Ph.D. Dissertation Defense Presentation

Boosted Feature Selection for Class Dedicated SVM and its Application in Fetal Health Prediction

Jinpyo Lee

1. Introduction

Motivation of Research

- SVM is high performance machine learning algorithm with advantages in high correct classification rate (CCR) and ability to avoid overfitting.

- This research intends to develop new feature selection / extraction and classification methodologies by overcoming high computational complexity on large-scale or multiclass data.

- For the purpose, this research searches for efficient algorithm by applying to Cardiotocography data, which is used in medical diagnosis of fetal state.

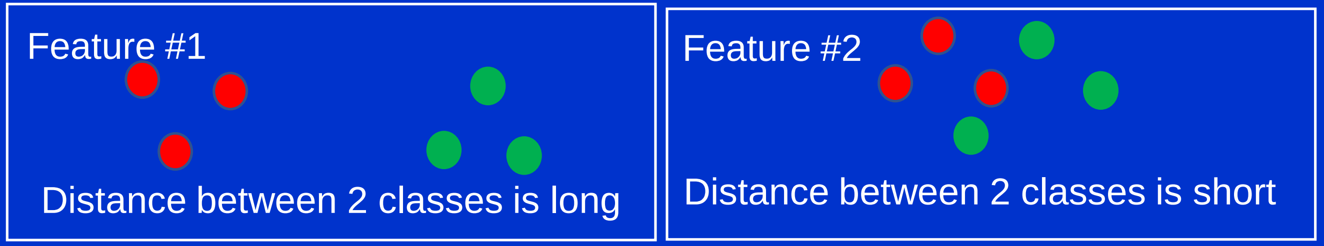
Motivation of Research – Boosted Feature Selection

- Feature ranking methodologies apply reasonable criteria to individual feature.

- Applying the criteria to all instances of individual features is inefficient.

- The instances can be divided into 2 groups:
  (1) Easy to classify (2) Hard to classify.

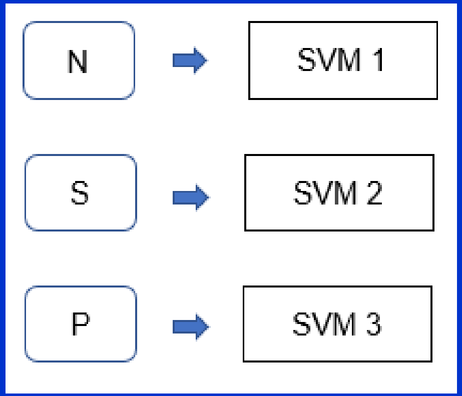
- Feature selection focusing on (2) group is more efficient methodology to increase classification performance on Cardiotocography data.



Example of features with long vs. short distance

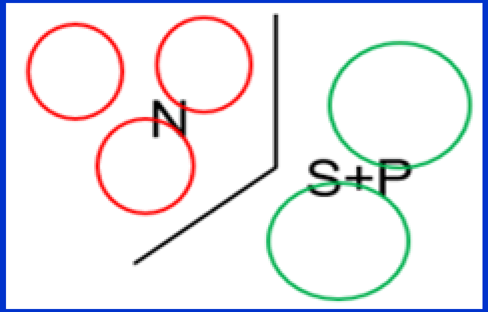
Motivation of Research – Class Dedicated SVM

- SVM is originally binary classifiers, developed for binary classification.
- Either One vs. One or One. vs. All classification architecture should be selected for multiclass classification.
- The classification performance of One vs. All classification is better even if it is more computationally expensive.
- One vs. All is favorable in feature selection depending on each binary classifications.



Motivation of Research – Feature Extraction

- Feature extraction can increase classification performance by creating newly extracted features.



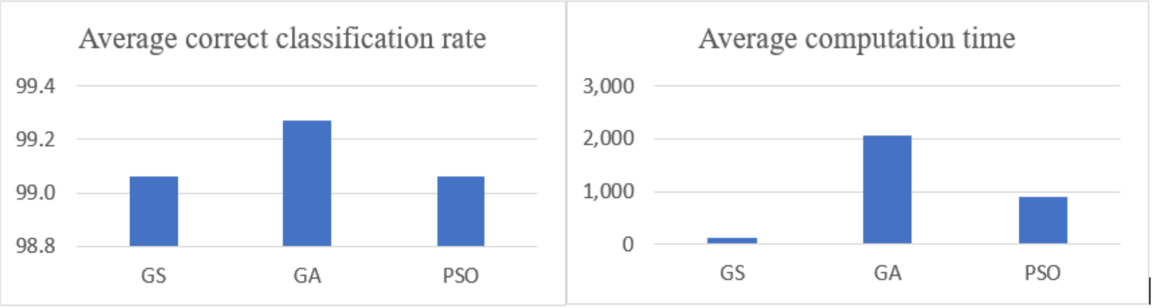
- New features extraction by clustering algorithm overcomes the disadvantage of one vs. all classification architecture, i.e., unbalanced number of instances in each class of binary classifications.

- New feature reconstruction by clustering algorithm resulted in improved sensitivity in literature of Cardiotocography data (Chamidah and Wasito, 2015)

Problem Description

The detailed information with three optimization methods.

The optimization method	Average accuracy (%)	Average computing time (sec)
GS	99.0583	116.52
GA	99.2708	2063.60
PSO	99.0625	908.45



Comparison of accuracy and computation time in 3 methods in Gao an Hou (2016) 's research

GS : SVM with Grid Search of parameters

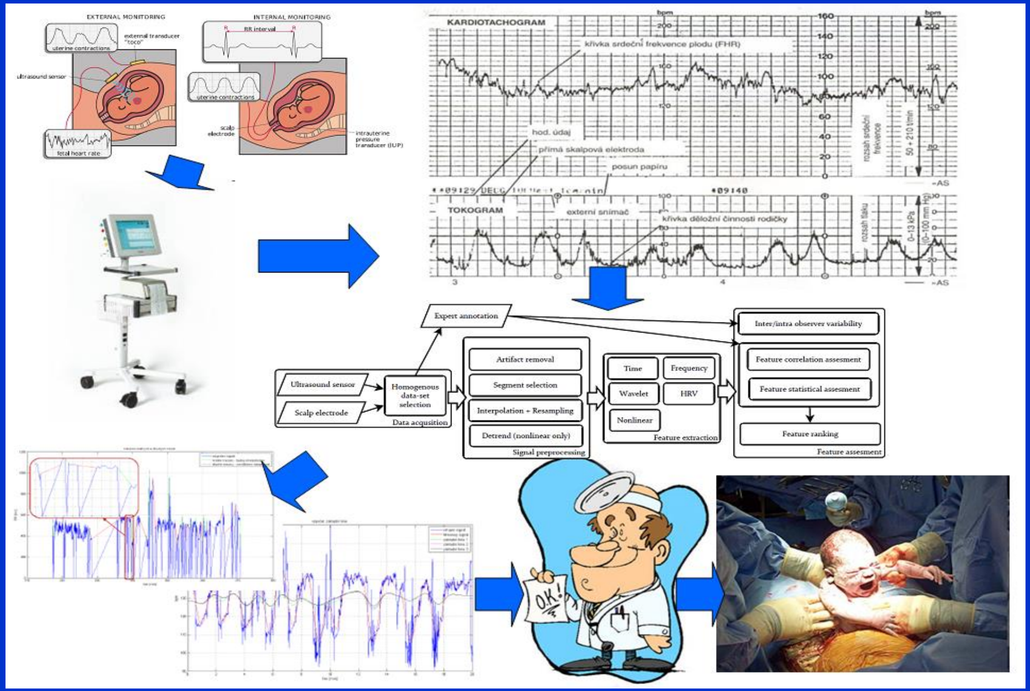
GA : Genetic Algorithm

PSO : Particle Swarm Optimization

Comparison of 3 optimization methods

- High complexity of SVM is due to parameter optimization of RBF by Grid Search.
- GS is more efficient than other optimization methods, GA or PSO with same CCR.
- This research intends to develop methodology for highest CCR and less complexity.

Problem Description



The flow of diagnosis activity using Cardiotocography data

- Cardiotocography data is used in diagnosing fetal status until delivery.
- 3-class which consists of normal, suspect and pathologic class
- High dimensionality with 21 features

Fetal Disease Description

The pathologic fetal state can be caused by the following disease conditions.

- Congenital malformation on bone, digestive organs etc.
- Hemolytic disease: Blood disorder in a fetus
- Hemorrhagic disease: Lack of blood clotting in a fetus
- Fetal obstructive uropathy: Abnormality in kidney or urine function.
- Low weight fetus: Fetus are not growing normally while pregnancy.

