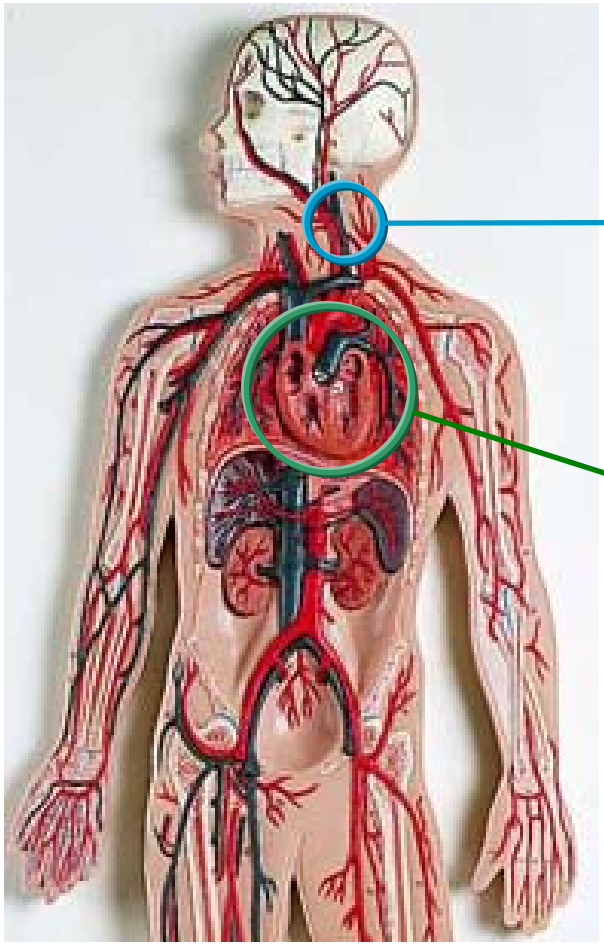
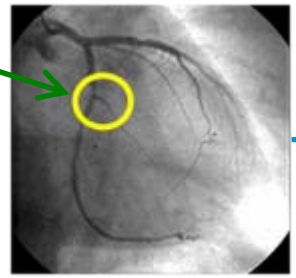
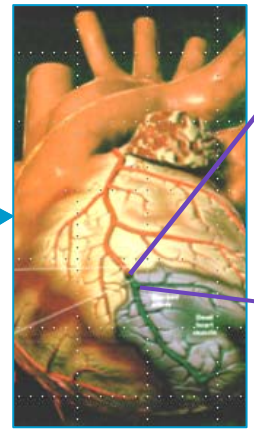
Coronary Artery Disease

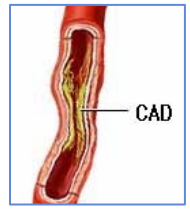


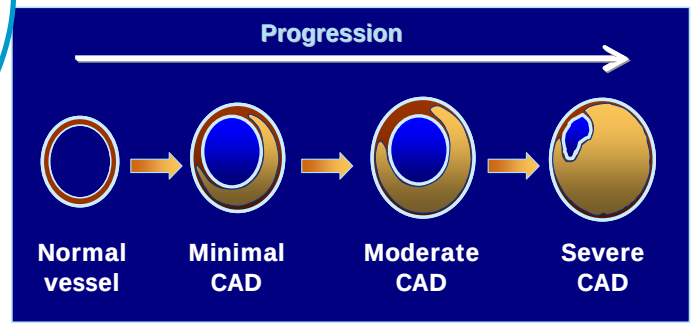
Carotid wall thickness

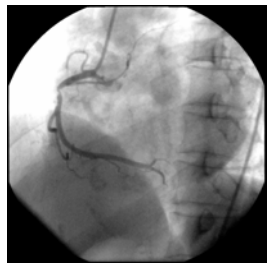


Coronary Arterial Wall

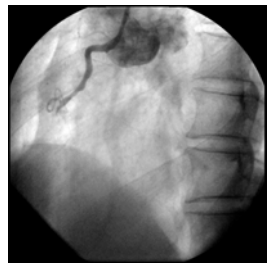








Angina Pectoris



Myocardial infraction

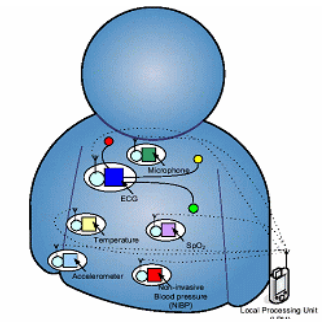
Sudden death of coronary artery disease

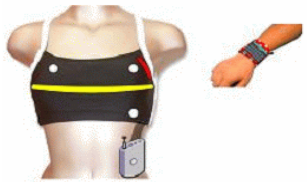
- The death rates that of from diseases in the a recent medical report are much higher than that of from accidents and disasters
  - Statistical reports show that 130(cancer), 77(cerebrovascular), and 37(cardiovascular) per 100,000 people in the Korean population have been dying as the result of the diseases every year
  - Increased carotid intima-media thickness (IMT) is associated with atherosclerosis risk factors and adverse cardiovascular outcomes

Our objective

- The first aim is to implement Korean reference standard database
  - Normal and abnormal reference and standard data such as multi-parametric features, clinical information
  - For different heart diseases, we develop diagnostic indexes and algorithms
- **The second aim is to develop biosignal data mining methodology for predicting and diagnosing the cardiovascular disease**
  - A multi-parametric measure of HRV, ST-segments and IMT measurements and
  - A suitable prediction method to enhance the reliability of medical treatment
  - We apply and analyze existing classification and statistical methods

Development of intelligent automatic diagnosis system prototype (KOSEF project)





Clinical/hospital information system

Emergency service

Internet

Bluetooth™ or WLAN

ECG signal Processing Server

Raw Data Cleaning Process

Biosignal Data Process

Internet

Hospital

Gateway

Data Mining Server

Server Server

Physician

intelligent automatic diagnosis system architecture

- iDB 2008 -

A Data Mining Approach and Framework of Intelligent Diagnosis System for Coronary Artery Disease Prediction

Keun Ho Ryu[1], Heon Gyu Lee[1], Wuon-Shick Kim[2]

[1](khryu, hglee@dblab.chungbuk.ac.kr), [2]wskim@kriss.re.kr

[1]Database/Bioinformatics Laboratory,

Chungbuk National University, Korea

[2]Korea Research Institute of Standards and Science, Korea

Outline

- **Introduction**
  - **Application of data mining to medical data**
  - **Motivation and objective**
- Background
  - ECG Analysis
  - ST-segments
  - Heart Rate Variability
  - Carotid arterial wall thickness
- Automated detection of Ischemic ECG beats
  - ST-segments feature extraction
  - CASE study1: Ischemia beat classification
- Heart Rate Variability
  - ECG processing, Linear/ Nonlinear features
  - Classification methods for diagnosing CAD
  - CASE study2 : AP and ACS diseases classification
- Dyslipidemia diagnosis methodology
  - Measuring Intima-Media thickness
  - Feature extraction from IMT image
  - CASE study3: dyslipidemia classification
- Discussion
- On going work
  - Compressed Patricia FP-Tree for Frequent Itemsets Mining
- Conclusion

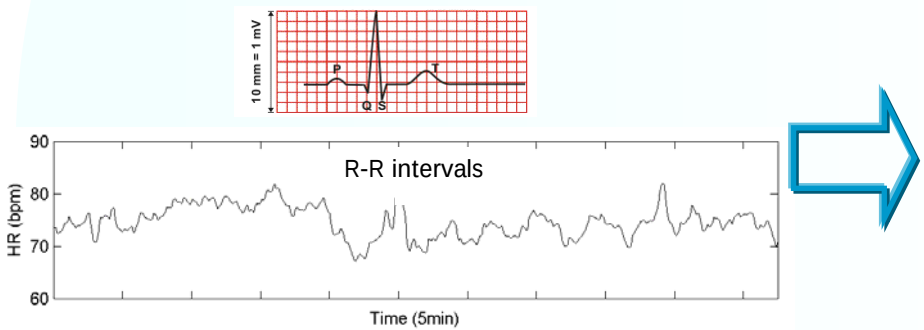
Introduction ...
: Why data mining?

- Modern medicine generates, almost daily, huge amounts of heterogeneous data
  - Medical data may contain signals like ECG, clinical information like temperature, cholesterol levels, etc., as well as the physician's interpretation
  - Those who deal with such data understand that there is a widening gap between data collection and data comprehension
  - As more and more medical procedures employ signal and imaging as preferred diagnostic tools

- **There is a need to develop methods for efficient mining in databases of signals**

Introduction ...
: Application of data mining to medical data

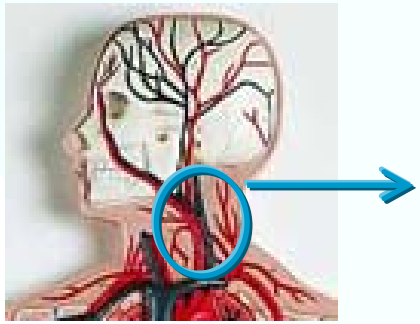
- Cardiovascular disease
  - Myocardial ischemia : ST-segments from ECG signal
  - Coronary artery disease: Multi-parametric features of Heart rate variability from ECG signal



Myocardial ischemia

Angina Pectoris

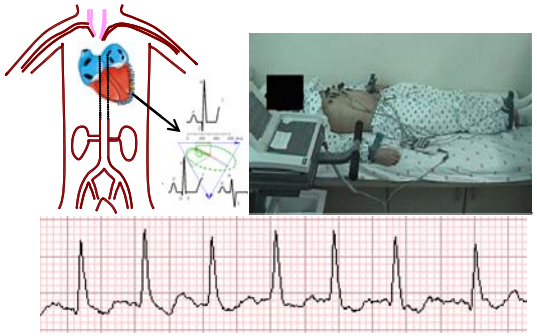
  - Carotid wall thickness (intima-media thickness)



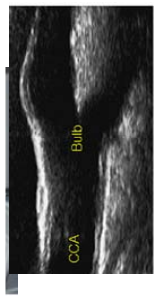
Atherosclerosis

A Data Mining Framework of Intelligent Heart Disease Diagnosis System

ECG/IMT Dataset



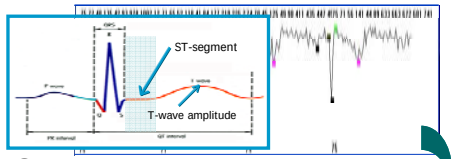
ECG Signals
Lead II Channel



Carotid Arterial Wall Thickness

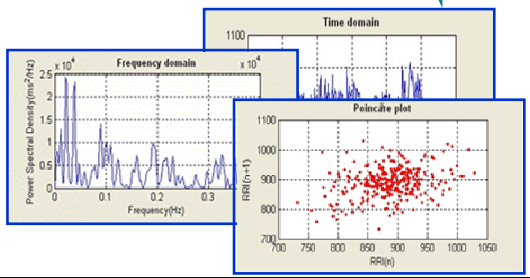
Multi-parametric Feature

QRS wave,
R-R interval detection

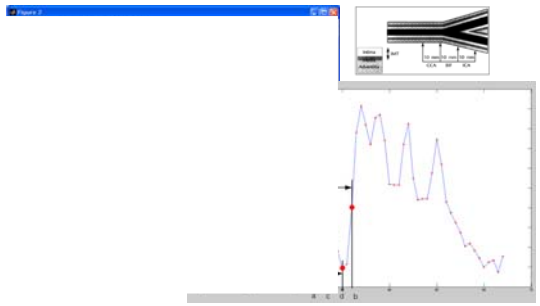


ST-seg.

