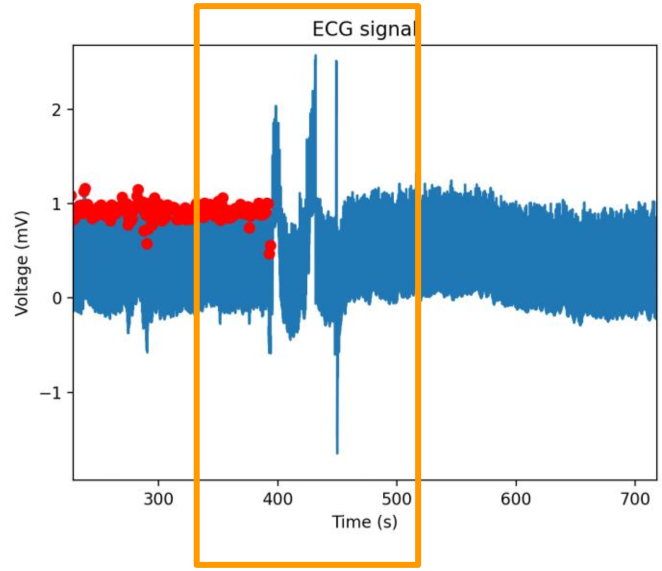
- Import raw ECG
- **Data balance**
- ECG denoising
- R-peaks detection
- Heart rate variability

① Fantasia database
- **40 healthy subjects**
- **120 min** ECG
- **250 Hz**
- 1 signal for each

② St Petersburg database
- **7 CAD patients**
- **30 min** ECG
- **257 Hz**
- 12 signals for each

Methods:

- **Upsample** the signals in Fantasia from **250Hz → 257Hz**
- Utilize **most data** provided to maximize efficiency of finalised diagnosis
- Segment data into **10 minute intervals** to ensure corrupted spikes of ECGs are not involved in the HRV calculations
- Utilize **overlapping techniques** to find appropriate intervals for HRV analysis



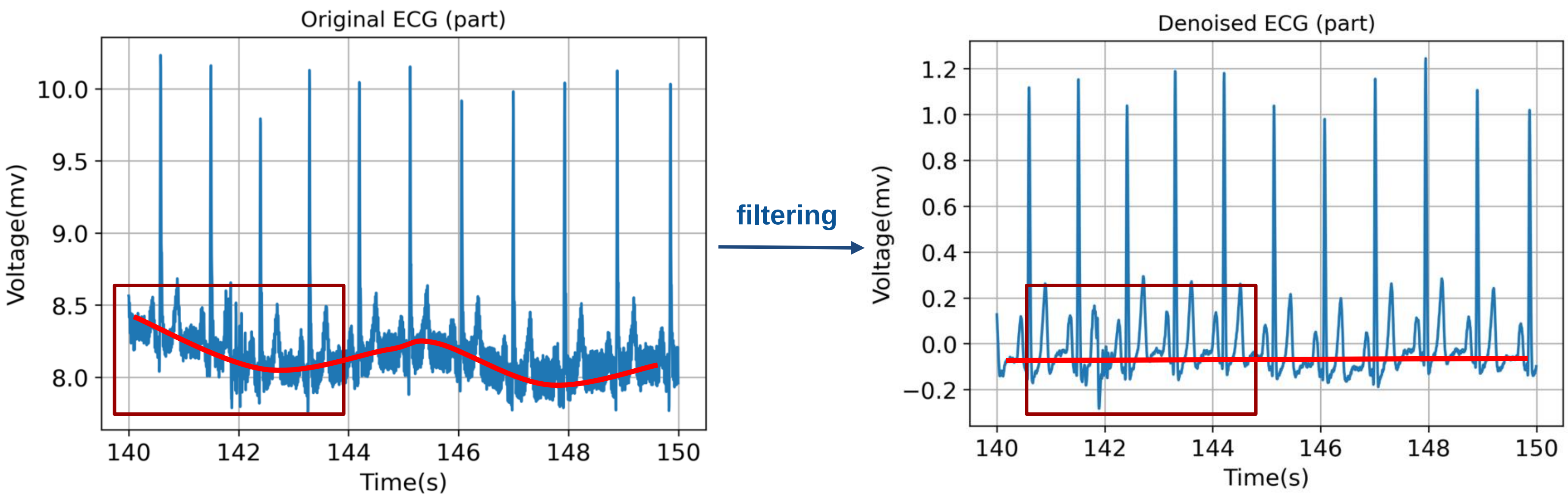
Sridhar, C., Acharya, U. R., Fujita, H., & Bairy, G. M. (2016, October). Automated diagnosis of Coronary Artery Disease using nonlinear features extracted from ECG signals. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 000545-000549). IEEE.