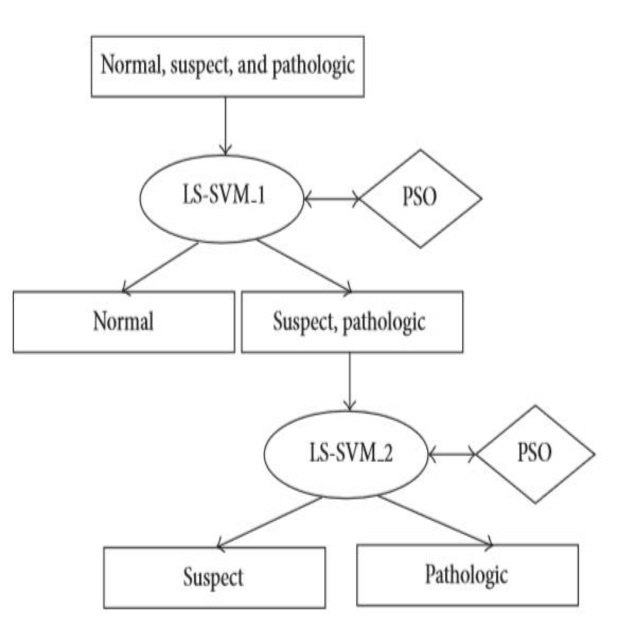
Problem Description

The details of 3 classes

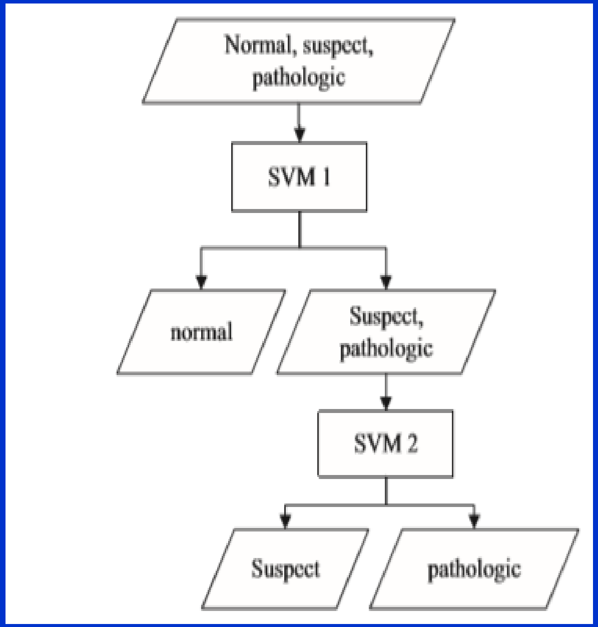
No.	Code	Fetal State	Number of Instances	Ratio
1	N	Normal	1,655	77.8%
2	S	Suspect	295	13.9%
3	P	Pathologic	176	8.3%
Sum			**2,126**	100.0%

- Classification performance in literature is low CCR 91.6%, sensitivity 0.852
- 14.8% of pathologic status is incorrectly classified.
- Low reliability and inefficiency of decision support system
- Additional medical examination and cost are needed.

Problem Description



Yilmaz and Kilikcier, 2013

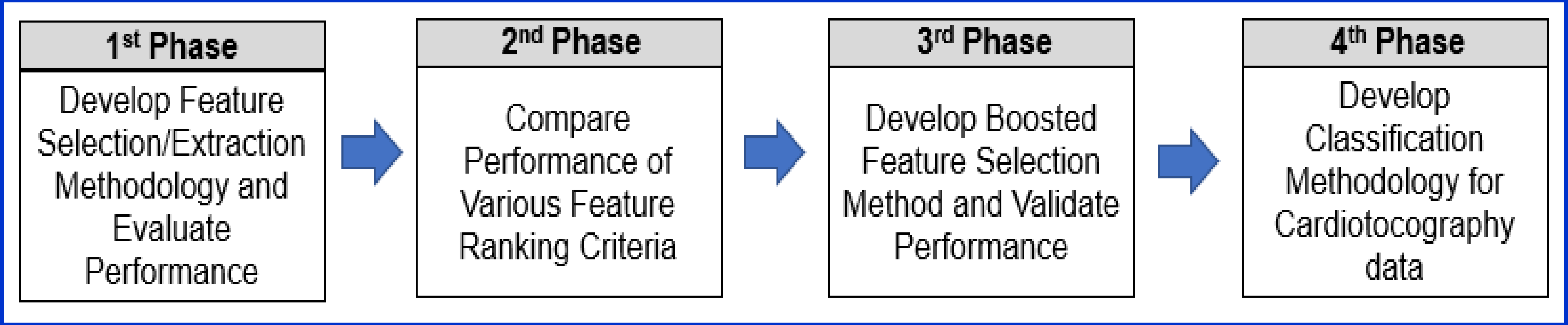


Chamidah and Wasito, 2015

- Binary Decision Tree (BDT) architecture has a limitation.
- The misclassifications from SVM 1 negatively affects performance of SVM 2.
- This research developed improved feature selection / extraction and classification methodology to increase the accuracy.

Research Framework

Four Phases of Research Flow



- 1st Phase : Develop feature ranking – PCA ensemble and validate performance
- 2nd Phase: Compare feature ranking criteria- LDA,PCA, Distance between classes
- 3rd Phase: Develop ranking method using distance among misclassified instances
- 4th Phase: Apply boosted feature selection, clustering and class-dedicated SVM.

Research Objectives

- Develop classification methodology by applying
  - Various ranking criteria (CCR, PCA, LDA, Distance between classes)
  - Various kernels (Linear, Polynomial, Sigmoid, Radial Basis Function)
- Develop efficient algorithm depending on feature type
- Develop feature selection / extraction methodology for Cardiotocography data
- Develop classification methodology for multiclass Cardiotocography data, overcoming the limitation in literature.

Research Contribution and Significance

- SVM ensemble achieves equivalent or higher CCR with less time complexity compared to literature.
- Developed boosting-based feature selection methodology, and evaluated the effectiveness.
- Developed improved classification methodology for Cardiotocography data and validated the effectiveness.
- Reliable and efficient decision support system to diagnose fetal status by predicting pathologic status accurately.

Research Uniqueness and Contribution

- This research developed efficient ensemble algorithm by kernels selection in SVM and its combination with feature selection / extraction methods.

- This research developed new feature selection methodology based on distance between two classes on misclassified instances from SVM.

- This research developed new feature extraction methodology from clustering by adjusting number of clusters for improved classification on multiclass data.

- This research used Class-dedicated classification architecture for Cardio-tocography data, contributing to building accurate decision support system.

2. Literature Review

Comparison between Feature Selection and Feature Extraction

- **Feature selection** : Process to select features which contribute most to prediction variable or output

- **Feature extraction**: Process to extract new features which are informative and not redundant

Method	Advantages	Disadvantages
Selection	Preserving data characteristics for interpretability	Discriminative power Lower shorter training times Reducing overfitting
Extraction	Higher discriminating power Control overfitting when it is unsupervised	Loss of data interpretability Transformation maybe expensive

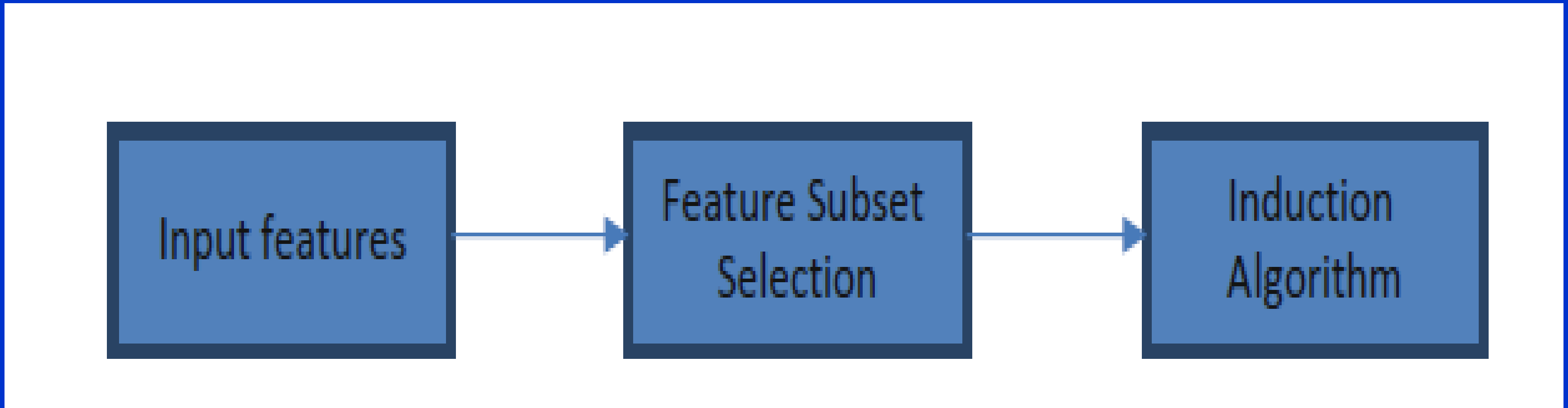
Comparison between Feature Selection and Extraction (Hira and Gillies, 2015)

