Machine Learning Models/Analysis for Heart Disease Detection

Hamid Vakilzadian, Ph.D.

Professor, Department of Electrical and Computer Engineering, UNL College of Engineering

Acknowledgement

This work in progress is partially funded by a grant from NU System Collaboration–Initiative team seed grant 21-1125-6011

Team members: Drs. O’Leary, Haider, Mathers, Azmeen, Vakilzadian

Some information in these slides are from the internet resources.

Outline

- The idea behind this work
  - Research for use of AI technology for early heart disease diagnosis.
  - Important for rural resident with little or no access to a specialists.

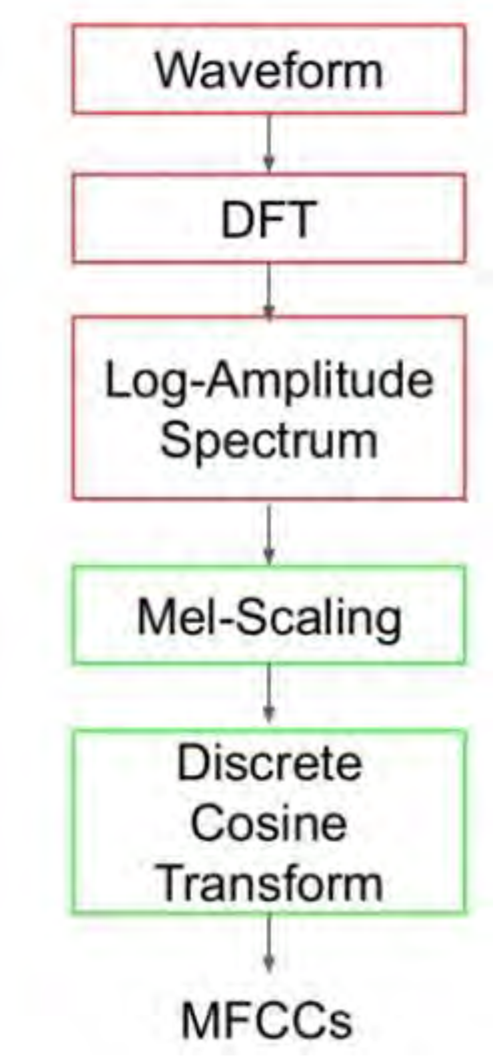
Presentation has two parts:

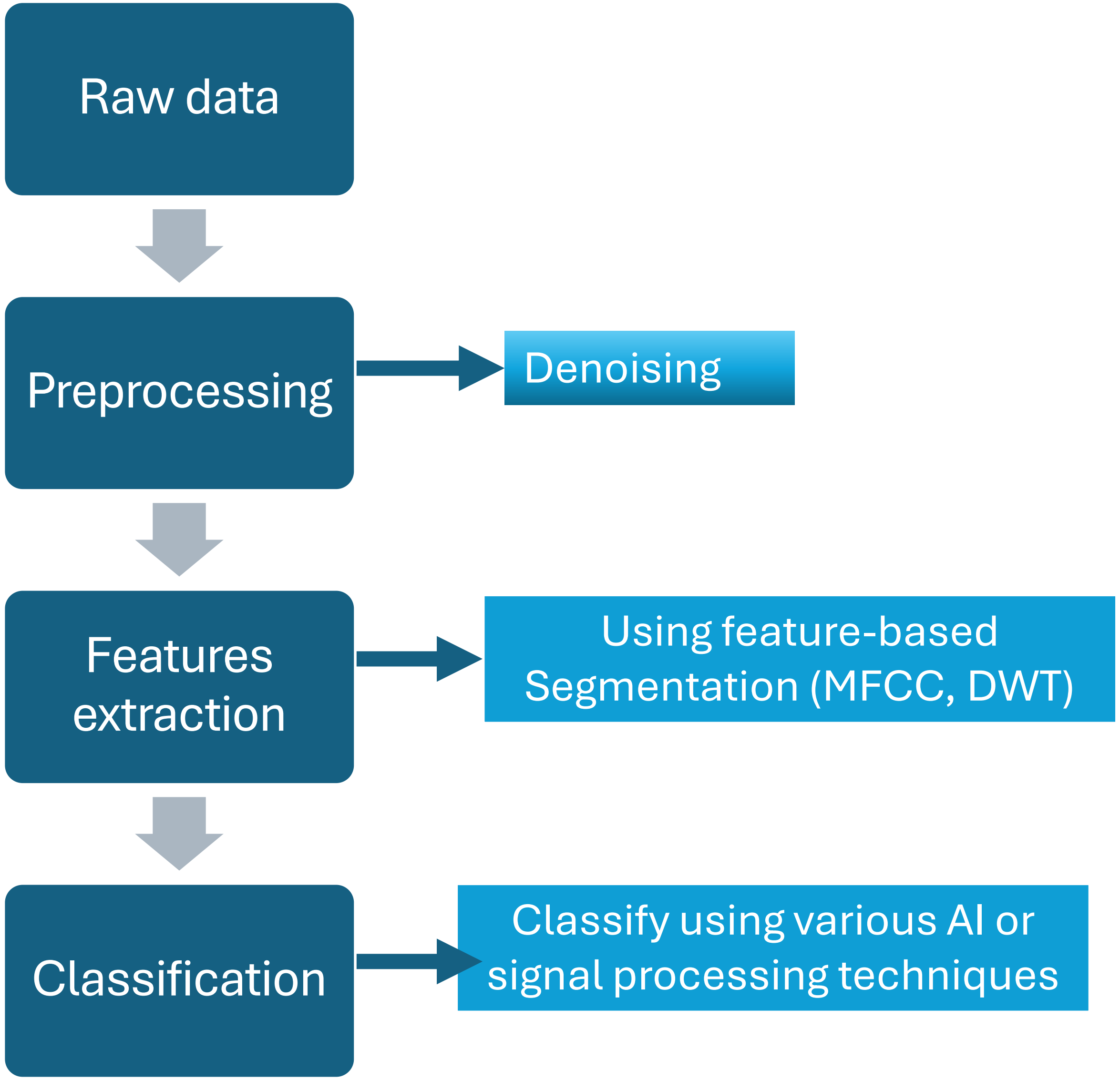
- Current work in heart sound analysis and AI – Work in progress
  - Feature-based segmentation of heart sounds (MFCC, DWT) to capture the characteristics of the sound such as tone, rhythm, etc.
  - Use the deep learning techniques (CNN, RNN, Transfer Learning) and compare their detection accuracy and capabilities with traditional ML techniques such as Decision tree classifiers.
- Classification and prediction of heart diseases using data mining techniques based on number of selected attributes
  - Such as Bayesian, PCA, Decision Tree, Support Vector Machine
  - Determine accuracy level of machine learning techniques based on the number of selected attributes such as physical activity, blood pressure, etc.

Long term goal of this work in research

Use the signal processing and AI techniques using the heart sound signals collected from the stethoscope to detect heart diseases early and reduce expensive tests.

Heart Sound Signal Analysis





Raw data

Preprocessing

Denoising

Features extraction

Using feature-based Segmentation (MFCC, DWT)

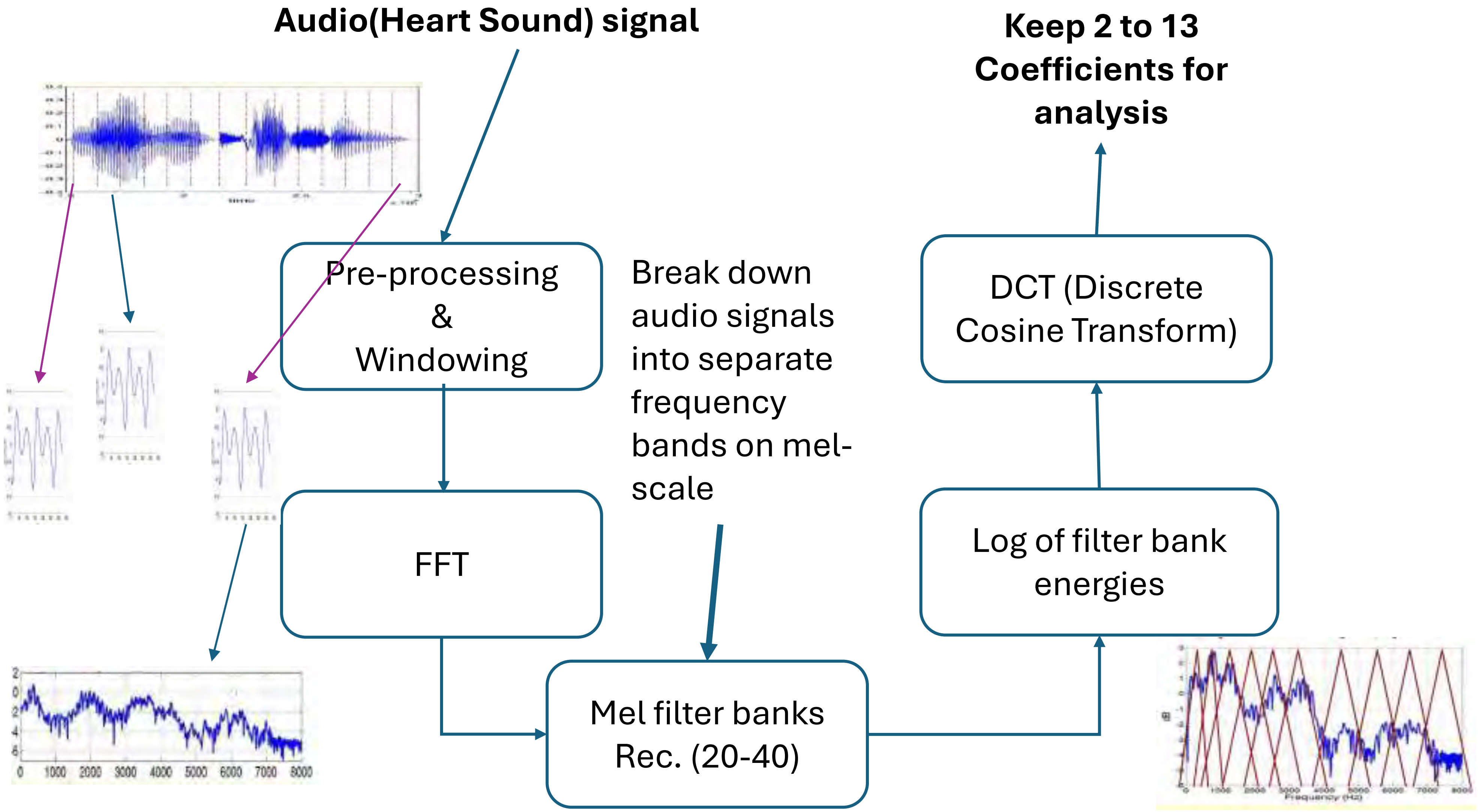
Classification

Classify using various AI or signal processing techniques

Mel-Scale and Mel-Frequency Cepstral Coefficient (MFCC)

- Heart sounds are audio signals; they need special tools for analysis
- Widely used in automatic speech recognition
- It is based on human perception to extract features from an audio signal
- Humans perceive frequencies on a log scale, so the Mel-Scale is used to better map the frequencies to how we would hear them.
- The scale was developed using a trial-and-error method through experiments
- Same Frequency bands (60-260 and 1568-1768) have different pitch
    - Therefore, the energy level of high frequency bands need to be adjusted for auto detection.

Mel Frequency Cepstral Coefficient (MFCC)



Audio(Heart Sound) signal

Pre-processing & Windowing

FFT

Break down audio signals into separate frequency bands on mel-scale

Mel filter banks Rec. (20-40)

Log of filter bank energies

DCT (Discrete Cosine Transform)

Keep 2 to 13 Coefficients for analysis

Convert frequencies to Mel-scale

1. Choose number of mel bands (20, 60, 90, 128 ?)

2. Construct mel filter banks →

$$m = 2595 \cdot log(1 + \frac{f}{500})$$

   - Convert lowest/highest frequency to mel
   - Create # bands equally spaced points