Data Collection (data denoising)

- Import raw ECG
- Data balance
- **ECG denoising**
- R-peaks detection
- Heart rate variability

ECG signals should be denoised before extracting R peaks

- **Baseline wander:** caused by respiration, movements or electrode contact
  - **Butterworth high-pass filter** (>0.3Hz)
- **Electromyographic noise:** caused by electrical activity of skeletal muscles
  - **Butterworth low-pass filter** (<15Hz)

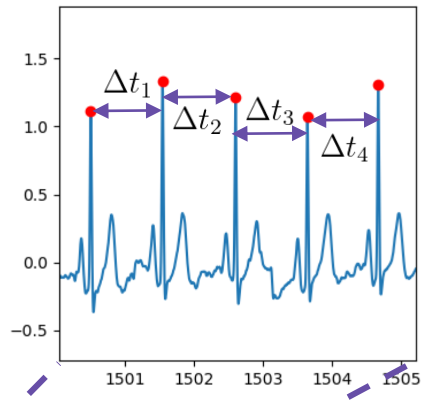


Maggio, A. C. V., Bonomini, M. P., Leber, E. L., & Arini, P. D. (2012). Quantification of ventricular repolarization dispersion using digital processing of the surface ECG. Advances in Electrocardiograms-Methods and Analysis.

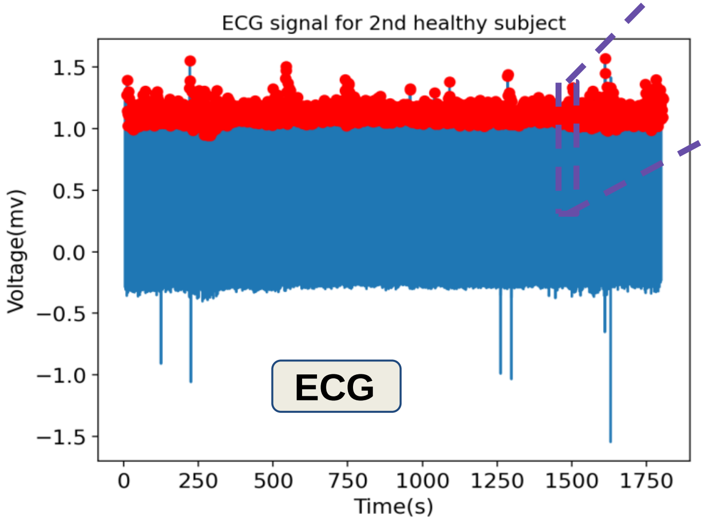
Data Collection (R-peaks detection and HRV)

R peaks are detected to derive the heart rate variability curve

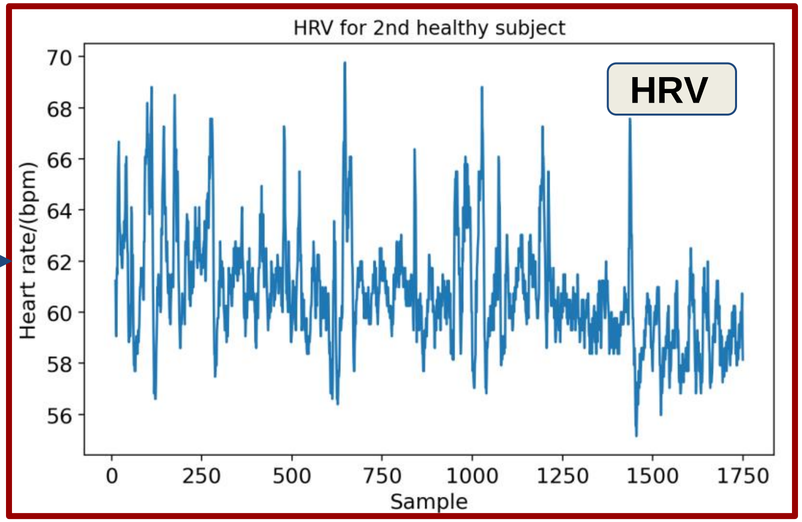
- Import raw ECG
- Data balance
- ECG denoising
- **R-peaks detection**
- **Heart rate variability**