Feature Selection

Three Major Categories in Literature

Filter, Wrapper, Embedded method

1. Filter method - Feature selection independent of the learning algorithm.



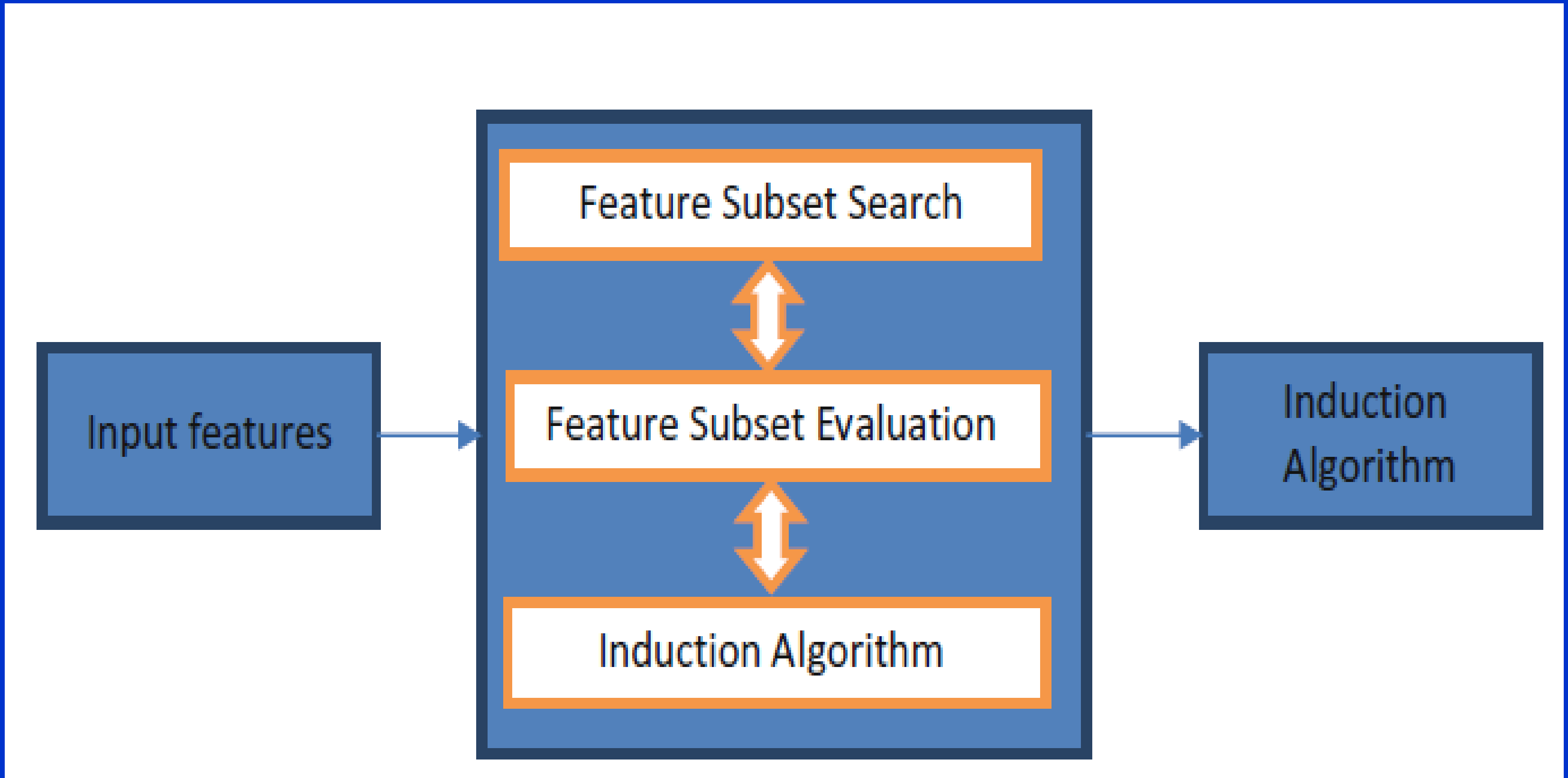
Example of Filter Method

- Variable Ranking - Sort or arrange features of data according to certain criteria. (Chang, Y.W et al. 2008)

- mRMR - Penalize the feature's redundancy and maximize relevance of a feature set for the class. (Peng, H. et al, 2005)

- Others: Relief-F, Chi Squared (CS), Gain Ratio (GR)

2. Wrapper method - Feedback from classifier is used to evaluate the quality of selected features.

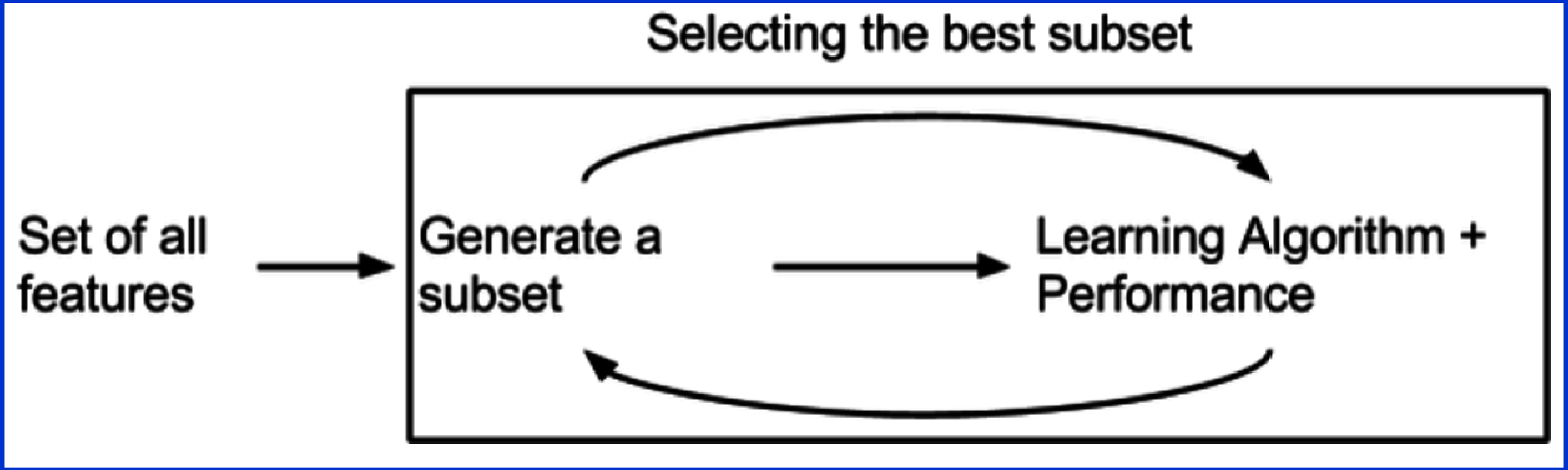


Example of Wrapper Method

- Particle Swarm Optimization (PSO)- Optimizes a problem by iteratively trying to improve a candidate solution
(Unler & Murat 2010,Yilmaz & Kilikcier, 2013)

- Genetic Algorithm (GA) - Search problems by using operator such as mutation, crossover and selection.
(Huang, C.L. et al. 2006, Zhang & Yang 2008, Ocak, H, 2012)

3. Embedded method - Construct feature subsets as part of building a classifier

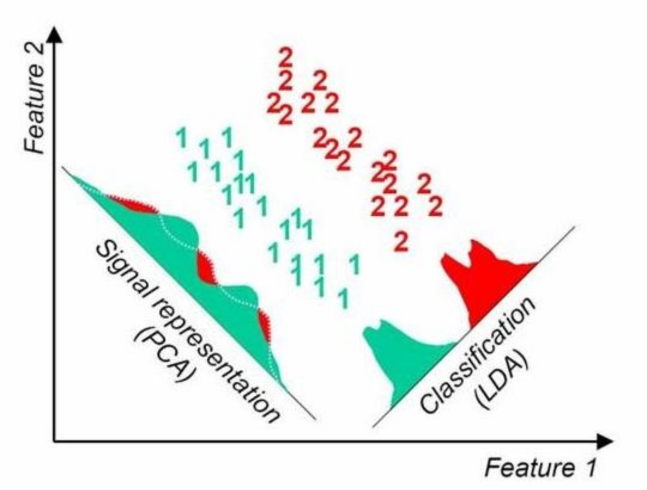


- Example: Genetic Algorithm + Iterated Local Search (Duval, 2009)