   - Convert points back to Hertz →

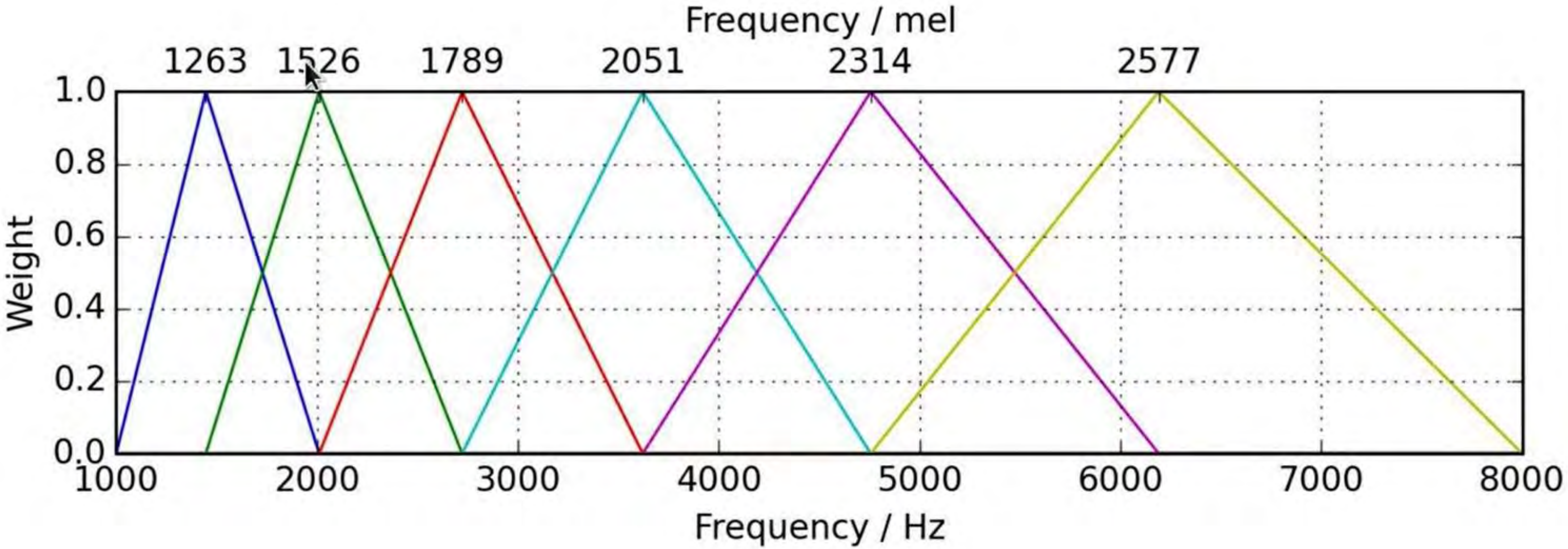
$$f = 700(10^{m/2595} - 1)$$

   - Round to nearest frequency bin

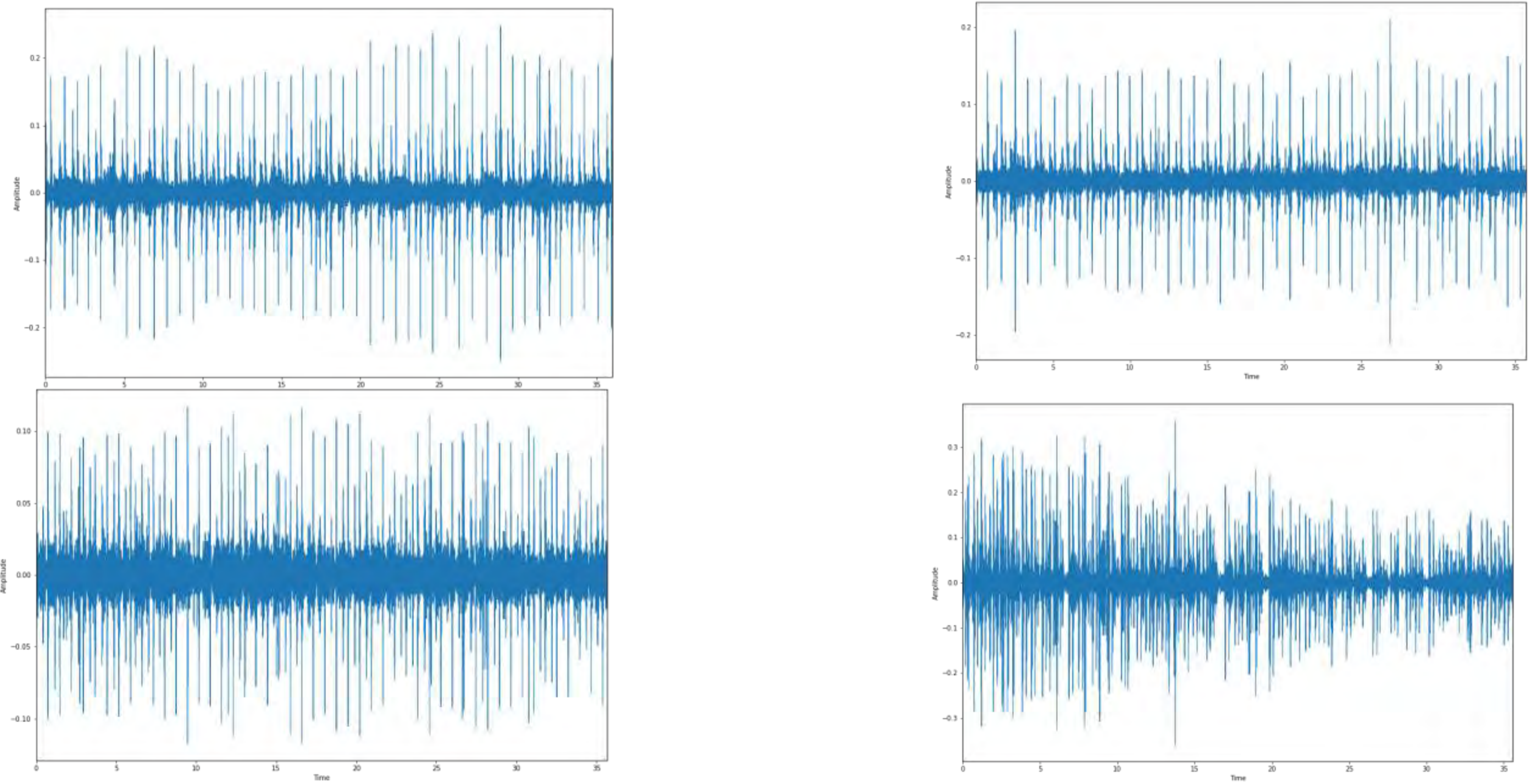
- Apply mel filter banks to spectrogram

Mel Filter Banks

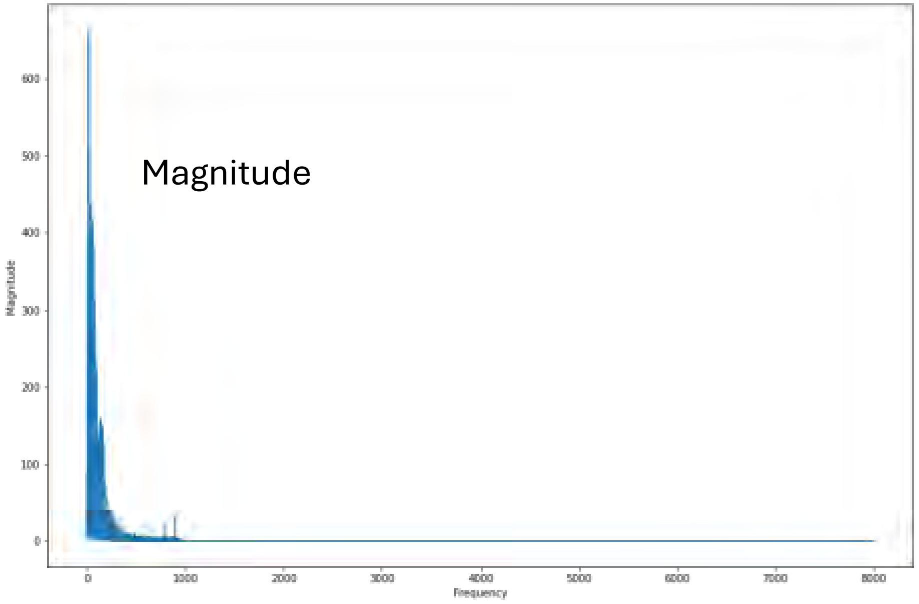
263 mel

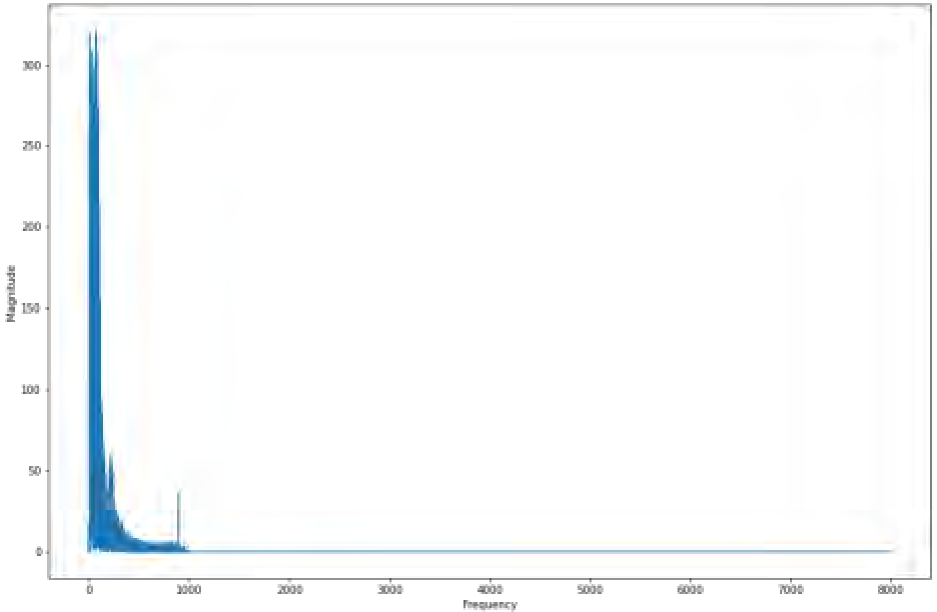


Abnormal phonocardiogram (PCG) (bottom) vs Normal PCG (top) [Waveform]

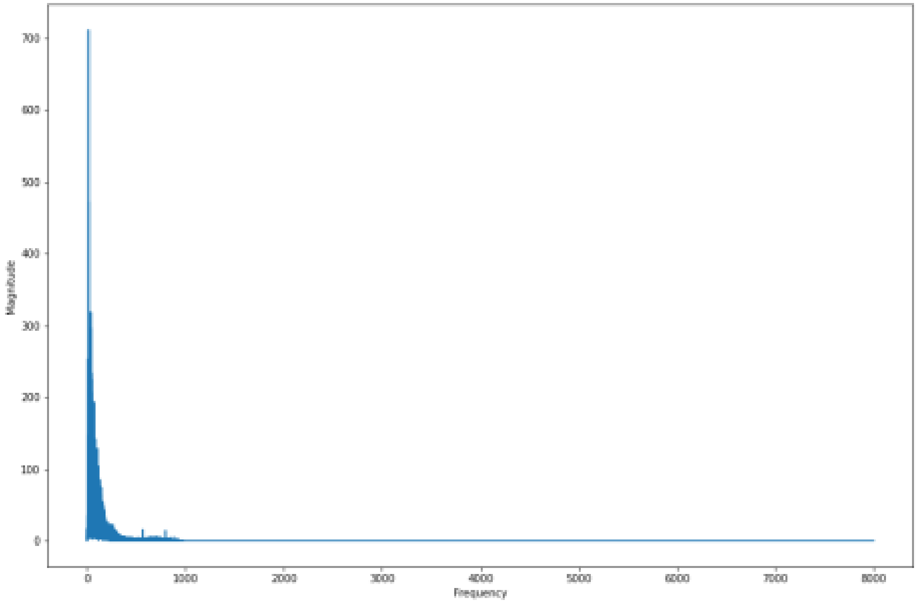


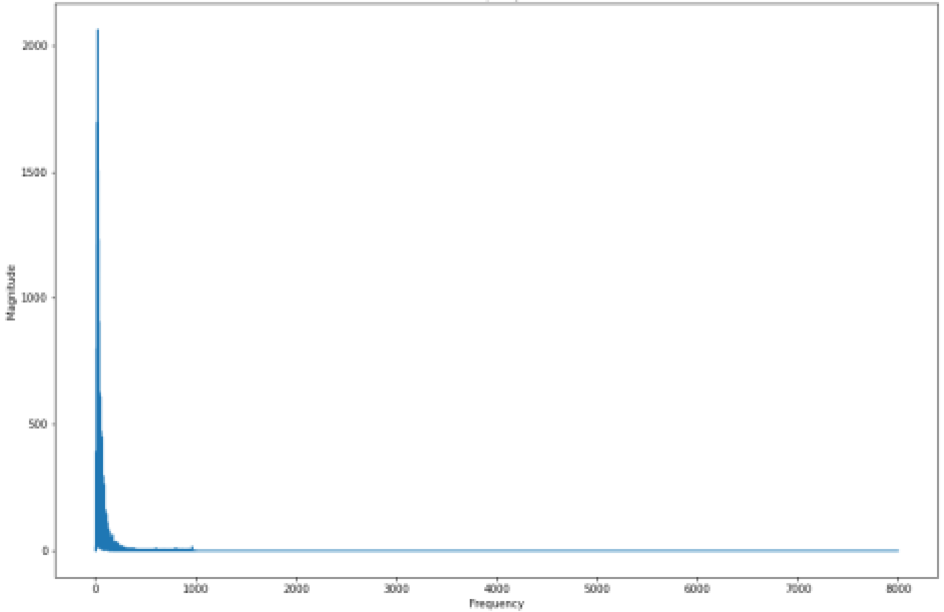
Abnormal PCG (bottom) vs Normal PCG (top) [Fast Fourier transform]