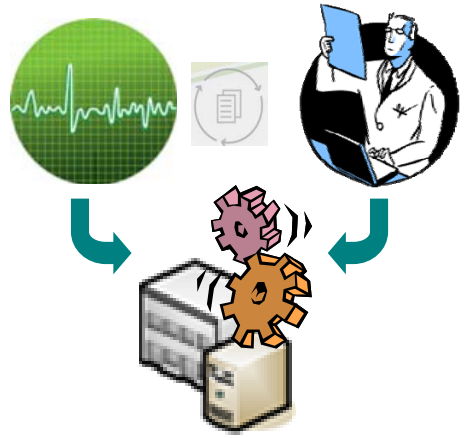
HRV Analysis



Measuring IMT

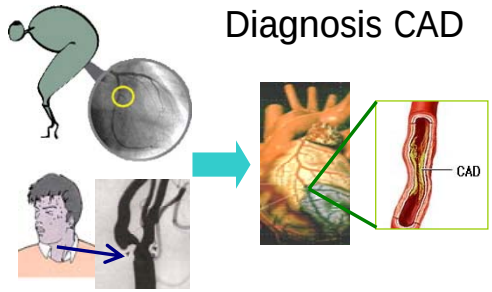


Heart Disease Diagnosis

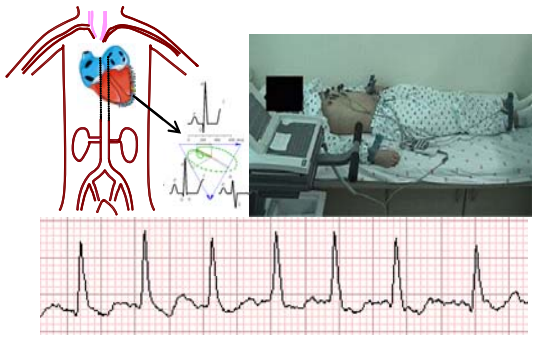


Learning and Prediction
using multi-parametric
features

Diagnosis CAD



ECG/IMT Dataset



ECG Signals
Lead II Channel

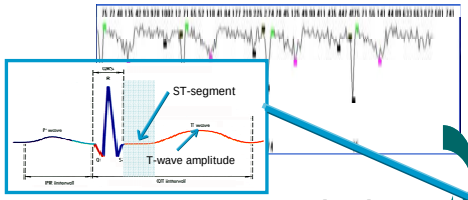
Carotid Arterial Wall Thickness

Step 1: Data Preparation

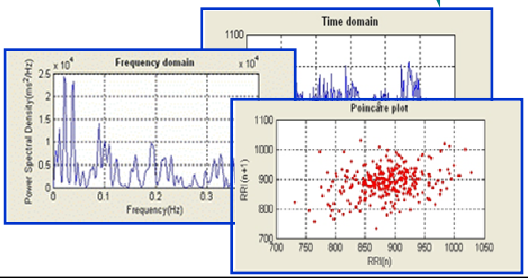
- Measure ECG Signals
- Scan Carotid arterial wall thickness
  : intima-media thickness
- Collect clinical information

Multi-parametric Feature

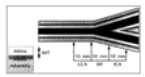
QRS wave,
R-R interval detection



ST-seg. HRV Analysis



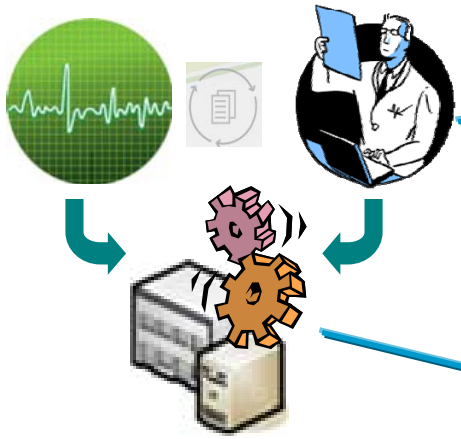
Measuring IMT



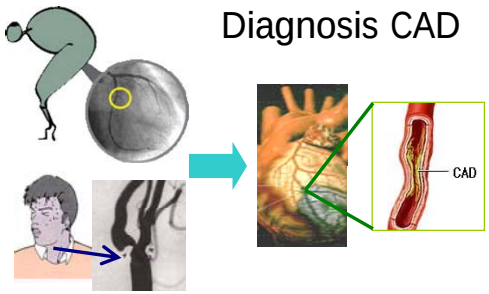
❖ Step 2: Diagnosis Feature extraction

- Detect RR intervals from ECG signals
- Extracting ST-segments features
- HRV analysis
  : Linear/nonlinear feature extraction
- Measure intima-media thickness separately

Heart Disease Diagnosis



Learning and Prediction using multi-parametric features



In our experiments

- Ischemia beat classification
- Angina Pectoris and Acute Coronary Syndrome classification
- dyslipidemia detection

➔ gave about 91%, 89%, 93% goodness of accuracy

❖ Step 3: Coronary Artery Disease Diagnosis

- build classifiers using multi-parametric features.
- classify the patients with coronary artery disease.
- evaluate the results of classification

Motivation
: Why need multi-parametric features

- Coronary heart disease is in danger of becoming the most fatal disease among Koreans because of **smoking** and the adoption of **westernized eating habits**
- It is very important social issue to detect early heart problems and predict accurate heart disease
- The various diagnosis indexes have different physiological meanings
  - There is not a master feature that explains the whole characteristics of diagnosis indexes at one time
  - Researchers must select to use among various features
  - Need multi-parametric feature with accuracy
    - It will be helpful to predict the heart disease

Objective

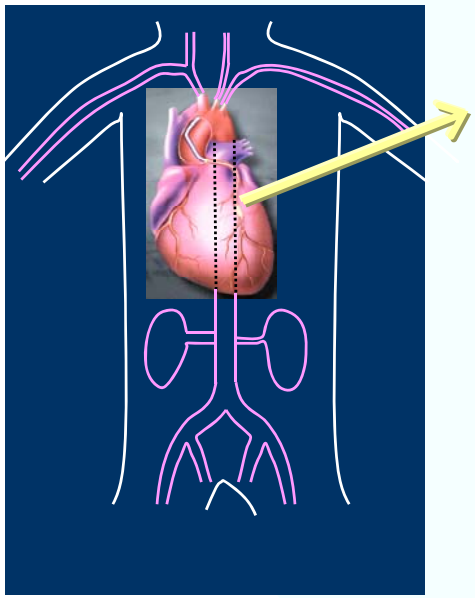
- A helpful diagnostic supplementary tool, using various features, that takes into consideration all possible diagnostic features
  - Linear and Nonlinear features of HRV, ST-segments, IMT

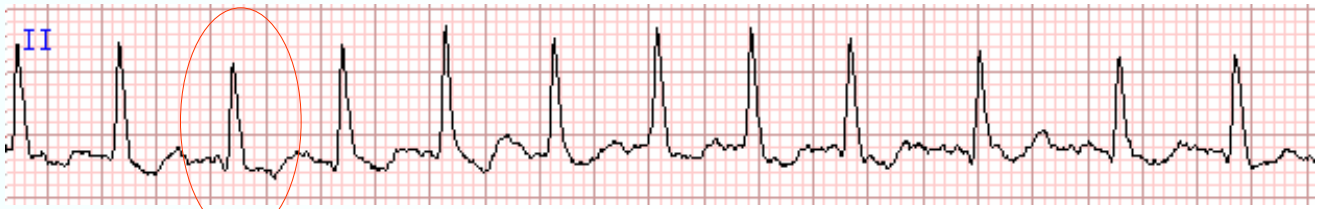
Outline

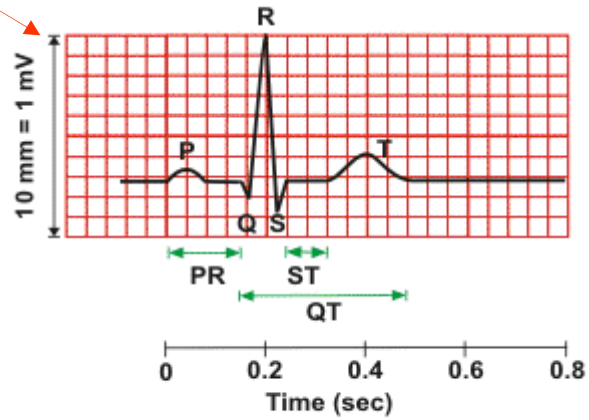
- Introduction
  - Application of data mining to medical data
  - Motivation and objective
- **Background**
  - **ECG Analysis**
  - **ST-segments**
  - **Heart Rate Variability**
  - **Carotid arterial wall thickness**
- Automated detection of Ischemic ECG beats
  - ST-segments feature extraction
  - CASE study1: Ischemia beat classification
- Heart Rate Variability
  - ECG processing, Linear/ Nonlinear features
  - Classification methods for diagnosing CAD
  - CASE study2 : AP and ACS diseases classification
- Dyslipidemia diagnosis methodology
  - Measuring Intima-Media thickness
  - Feature extraction from IMT image
  - CASE study3: dyslipidemia classification
- Discussion
- On going work
  - Development of Contrast data mining and frequent pattern mining methods
- Conclusion

Background
: BioSignals (ECG:Electrocardiogram)

- The most used signal in clinical practice is the Electrocardiogram (ECG)
  - ECG is test that measures a heart's electrical activity
  - ECG provides clinical information about the heart's status





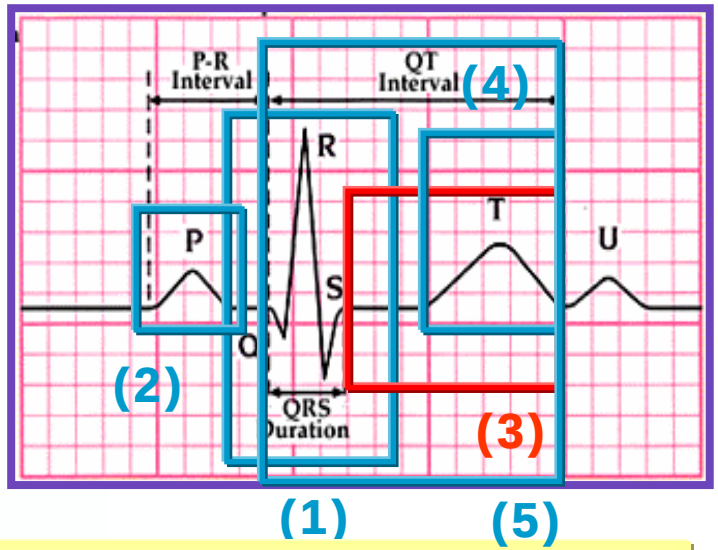


P wave (0.08 - 0.10 s) QRS (0.06 - 0.10 s)

Richard E. Klabunde

Background …
: A basic approach to ECG analysis

- Simply a guide to understanding how clinicians identify abnormalities in the ECG
  1. Identify the QRS complexes
  2. Identify the P waves
  3. Examine ST-T segments: are there abnormalities (such as elevation or depression)?
  4. Examine the T wave
  5. Examine the QT interval

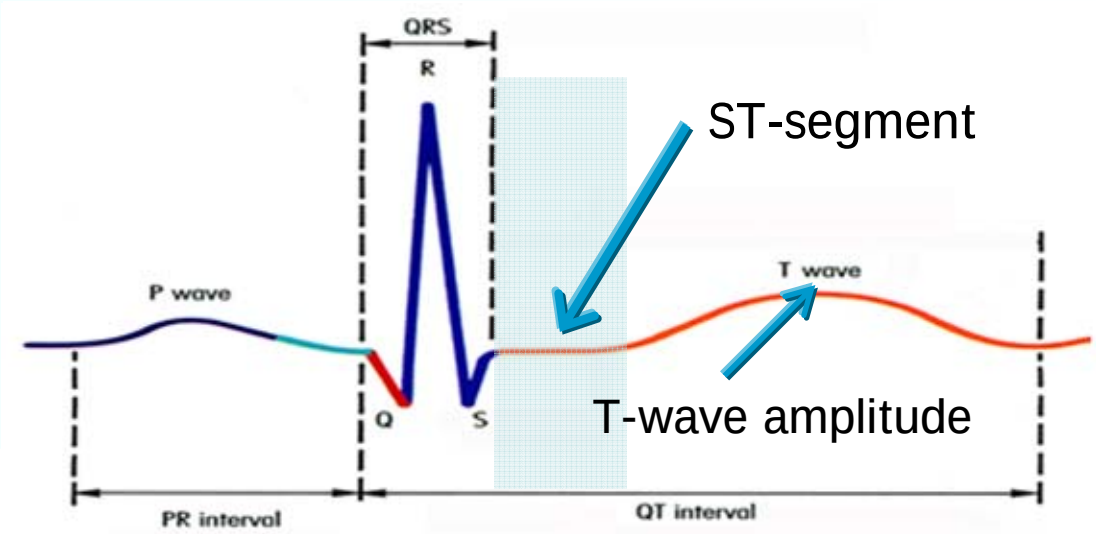


Is the abnormality suggestive of Ischemia or infarction?