Feature Extraction

1. **Principal Component Analysis (PCA)** - Extracts uncorrelated features in smaller dimensions. Unsupervised method.
   (Zhai,G. et al. 2015, Gao, X. et al. 2016)

2. **Linear Discriminant Analysis (LDA)** - Reduces the dimension by maximizing the ratio of between-class scatter to within-class scatter
   (Safo & Ahn, 2016, Silva, A et al. 2016, Uncini, A et al. 2017)

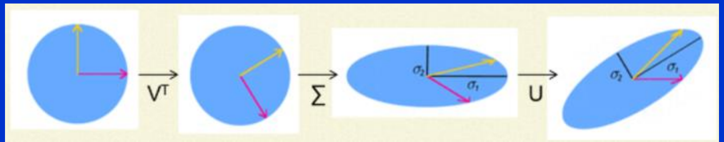


Feature Extraction

3. **Canonical Correlation Analysis (CCA)** - Linear combinations of 2 vectors which have maximum correlation with each other.
   (Wang, Z. et al. 2007, Shen, C. et al. 2014)

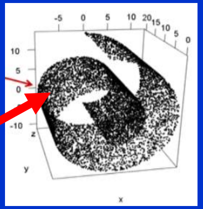
4. **ISOMAP** – Nonlinear dimensionality reduction method
   (Park, H. 2012, Bu,Y. et al. 2014)

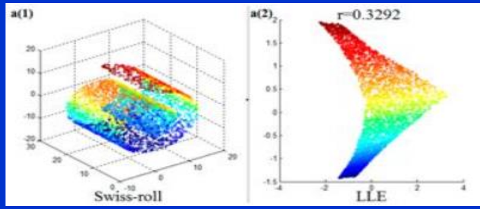
$$A=U\Sigma V^T$$



5. **Locally Linear Embedding (LLE)** – A method similar to ISOMAP, but more efficient. (Liu, X. et al. 2013, Coy,B. 2012)

Manifold



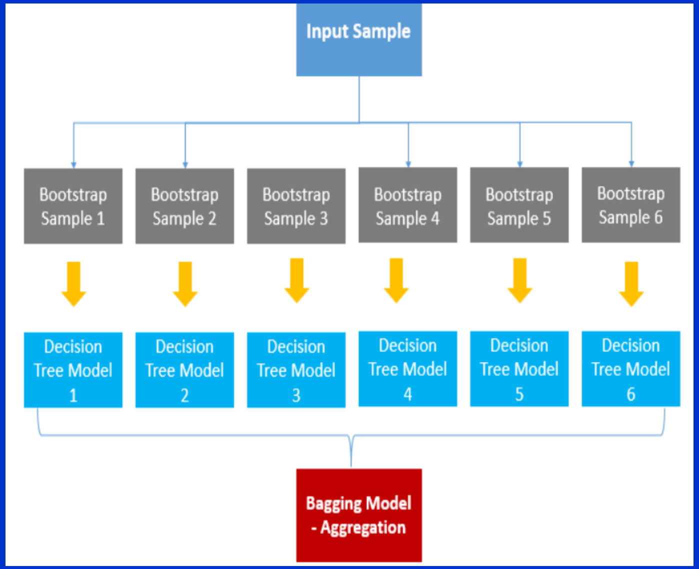


Classification

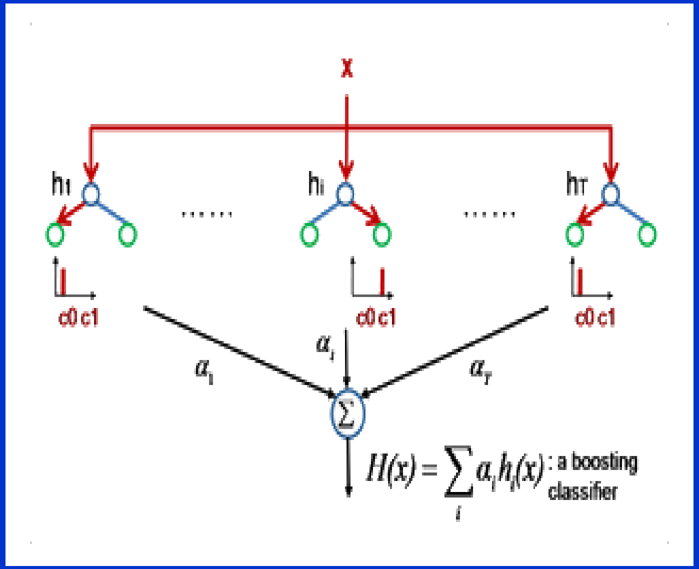
- **Ensemble** – Results in CCRs with higher reliability compared to single model.

  1. **Bagging:** training identical model by restored random sampling data

  2. **Boosting:** assign weighted value on model which solved the difficult problem.



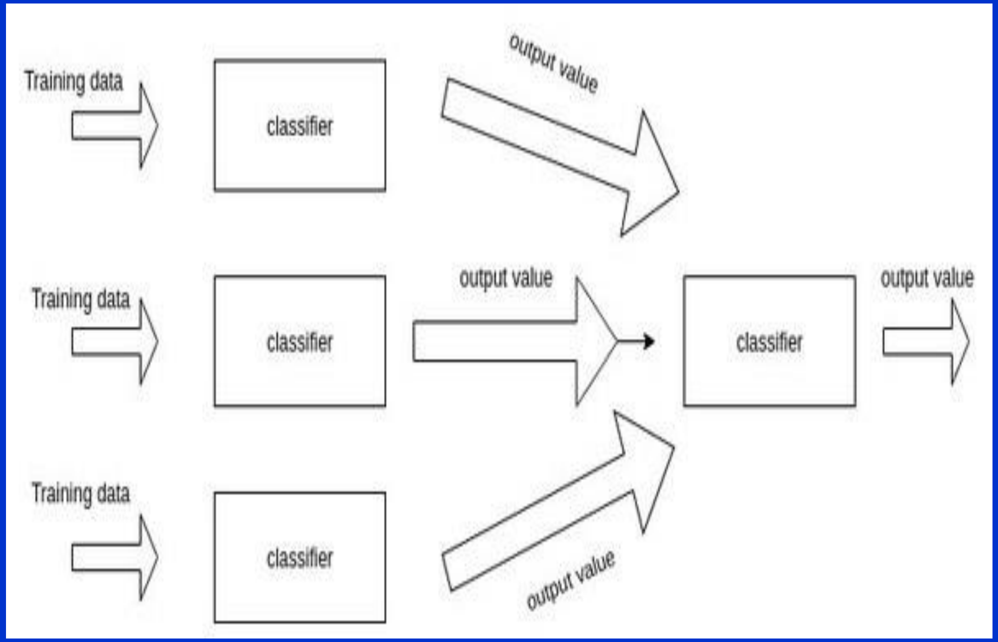
Bagging



Boosting

Classification

3. **Stacking :** create model for best performance by combining different models.



Stacking

Ensemble method is used in the literature, Huang et al. 2017, Pujari & Gupta, 2012, Zhang & Yang, 2008, etc.

Summary of Literature on Feature Selection/Extraction and Ensemble

- Wrapper method has an advantage of higher CCR compared to filter.
- PCA is effective in dimensional reduction and noise elimination.
- Ensemble methods are effective in increasing performance of model.
- The simultaneous use of various kernels has been researched.
- Lack of literature researching the advantage and disadvantage of kernels.
- Classification performance and complexity are in trade-off relation.

Literature on Feature Selection / Classification of Cardiotocography Data

- LS-SVM and Particle Swarm Optimization by using BDT architecture

  (Yilmaz & Kilikcier, 2013)

  - Overall CCR 91.62%, sensitivity 0.767, specificity 0.969.