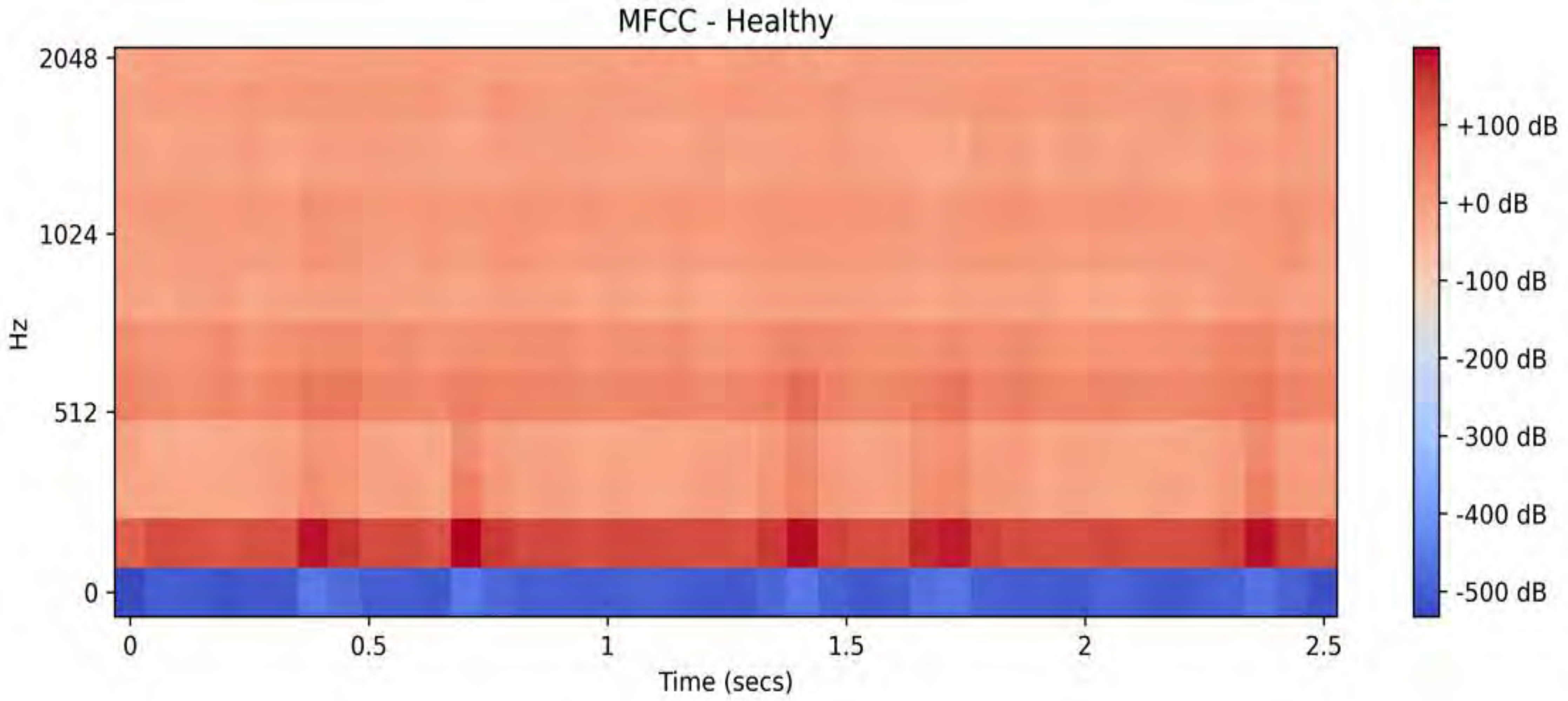


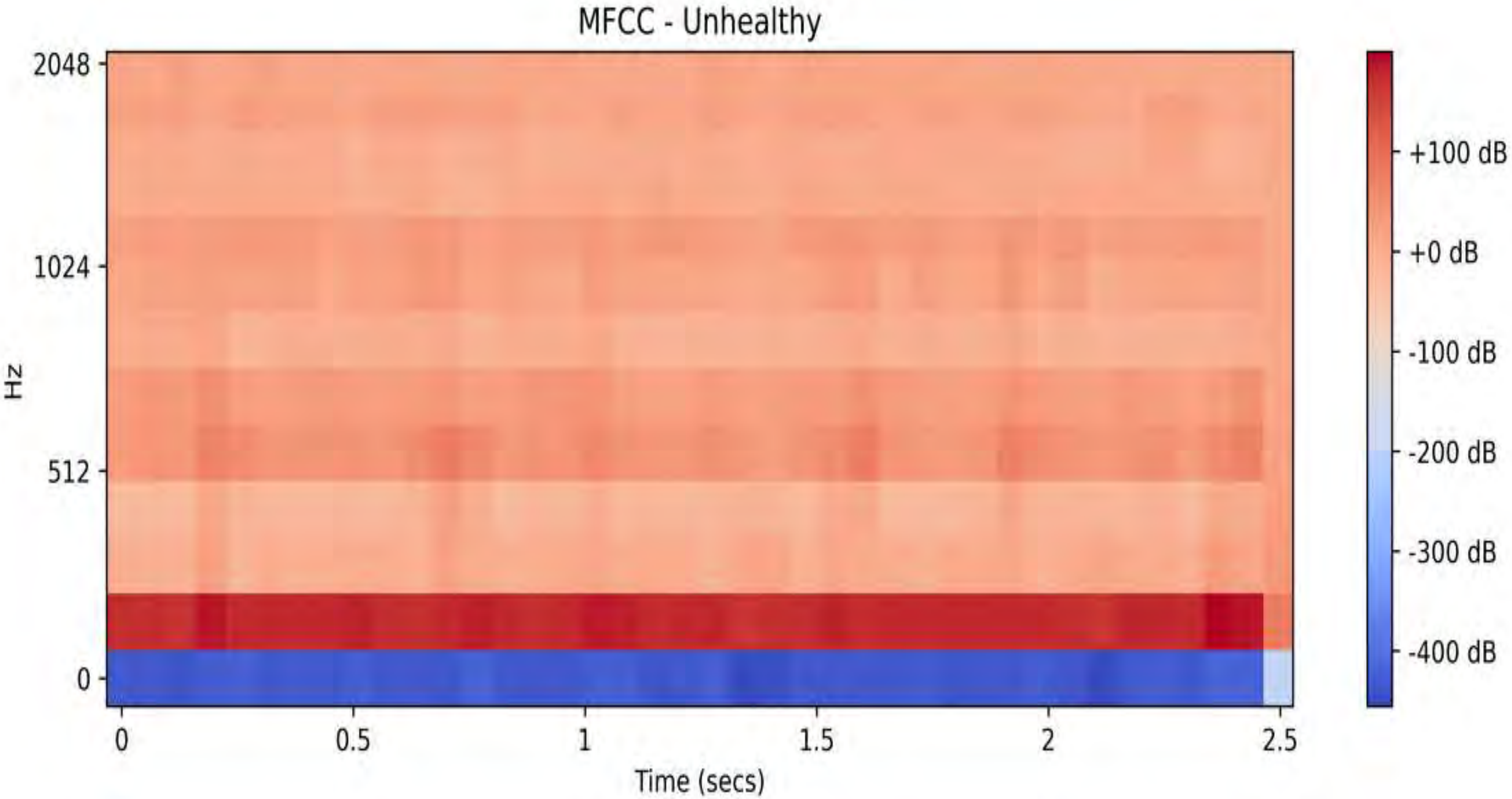


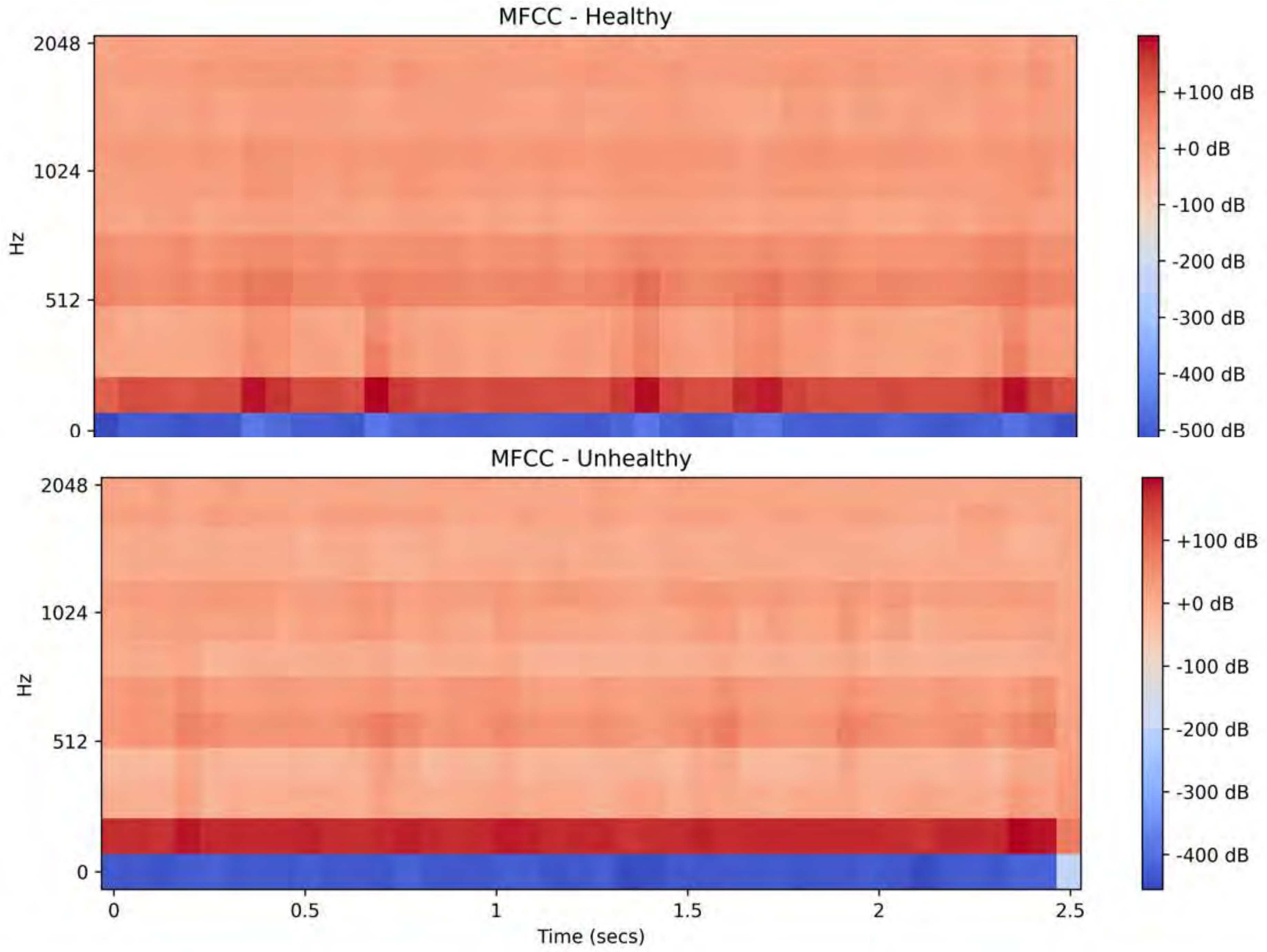
Frequency



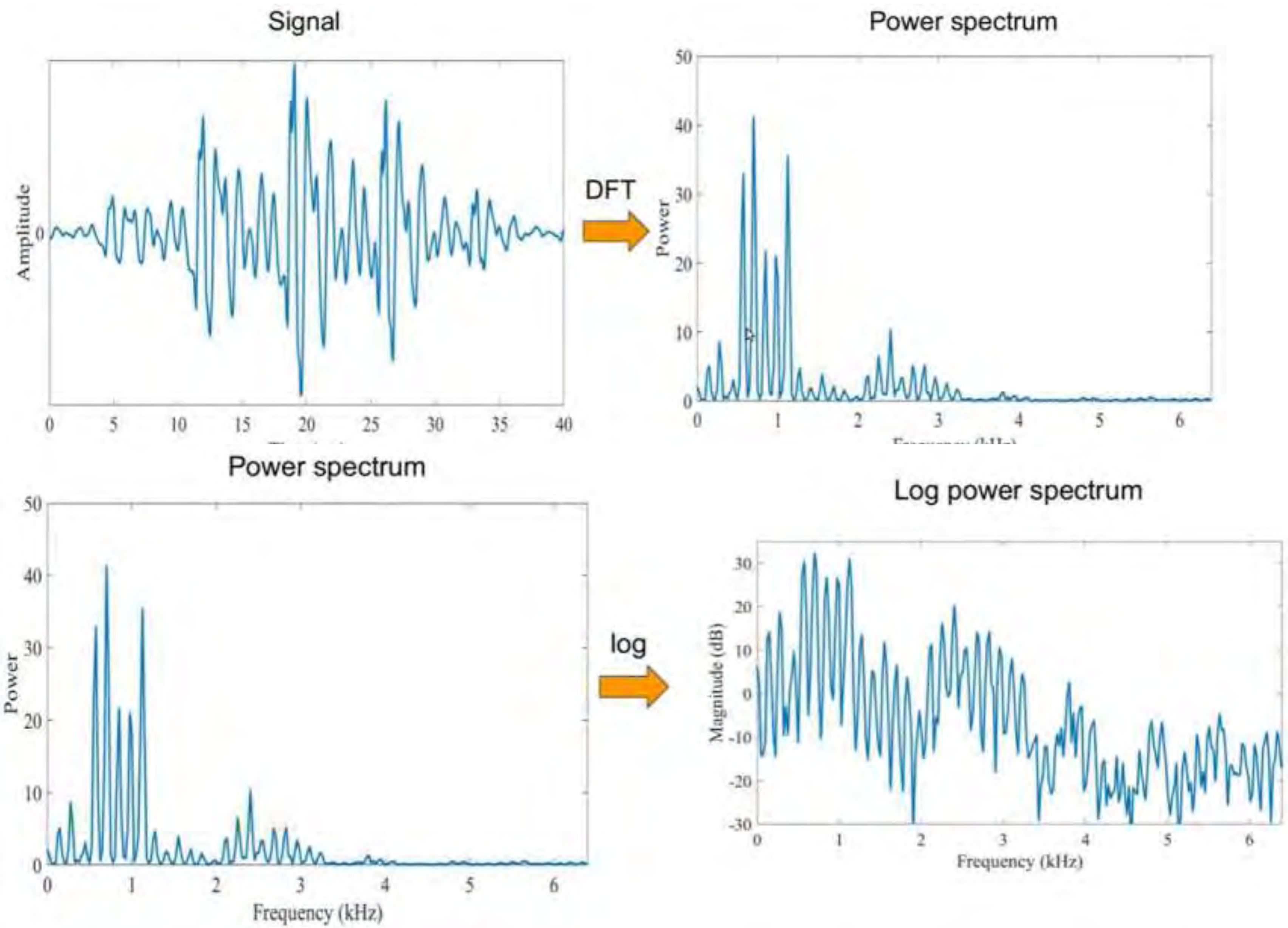



MFCC - Healthy
2048
1024
512
0
Hz
0
0.5
1
1.5
2
2.5
Time (secs)
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
-500 dB


MFCC - Unhealthy
2048
1024
512
0
Hz
0
0.5
1
1.5
2
2.5
Time (secs)
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB

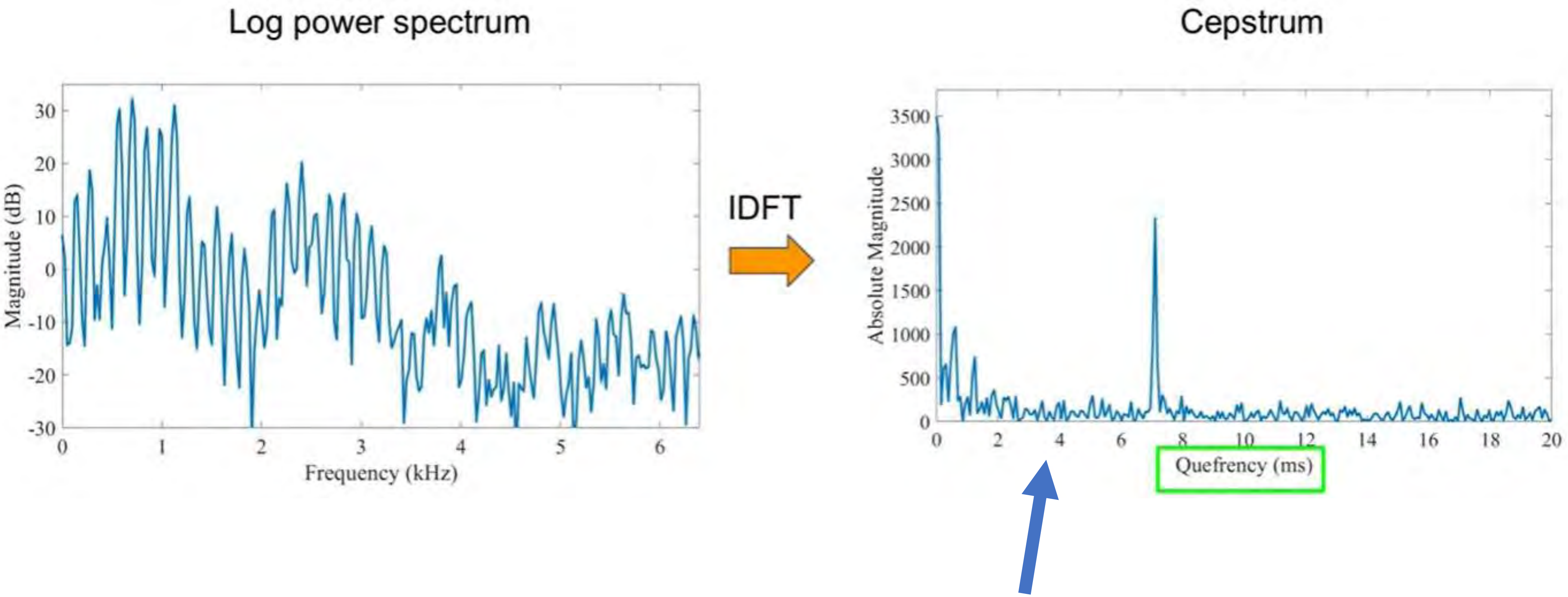

MFCC - Healthy
2048
1024
512
0
Hz
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
-500 dB
MFCC - Unhealthy
2048
1024
512
0
Hz
+100 dB
+0 dB
-100 dB
-200 dB
-300 dB
-400 dB
0
0.5
1
1.5
2
2.5
Time (secs)

Visualizing the Cepstrum



The energy of a time signal is distributed in the frequency domain.

Mel-Spectrograms



Log power spectrum

Cepstrum

Shows the parodic structure in frequency spectra.

Such effects are related to noticeable echoes or reflections in the signal, or to the occurrence of harmonic frequencies.

Classification of heart diseases using data mining techniques based on number features

- Used processed Hungarian medical dataset from UC-Irvine
- Data set has 76 attributes and 294 instances (data points)

- Characterized using 4 values:
  - Value 1: typical angina – 11 instances
  - Value 2: atypical angina – 106 instances
  - Value 3: non-angina pain – 54 instances
  - Value 4: asymptomatic – 123 instances

1) Selected different subsets of attributes to assess the accuracy in presence of heart disease
2) Determined the risk level and accuracy level of the selected attributes using ML approach
3) Developed a decision-making model for different risk levels with six different classifier models

- We used
  - Supervised learning - requires correctly labeled data with the correctly labeled results to train the algorithms before using.

  - Non-supervised method – used principal component analysis to reduce the dimensionality of the relevant data for selecting the attributes

Analysis

- First conducted a preliminary analysis to know complexity of data for classification and establish a baseline for the classification.

- Selected 5 different classifiers
- **Baseline classifier** – basic classifier that provides the majority class
  - Has no predictability power but is useful for establishing a baseline performance to serve as a benchmark.
  - Provides a sense of whether attributes are informative or not.
- **Bayesian classifier** –works on the principle of probability for the classification [12].
- **Decision tree** – uses tree data structure for the classification [6].
- **Nearest neighbor** – uses the knowledge of nearest data points to predict the output [11]
- **Support vector machine** – uses a separation line in a hyperplane such that there is a maximum separation between the different sets of data [13].

MLC: Bayesian (Multiclass)

- A classifier based on the Bayes probability algorithm
- Minimizes the probability of miscalculation

Strength

- Simple to implement and computationally is light
- Is a linear algorithm and does not involve iterative calculations.
- Is surprisingly accurate for a large set of problems, although its assumptions are not valid in most cases Ex: predictors are conditionally independent, or unrelated to any of the other feature in the model.
- Can be used to construct multi-layer decision trees with a Bayes classifier at every node.

Weakness

- Sensitive to the set of selected categories, which must be exhaustive.
- Accuracy can dramatically reduce if categories are overlapping or there are unknown categories

MLC: Decision Tree (multiclass)

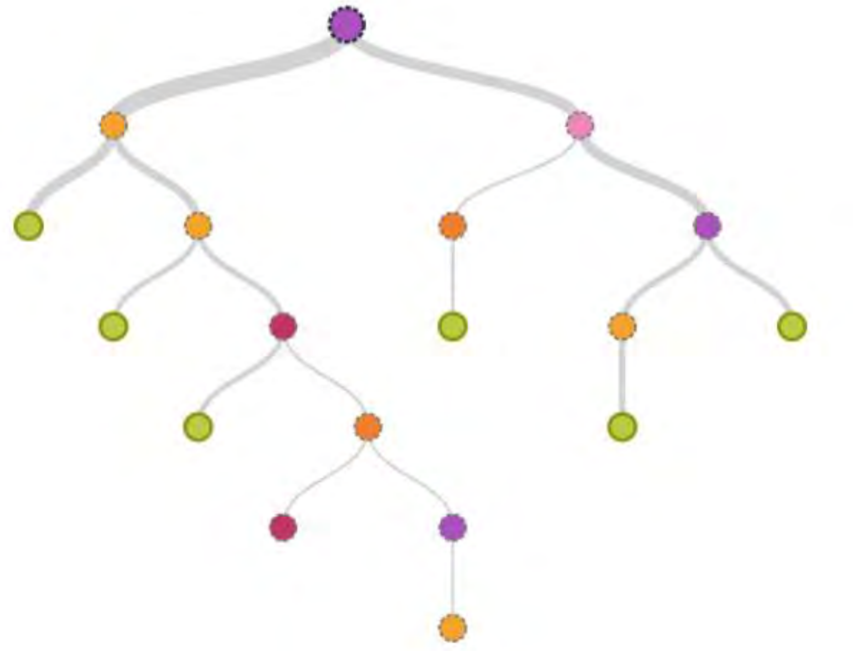
- Construct a tree with a set of “if-then” rules to classify data points.

- The “knowledge” learned by a decision tree during training is directly formulated into a hierarchical structure. This structure displays the knowledge in such a way that it can be easily understood, even by non-experts.

- Attributes at the top of the tree have a larger impact on the classification decision. The training process continues until it meets a termination condition.

Strengths - Able to model complex decision processes, and interpretation of results is very intuitive

Weaknesses - Can easily overfit the data by over-growing a tree with branches that reflect outliers in the data set.