R peaks are automatically identified in real-time using the adaptive thresholds algorithm proposed by Lourenço et al. (2012)



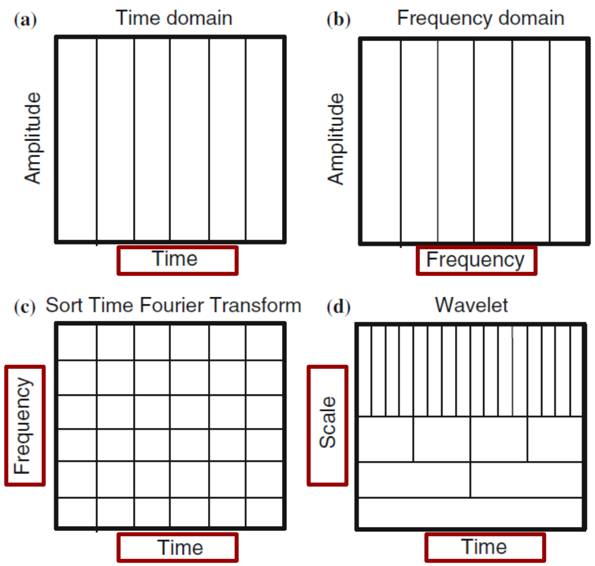
$$HR_i = \frac{60}{\Delta t_i}$$



Lourenço, A., Silva, H., Leite, P., Lourenço, R., & Fred, A. L. (2012, February). Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics. Biosignals (pp. 49-54).

Feature extraction

- **Feature extraction: time domain features, frequency domain features, and time-frequency domain features**



  - **Time domain**: mean values, standard deviations. (May miss subtle information)
  - **Frequency domain**: power distribution in the frequency domain. (Cannot account for non-stationary characteristics)
  - **Time-frequency domain**: short-time Fourier transform has fixed windows size; wavelet transform can achieve good resolution in both low and high frequency

Kuyuk, H. S. (2015). On the use of Stockwell transform in structural dynamic analysis. Sadhana, 40(1), 295-306.

Feature extraction (linear features)

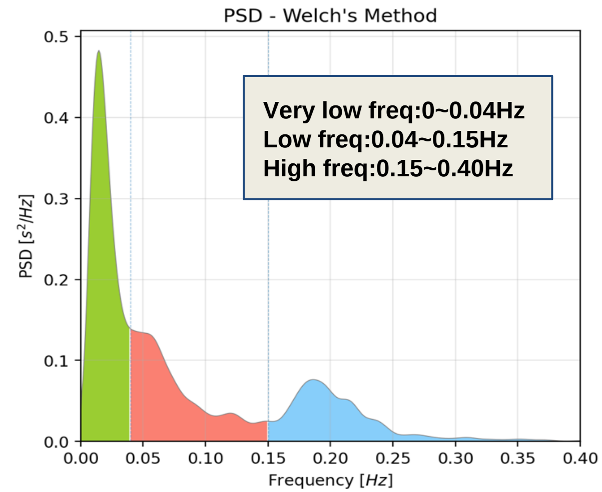
Time domain

Frequency domain

Time-freq domain

- **Features from the time domain**
  - **Mean** of heartbeat durations
  $$\bar{\Delta}_t = \sum_{j=1}^{n} \Delta_{t_i}$$
  - **Standard deviation** of heartbeat durations
  $$SD = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(\Delta_{t_j} - \overline{\Delta_t}\right)^2}$$
  - **Standard deviation** of heartbeat duration **differences**
  $$SDSD = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(\Delta(\Delta_{t_j}) - \overline{\Delta(\Delta_t)}\right)^2}$$

- **Features from the frequency domain**
  - **Total power**: detecting abnormal autonomic activity
  - **Low frequency power**: sympathetic modulation
  - **High frequency power**: parasympathetic modulation
  - **LF/HF ratio**: sympathetic/parasympathetic balance