- Evaluation of fetal well-being by using SVM and Genetic Algorithm (GA) on 2-class (only normal & pathologic) data

  (Ocak, H ,2013)

Literature on Feature Selection / Classification of Cardiotocography Data

- Hybrid K-means and SVM in classification of fetal state by using BDT architecture and feature extraction (Chamidah & Wasito, 2015)

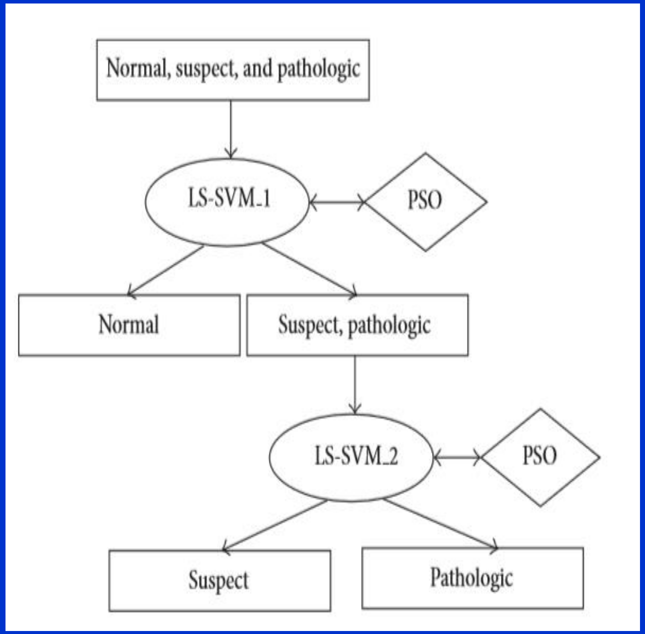
  - CCR, 90.6%. sensitivity 0.852, specificity 0.912. Seven extracted features used.

- Hybrid K-means and SVM for breast cancer diagnosis (Zheng, B et al. 2014)

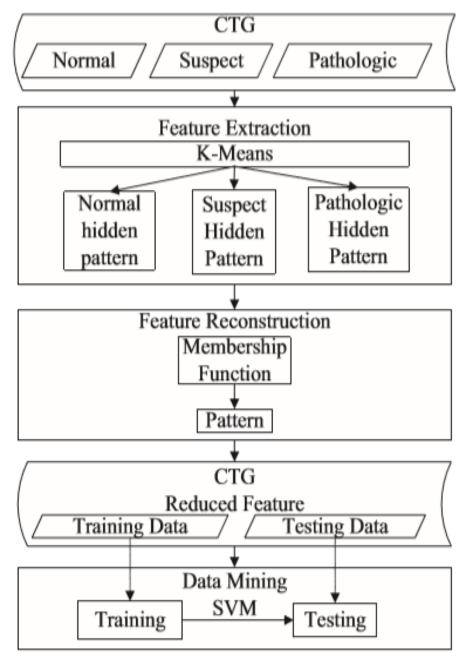
  - Reduced computation time by maintaining the highest accuracy in literature.

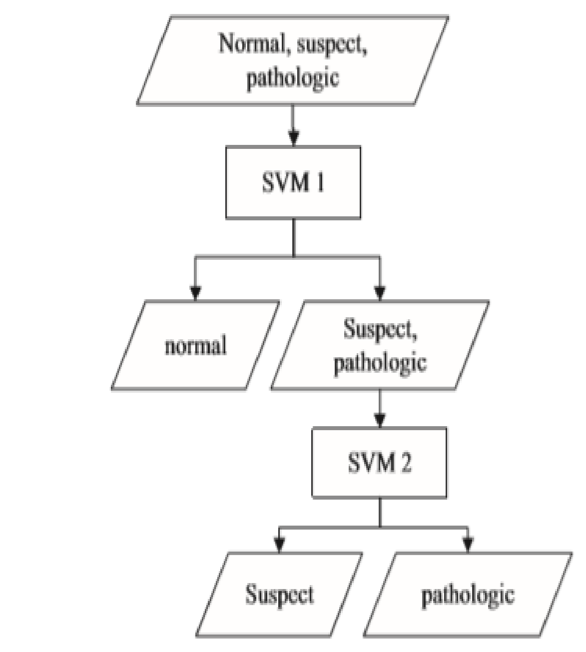
Literature on Feature Selection / Classification of Cardiotocography Data

- In literature of 3-class Cardiotocography data, Class-dedicated architecture has not been researched.
- Clustering algorithm has not been used with class-dedicated SVM.



The architecture of method
in Yilmaz and Kilikcier, 2013





Research method
in Chamidah and Wasito, 2015

Literature Review on Performance Criteria

- There are 2 kinds of literatures :   2-class data vs. 3- class data

1) Literature on 2-Class Cardiotocography data

- The definition of sensitivity and specificity has been used in literature.

  (Krupa, N et al. 2011, Ocak, 2013)

- Only the literature which used 2-class (Normal & Pathologic) Cardiotocography data used the terms ‘sensitivity’ and ‘specificity’.

- The 2-class classification methodology is not applicable to actual diagnosis activity because it distorts actual patterns of all patients.

Literature Review on Performance Criteria

2) Literature on 3-Class Cardiotocography data

- The literature of this category calculated the correctly classified ratio per each class, not referring them as sensitivity or specificity.
- The term ‘The CCR of class 2’ , which is used in this dissertation, has not been used in literature.
- However, it is the same as ‘the percentage of suspect data points which are correctly classified as suspect’. (Yilmaz & Kilikcier, 2013)

Gaps in Literature and Methodologies in This Research

- 6 Gaps in literature and methodologies in this research are summarized.

No.	Referenced literature	Gaps	The proposed methodology in this research
1	Zhai, G. et al. (2015) Gao, X. et al. (2016) Maldonado, S. et al. (2009) Li and Sun (2011)	Feature classification rate ranking method and PCA were not used simultaneously.	Feature classification rate ranking method and PCA were used complementarily to merge the advantages of both algorithms.
2	Li and Sun (2011) Abdiansah, A. et al. (2015) Gao, X. et al. (2016) Lee, S.B. et al. (2017)	Reducing instances of training data to reduce computation time for grid search, was not used.	This research reduced instances of training data to reduce computation time for grid search.

Gaps in Literature and Methodologies in This Research

No.	Referenced literature	Gaps	The proposed methodology in this research
3	Huang & Wang (2006) Chang & Lin (2008) Maldonado, S. et al. (2009) Chen, G. et al. (2015) Lin, X. et al. (2018)	Searching for algorithms depending on feature type was not used.	This research searched for algorithms depending on feature type to reduce the computation time further.
4	Wang, Z. et al. (2007) Bhavsar, H. et al. (2012) Wang, Z. et al. (2014) Abdiansah, A. et al. (2015) Gao, X. et al. (2016) Huang, M.W. et al. (2017)	Various options regarding the choice between the correct classification rate and computation time were not provided.	This research provided various algorithms with different correct classification rate and computation time

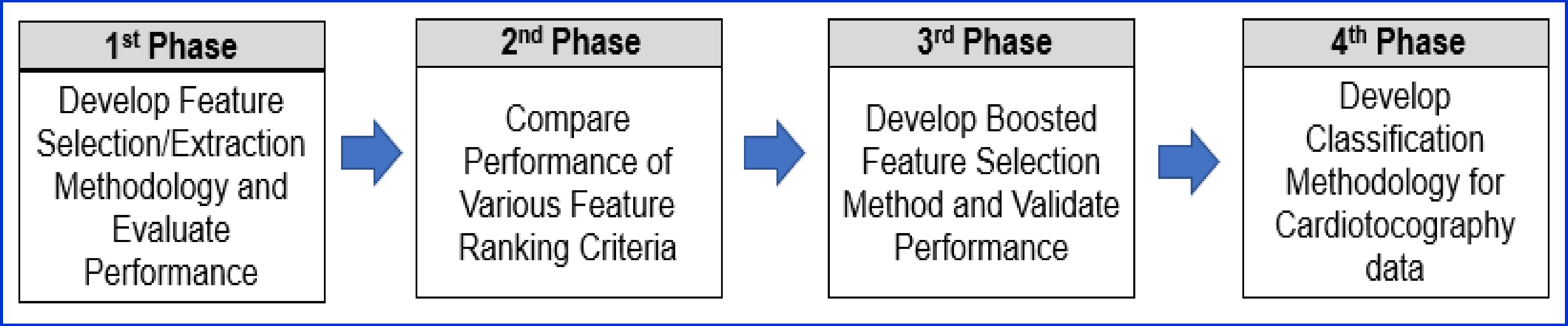
Gaps in Literature and Methodologies in This Research

No.	Referenced literature	Gaps	The proposed methodology in this research
5	Chang, Y.W. et al. (2008) Maldonado, S. et al. (2009) Ocak, H (2012) Ocak, H (2013) Chamidah & Wasito (2015) Yilmaz & Kilikcier (2013) Wang & You (2013)	Boosted feature selection methodology by using wrapper method based on sorting according to the distance between classes among misclassified instances, has not been used.	This research used boosted feature selection methodology and proved the effectiveness by applying to other data and classifiers.
6	Ocak, H (2012) Ocak, H (2013) Chamidah & Wasito (2015) Yilmaz & Kilikcier (2013)	Class-dedicated SVMs have not been applied to the classification of 3-class Cardiotocography data.	This research developed class-dedicated classification architecture to increase the performance of classification methodology for Cardiotocography data.