- ECG analysis is a routine part of any complete medical evaluation, due to the heart's essential role in human health and disease

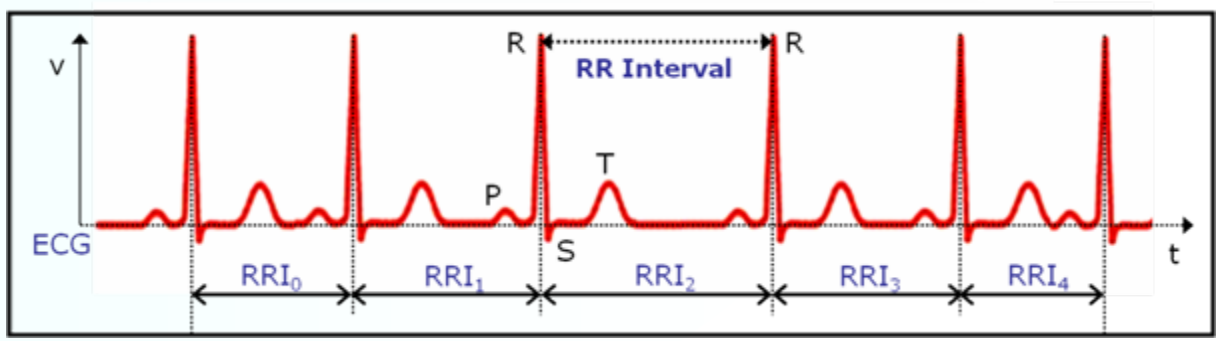
Background ...
: ST-segments and T amplitude

- Several techniques the evaluate the ST-segment changes and the T-wave alterations by different methodologies have been used to ischemic beat detection
  - Parametric modeling, wavelet theory, set of rules, neural nets, decision analysis, genetic algorithms
- It is clinically important if elevated or depressed as it can be a sign of ischemia or infarction



Background ...
: Heart Rate Variability

- HRV (Heart Rate Variability) has been used to assess autonomic control of the heart under physiological and pathological conditions
  - Control of Heart Rate is known to be affected by the sympathetic and parasympathetic system
- HRV is used as a clinical tool predicting heart function in both health and disease
- HRV is based on analysis of RR intervals in ECG signals
  - *RRIs (RR intervals)* are the series of time intervals between heartbeats



Background ...
: Heart Rate Variability

- HRV is one of the most promising indications of autonomic activity (sympathetic / parasympathetic activity)
  - Reduced parasympathetic activity has been reported in patients with CAD (Coronary Artery Disease)
  - This reduction in parasympathetic activity evaluated by HRV analysis
- Two properties of HRV analysis
  - Linear analysis : time- and frequency-domain
  - Nonlinear analysis : complexity and irregularity

Background ...
: Heart Rate Variability

- Linear property : Time- / Frequency- domain
  - Linear time domain have been employed in quantification of the overall variability of HR
  - Linear frequency domain variables provide markers of the heart regulation

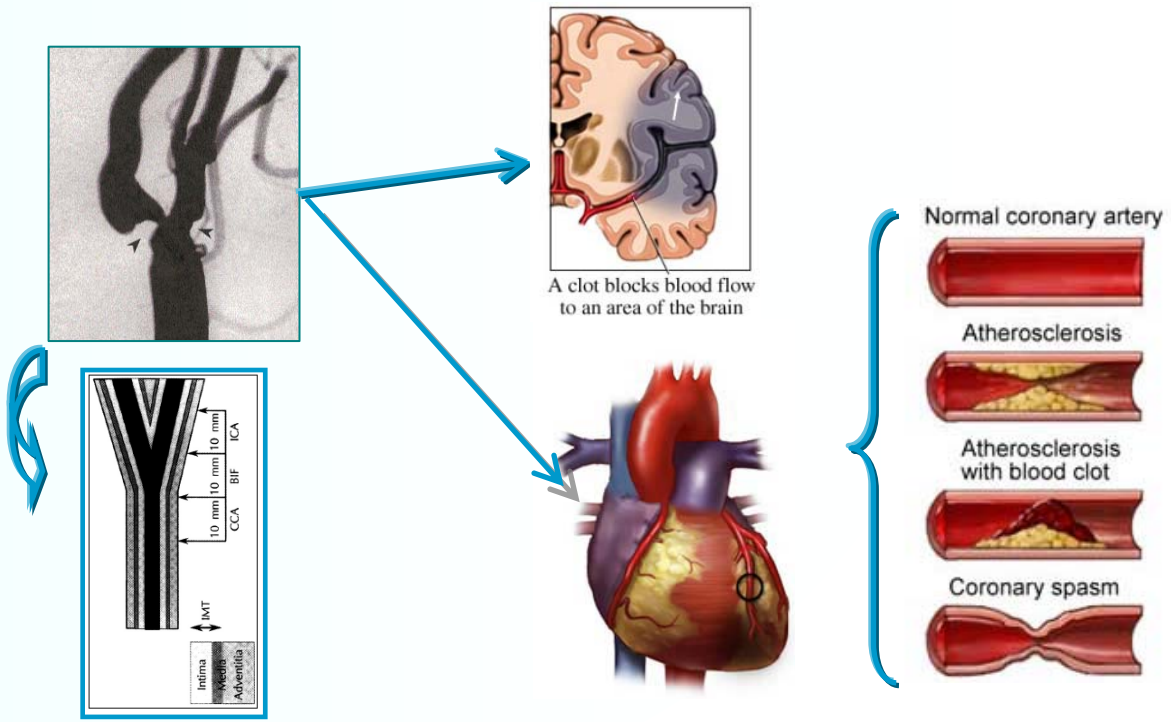
- Nonlinear property
  - The heart rate has complexity which reflect the healthy condition in a living body
  - The complexity of the human physiological system is reduced in bad health, but increased in good health
  - This complexity can be analyzed by various nonlinear methods

Background ...
: Effect of 3 recumbent positions on nervous system using HRV

- There is a relationship between recumbent position and HRV
  - In linear analysis, parasympathetic activity is increased and sympathetic activity is decreased in the right position
  - In nonlinear analysis, complexity are increased in right position
  - Among 3 recumbent positions (supine, right, left), the right position can enhance the parasympathetic activity
- Right recumbent position is recommended in patient to recover the depressed parasympathetic activity
- Because of the effect of the recumbent position on HRV, we considered that it is worthwhile to include Linear/Nonlinear properties of HRV for three positions

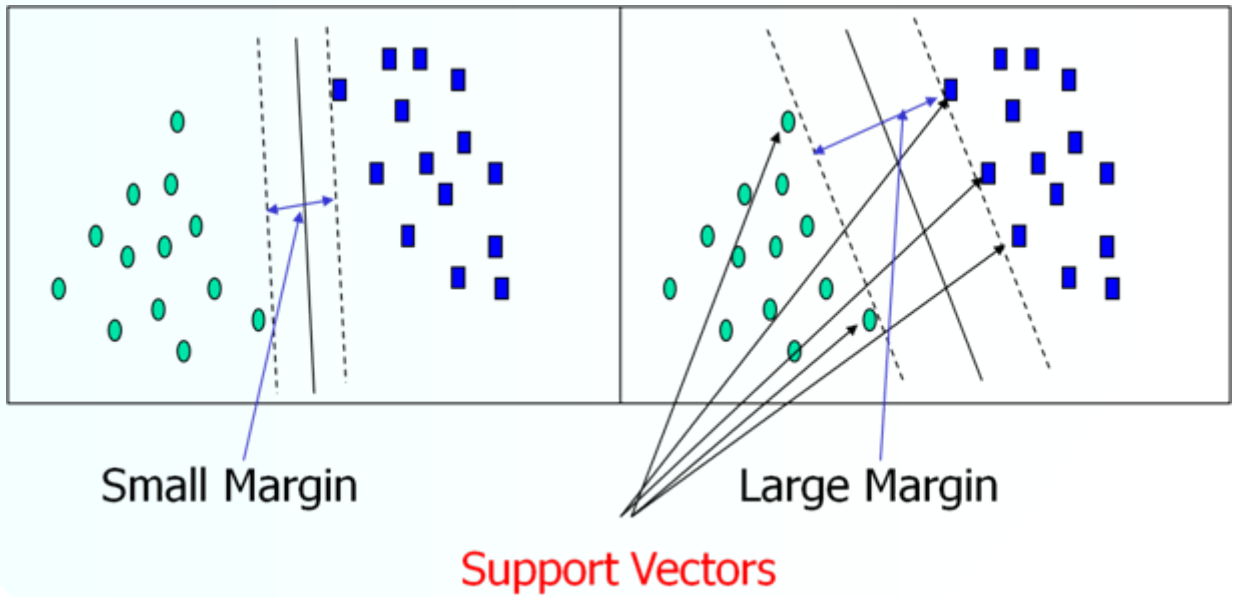
Background ...
: Carotid arterial wall thickness

- The appearance of atherosclerosis has been highly associated with incidence of coronary heart disease
- the severity of carotid IMT is an independent predictor of coronary events such as myocardial infarction [American Heart Association]
  - Carotid IMT consists of intima thickness and media thickness



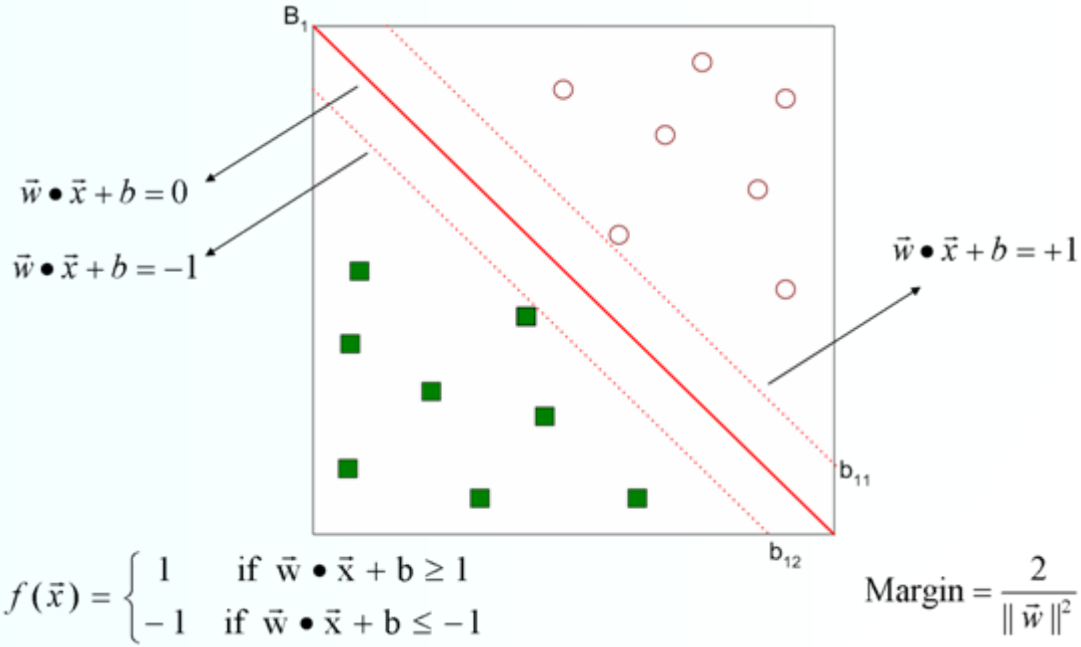
Background
: Classification for Diagnosing Disease (SVM)