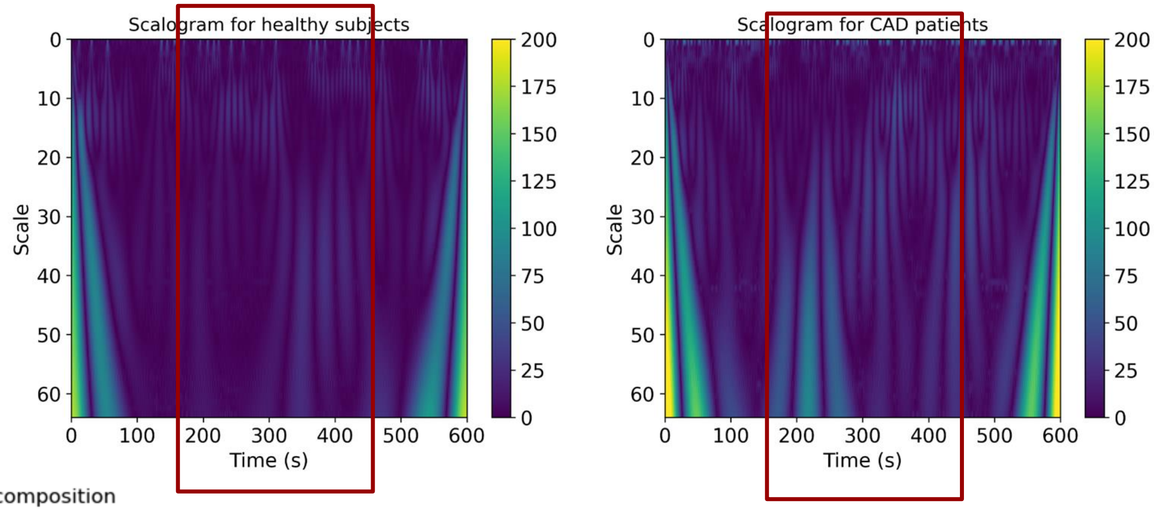
Lee, H. G., Noh, K. Y., & Ryu, K. H. (2008, May). A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In 2008 International conference on biomedical engineering and informatics (Vol. 1, pp. 200-206). IEEE..

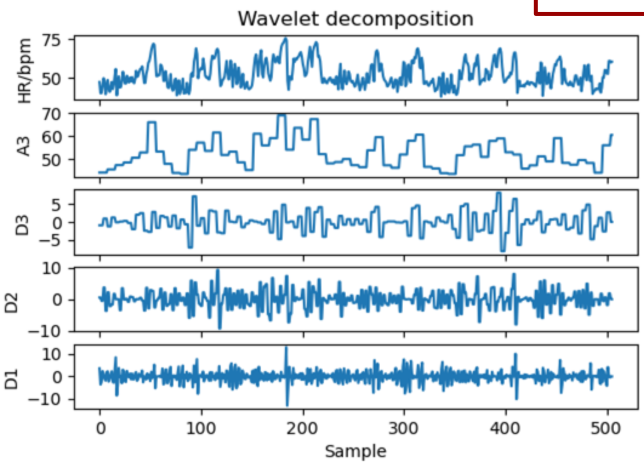
Feature extraction (nonlinear features)

Time domain
Frequency domain
Time-freq domain

Wavelet decomposition of HRV (multi-resolution)

Wavelet: $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$ ***a*: scaling parameter; *b*: translating parameter**





Nonlinear feature extraction for coefficients of each level

1. **Shannon entropy:** **measure of data uncertainty and variability**
2. **approximation entropy:** **quantify the amount of regularity and the unpredictability of fluctuations**
3. **sampling entropy:** **assess complexities of physiological signals**

Giri, D., Acharya, U. R., Martis, R. J., Sree, S. V., Lim, T. C., VI, T. A., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-Based Systems, 37, 274-282.