- A way to deal with overfitting is pruning the model from growing unnecessary branches (pre-pruning), or removing them after the tree is grown (post-pruning).

MLC: K-Nearest Neighbor (multiclass)

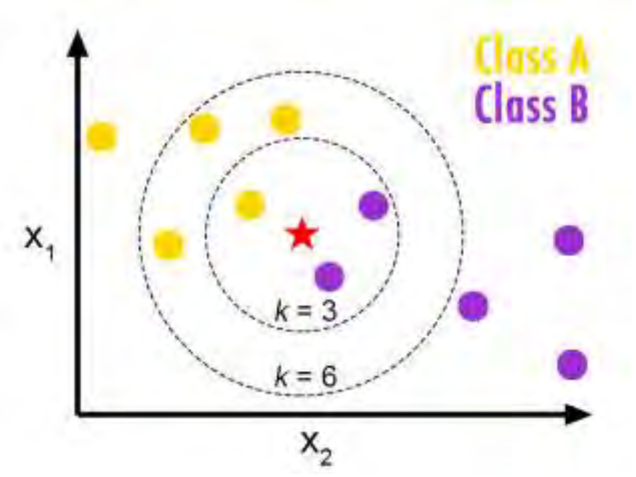
- Knowledge of nearest data points in training set is used to predict the output.
- The new data point is assigned the class most commonly found among its neighbors.
- The algorithm makes no assumptions on the underlying data and it does not pre-train (all training data is used during classification).

Strengths: Very simple to implement and understand

- Highly effective for many classification problems, especially with small number of features or input variables.

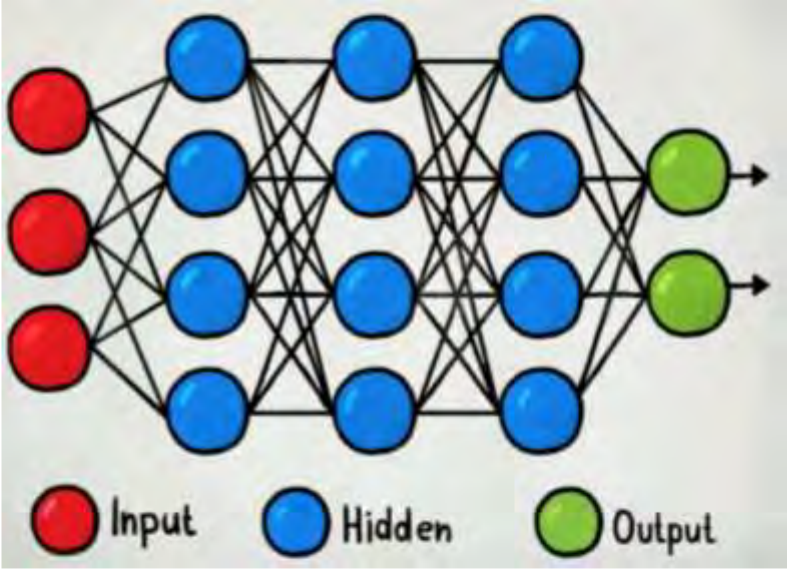
Weaknesses: KNN's accuracy is not comparable to that of supervised learning

- Not suitable for high dimensionality problems.
- Computationally intensive, especially with a large training set.



Simple and Complex NN Classifier (binary or multiclass)

- By constructing multiple layers of neurons, each of which receives part of the input variables, and then passes on its results to the next layers,
- The network can learn very complex functions.
- Theoretically, a neural network is capable of learning the shape of just any function, given enough computational power.



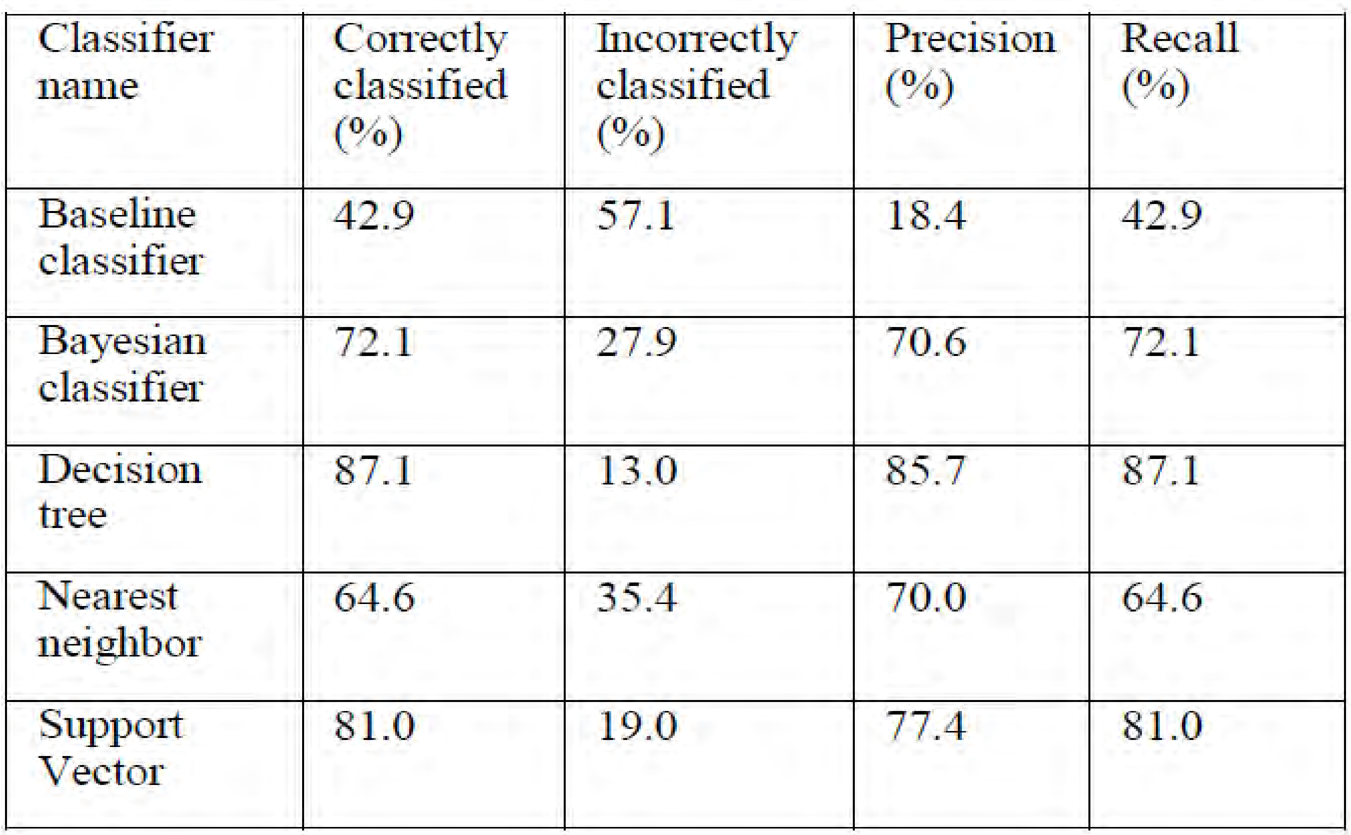
Strengths

- Very effective for high dimensionality problems
- Can deal with complex relations between variables
- Powerful tuning options to prevent over- and under-fitting.

Weaknesses

- Complex and difficult to implement.
- Non-intuitive, requires expertise to tune. In some cases, requires a large training set to be effective.

TABLE I Preliminary analysis of data



Classifier name	Correctly classified (%)	Incorrectly classified (%)	Precision (%)	Recall (%)
Baseline classifier	42.9	57.1	18.4	42.9
Bayesian classifier	72.1	27.9	70.6	72.1
Decision tree	87.1	13.0	85.7	87.1
Nearest neighbor	64.6	35.4	70.0	64.6
Support Vector	81.0	19.0	77.4	81.0

Used WEKA classifier[10].

- Baseline classifier is used to determine nature of the data separation
- Precision measures the fraction of classified heart disease instances that are **correct** **Precision = true positive/ (true positive + false positive.**
- Recall measures the fraction of actual heart disease instances which are **correctly predicted** (**Recall - true positive/(true positive + false negative**) [18].

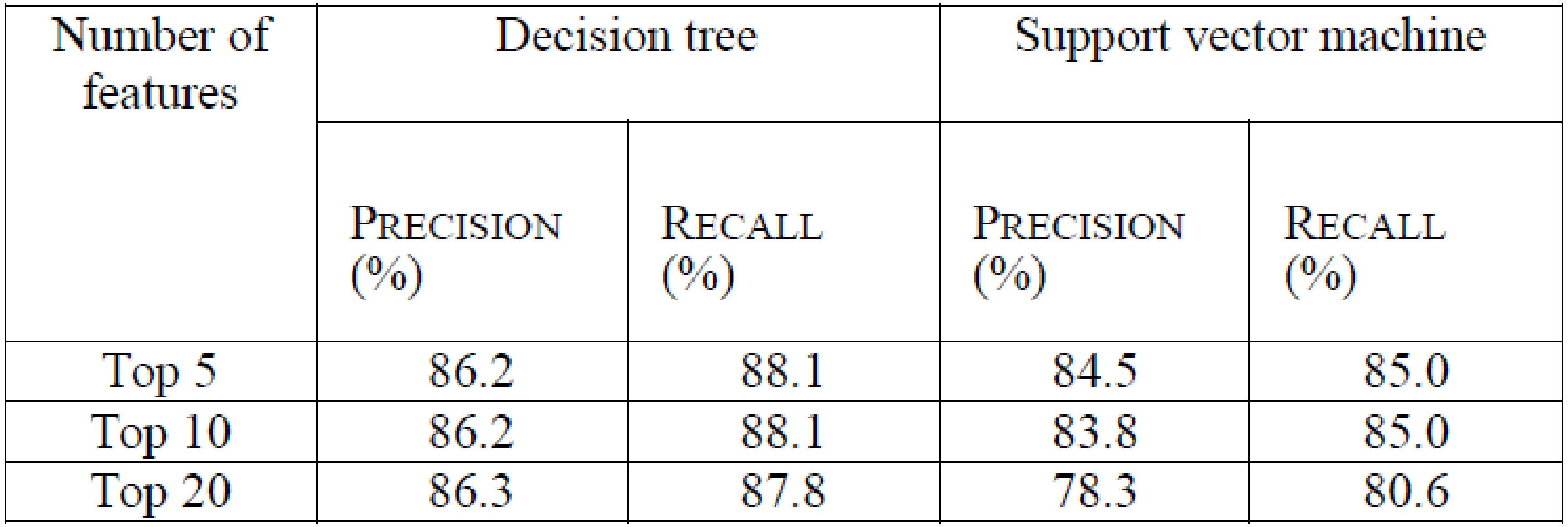
Attribute Selection

- The purpose was:
  1. To remove attributes that do not contribute to the improvement of the classification
  2. To speed up the classification process.

- Entropy is used to determine how much each attribute contributes to the classification
  1. Entropy - a metric to quantify the average amount of information in a dataset
  2. Information gain - a measure to determine which feature should be used to split the data at each internal node of the decision tree
- Compared the classification based on 5, 10, and 20 attributes

TABLE II

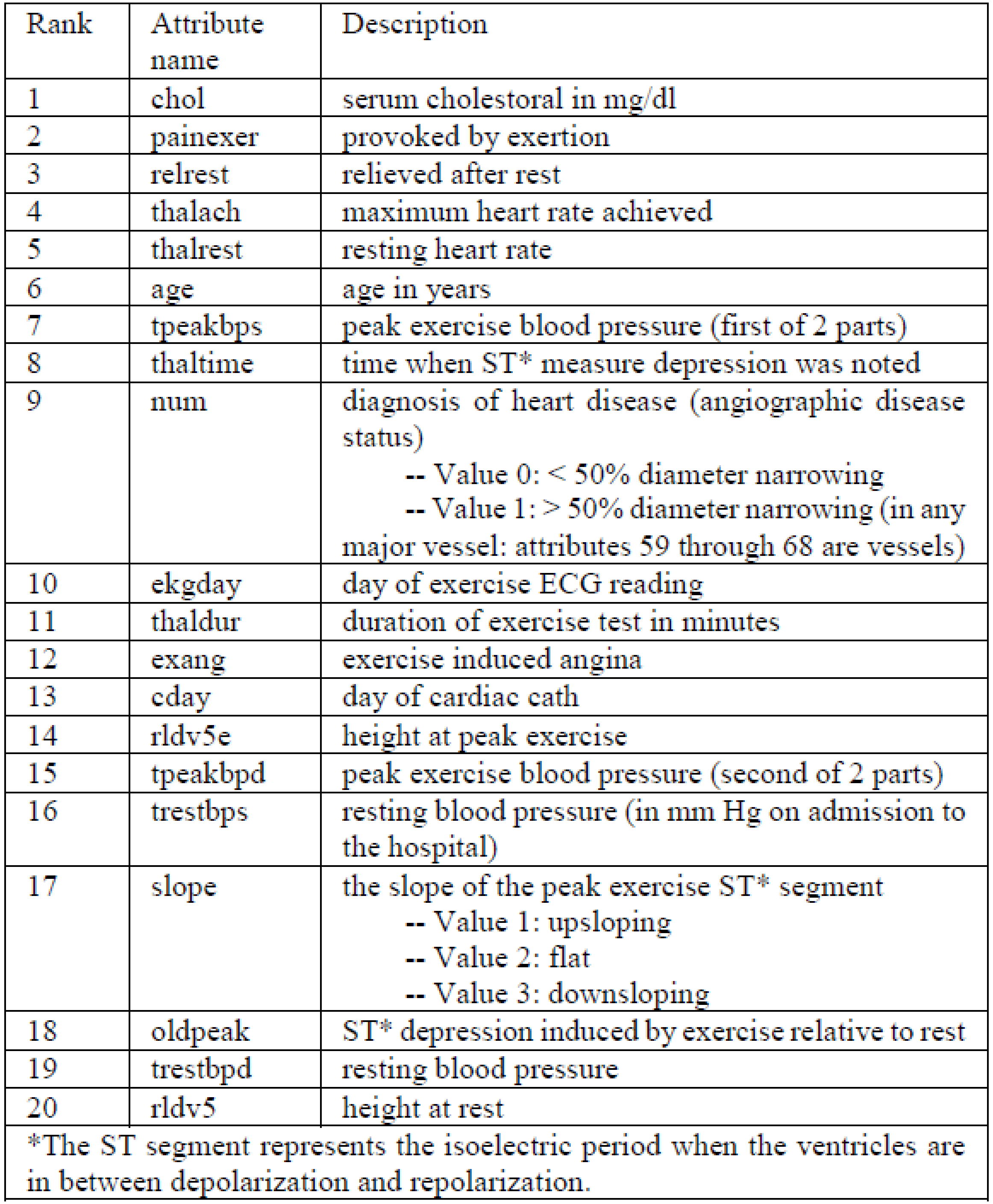
ENTROPY BASED PRELIMINARY CLASSIFICATION RESULTS



Number of features	Decision tree		Support vector machine	
	PRECISION (%)	RECALL (%)	PRECISION (%)	RECALL (%)
Top 5	86.2	88.1	84.5	85.0
Top 10	86.2	88.1	83.8	85.0
Top 20	86.3	87.8	78.3	80.6