- SVM (Support Vector Machine)
  - A new classification method for both linear and nonlinear data
  - It uses a nonlinear mapping to transform the original training data into a higher dimension



Background ...
: Classification for Diagnosing Disease (SVM)

- With the new dimension, it searches for the linear optimal separating *hyperplane* (i.e., “decision boundary”)
- SVM finds this hyperplane using *support vectors* (“essential” training tuples) and *margins* (defined by the support vectors)

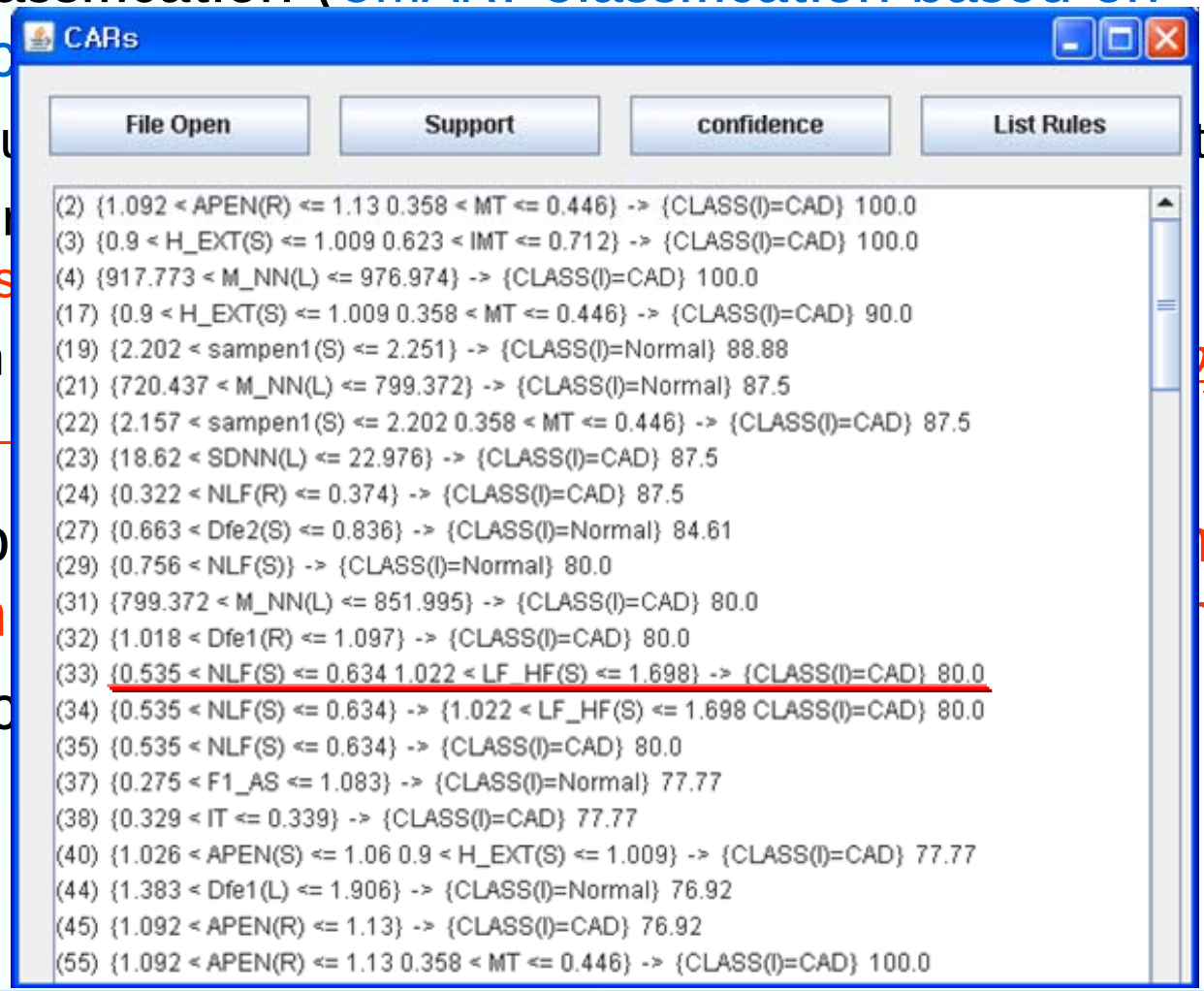


Background
: Classification for Diagnosing Disease (MDA)

- Multiple Discriminant Analysis is an analysis of dependence method that is a special case of canonical correlation
- With more than two groups, there will potentially be more than one discriminant function that can be used to explain the differences among groups
  - For example, if we want to discriminate among three groups, two canonical discriminant functions will be derived
  - The first discriminant function separates group 1 from groups 2 and 3, and the second discriminant function separates group 2 from group 3.
- We can obtain classification function for prediction through the discriminant analysis
- Classification function is generated for each group
  - If new case we have to classify comes into existence, this subject will belong to a group that has the highest value of classification function

Background ...
: Classification for Diagnosing Disease (CMAR)

- Associative classification (CMAR: Classification based on multiple assoc
  - Association ru
  - Search for str (conjunctions
  - Classification $p_2 \ldots \wedge p_l \rightarrow$
- It may overco tree induction
- CMAR classific



Background ...
: Classification for Diagnosing Disease (Bayes.)

- Let *D* be a training set of tuples and their associated class labels, and each tuple is represented by an *n-D* attribute vector $X = (x_1, x_2, ..., x_n)$
- Suppose there are m classes $C_1, C_2, ..., C_m$.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|X)$
- This can be derived from Bayes' theorem

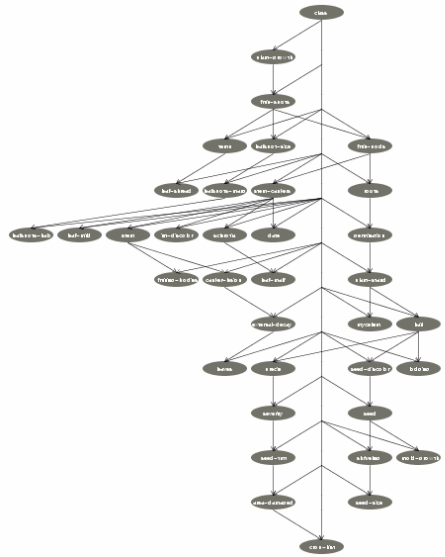
$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- A simplified assumption: attributes are conditionally independent -> Naïve Bayesian

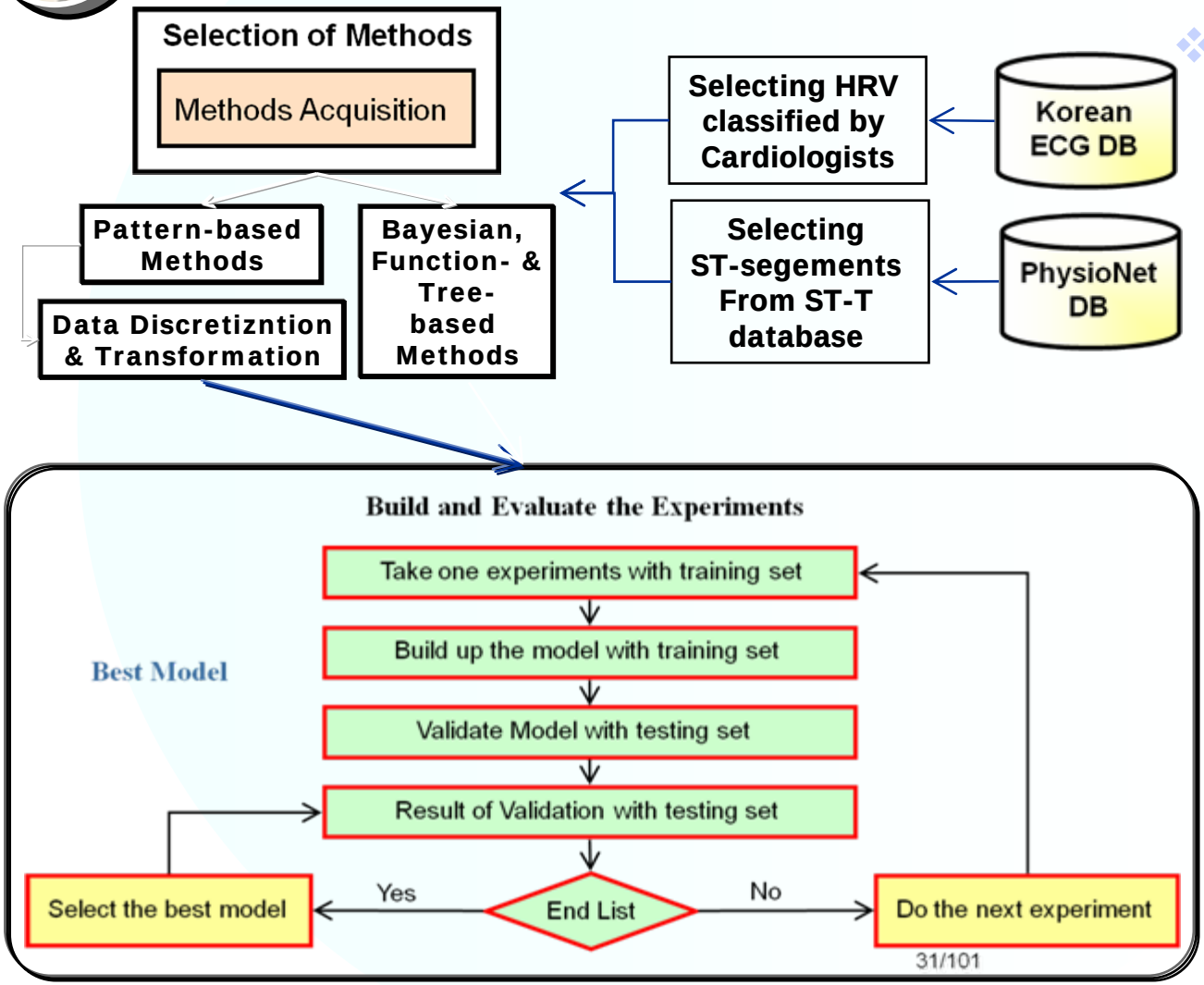
$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ... \times P(x_n|C_i)$$

Background …
: Classification for Diagnosing Disease (Bayes.)