3. Methodology

Research Framework

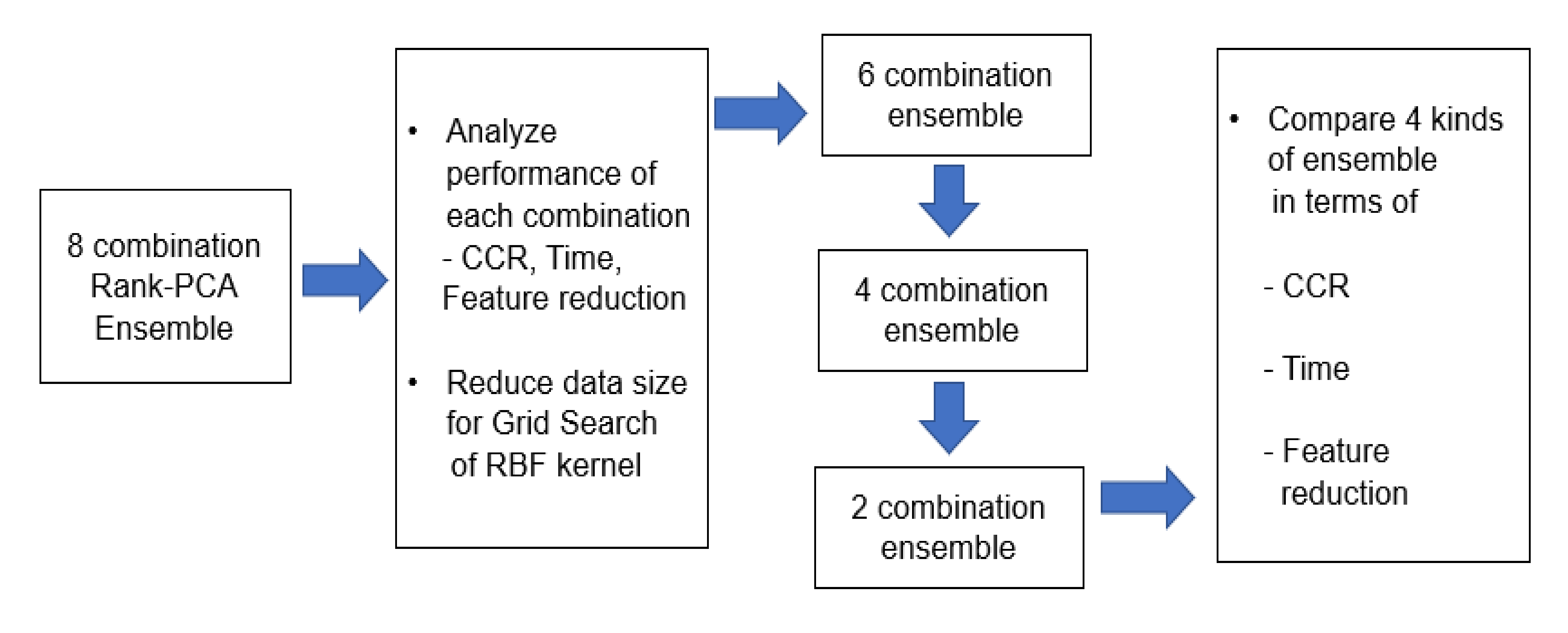
Four Phases of Research Flow



- 1st Phase : Develop feature ranking – PCA ensemble and validate performance
- 2nd Phase: Compare ranking criteria i.e, LDA,PCA, Distance between classes
- 3rd Phase: Develop ranking method using distance among misclassified instances
- 4th Phase: Apply boosted feature selection, clustering and class-dedicated SVM.

1st Phase: Develop Feature Selection / Extraction Methodology

Developing 4 kinds of Feature ranking-PCA ensemble algorithms



- Develop and evaluate performance of feature ranking – PCA ensemble algorithms by composing different combinations with kernels.

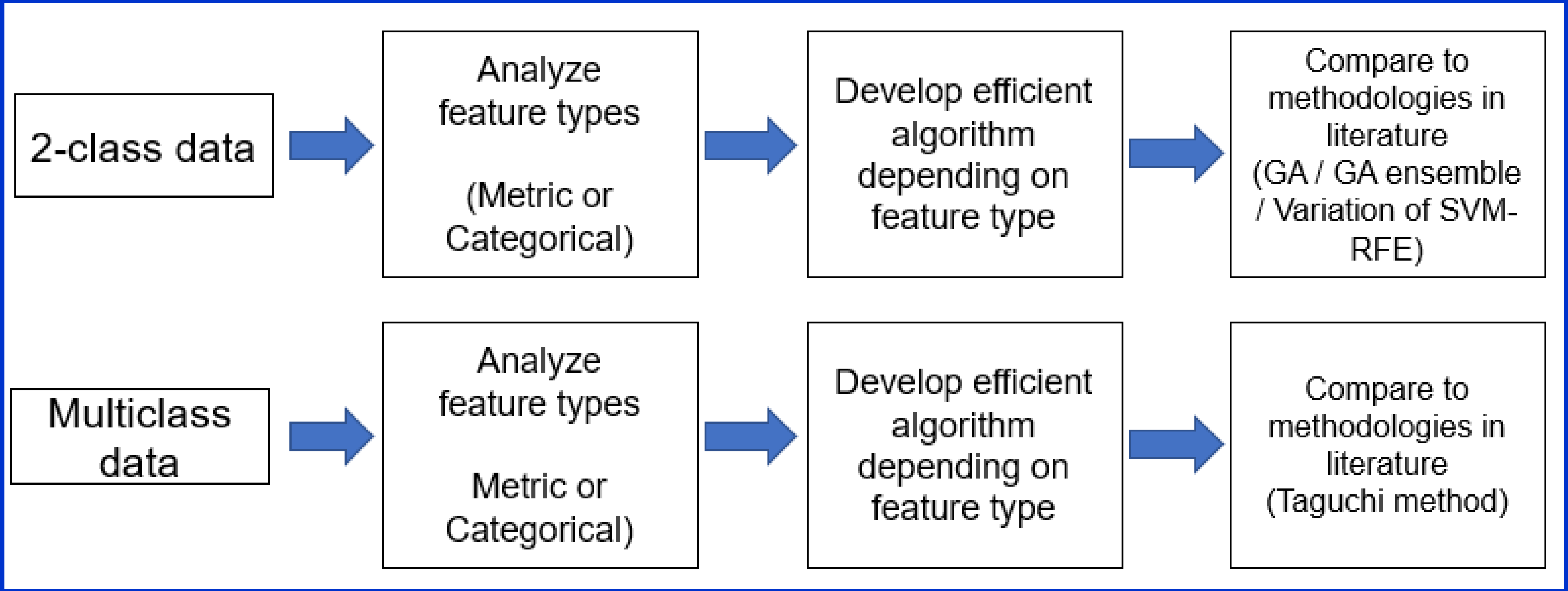
4 Kinds of Rank-PCA Ensemble Algorithms

Feature selection or extraction	Feature ranking				PCA			
Kernel in SVM	Polynomial	Sigmoid	Radial	Linear	Polynomial	Sigmoid	Radial	Linear
8 Combinations	O	O	O	O	O	O	O	O
6 Combinations	O	O		O	O	O		O
4 Combinations	O		O	O			O	
2 Combinations				O				O

- 8 Combinations - includes all kernels to maximize the CCR regardless of time.
- 6 Combinations - excludes the most time-consuming RBF, degrading CCR.
- 4 Combinations – includes RBF, and exclude other kernels not contributing to CCR.
- 2 Combinations - includes only linear SVMs, pursuing the least time complexity.