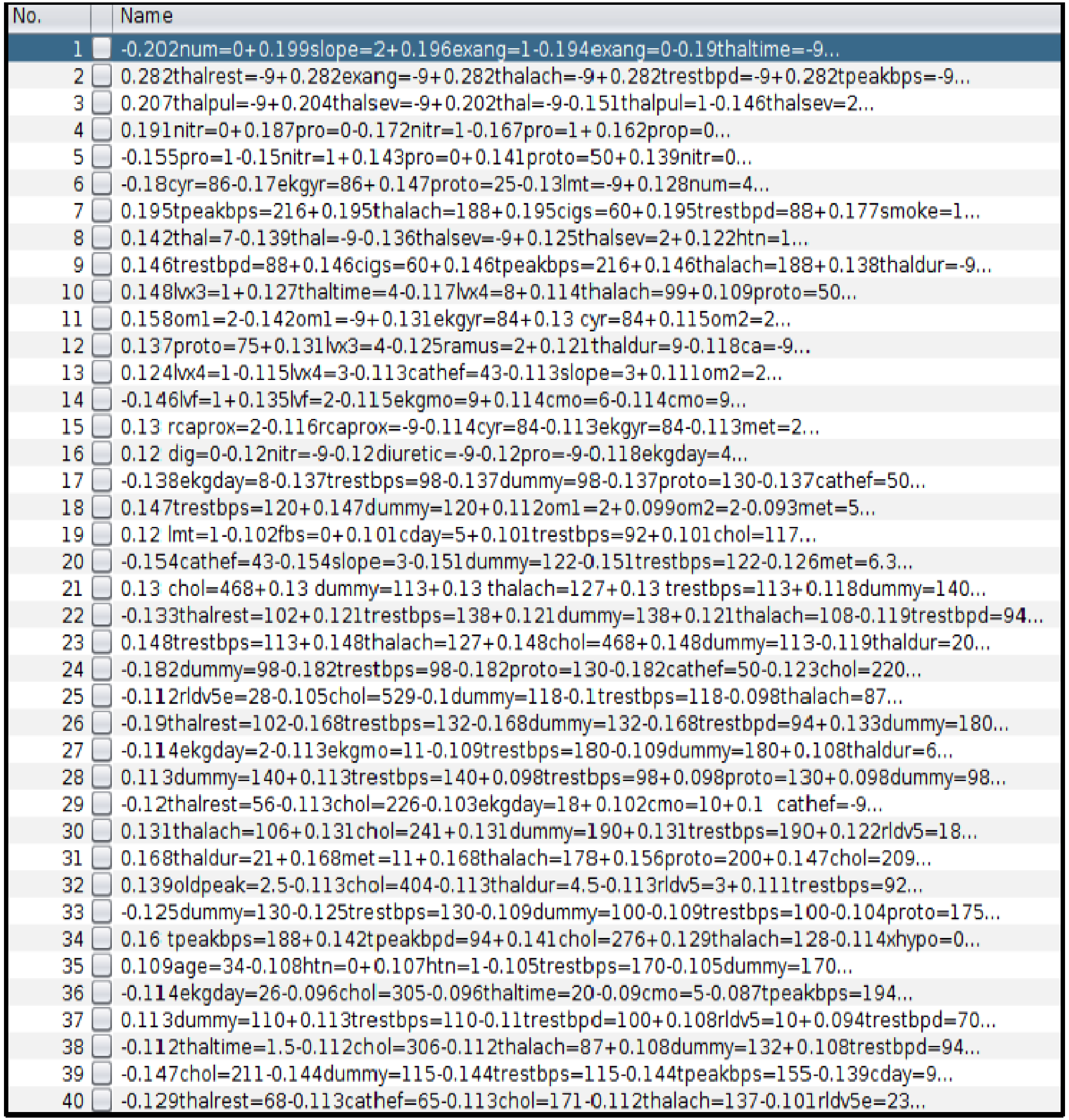
Attributes were selected based on the contribution of each attribute to the classification, we used entropy.

TABLE III Top 20 Feature attributes Based on PCA



Rank	Attribute name	Description
1	chol	serum cholestoral in mg/dl
2	painexer	provoked by exertion
3	relrest	relieved after rest
4	thalach	maximum heart rate achieved
5	thalrest	resting heart rate
6	age	age in years
7	tpeakbps	peak exercise blood pressure (first of 2 parts)
8	thaltime	time when ST* measure depression was noted
9	num	diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)
10	ekgday	day of exercise ECG reading
11	thaldur	duration of exercise test in minutes
12	exang	exercise induced angina
13	cday	day of cardiac cath
14	rldv5e	height at peak exercise
15	tpeakbpd	peak exercise blood pressure (second of 2 parts)
16	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
17	slope	the slope of the peak exercise ST* segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
18	oldpeak	ST* depression induced by exercise relative to rest
19	trestbpd	resting blood pressure
20	rldv5	height at rest

*The ST segment represents the isoelectric period when the ventricles are in between depolarization and repolarization.

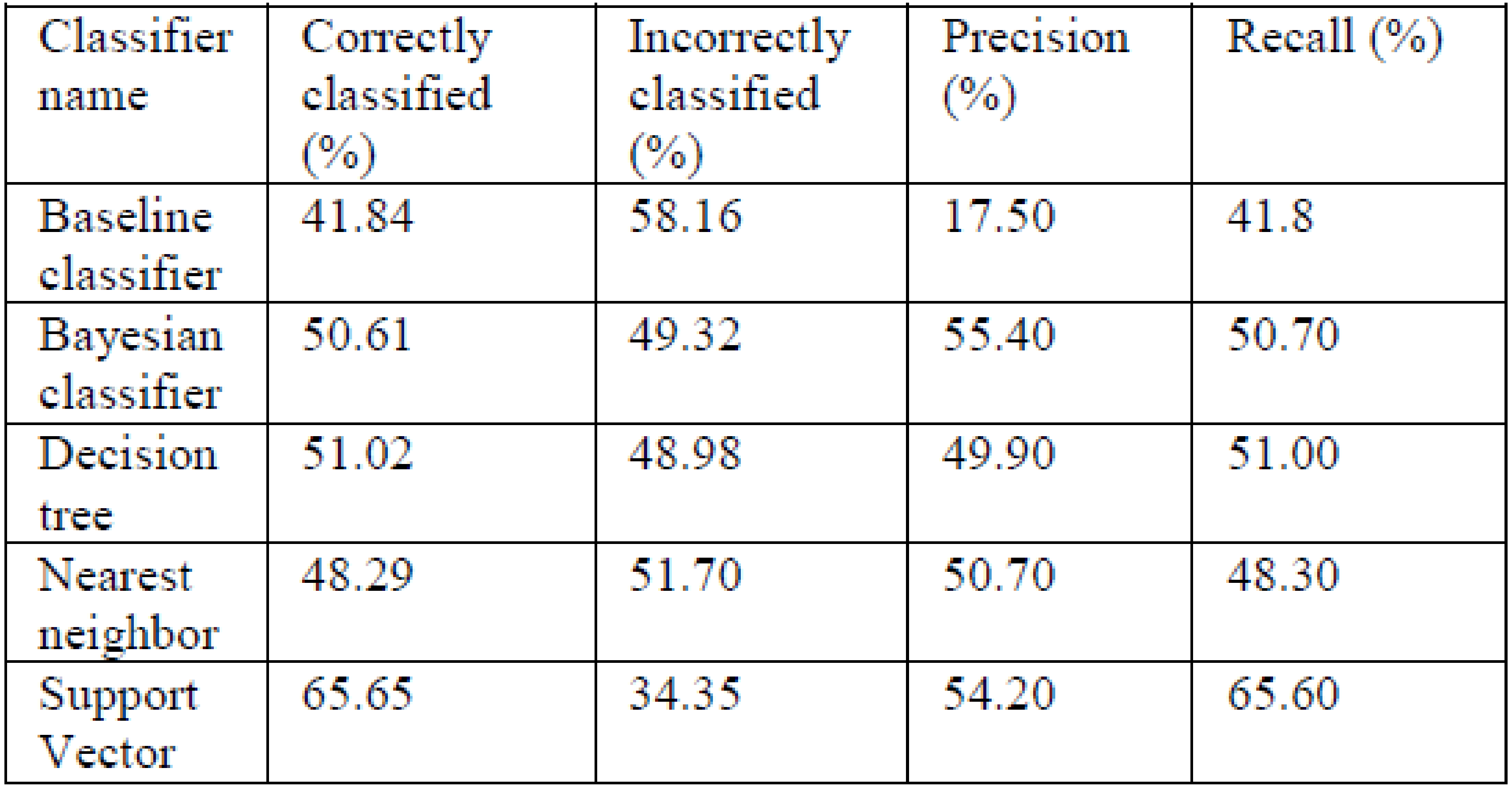


No.		Name
1	□	-0.202num=0+0.199slope=2+0.196exang=1-0.194exang=0-0.19thaltime=-9...
2	□	0.282thalrest=-9+0.282exang=-9+0.282thalach=-9+0.282trestbpd=-9+0.282tpeakbps=-9...
3	□	0.207thalpul=-9+0.204thalsev=-9+0.202thal=-9-0.151thalpul=1-0.146thalsev=2...
4	□	0.191nitr=0+0.187pro=0-0.172nitr=1-0.167pro=1+0.162prop=0...
5	□	-0.155pro=1-0.15nitr=1+0.143pro=0+0.141proto=50+0.139nitr=0...
6	□	-0.18cyr=86-0.17ekgyr=86+0.147proto=25-0.13lmt=-9+0.128num=4...
7	□	0.195tpeakbps=216+0.195thalach=188+0.195cigs=60+0.195trestbpd=88+0.177smoke=1...
8	□	0.142thal=7-0.139thal=-9-0.136thalsev=-9+0.125thalsev=2+0.122htn=1...
9	□	0.146trestbpd=88+0.146cigs=60+0.146tpeakbps=216+0.146thalach=188+0.138thaldur=-9...
10	□	0.148lvx3=1+0.127thaltime=4-0.117lvx4=8+0.114thalach=99+0.109proto=50...
11	□	0.158om1=2-0.142om1=-9+0.131ekgyr=84+0.13 cyr=84+0.115om2=2...
12	□	0.137proto=75+0.131lvx3=4-0.125ramus=2+0.121thaldur=9-0.118ca=-9...
13	□	0.124lvx4=1-0.115lvx4=3-0.113cathef=43-0.113slope=3+0.111om2=2...
14	□	-0.146lvf=1+0.135lvf=2-0.115ekgmo=9+0.114cmo=6-0.114cmo=9...
15	□	0.13 rcaprox=2-0.116rcaprox=-9-0.114cyr=84-0.113ekgyr=84-0.113met=2...
16	□	0.12 dig=0-0.12nitr=-9-0.12diuretic=-9-0.12pro=-9-0.118ekgday=4...
17	□	-0.138ekgday=8-0.137trestbps=98-0.137dummy=98-0.137proto=130-0.137cathef=50...
18	□	0.147trestbps=120+0.147dummy=120+0.112om1=2+0.099om2=2-0.093met=5...
19	□	0.12 lmt=1-0.102fbs=0+0.101cday=5+0.101trestbps=92+0.101chol=117...
20	□	-0.154cathef=43-0.154slope=3-0.151dummy=122-0.151trestbps=122-0.126met=6.3...
21	□	0.13 chol=468+0.13 dummy=113+0.13 thalach=127+0.13 trestbps=113+0.118dummy=140...
22	□	-0.133thalrest=102+0.121trestbps=138+0.121dummy=138+0.121thalach=108-0.119trestbpd=94...
23	□	0.148trestbps=113+0.148thalach=127+0.148chol=468+0.148dummy=113-0.119thaldur=20...
24	□	-0.182dummy=98-0.182trestbps=98-0.182proto=130-0.182cathef=50-0.123chol=220...
25	□	-0.112rldv5e=28-0.105chol=529-0.1dummy=118-0.1trestbps=118-0.098thalach=87...
26	□	-0.19thalrest=102-0.168trestbps=132-0.168dummy=132-0.168trestbpd=94+0.133dummy=180...
27	□	-0.114ekgday=2-0.113ekgmo=11-0.109trestbps=180-0.109dummy=180+0.108thaldur=6...
28	□	0.113dummy=140+0.113trestbps=140+0.098trestbps=98+0.098proto=130+0.098dummy=98...
29	□	-0.12thalrest=56-0.113chol=226-0.103ekgday=18+0.102cmo=10+0.1 cathef=-9...
30	□	0.131thalach=106+0.131chol=241+0.131dummy=190+0.131trestbps=190+0.122rldv5=18...
31	□	0.168thaldur=21+0.168met=11+0.168thalach=178+0.156proto=200+0.147chol=209...
32	□	0.139oldpeak=2.5-0.113chol=404-0.113thaldur=4.5-0.113rldv5=3+0.111trestbps=92...
33	□	-0.125dummy=130-0.125trestbps=130-0.109dummy=100-0.109trestbps=100-0.104proto=175...
34	□	0.16 tpeakbps=188+0.142tpeakbpd=94+0.141chol=276+0.129thalach=128-0.114xhypo=0...
35	□	0.109age=34-0.108htn=0+0.107htn=1-0.105trestbps=170-0.105dummy=170...
36	□	-0.114ekgday=26-0.096chol=305-0.096thaltime=20-0.09cmo=5-0.087tpeakbps=194...
37	□	0.113dummy=110+0.113trestbps=110-0.11trestbpd=100+0.108rldv5=10+0.094trestbpd=70...
38	□	-0.112thaltime=1.5-0.112chol=306-0.112thalach=87+0.108dummy=132+0.108trestbpd=94...
39	□	-0.147chol=211-0.144dummy=115-0.144trestbps=115-0.144tpeakbps=155-0.139cday=9...
40	□	-0.129thalrest=68-0.113cathef=65-0.113chol=171-0.112thalach=137-0.101rldv5e=23...

Figure 2. PCA based top 40 attributes

TABLE IV

PCA BASED PRELIMINARY CLASSIFICATION RESULTS

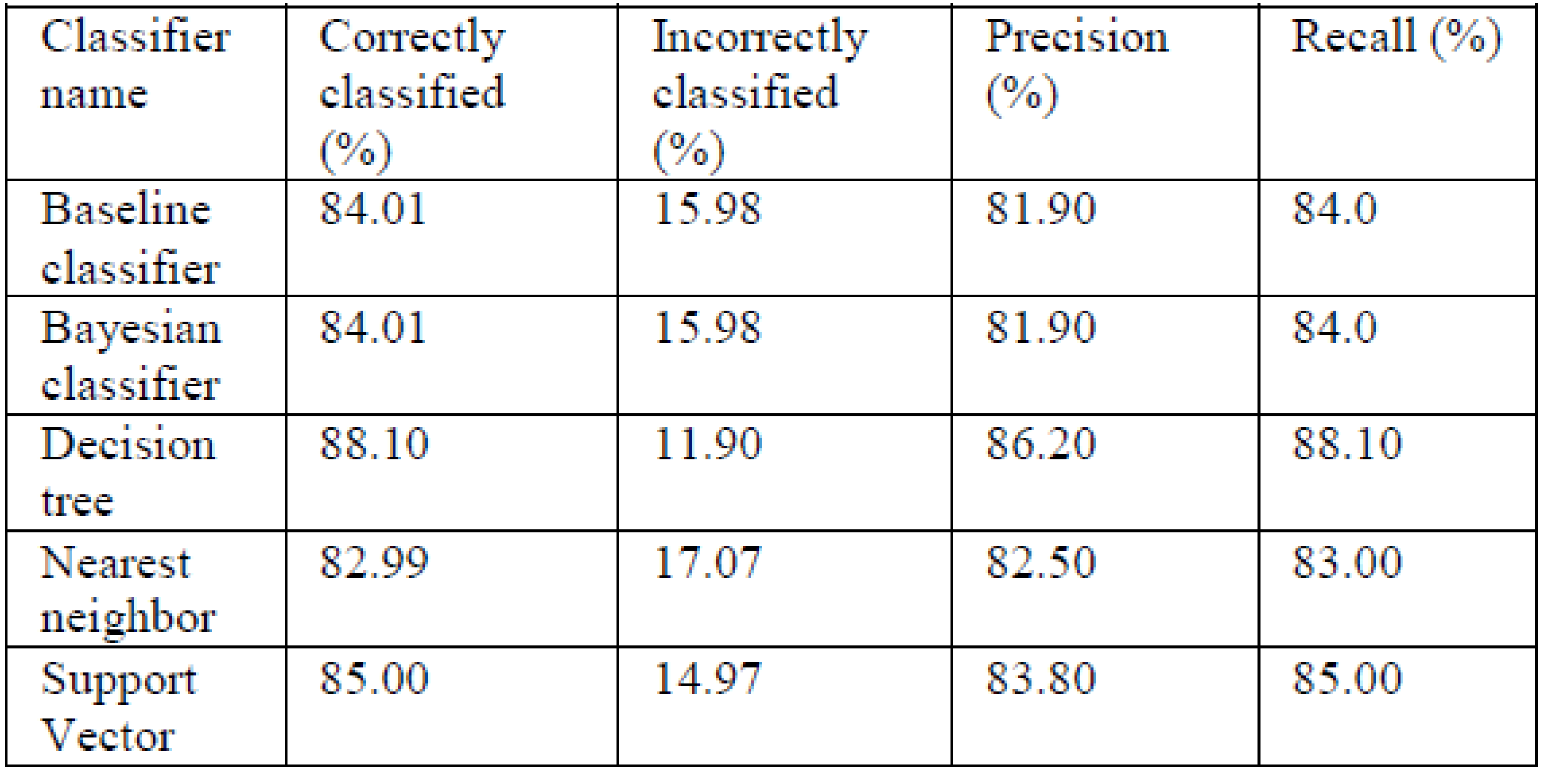


Classifier name	Correctly classified (%)	Incorrectly classified (%)	Precision (%)	Recall (%)
Baseline classifier	41.84	58.16	17.50	41.8
Bayesian classifier	50.61	49.32	55.40	50.70
Decision tree	51.02	48.98	49.90	51.00
Nearest neighbor	48.29	51.70	50.70	48.30
Support Vector	65.65	34.35	54.20	65.60

- Precision measures the fraction of classified heart disease instances that are **correct.**
- Recall measures the fraction of actual heart disease instances which are **correctly predicted**.