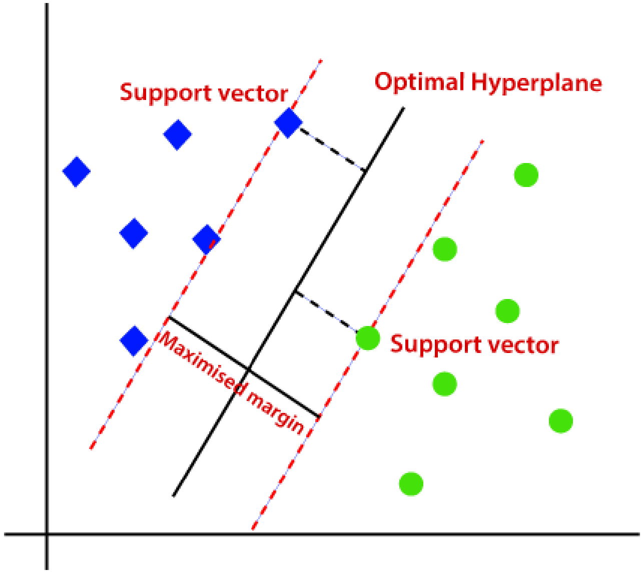
Feature extraction (summary)

- **Number of samples: 394**
  - Healthy subjects (197); CAD patients (197)
- **Number of dimensions: 19**
  - Time domain (3)+ frequency domain (4)+ time-frequency domain(12)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Mean	SD	SDSD	TP	LP	HP	LP/HP	SH_ca3	SH_cd3	SH_cd2	SH_cd1	AP_ca3	AP_cd3	AP_cd2	AP_cd1	SA_ca3	SA_cd3	SA_cd2	SA_cd1
2	0.988997	0.030015	0.00887	0.000769	83.66918	16.33082	5.123393	6.247928	6.247928	7.238405	7.676099	0.625126	0.616317	0.757336	0.963932	1.498212	2.031432	2.055344	2.058027
3	1.043633	0.041364	0.06454	0.001224	10.36627	89.63373	0.115651	6.169925	6.169925	7.159871	7.604802	0.794325	0.660468	0.634127	0.43662	1.543298	0.703912	0.551263	0.385133
4	0.97909	0.033975	0.02516	0.000709	52.74657	47.25343	1.116249	6.266787	6.266787	7.257388	7.861867	0.628875	0.766807	0.931206	0.857989	1.824549	1.312186	1.020835	0.825191
5	1.149948	0.085139	0.045644	0.006396	71.99376	28.00624	2.570633	6.022368	6.022368	7.022368	7.814469	0.473153	0.73102	0.974973	0.702723	1.89712	1.609438	1.162379	0.512079
6	1.050118	0.054175	0.045234	0.000692	59.44827	40.55173	1.465986	6.149747	6.149747	7.121578	7.414695	0.804262	0.487661	0.141516	0.472958	1.133704	0.473416	0.083057	0.331246
7	1.175074	0.034465	0.044577	0.000824	16.82538	83.17462	0.20229	6	6	6.988685	7.828232	0.651789	0.786937	0.94714	0.977546	1.94591	1.149165	1.214838	0.956673
8	0.982733	0.036619	0.035794	0.000925	38.82031	61.17969	0.634529	6.247928	6.247928	7.247928	7.997799	0.453836	0.630304	0.984182	1.030954	1.757858	1.362197	1.555949	1.037424
9	0.835155	0.034753	0.023557	0.000883	55.67778	44.32222	1.256205	6.491853	6.491853	7.483816	8.246775	0.525545	0.613402	0.851395	1.332363	1.642228	2.70805	1.883875	1.750896
10	1.420308	0.187411	0.21935	0.030181	26.85718	73.14282	0.367188	5.72792	5.72792	6.714246	7.695198	0.366119	0.415628	0.86342	0.818189	1.481605	2.079442	1.055871	0.647255
11	0.87096	0.046267	0.023521	0.001754	78.83501	21.16499	3.724783	6.426265	6.426265	7.39862	8.193116	0.475906	0.520179	0.909385	1.366715	1.863218	1.966113	1.762981	1.555626
12	0.923661	0.077579	0.115201	0.002566	20.07883	79.92117	0.251233	6.33985	6.33985	7.33985	8.135609	0.676634	0.664116	1.055733	0.393646	1.53393	1.127433	1.317221	0.300606
13	1.0993	0.031307	0.018108	0.000738	42.30768	57.69232	0.733333	6.087463	6.087463	7.087463	7.894792	0.420139	0.508159	0.782011	1.143295	1.276293	2.427748	1.884541	1.788668
14	1.069844	0.045593	0.05929	0.001692	22.56433	77.43567	0.291394	6.129283	6.129283	7.129283	7.877688	0.572485	0.739216	0.898018	0.767253	1.091177	0.82198	0.918131	0.718113
15	1.042116	0.029774	0.015199	0.000644	22.91061	77.08939	0.297195	6.169925	6.169925	7.169925	7.998874	0.564915	0.483814	0.896516	1.147272	1.424035	2.456736	2.063003	1.981767
16	0.792231	0.028867	0.033696	0.00042	64.16743	35.83257	1.790757	6.548803	6.569856	7.526502	7.666494	0.590982	0.770653	1.098751	0.687393	1.878771	1.575536	1.636963	0.581358
17	1.202193	0.037432	0.014952	0.001174	54.09283	45.90717	1.178309	5.954196	5.954196	6.954196	7.834634	0.272961	0.369745	0.669818	1.042194	2.70805	2.442347	2.062634	2.029845
18	0.997519	0.062197	0.091654	0.001225	17.59704	82.40296	0.213549	6.228819	6.228819	7.215485	7.982904	0.694595	0.835018	0.93465	0.437152	1.544197	1.033416	1.035868	0.356404
19	1.138244	0.047797	0.055176	0.002248	22.13626	77.86374	0.284295	6.044394	6.044394	7.033423	7.70604	0.453467	0.54899	1.088159	0.542008	1.189584	1.658228	1.360652	0.446735
20	0.795121	0.047011	0.056553	0.002813	43.81621	56.18379	0.779873	6.554589	6.554589	7.543951	7.823754	0.550644	0.807687	1.061406	1.15084	1.597603	1.397105	1.403586	1.141908
21	0.790692	0.06716	0.049228	0.004022	37.42589	62.57411	0.598105	6.569856	6.569856	7.562242	8.414163	0.663948	0.671024	0.901162	1.192052	1.867745	1.996554	2.03816	1.662101