- Bayesian belief network allows a subset of the variables conditionally independent
  - A graphical model of causal relationships
  - Represents dependency among the variables
  - Gives a specification of joint probability distribution
    - Nodes: random variables
    - Links: dependency
  - Network search algorithm
    - Tree Augmented Naïve Bayes (TAN)



Classifiers Evaluation Methodology



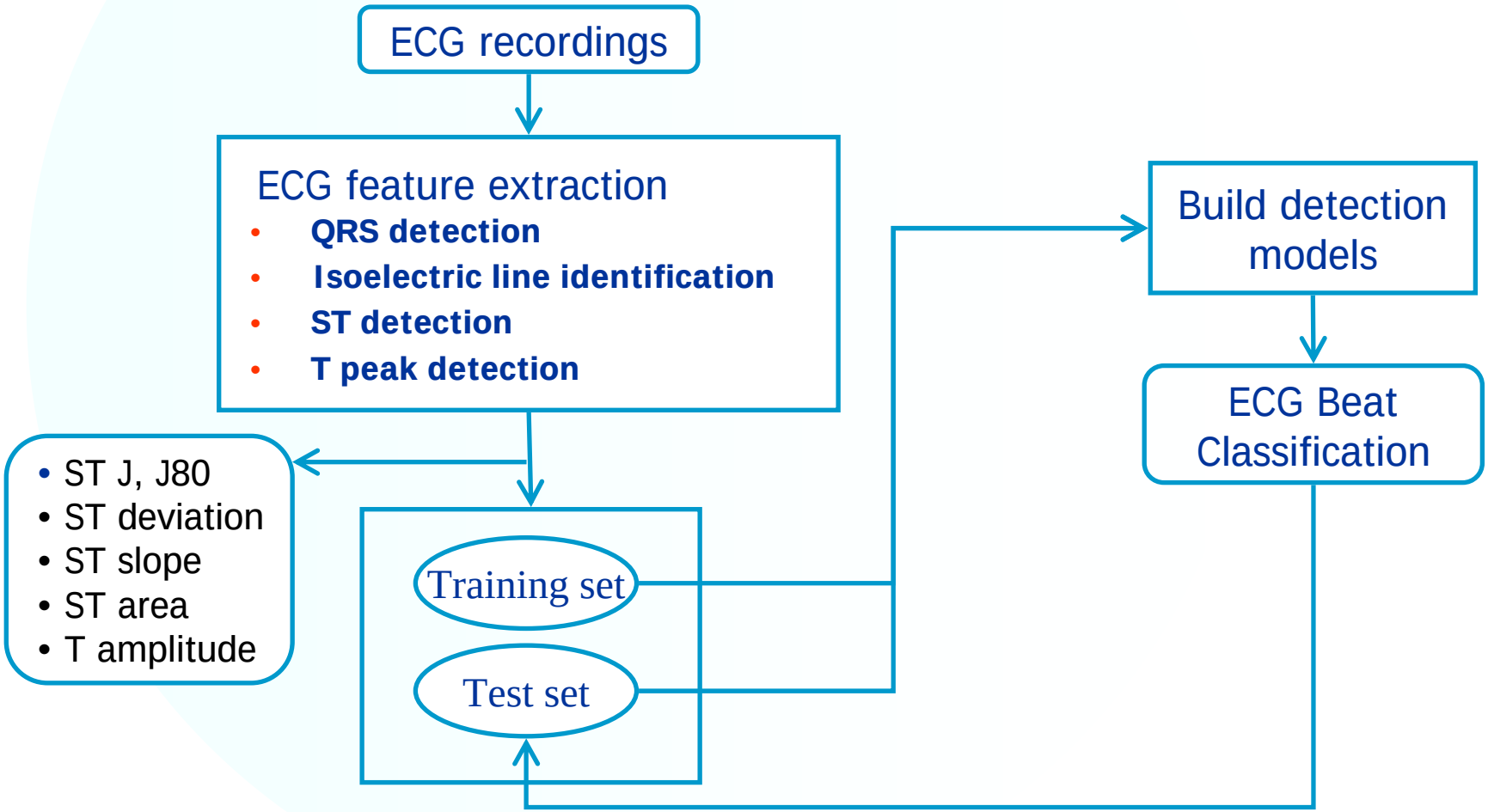
- Flowchart of the overall ischemic beat and patients with CAD classification methodology

Outline

- Introduction
  - Application of data mining to medical data
  - Motivation and objective
- Background
  - ECG Analysis
  - ST-segments
  - Heart Rate Variability
  - Carotid arterial wall thickness
- **Automated detection of Ischemic ECG beats**
  - **ST-segments feature extraction**
  - **CASE study1: Ischemia beat classification**
- Heart Rate Variability
  - ECG processing, Linear/ Nonlinear features
  - Classification methods for diagnosing CAD
  - CASE study2 : AP and ACS diseases classification
- Dyslipidemia diagnosis methodology
  - Measuring Intima-Media thickness
  - Feature extraction from IMT image
  - CASE study3: dyslipidemia classification
- Discussion
- On going work
  - Development of Contrast data mining and frequent pattern mining methods
- Conclusion

Automated detection of Ischemic beats

- Flowchart of the overall ischemic beat classification methodology

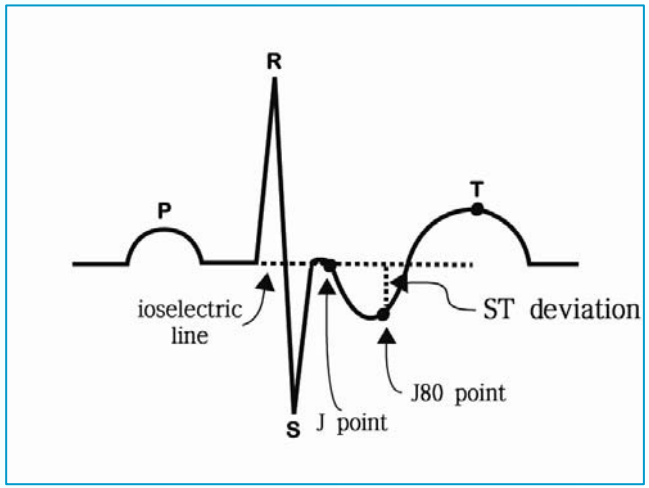


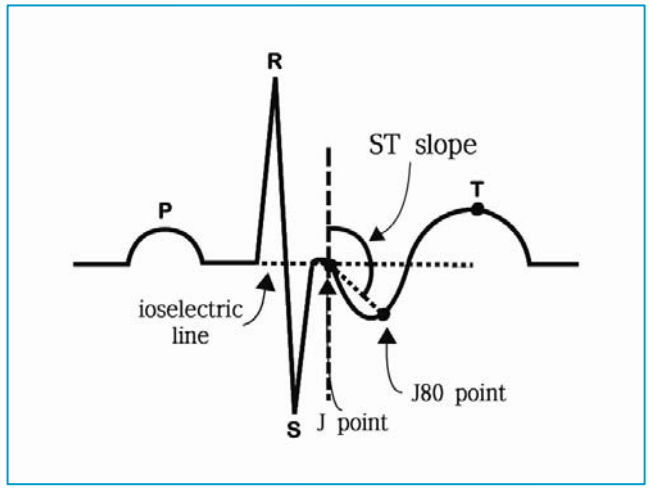
ST-segments feature extraction

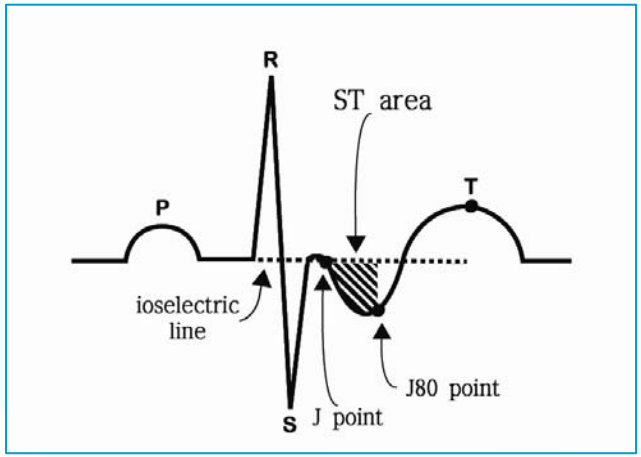
- Location of the J point was detected
  - If (heart rate ≤ 120 beats/min) then J80
  - If (heart rate > 120 beats/min) then J60

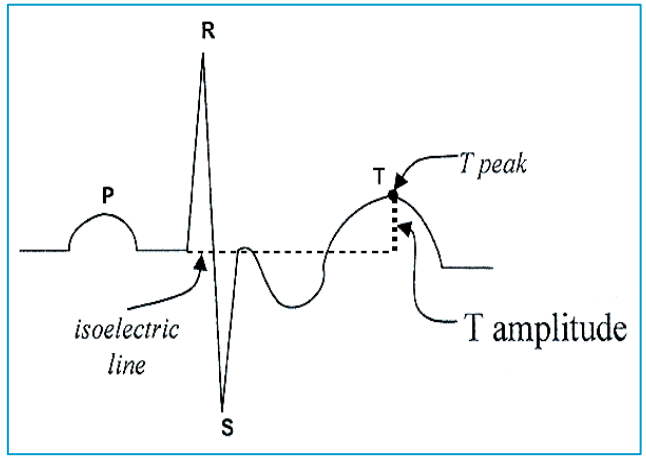
- ST-segment deviation: refer to the amplitude deviation of the ST-segment from the isoelectric line (the line defining the level of zero amplitude)
- ST-segment slope: the slope of the line connecting the J and J80 (or J60) points
- ST-segment area: the area between the ECG trace, the isoelectric line and the points J and J80
- T-wave amplitude: the amplitude deviation of the T-wave peak from the isoelectric line

ST-segments feature extraction ...



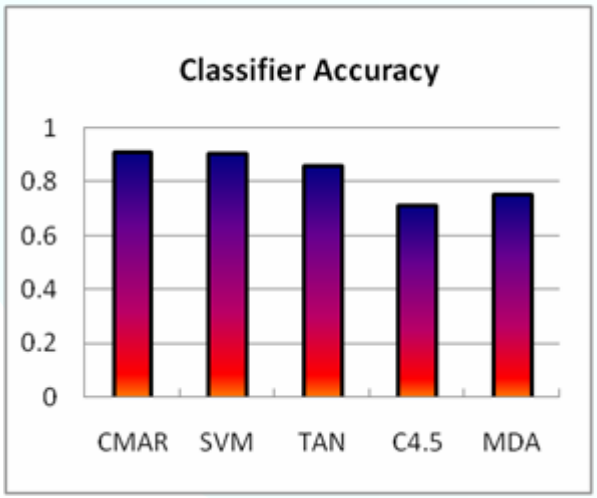






CASE study – I
: Ischemia beat classification

- Data generation
  - Download ST-segments data from *PhysioNet database* (http://physionet.org)
  - Total # of data : Ischemia(204), Normal(198)
  - Selected features for our experiment
    - ***ST-deviation, Slope, Area, T amplitude, Class* (Ischemia/Normal)**
- Experimental results