TABEL V

HEART DISEASE DETECTION RESULTS USING 10-FOLD CROSS-VALIDATION



Classifier name	Correctly classified (%)	Incorrectly classified (%)	Precision (%)	Recall (%)
Baseline classifier	84.01	15.98	81.90	84.0
Bayesian classifier	84.01	15.98	81.90	84.0
Decision tree	88.10	11.90	86.20	88.10
Nearest neighbor	82.99	17.07	82.50	83.00
Support Vector	85.00	14.97	83.80	85.00

- 9 parts used for training and part 10 for testing.
- This process is repeated until all the parts are tested.
- The results in each iteration is averaged to get the final classification accuracy.
- Each iteration is independent (that is knowledge of previous training is not retained for the next iteration)

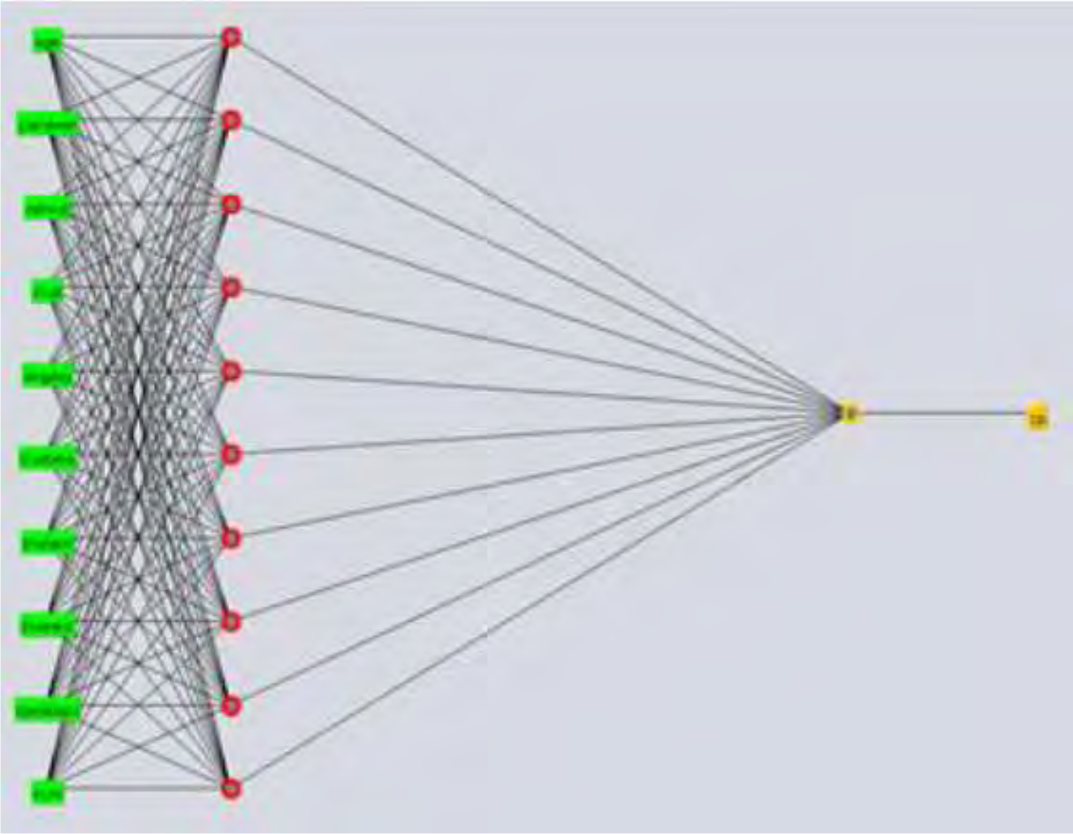


Fig. 3. One hidden layer neural network configuration

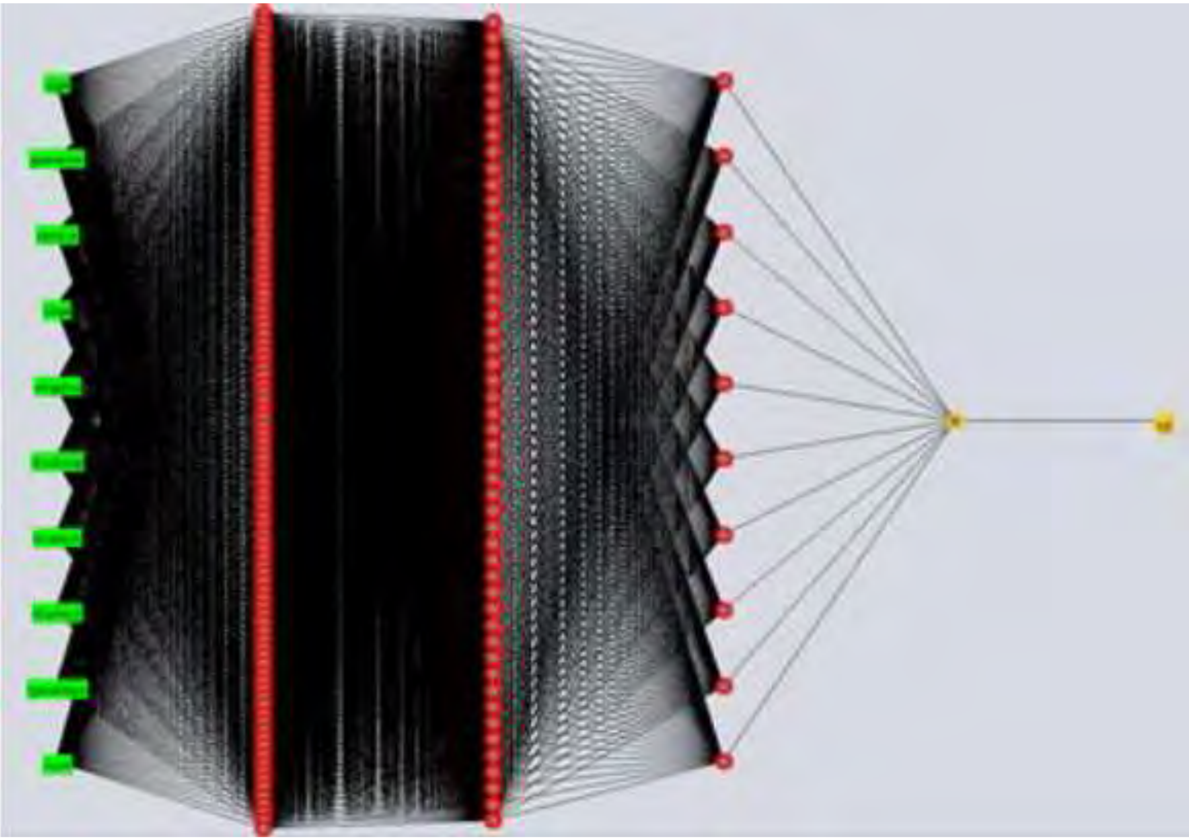
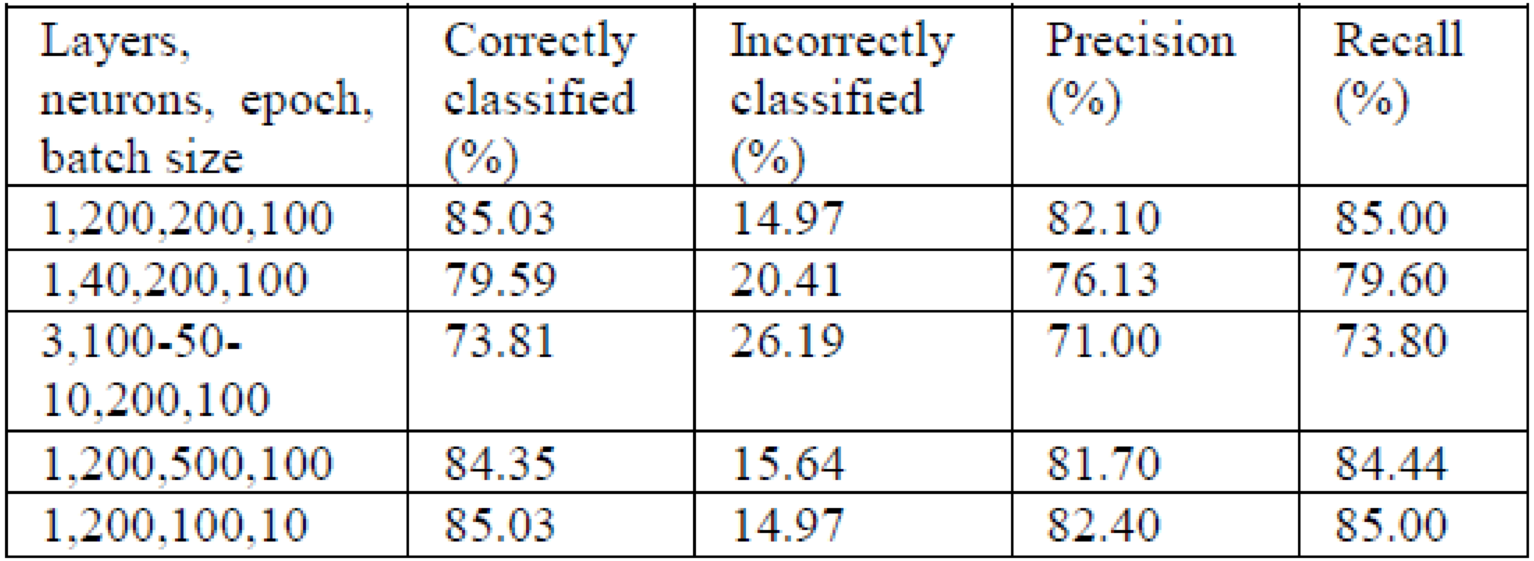


Fig. 4. Three hidden layer neural network configuration

TABLE VI

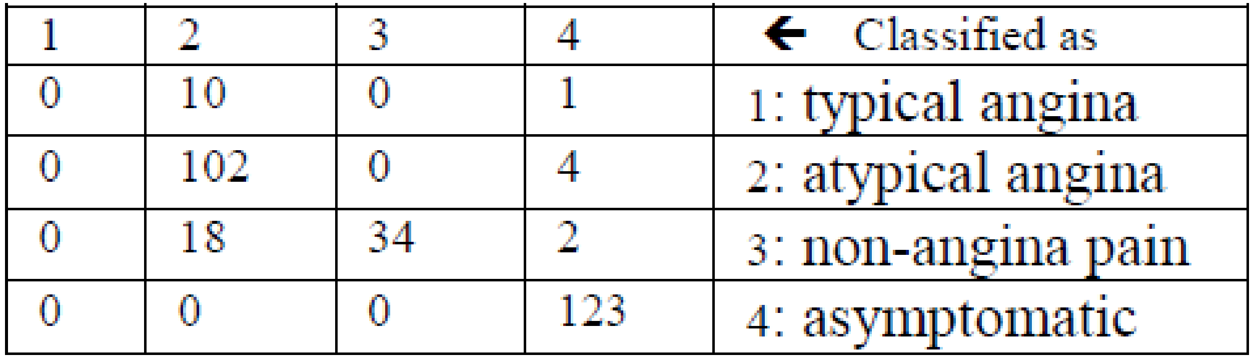
NEURAL NETWORK BASED CLASSIFICATION RESULTS WITH TOP 20 ATTRIBUTES



Layers, neurons, epoch, batch size	Correctly classified (%)	Incorrectly classified (%)	Precision (%)	Recall (%)
1,200,200,100	85.03	14.97	82.10	85.00
1,40,200,100	79.59	20.41	76.13	79.60
3,100-50-10,200,100	73.81	26.19	71.00	73.80
1,200,500,100	84.35	15.64	81.70	84.44
1,200,100,10	85.03	14.97	82.40	85.00

Started with default values from WEKA, then increased the values of each parameter until there is change and stopped tuning it if there was a decrease in classification performance.

TABLE VII CONFUSION MATRIX FOR DECISION TREE RESULTS



1	2	3	4	← Classified as
0	10	0	1	1: typical angina
0	102	0	4	2: atypical angina
0	18	34	2	3: non-angina pain
0	0	0	123	4: asymptomatic

Value 1: typical angina – 11 instances
Value 2: atypical angina – 106 instances
Value 3: non-angina pain – 54 instances
Value 4: asymptomatic – 123 instances

Reason for misclassifications: May be related to particular age group! The dataset had person ages ranging from 28 to 66.

Misclassified 11 instances of typical angina category did not have any special age group, they were random.

Conclusion

- Using WEKA analyzed processed Hungarian medical data from UCI

- It was shown that the decision tree and support vector machine models achieved correct classification rates of 88.1% and 88.5% with precision rates of 86.20% and 83.80%, respectively.

- This is about 10% or better correct classification compared to the other methods. The classification accuracy of the neural network with 100 to 200 neurons varied between 73.81% to 85.03%.

Thank you!

Data Set Information:

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to
this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

To see Test Costs (donated by Peter Turney), please see the folder "Costs"

Attribute Information:

Only 14 attributes used:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

1 id: patient identification number
2 ccf: social security number (I replaced this with a dummy value of 0)
3 age: age in years
4 sex: sex (1 = male; 0 = female)
5 painloc: chest pain location (1 = substernal; 0 = otherwise)
6 painexer (1 = provoked by exertion; 0 = otherwise)
7 relrest (1 = relieved after rest; 0 = otherwise)
8 pncaden (sum of 5, 6, and 7)
9 cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)
11 htn
12 chol: serum cholestoral in mg/dl
13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14 cigs (cigarettes per day)
15 years (number of years as a smoker)
16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17 dm (1 = history of diabetes; 0 = no such history)
18 famhist: family history of coronary artery disease (1 = yes; 0 = no)

19 restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
20 ekgmo (month of exercise ECG reading)
21 ekgday(day of exercise ECG reading)
22 ekgyr (year of exercise ECG reading)
23 dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)
24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27 diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
28 proto: exercise protocol
1 = Bruce
2 = Kottus
3 = McHenry
4 = fast Balke
5 = Balke
6 = Noughton

7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)
8 = bike 125 kpa min/min
9 = bike 100 kpa min/min
10 = bike 75 kpa min/min
11 = bike 50 kpa min/min
12 = arm ergometer
29 thaldur: duration of exercise test in minutes
30 thaltime: time when ST measure depression was noted
31 met: mets achieved
32 thalach: maximum heart rate achieved
33 thalrest: resting heart rate
34 tpeakbps: peak exercise blood pressure (first of 2 parts)
35 tpeakbpd: peak exercise blood pressure (second of 2 parts)
36 dummy
37 trestbpd: resting blood pressure
38 exang: exercise induced angina (1 = yes; 0 = no)
39 xhypo: (1 = yes; 0 = no)
40 oldpeak = ST depression induced by exercise relative to rest
41 slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
42 rldv5: height at rest

43 rldv5e: height at peak exercise
44 ca: number of major vessels (0-3) colored by flourosopy
45 restckm: irrelevant
46 exerckm: irrelevant
47 restef: rest raidonuclid (sp?) ejection fraction
48 restwm: rest wall (sp?) motion abnormality
0 = none
1 = mild or moderate
2 = moderate or severe
3 = akinesis or dyskmem (sp?)
49 exeref: exercise radinalid (sp?) ejection fraction
50 exerwm: exercise wall (sp?) motion
51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
52 thalsev: not used
53 thalpul: not used
54 earlobe: not used
55 cmo: month of cardiac cath (sp?) (perhaps "call")
56 cday: day of cardiac cath (sp?)
57 cyr: year of cardiac cath (sp?)
58 num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

(in any major vessel: attributes 59 through 68 are vessels)

59 lmt

60 ladprox

61 laddist

62 diag

63 cxmain

64 ramus

65 om1

66 om2

67 rcaprox

68 rcadist

69 lvx1: not used

70 lvx2: not used

71 lvx3: not used

72 lvx4: not used

73 lvf: not used

74 cathef: not used

75 junk: not used

76 name: last name of patient (I replaced this with the dummy string "name")

References

- http://www.lc.leidenuniv.nl/lc/web/2005/160/info.php3?wsid=160 (Workshop “Model OrderReduction, Coupled Problems and Optimization”)
- http://web.mit.edu/mor/ (Model Order Reduction website at MIT)
- http://www.imtek.de/simulation/index.php?page=http://www.imtek.uni-freiburg.de/simulation/benchmark/ (Oberwolfach Model Reduction Benchmark Collection)
- Model order reduction page at Institut für Automatisierungstechnik, University of Bremen.
- A very big collection of control-related aricles and theses of the Control Group at the University of Cambridge, UK.
- Collection of the Model Order Reduction benchmarks for linear and nonlinear problems at the University of Freiburg, Germany.
- Another benchmark collection for model reduction from the Niconet web site.
- Course material for "Dynamic systems and control" (6.241) course at MIT; essential for understanding dynamic systems theory.