Classification using support vector machine (SVM)

- **Scale input features** before classification
- Label the outputs for **healthy** subjects as **1**; **CAD** patients as **0**
- **Support vector machine** (SVM) is used to classify healthy and CAD patients
  - SVM intends to find an **optimal hyperplane** that has **largest distance** to **support vectors**
  - SVM works well for problems with **small datasets**
  - **Linear kernel** is used in the current analysis



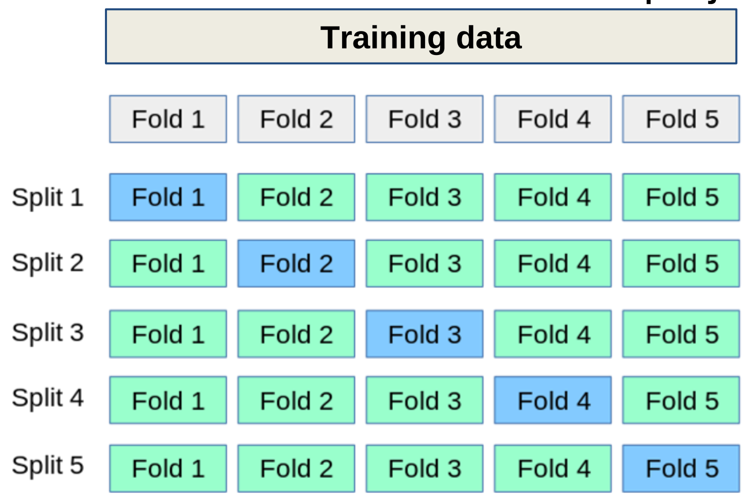
Scale inputs and define SVM model

```
from sklearn import svm
scaler = preprocessing.StandardScaler().fit(X)
joblib.dump(scaler, 'data_scaler.pkl')
X = scaler.transform(X)
clf = svm.SVC(kernel='linear', C=1000, random_state=0)
```

https://www.analyticsvidhya.com/blog/2021/05/5-classification-algorithms-you-should-know-introductory-guide/