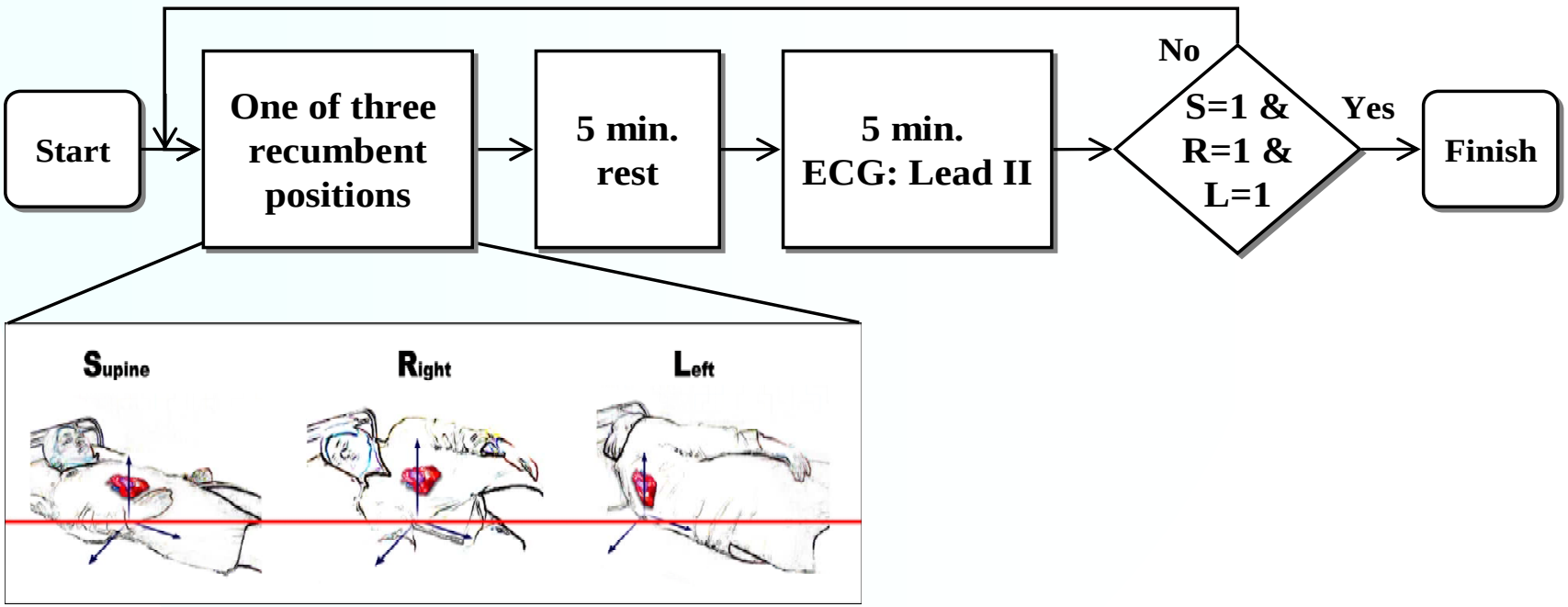
Classifier	TP	FP	Precision	Recall	Class
CMAR	0.924 0.878	0.122 0.076	0.885 0.919	0.924 0.878	Ischemia Normal
SVM	0.909 0.904	0.096 0.091	0.909 0.904	0.909 0.904	Ischemia Normal
TAN	0.818 0.766	0.234 0.182	0.786 0.8	0.818 0.766	Ischemia Normal
C4.5	0.768 0.84	0.16 0.232	0.835 0.775	0.768 0.84	Ischemia Normal
MDA	0.838 0.723	0.277 0.162	0.761 0.81	0.838 0.723	Ischemia Normal

Outline

- Introduction
  - Application of data mining to medical data
  - Motivation and objective
- Background
  - ECG Analysis
  - ST-segments
  - Heart Rate Variability
  - Carotid arterial wall thickness
- Automated detection of Ischemic ECG beats
  - ST-segments feature extraction
  - CASE study1: Ischemia beat classification
- **Heart Rate Variability**
  - **ECG processing, Linear/ Nonlinear features**
  - **Classification methods for diagnosing CAD**
  - **CASE study2 : AP and ACS diseases classification**
- Dyslipidemia diagnosis methodology
  - Measuring Intima-Media thickness
  - Feature extraction from IMT image
  - CASE study3: dyslipidemia classification
- Discussion
- On going work
  - Development of Contrast data mining and frequent pattern mining methods
- Conclusion

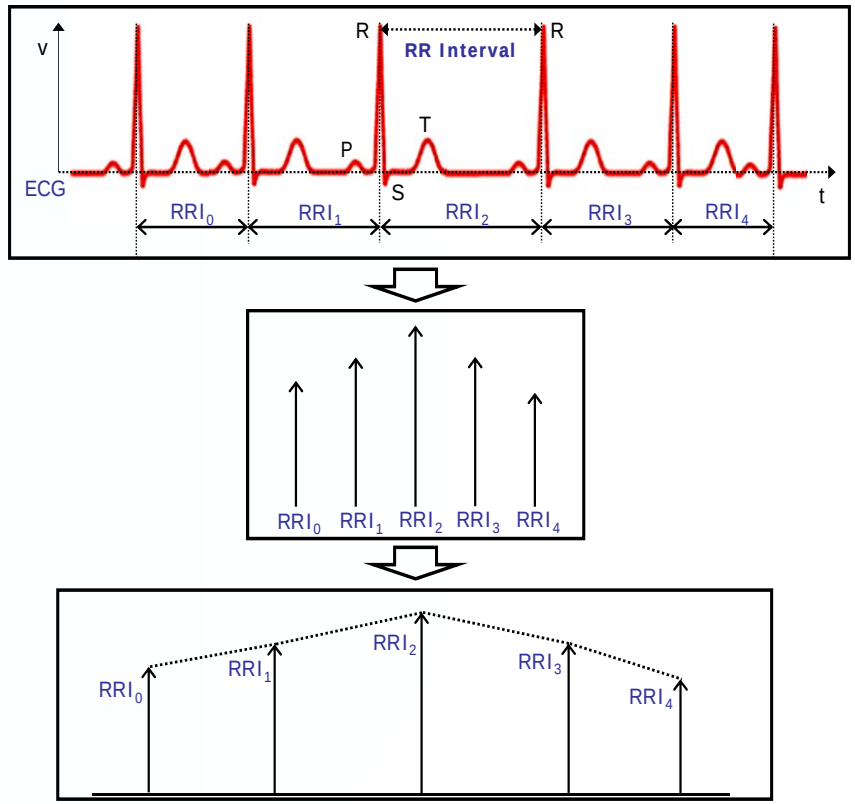
HRV analysis
: ECG processing

- Measuring ECG for each 3 position
  - The ECG signals are recorded during 5min. for each 3 recumbent position (Supine, Right, Left)



HRV analysis
: ECG processing ...

- Time series of RR intervals
  - RR intervals are the series of time intervals between heartbeat
  - The ECG are retrieved to measure the RR intervals by using S/W for the detection of R waves : Extract the R-peaks from the ECG recordings
  - RR intervals time series were re-sampled at rate of 4Hz to obtain power spectral density (frequency domain analysis)



HRV analysis
: Linear Feature Extraction

- Time domain of Linear Features
  - Simple time domain features include :
    1. RRm: the mean RR intervals
    2. SDRR: the standard deviation of all RR intervals
    3. SDSD: the standard deviation of differences between RR intervals
- Frequency domain of Linear Features
  - After Calculating the heart rate, performed the power spectrum analysis
  - Define areas of spectral peaks as follows:
    1. Total Power (TP): 0~0.4Hz
    2. Very Low Frequency (VLF) power: 0~0.04Hz
    3. Low Frequency (LF) power: 0.04~0.15Hz
    4. High Frequency (HF) power: 0.15~0.4Hz.
    5. Ratio LF/HF : LF to HF ratio
    6. Normalized LF (nLF): $$nLF = \frac{(TP - VLF)}{LF} \times 100$$
    7. Normalized HF (nHF): $$nHF = \frac{(TP - VLF)}{HF} \times 100$$

HRV analysis : Nonlinear Feature Extraction

Poincare Plots

- The Poincare Plots is constructed by plotting each RR interval against the previous one
- SD1: the standard deviation of the distances of points from y=x axis
- SD2: the standard deviation of the distances of points from y= -x + RR_{avg}
- SD2/SD1: the ratio SD2/SD1 is the measure of heart activity

Approximate Entropy

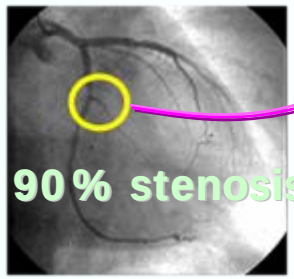
- The ApEn describes the complexity and irregularity of the signal
- The value of ApEn is low in regular time series (bad health) / high in complex irregular ones (good health)

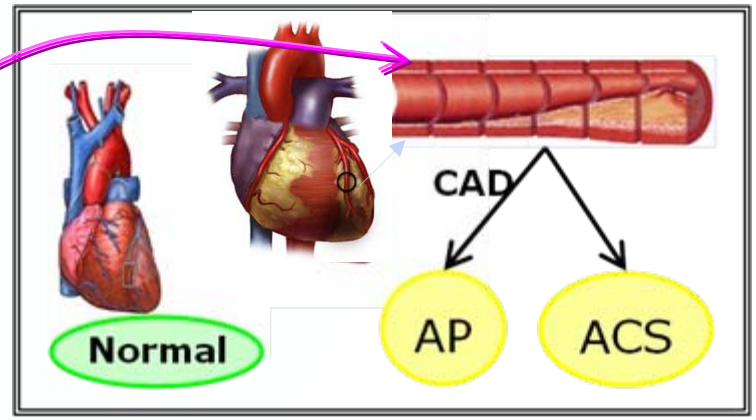
HRV analysis
: Nonlinear Feature Extraction ...

- Hurst Exponent, H
  - H is the measure of the smoothness of a time series based on asymptotical behavior of the rescaled range of the process
  - H = 1 : regular behavior (bad health)
  - 0.5 < H < 1.0 : normal (good health : 0.7)
  - H = 0.5 : random behavior

CASE study – II
: AP and ACS classification

- Real ECG datasets : Patients in five university hospitals
  - Patients with stenosis by using angiography
    - Luminal narrowing >= 50% : CAD Group
    - The others: Normal Group
  - CAD group was divided into two group by cardiologists, AP (Angina Pectoris) and ACS (Acute Coronary Syndrome)



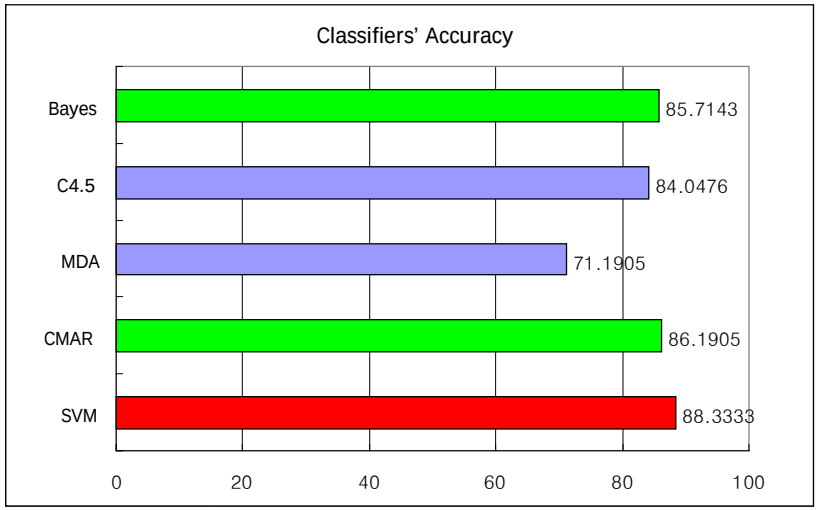


- Collect 193 ECG data from our ECG DB
  - 99 patients with CAD group and 94 normal group were studied

CASE study – II ...
: Classifiers Evaluation Results

- Analysis : Use of multi-parametric features of HRV

Classifier	TP rate	FP rate	Precision	Recall	F-measure	Class
SVM	**0.925**	**0.103**	**0.933**	**0.925**	**0.929**	**AP**
	0.820	**0.072**	**0.781**	**0.820**	**0.800**	**Control**
	0.815	**0.025**	**0.855**	**0.815**	**0.835**	**ACS**
Bayesian (TAN)	0.957	0.200	0.881	0.957	0.917	AP
	0.730	0.044	0.839	0.730	0.781	Control
	0.692	0.031	0.804	0.692	0.744	ACS
MDA	0.792	0.255	0.828	0.792	0.810	AP
	0.520	0.066	0.712	0.520	0.601	Control
	0.692	0.163	0.437	0.692	0.536	ACS
C4.5	0.918	0.182	0.886	0.918	0.902	AP
	0.780	0.059	0.804	0.780	0.792	Control
	0.631	0.051	0.695	0.631	0.661	ACS
CMAR	0.945	0.164	0.899	0.945	0.922	AP
	0.740	0.044	0.841	0.740	0.787	Control
	0.692	0.054	0.703	0.692	0.698	ACS