1. D. Aha, D. Kibler, "Instance-based prediction of heart-disease presence with the Cleveland database", Technical Report, University of California, Irvine, Department of Information and Computer Science, Number ICS-TR-88-07, 1988.

2. D. Aha, D. Kibler, "Instance-based learning algorithms", Machine Learning. 6:37-66, 1991).

3. C.L. Blake and C.J. Merz, "UCI Repository of machine learning databases", Irvine, CA, 1998. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/heart+disease.

4. G. Dangi et al, "A Smart Approach to Diagnose Heart Disease through Machine Learning and Springleaf Marketing Response", IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2016), December 23-25, 2016.

5. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, International application of a new probability algorithm for the diagnosis of coronary artery disease, American Journal of Cardiology, 64, pp. 304-310, 1989.

6. B. H. Edmonds, "Using localised 'Gossip' to structure distributed learning, centre for policy modelling, Proceedings of the joint symposium on socially inspired computing engineering with social metaphors", AISB 127–134, University of Hertfordshire, Hatfield, UK, 2005.

1. A. L. Eikendal, K.A, Groenewegen, M.L. Bots, S.A. Peters, C.S. Uiterwaal, and H.M. den Ruijter, “Relation Between Adolescent Cardiovascular Risk Factors and Carotid Intima-Media Echogenicity in Healthy Young Adults: The Atherosclerosis Risk in Young Adults (ARYA) Study”, Journal of the American Heart Association, 2016.

2. L. Fiorini, R. Esposito, M. Bonaccorsi, C. Petrazzuolo, F. Saponara, R. Giannantonio, G. De Petris, P. Dario, “Enabling personalised medical support for chronic disease management through a hybrid robot-cloud approach”, Springer, 2016.

3. J.H. Gennari, P. Langley, D. Fisher, “Models of incremental concept formation”. Artificial Intelligence, 40, pp. 11-61, 1989.

4. M.A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update” SIGKDD Explorations, Volume 11, Issue 1, 2009.

5. M.A. Hall, G. Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining”, IEEE transactions on knowledge and data engineering, vol.15, no.3,May/June, 2003.

6. G. H. John, P. Langley, “Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence”, San Mateo, 338-345, 1995.

1. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, “Improvements to Platt's SMO Algorithm for SVM Classifier Design”, Neural Computation. 13(3):637-649, 2001.

2. J. Nahar, T. Imam, K. S. Tickle, Y. P. Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, In Expert Systems with Applications”, Volume 40, Issue 1, 2013.

3. H. Nichols, “The top 10 leading causes of death in the united states”, Medical News Today, 2017 [Online]. Available: https://www.medicalnewstoday.com/articles/282929.php

4. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space", Philosophical Magazine. 2 (11): 559–572. doi:10.1080/14786440109462720, 1901

5. D. Portugal, et al, “SocialRobot: An Interactive Mobile Robot for Elderly Home Care”, IEEE/SICE International Symposium on System Integration (SII), Dec, 2015.

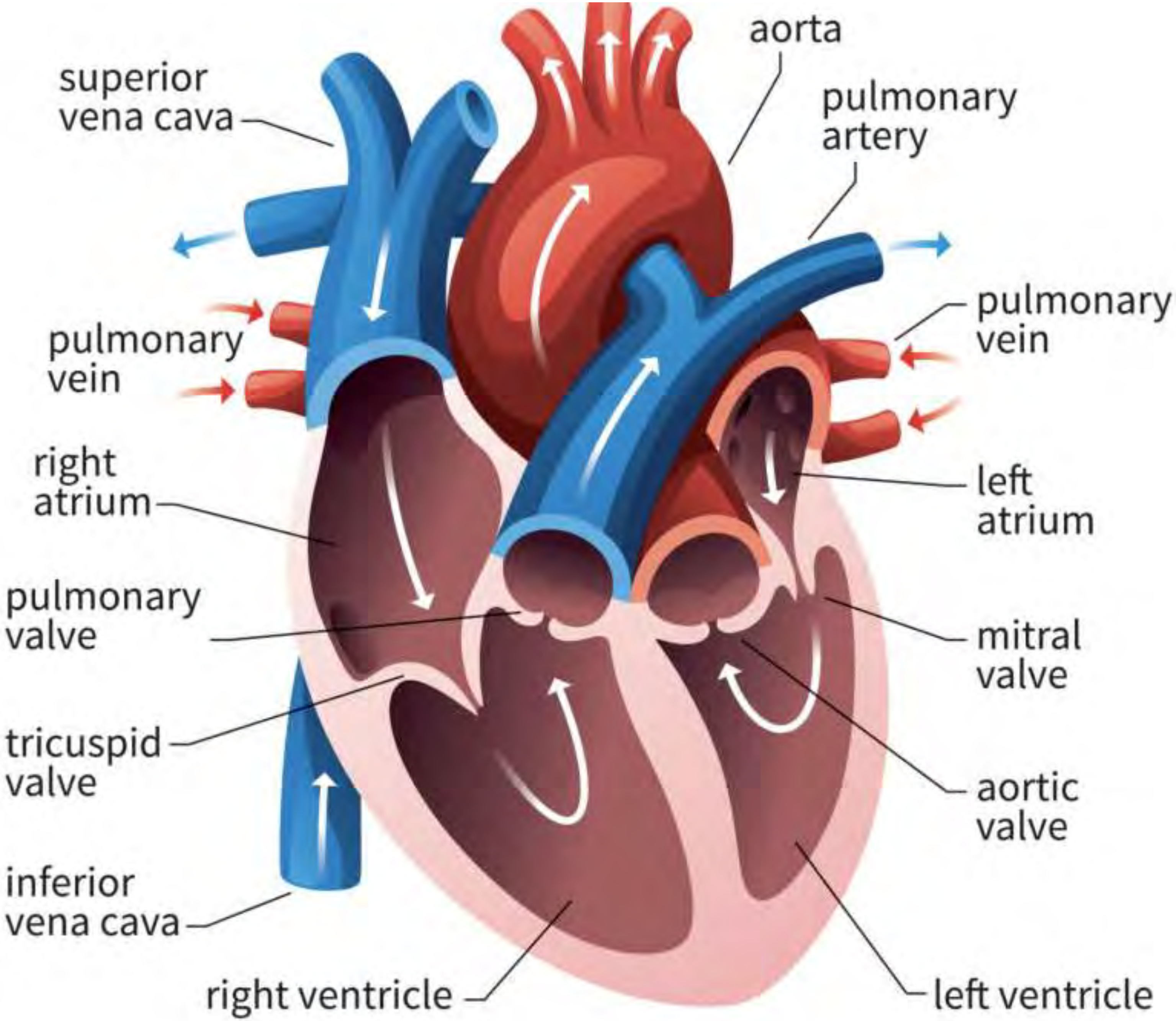
6. D.M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", 2011.

1. R. Quinlan, “Programs for Machine Learning”, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
2. M. Shamim Hossain, “Cloud-Supported Cyber–Physical Localization Framework for Patients Monitoring”, IEEE SYSTEMS JOURNAL, VOL. 11, NO. 1, March, 2017.
3. Y. Zhang, M. Qiu, C.W. Tsai, M.M. Hassan and A. Almari, “Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data”, IEEE SYSTEMS JOURNAL, March, 2017.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization

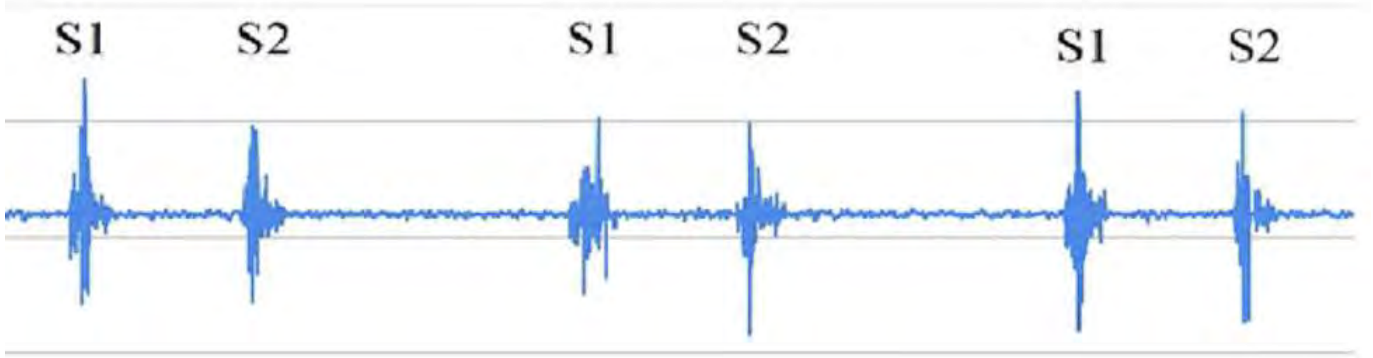
Waikato Environment for Knowledge Analysis

Weka is an open-source software under the GNU General Public License System. It was developed by the Machine Learning Group, University of Waikato, New Zealand. Although named after a flightless New Zealand bird, 'WEKA' stands for Waikato Environment for Knowledge Analysis.

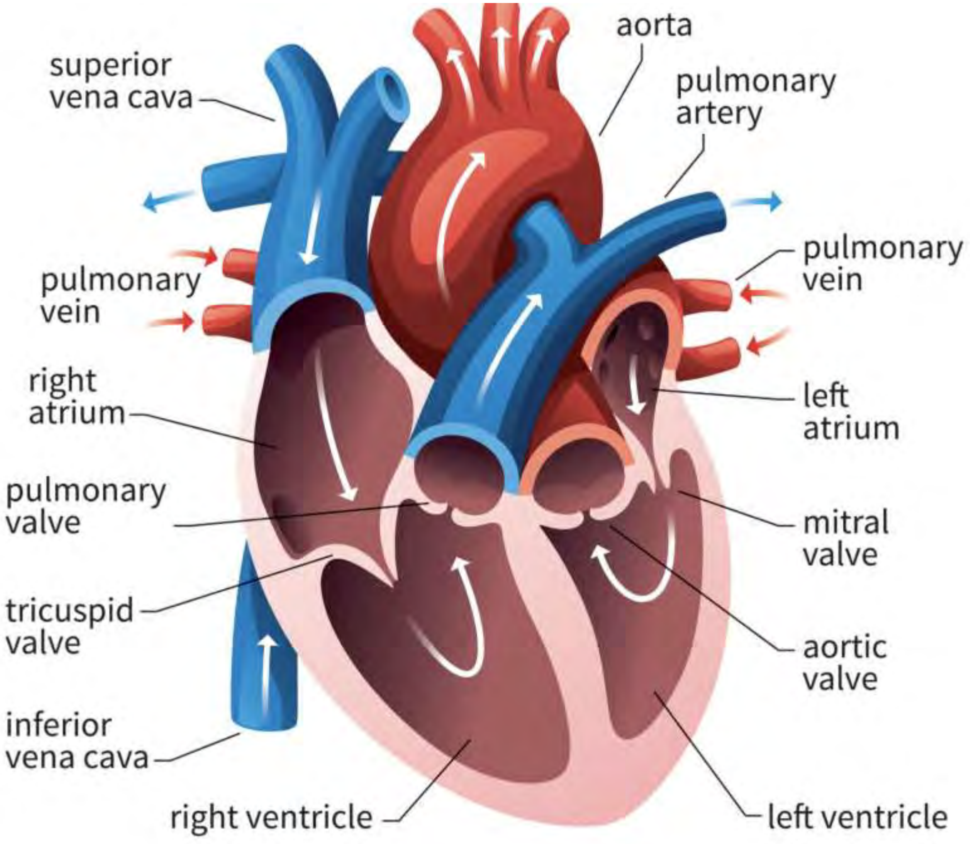

aorta
superior vena cava
pulmonary artery
pulmonary vein
pulmonary vein
right atrium
left atrium
pulmonary valve
mitral valve
tricuspid valve
aortic valve
inferior vena cava
right ventricle
left ventricle

Normal Heart Sounds – S1 and S2

- "Lub Dub.....Lub Dub..."
- No or little Murmurs
- S1...Systole...S2...Diastole



- The closing of Tricuspid and Mitral valves makes the "Lub" or S1.
- The closing of Aortic and Pulmonary valves makes the or S2.



Normal Heart Sounds

Frequency: 20 to 200 Hz.

Resting heart rate for adults: 60 to 100 beats per minute.

Although there's a wide range of normal, an unusually high or low heart rate may indicate an underlying problem.

Heart Sounds – S3

- A rare, low-frequency vibration (10-50 Hz) that occurs after the normal two heart sounds, "lub-dub".
- A brief sound during early diastole, around a third of the way in.
- Difficult to hear because usually it is very soft and can be masked by other sounds, such as lung or abdominal noise, or tightening of the chest wall muscles.
- It's also only audible over a small area of the chest wall.
- A rapid rush of blood from the atrium into the ventricle as it starts relaxing. This may be a normal sound in some people but in people with heart conditions, S3 may indicate heart failure.
- The fourth is a low-intensity sound heard just before S1 in the cardiac cycle. The sudden slowing of blood flow by the ventricle as the atrium contracts causes this sound, which may be a sign of heart disease.

Heart Sounds – S4 (atrial or presystolic gallop)

- A low-intensity (low-pitched) sound heard in late diastole just before S1 in the cardiac cycle.

- The sudden slowing of blood flow by the ventricle as the atrium contracts causes this sound, which may be a sign of heart disease

- Caused by the atria contracting against resistance to fill the ventricle, which can be due to a stiff or hypertrophic (enlarge) ventricle.