Classification using K-fold cross-validation

- **5-fold cross validation** is used in the project



Dataset is **divided** into 5-subsets

Each split is divided into **train+validation**

https://scikit-learn.org/stable/modules/cross_validation.html

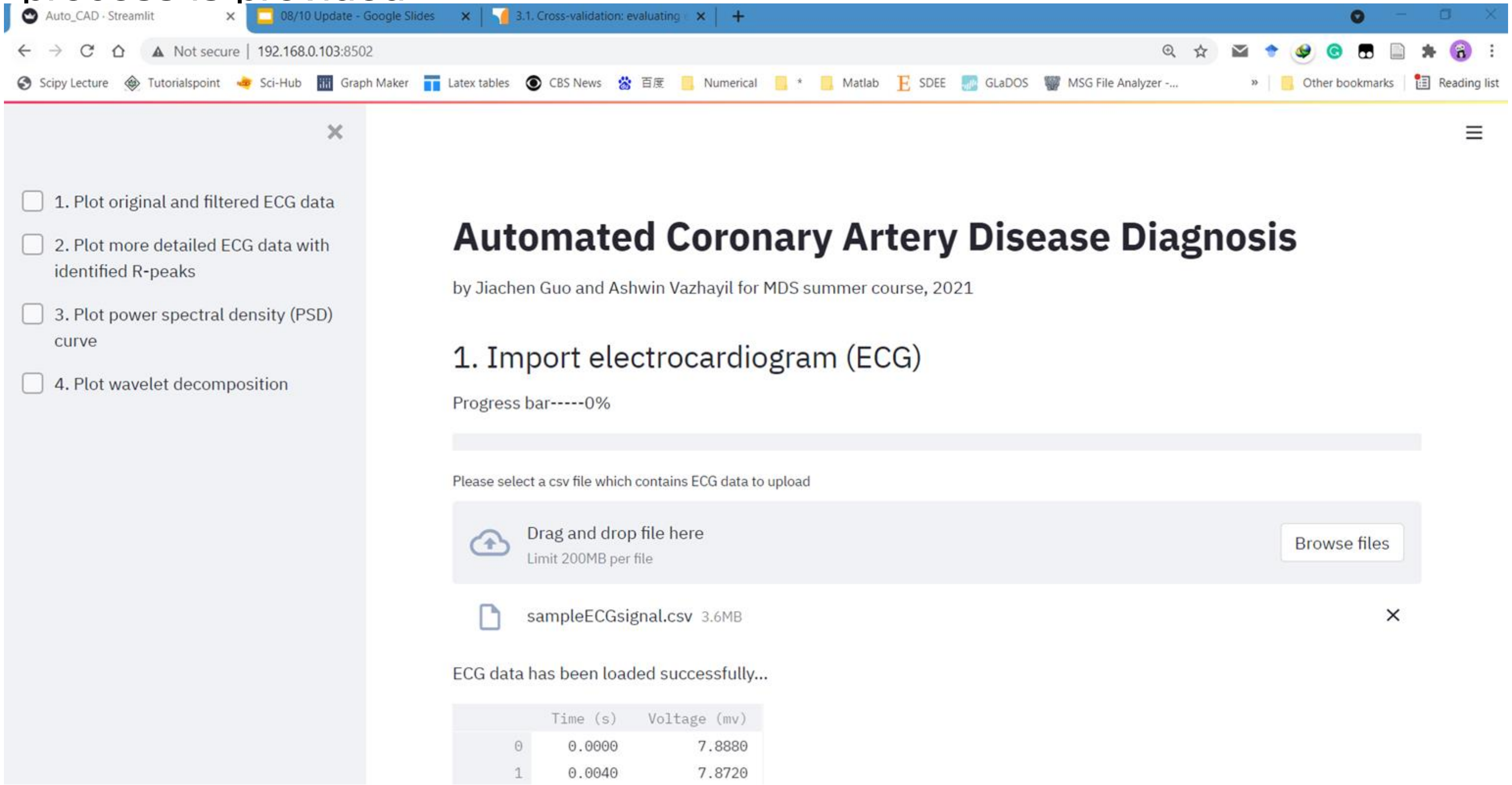
- **Training results** for the current model

```
In [20]: runfile('F:/MDS/svmTraining.py', wdir='F:/MDS')
Fold:  1, Accuracy: 0.962, F1 score 0.961
Fold:  2, Accuracy: 0.949, F1 score 0.947
Fold:  3, Accuracy: 0.975, F1 score 0.974
Fold:  4, Accuracy: 0.987, F1 score 0.987
Fold:  5, Accuracy: 0.910, F1 score 0.914

Cross-Validation accuracy: 0.957 +/- 0.026
Cross-Validation F1 score: 0.957 +/- 0.025
```