	Classifier	Class Label	Predicted Class		
			AP	Control	ACS
Actual Class	SVM	AP	**92.55%**	5.88%	1.57%
		Control	13%	**82%**	5%
		ACS	6.15%	12.31%	**81.54%**
	CMAR	AP	**94.51%**	1.57%	3.92%
		Control	17%	**74%**	9%
		ACS	15.38%	15.38%	**69.24%**

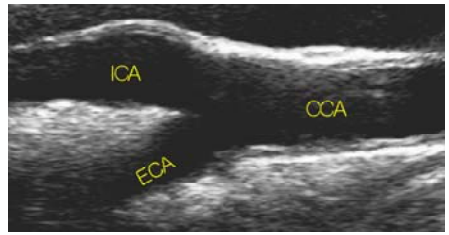
Confusion matrix

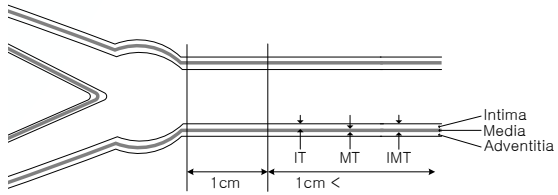
Outline

- Introduction
  - Application of data mining to medical data
  - Motivation and objective
- Background
  - ECG Analysis
  - ST-segments
  - Heart Rate Variability
  - Carotid arterial wall thickness
- Automated detection of Ischemic ECG beats
  - ST-segments feature extraction
  - CASE study1: Ischemia beat classification
- Heart Rate Variability
  - ECG processing, Linear/ Nonlinear features
  - Classification methods for diagnosing CAD
  - CASE study2 : AP and ACS diseases classification
- **Dyslipidemia diagnosis method**
  - **Measuring Intima-Media thickness**
  - **Feature extraction from IMT image**
  - **CASE study3: dyslipidemia classification**
- Discussion
- On going work
  - Development of Contrast data mining and frequent pattern mining methods
- Conclusion

Measuring Intima-Media thickness: KRISS

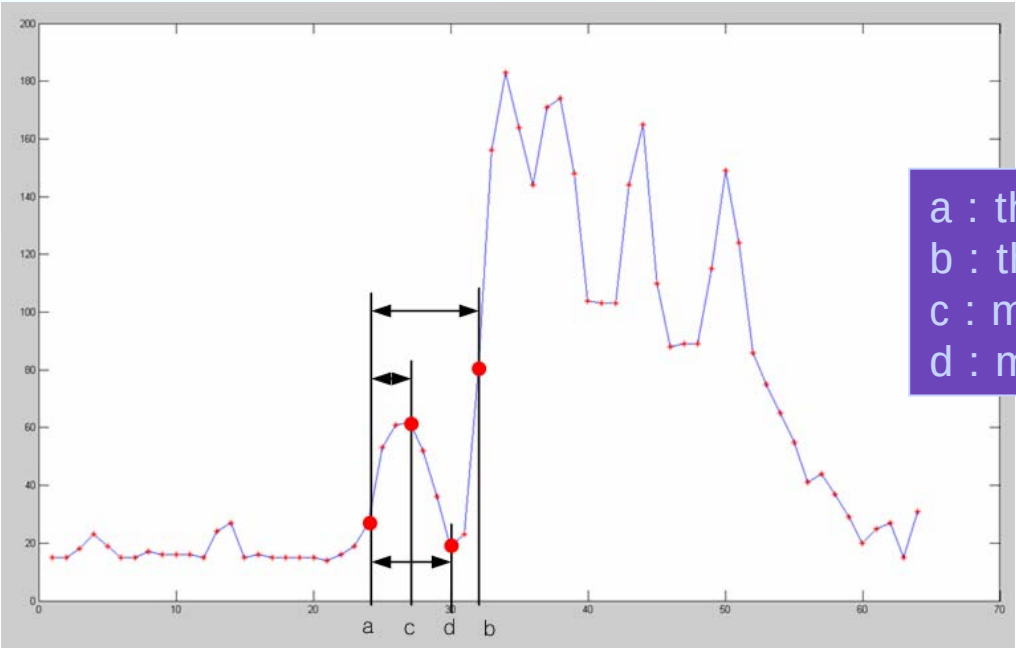
- IMT can be measured by analyzing the ultrasound images of Common Carotid Artery
  - measure the IMT at the far wall of a designated 1cm length of the right common carotid
  - calculate the average values over 200 points





Measuring Intima-Media thickness ... : KRISS

- CCA (Common Carotid Arterial) scanning with a high-resolution ultrasound system is done
  - Exam a cross-section of carotid artery
  - Detect four points (a, b, c, d)

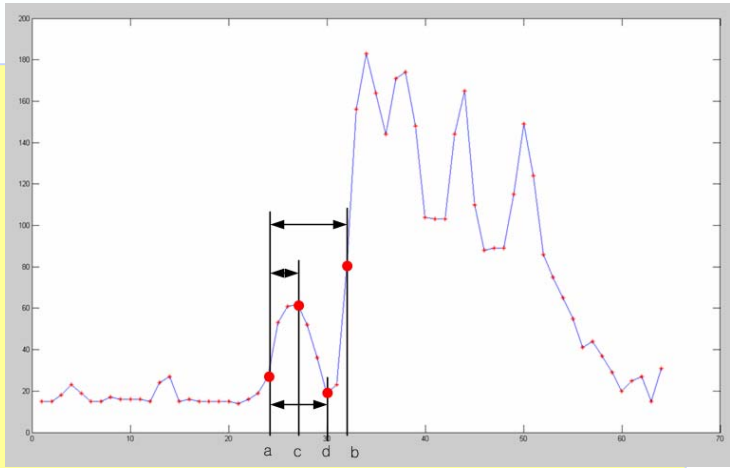


a : the start point of intima
b : the start point of adventitia
c : maximal point between a and b
d : minimal point between a a and b

Cross-sectional view of carotid artery wall

Feature extraction from IMT image: KRISS

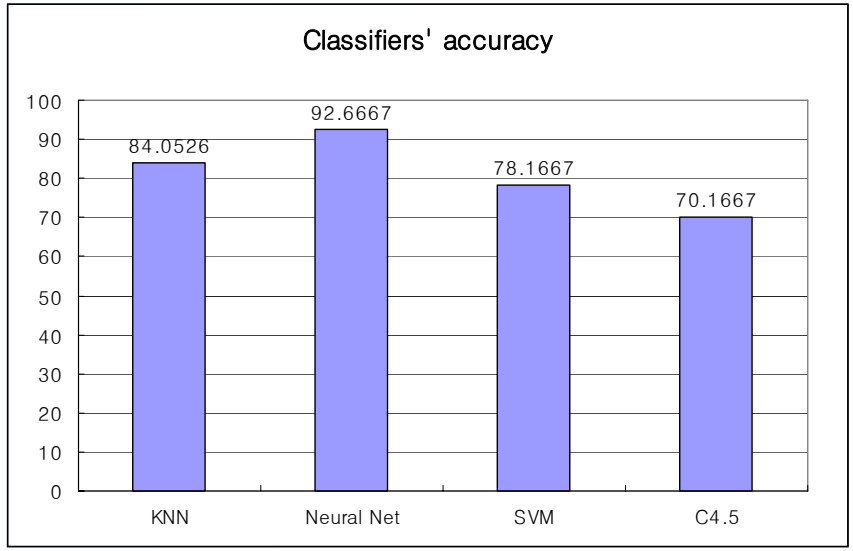
- In order to classify patients with dyslipidemia from normal group, total 15 features are extracted from ultrasound image



1. Fv01 : Distance between *a* and *b*
2. Fv02 : Area of Fv01[sum(*a*:*b*)]
3. Fv03 : Value of *c*
4. Fv04 : Distance between a and c
5. Fv05 : Area of Fv04 [sum(*a*:*c*)]
6. Fv06 : Value of *d*
7. Fv07 : Distance between a and d
8. Fv08 : Area of Fv07[sum(*a*:*d*)]
9. Fv09 : Gradient of straight line from value point of *a* to value point of *c*
10. Fv10 : Gradient of straight line from value point of *c* to value point of *d*
11. Fv11 : Gradient of straight line from value point of *a* to value point of *b*
12. Fv12 : Angle *acd*
13. Fv13 : Angle *cdb*
14. Fv14 : Value of *c*-value of *a*
15. Fv15 : Value of *c*-value of *d*

CASE study – III
: dyslipidemia classification using IMT features

- Collect 600 real IMT dataset
  - 328 patients with dyslipidemia and 272 patients with normal group were studied
  - Performance evaluation: *Accuracy and confusion matrix*
  - Result: suitable method for predicting patients with dyslipidemia : *Neural Net*



	Classifier	Class Label	Predicted Class: Dyslipidemia	Predicted Class: Normal
Actual Class	Neural Net	Dyslipidemia	**299**	29
		Normal	15	**257**
	SVM	Dyslipidemia	**271**	57
		Normal	48	**224**
	KNN	Dyslipidemia	**261**	67
		Normal	64	**208**
	C4.5	Dyslipidemia	**253**	75
		Normal	104	**168**

Discussion

- In this work, in order to predict and diagnose cardiovascular disease such as ischemia, coronary artery disease (AP, ACS)
  - Used ST-segments features of ECG
  - Suggested a possibility that multi-parametric feature of HRV may be helpful to make a diagnosis the heart disease
  - Extract useful features from carotid artery wall image
- Also, we applied data mining techniques (classification methods and association rules) to enhance the reliability of the diagnosis and detect early symptoms of heart problems

On going work

- Development of Contrast data mining and frequent pattern mining methods
  - Contrast data mining : Emerging Pattern Mining-Based Methodology for Automated Diagnosis of Cardiovascular Diseases
  - Frequent pattern mining : Compressed Patricia FP-Tree for Frequent Itemsets Mining

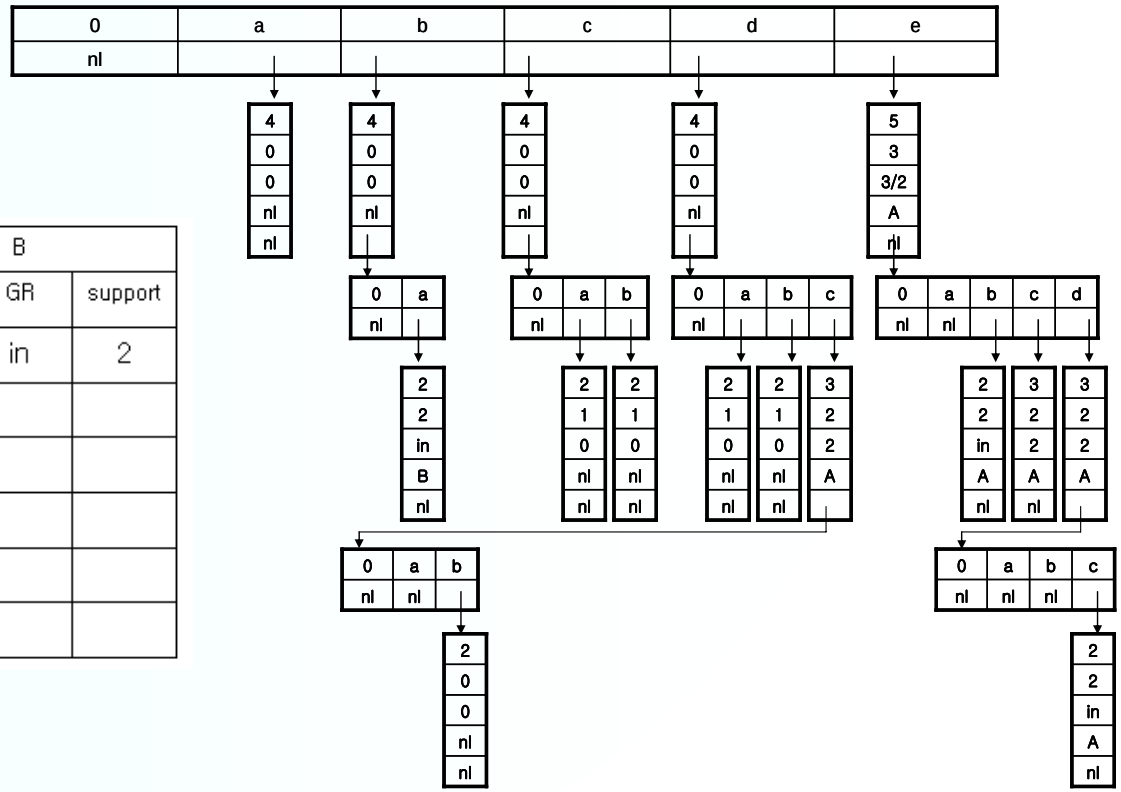
Emerging Pattern Mining-Based Methodology