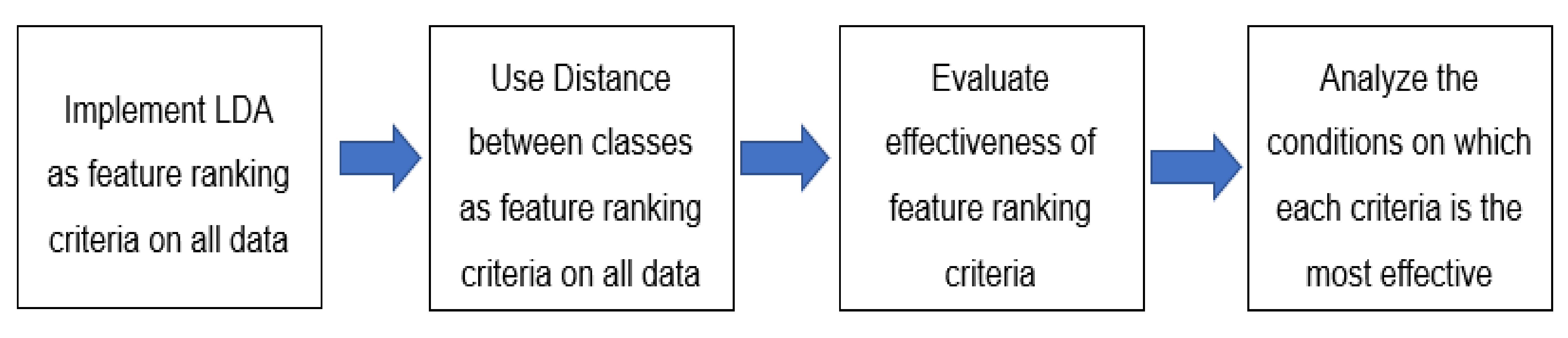
1st **Phase:** Develop Feature Selection / Extraction Methodology

Developing efficient algorithms depending on feature type



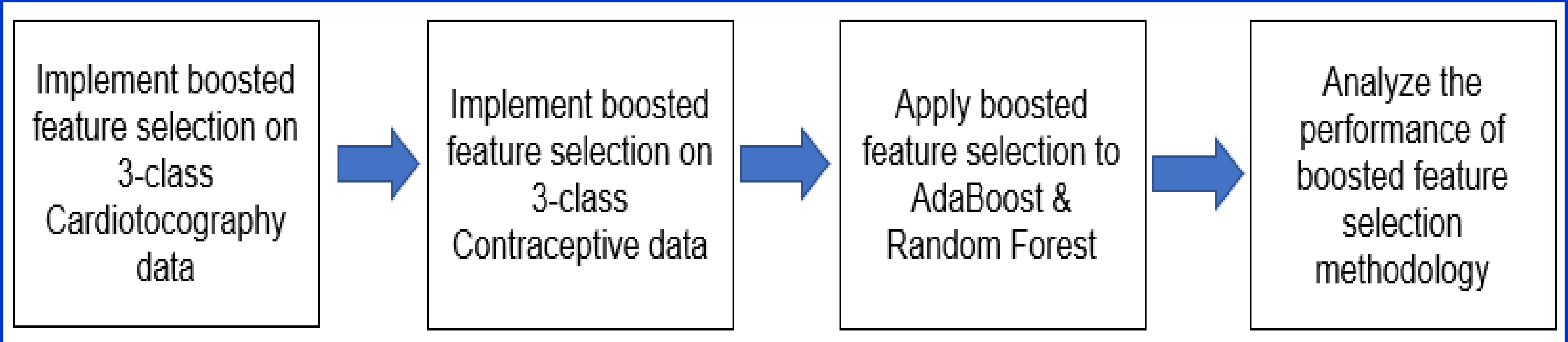
- Develop efficient ensemble depending on number of classes or feature type
- Finally, compare performance on the same data in literature.

2nd Phase: Comparing Performance of Various Feature Ranking Criteria



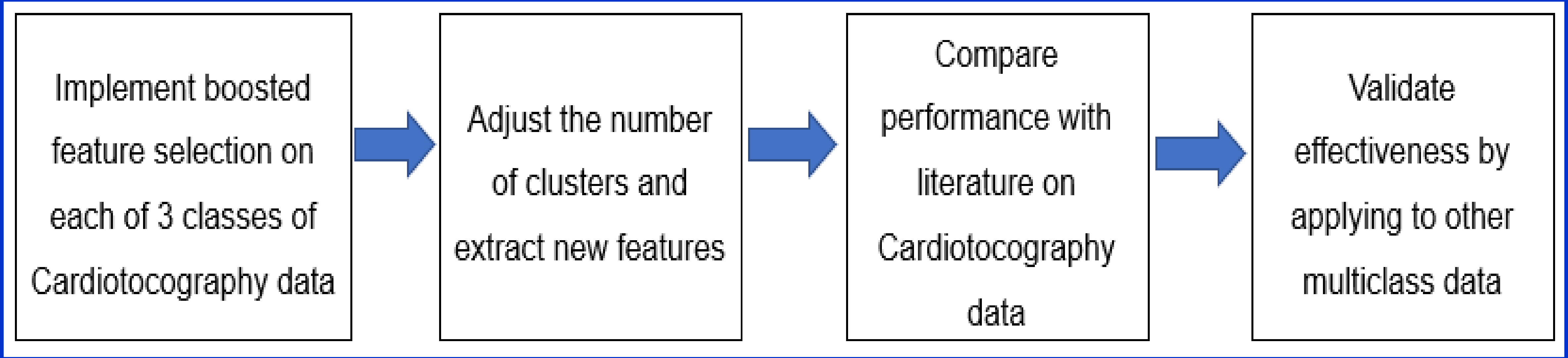
- Compare the effectiveness of various of feature ranking criteria and PCA.
- Analyze the condition on which each method is the most effective.

3rd Phase: Developing Boosted Feature Selection and Validating Performance



- Boosted Feature Selection uses the distance between classes only on misclassified instances, as feature ranking criteria.
- Apply to Cardiotocography, other data and other classifiers

4th Phase: Developing Classification Methodology for Cardiotocography Data



- Implement boosted feature selection, K-means clustering, class-dedicated SVM
- Compare performance with literature on the same conditions.
- Validate effectiveness by applying to different multiclass data

Data Preparation – 2-class and Multiclass data

No.	Data	Number of Classes	Number of instances	Number of features
1	Parkinson disease	2	195	22
2	Sonar	2	208	60
3	Heart disease	2	270	14
4	Ionosphere	2	351	33
5	Breast cancer (diagnostic)	2	569	30
6	Breast cancer	2	683	9
7	Australian credit card	2	690	14
8	Indian diabetes	2	768	8
9	German credit card	2	1,000	20
10	NBA rookie	2	1,340	19

No.	Data	Number of Classes	Number of instances	Number of features
1	Zoo	7	101	16
2	Iris	3	150	4
3	Soybean	15	266	35
4	Dermatology	6	358	34
5	Vehicle	4	846	18
6	Flare	6	1,389	12
7	Contraceptive	3	1,473	9

- 17 data (10 2-class & 7 multi) of different instances and features are prepared.
- Omissions in data sets were deleted and all features were normalized.

Data preparation – Cardiotocography data

- 2,126 instances, 21 features and 3 classes (Fetal state)
- 17 metric and 4 categorical features
- Outliers are not found in the normalized feature values.

The distribution of 3 classes

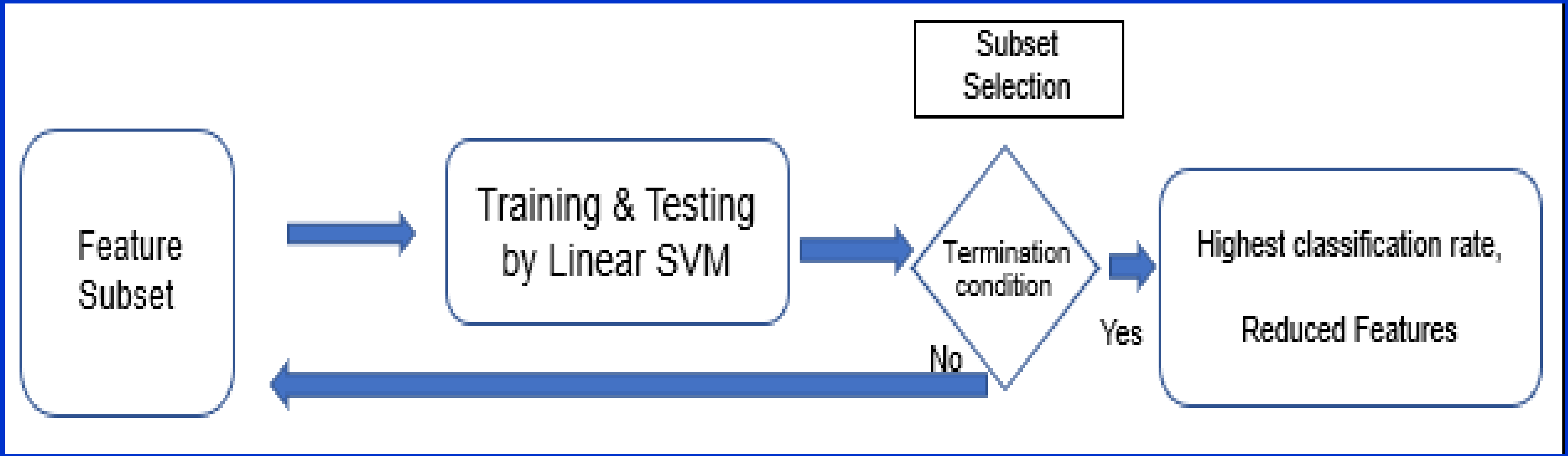
No.	Code	Fetal State	Number of Instances	Ratio
1	N	Normal	1,655	77.8%
2	S	Suspect	295	13.9%
3	P	Pathologic	176	8.3%
Sum			2,126	100.0%