System and Design (facilitating CAD diagnosis)

- **Inputs** are the **raw ECG signals** from a standard ECG recording
- **Outputs** are the **diagnosis result** (**healthy** or **CAD**)
- An **interactive mini-app** which facilitate the physicians with the CAD diagnosis process is provided



Conclusion

- **Mechanistic data science** approach is adopted to find the internal relationships between **physical heartbeats** and **coronary artery disease**
- **Data balance** for the different classes is achieved while maintaining the integrity of the original data
- **Features** extracted from **time domain, frequency domain and time-frequency domain** can effectively represent the characteristics of patients with coronary artery disease
- **SVM** is used as the classifier and **5-fold cross validation** is carried out in the training process. **High accuracy** (0.957 ± 0.026) is achieved, which manifests the effectiveness of the proposed method
- An **automated diagnosis mini-app** is developed to assist clinical diagnosis of coronary artery disease

Appendix

- **Terminology used in the presentation**
    - **MDS:** mechanistic data science
    - **CAD**: coronary artery disease
    - **ECG**: electrocardiogram
    - **HRV**: heart rate variability
    - **R peak**: the maximal point for R signal in QRS complex
    - **Denoising**: signal processing method used to filter certain wavelengths (noise) to obtain useful information
    - **Power spectral density (PSD):** a quantity describes how power of a signal is distributed over frequency
    - **Wavelet Transform**: a multiresolutional transform to convert functions from time based domain to time-frequency domain
    - **Support vector machine (SVM):** a supervised machine learning algorithm to find a decision boundary to classify data
    - **K-fold cross validation**: a procedure used to estimate the skill of the model on new data

References

[1]Giri, D., Acharya, U. R., Martis, R. J., Sree, S. V., Lim, T. C., VI, T. A., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform. Knowledge-Based Systems, 37, 274-282.

[2]Dua, S., Du, X., Sree, S. V., & VI, T. A. (2012). Novel classification of coronary artery disease using heart rate variability analysis. Journal of Mechanics in Medicine and Biology, 12(04), 1240017.

[3]Sridhar, C., Acharya, U. R., Fujita, H., & Bairy, G. M. (2016, October). Automated diagnosis of Coronary Artery Disease using nonlinear features extracted from ECG signals. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 000545-000549). IEEE.

[4]Xie, L., Li, Z., Zhou, Y., He, Y., & Zhu, J. (2020). Computational diagnostic techniques for electrocardiogram signal analysis. Sensors, 20(21), 6318.

[5]Sharan Yadav, G., Yadav, S., & Prachi, P. (2013). Time and Frequency Exploration of ECG Signal. International Journal of Computer Applications, 67(4), 5-8.

[6]Maggio, A. C. V., Bonomini, M. P., Leber, E. L., & Arini, P. D. (2012). Quantification of ventricular repolarization dispersion using digital processing of the surface ECG. Advances in Electrocardiograms-Methods and Analysis.

[7]Karimi, M., Amirfattahi, R., Sadri, S., & Marvasti, S. A. (2005). Noninvasive detection and classification of coronary artery occlusions using wavelet analysis of heart sounds with neural networks.

[8]Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., & Chua, C. K. (2017). Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. Knowledge-Based Systems, 132, 62-71.

[9]Acharya, U. R., Faust, O., Sree, V., Swapna, G., Martis, R. J., Kadri, N. A., & Suri, J. S. (2014). Linear and nonlinear analysis of normal and CAD-affected heart rate signals. Computer methods and programs in biomedicine, 113(1), 55-68.

[10]Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation, 101(23), e215-e220.

[11]Lourenço, A., Silva, H., Leite, P., Lourenço, R., & Fred, A. L. (2012, February). Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics. Biosignals (pp. 49-54).

Thank you!