- Propose a special type of Emerging Patterns (IEPs) and show that they are high quality patterns for building accurate classifiers
- Generalize the interestingness measure for EPs, including the min. support, min growth rate, the subset relationship between EPs and correlation based on common statistical measure (chi-square value)
- Develop tree-based pattern growth algorithms for mining only those interesting EPs

Emerging Pattern Mining-Based Methodology … : T-tree & P-tree structure

- Emerging Patterns (EPs) are contrast patterns between two classes of data whose support changes significantly between the two classes
  - In this study, we extend EPs mining algorithm using TFP algorithm

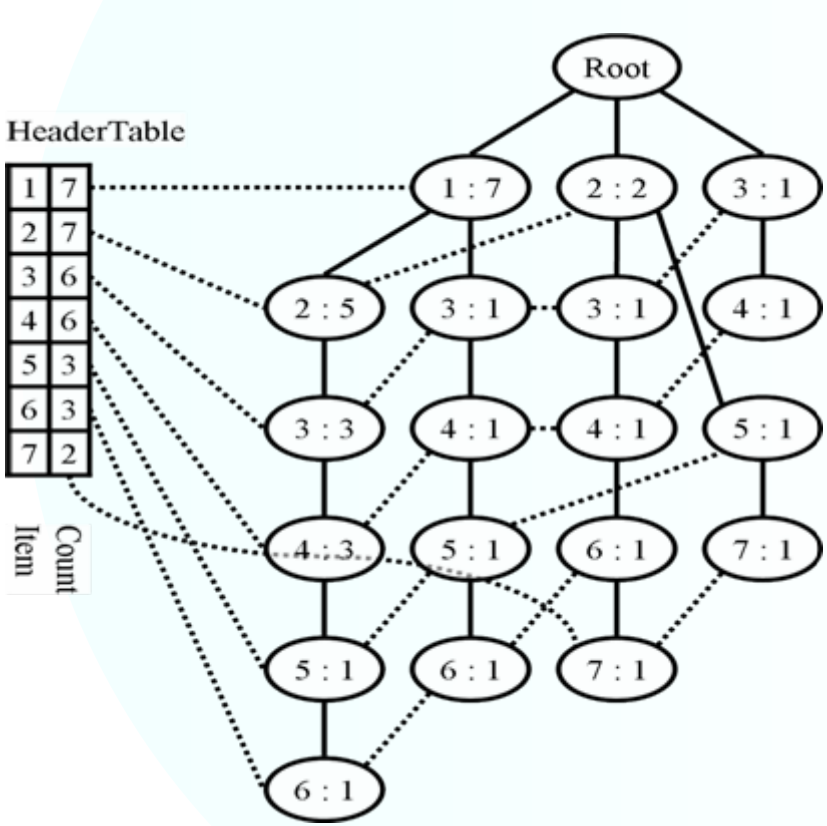
A			B		
itemset	GR	support	itemset	GR	support
e	3/2	3	ab	in	2
cd	2	2			
be	in	2			
ce	2	2			
de	2	2			
cde	in	2			

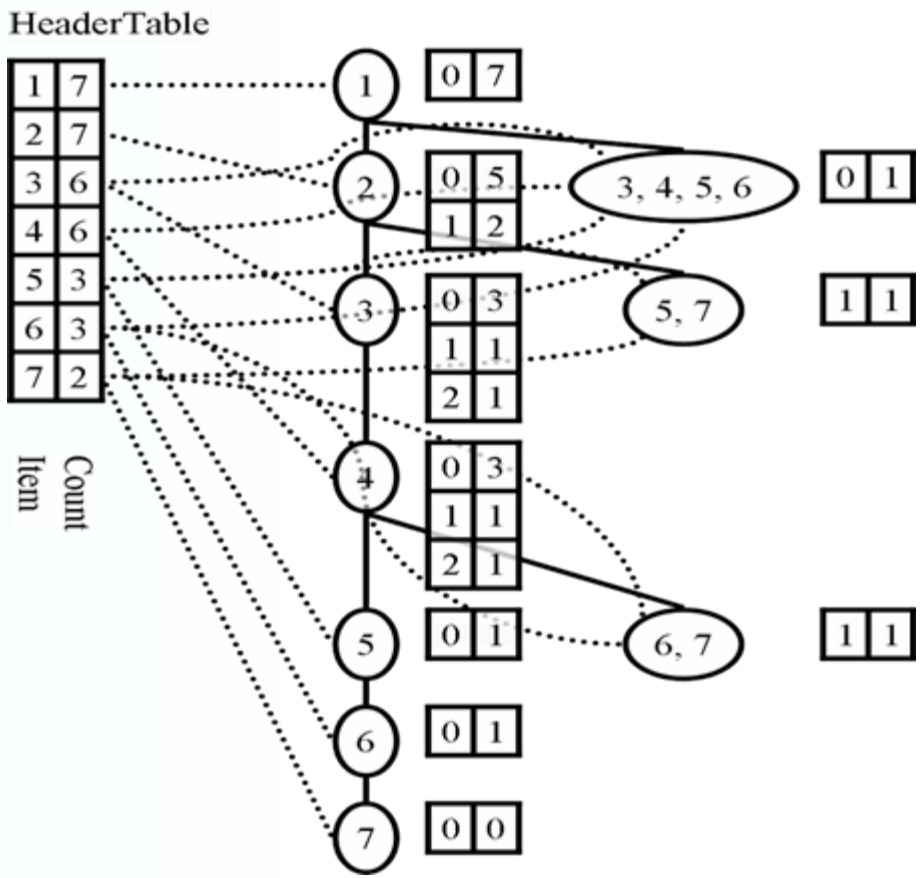


Compressed Patricia FP-Tree for Frequent Itemsets Mining

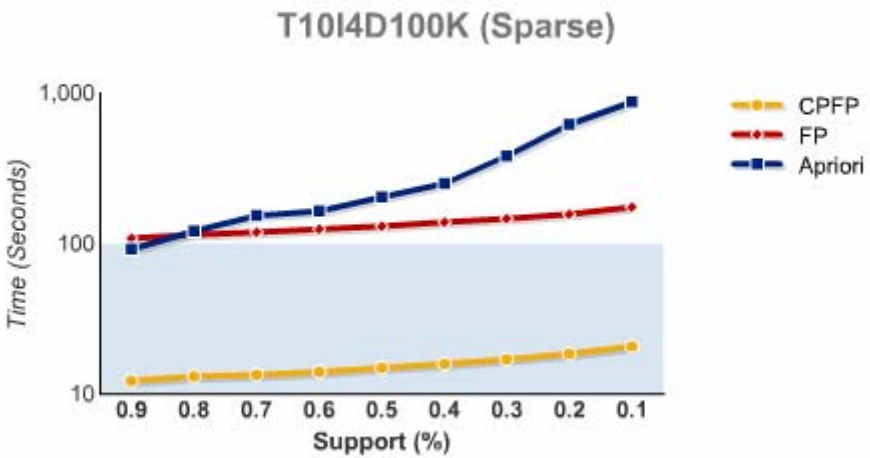
- Compressed Patricia FP-Tree (CPFP-Tree)
  - New data structure
  - Can find fast frequent itemsets
  - Sparse frequent itemsets
  - Dense frequent itemsets
- CPFP-Tree applications
  - Finding frequent itemsets
  - Association rules mining & other mining
  - System with limited memory
  - Sparse & dense frequent itemsets

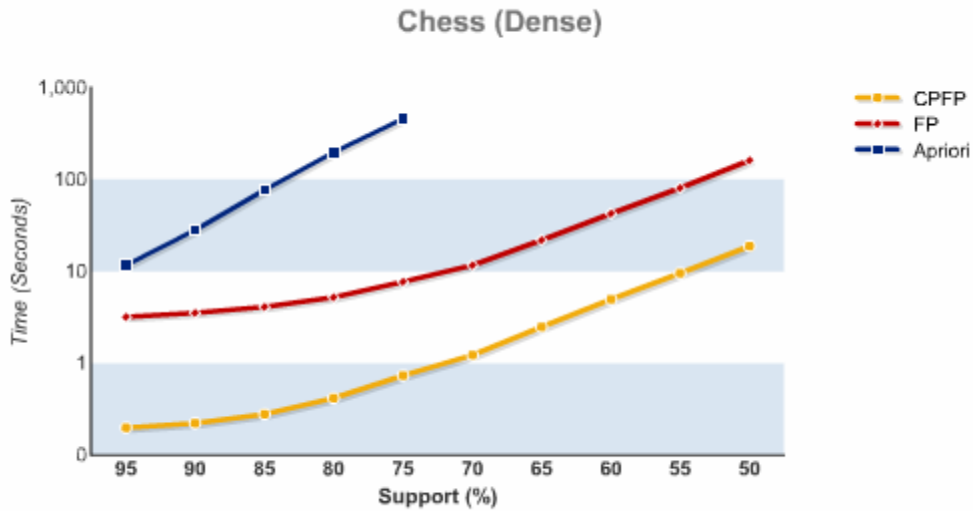
CPFP-Tree: example





CPFP-Tree: performance evaluation





Conclusion

- In this study,
  - We proposed data mining framework for coronary heart disease perdition
  - Automated diagnosis methodology using ST-segments, HRV features and IMT features
  - Our experiments showed that Function -, Pattern-based classifiers were the most accurate in comparison with all other classifiers for diagnosing heart diseases
- On going work,
  - We proposed emerging pattern mining-based Methodology for Automated Diagnosis of Cardiovascular Diseases
    - Extend emerging pattern based classifier – tree-based
    - Introduce the interestingness measure, Chi-square
    - Directly generate EPs and remove uninformative patterns
  - We developed a novel fast algorithm for discovering frequent itemsets
    - CPFP-Tree is efficient and scalable
    - Dense and Sparse frequent itemsets mining tasks
    - CPFP-Tree is faster than the Apriori and FP Growth algorithm

Thank You
Any Question?

- This work was supported by
  - The Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government(MOST) (R01-2007-000-10926-0)
  - The Korea Research Foundation Grant funded by the Korean Government(MOEHRD)" (The Regional Research Universities Program/Chungbuk BIT Research-Oriented University Consortium)
  - a grant from the Personalized Tumor Engineering Research Center(PTERC) and the Korea Science and Engineering Foundation(KOSEF).