- Is Sound of vibrations within the ventricle, and is often accompanied by increased resistance to ventricular filling

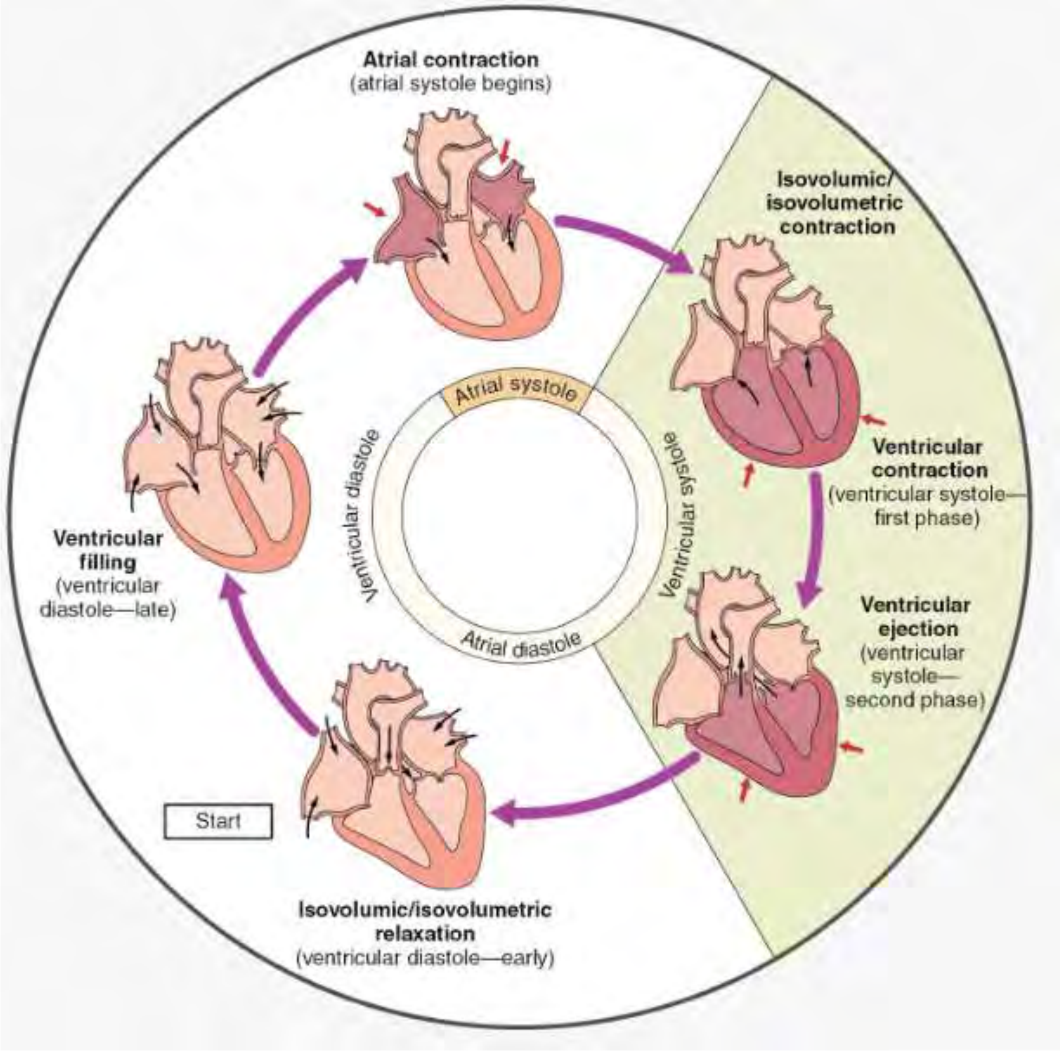
Cardiac Cycle – Different Phases

- A series of events (pressure changes) that occurs during the complete heartbeat that result in the movement of blood (the contraction and the relaxation of both atria and ventricle) through different chambers of the heart and the body.

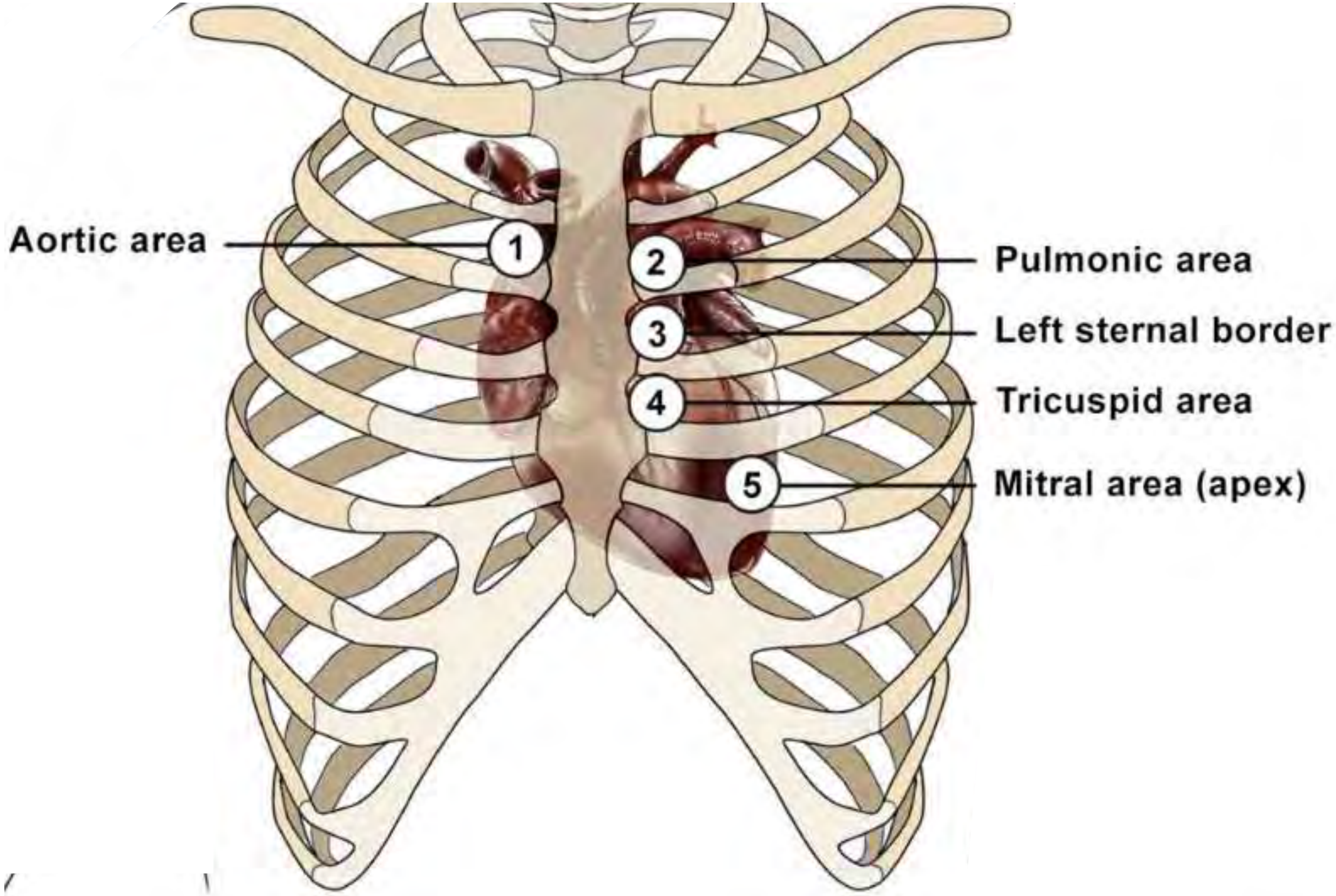
- A complete heart cycle includes the heart filling with blood and the blood being pumped out

Includes 2 phases:

- Heart muscles relax to fill ventricles (1st and 2nd diastole)
- Heart muscle contract to push the blood to arteries (systole)
- Sinoatrial Node – Send wave for atria to contract



Where are heart sounds taken?



$$\text{Entropy} = -\sum_{i}^{n} log_2(P_i)$$

here n= *number of features*
i *= feature*
P = Probability of i

Entropy is given by the equation (2), where p is the probability distribution of a certain feature attribute that is calculated over all the n attributes. In our case, n, the number of attributes, is 74.

- **Mel filter bank, returned as an M-by-N matrix, where:**

1. M is the number of bands, which is determined by the Auto-determine number of bands and Number of bands parameters.
2. N is the number of points in the spectrum. If you select Design one-sided filter bank, then N is equal to ceil (NFFT/2) , where NFFT is the FFT length.

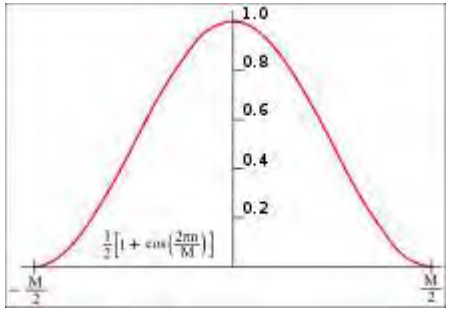
-

Data Segmentation - Mel Frequency Cepstral Coefficients (MFCCs)

- Convert the signals to Mel Frequency Cepstral Coefficients (MFCCs) to capture features in speech recognition or music information retrieval application for comparison of the characteristics of a sound like tone, rhythm, quality of sound.

- Signal bands in high frequencies have less energy compared to the same band in low frequencies, needed to boost the energy level for high frequency bands. The pre-emphasis is done by the first order high-pass filter.

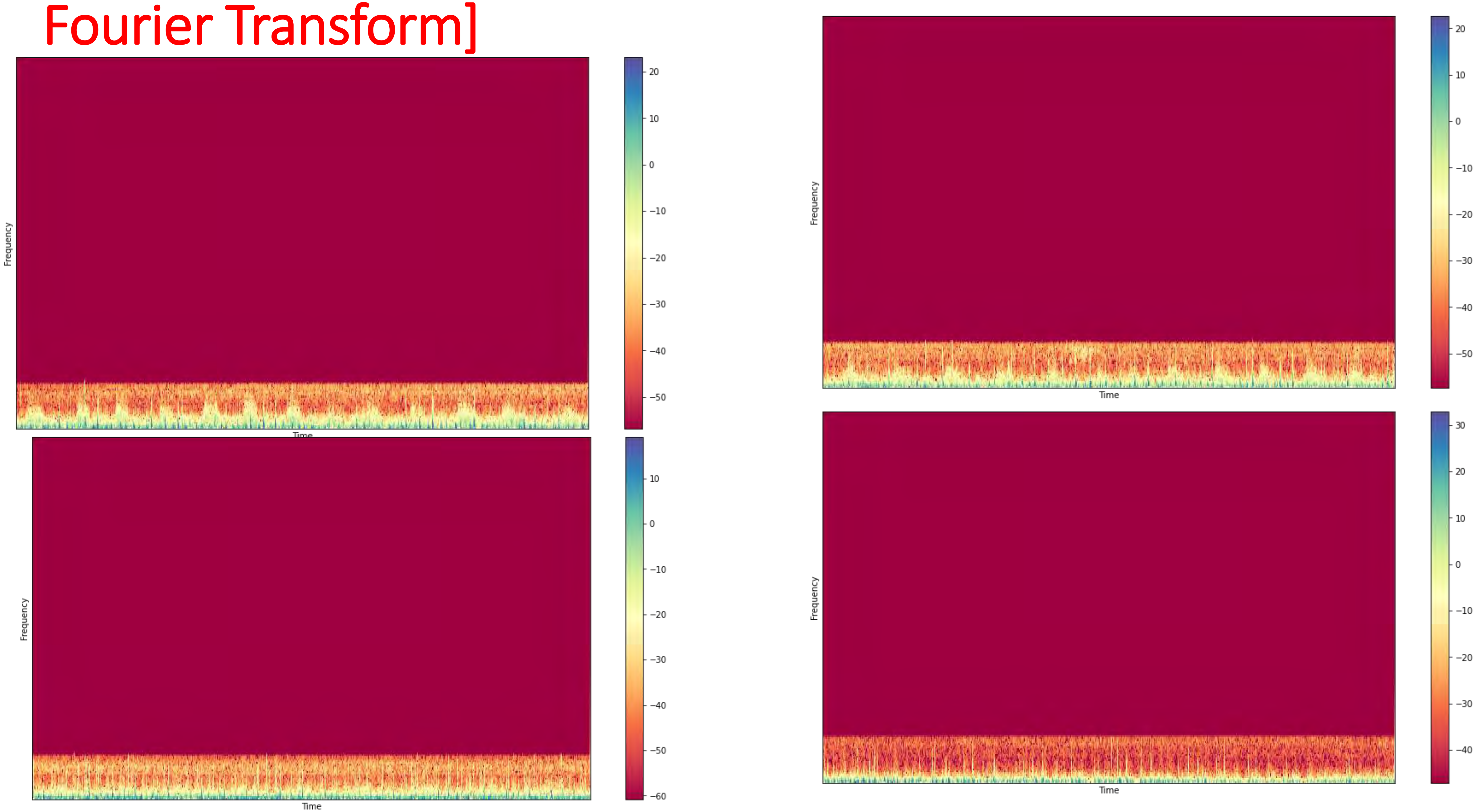
- Applied a Hamming window to the frame to reduce the effects of sounds (speech) at the edges of the window, which is useful in obtaining a smooth spectral representation.



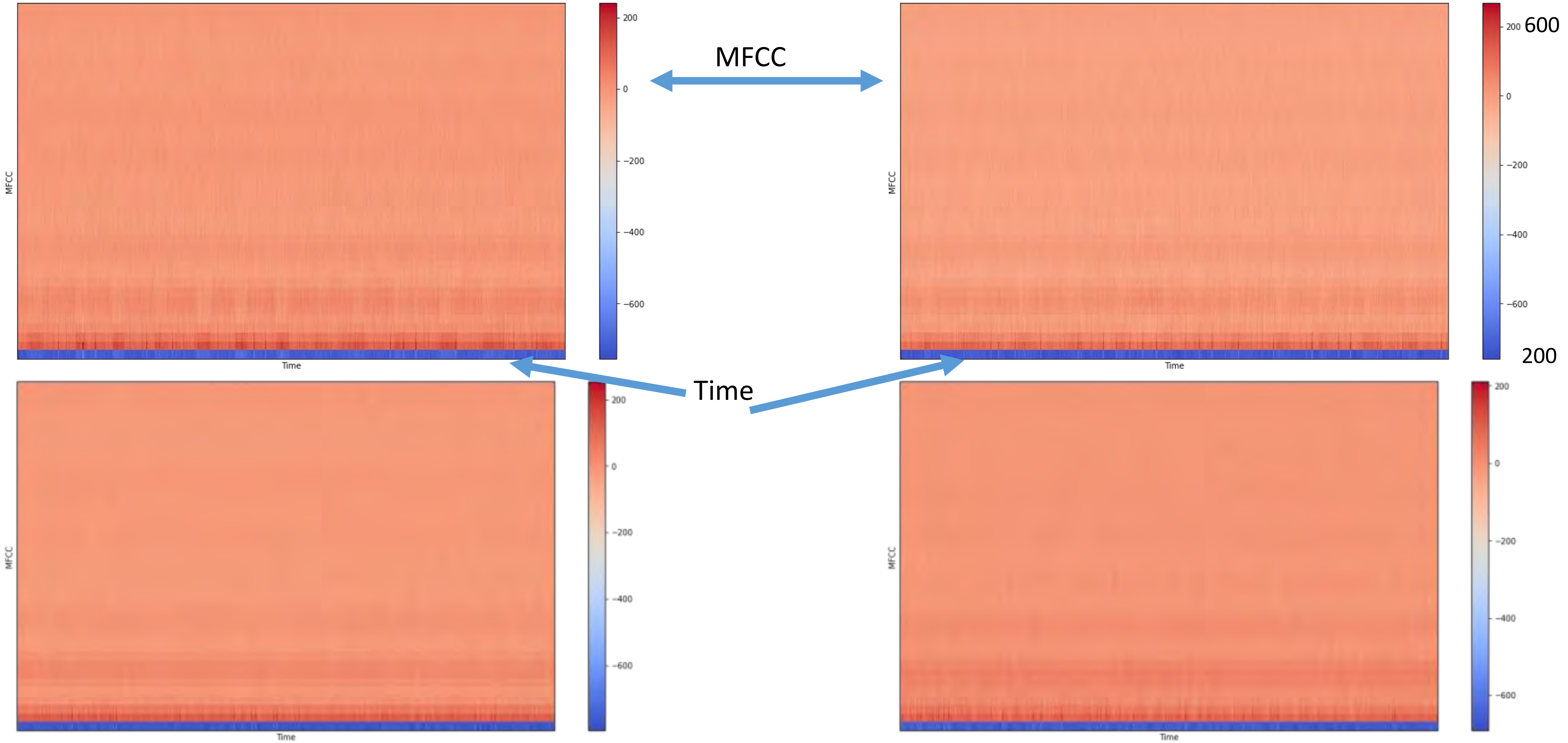
$$C(x(t)) = F^{-1}[log(F[x(t)])]$$

- MFCC model takes the first 12 coefficients of the signal after applying the IDFT (The Inverse Discrete Fourier Transform) operations. The 12 coefficients take the energy of the signal sample as the feature. It will help in identifying the phones. The formula for the energy of the sample is given below

Abnormal PCG (bottom) vs Normal PCG (top) [Short-time Fourier Transform]



Abnormal PCG (bottom) vs Normal PCG (top) [Mel Frequency Cepstral Coefficient (MFCC)]



Machine Learning Classifiers (MLC) – Baseline Classifier

- Useful to establish a baseline for performance comparison of advanced models.
- Provides a minimum requirement for any useful model.
- Decision-making: If the baseline model achieves satisfactory performance, no need for additional time and resources for building more complex model
- Provides a measure of average current level of performance that future performance measures are compared to test if performance is really improving.
- Two Common approaches for creating Baseline model for classification
  - Majority class – where most frequent class in data is predicted
  - Random classifier -Randomly assigns class labels based on the class distribution in the data.
- It provides a sense of whether or not the attributes are informative.
- Has no predictability power

Mel-Scale and Mel-Frequency Cepstral Coefficient (MFCC)

- The audio signals are divided into equal frames between 20-40 ms.

- A number of filter banks are selected to mimic human ear sensitivity.