The description and type of features

No.	Name of Feature	Detail Description	Feature Type (M: Metric C: Categorical)
1	LB	FHR base line (beats per minute)	M
2	AC	Number of accelerations per second	M
3	FM	Number of fetal movements per second	M
4	UC	Number of uterine contractions per second	M
5	DL	Number of light decelerations per second	M
6	DS	Number of severe decelerations per second	C
7	DP	Number of prolonged decelerations per second	C
8	ASTV	Percentage of time with abnormal short term variability	M
9	MSTV	Mean value of short term variability	M
10	ALTV	Percentage of time with abnormal long term variability	M
11	MLTV	Mean value of long term variability	M
12	Width	Width of FHR histogram	M
13	Min	Minimum of FHR historgram	M
14	Max	Maximum of FHR histogram	M
15	Nmax	Number of histogram peaks	M
16	Nzeros	Number of histogoram zeros	C
17	Mode	Histogram mode	M
18	Mean	Histogram mean	M
19	Median	Histogram median	M
20	Variance	Histogram variance	M
21	Tendency	Histogram tendency	C

Feature Ranking Method

Order of input	Combination of features
1	Top1
2	Top1, Top2
3	Top1, Top2, Top3
4	Top1, Top2, Top3, Top4
5	Top1, Top2, Top3, Top4, Top5
6	Top1, Top2, Top3, Top4, Top5, Top6
7	Top1, Top2, Top3, Top4, Top5, Top6, Top7
8	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8
9	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8, Top9

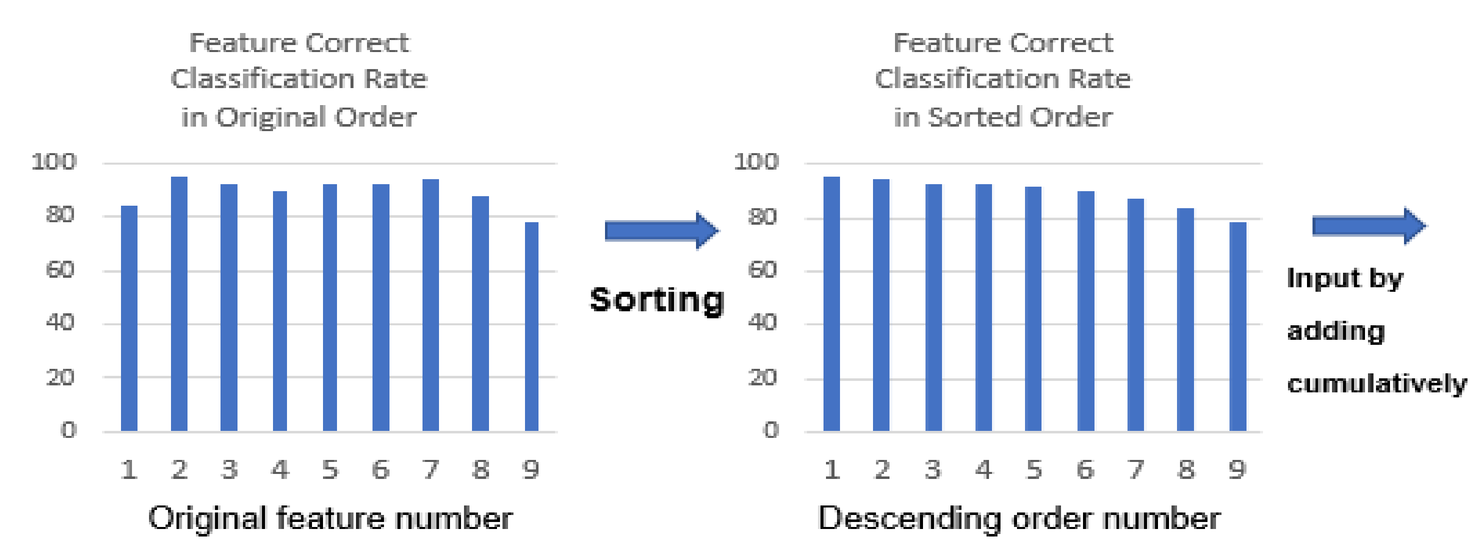


Criteria: - Same classifier
- LDA
- Distance between classes

- The possible combinations of feature increases exponentially as number of feature increases.
- Ranking method is effective in feature selection if appropriate criteria is applied.

Feature Ranking Method (referred as ‘Rank’) by CCR from SVM

1. Calculate CCR of each feature by SVM and sort features accordingly.
2. Searching for parameters by cumulatively adding top-ranked features.
3. Test by using the optimized parameters and obtain CCR.



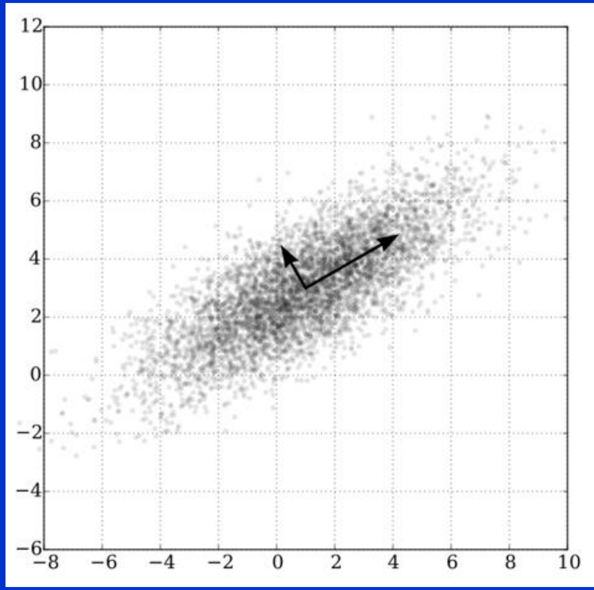
Order of input	Combination of features
1	Top1
2	Top1, Top2
3	Top1, Top2, Top3
4	Top1, Top2, Top3, Top4
5	Top1, Top2, Top3, Top4, Top5
6	Top1, Top2, Top3, Top4, Top5, Top6
7	Top1, Top2, Top3, Top4, Top5, Top6, Top7
8	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8
9	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8, Top9

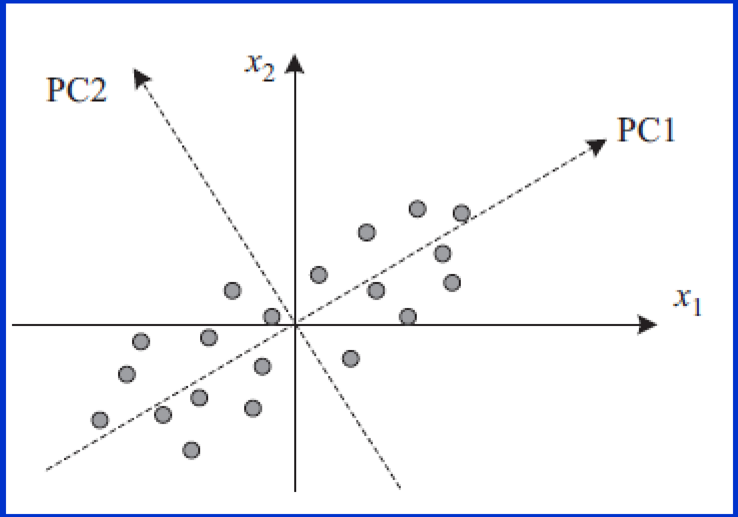
- Advantage: The characteristics of original feature can be used
- Disadvantage: Multicollinearity among features decreases CCR

Principal Component Analysis (PCA)

- Convert correlated features into linearly uncorrelated features, i.e., principal components (PC).

- Extracts features into smaller dimensions.

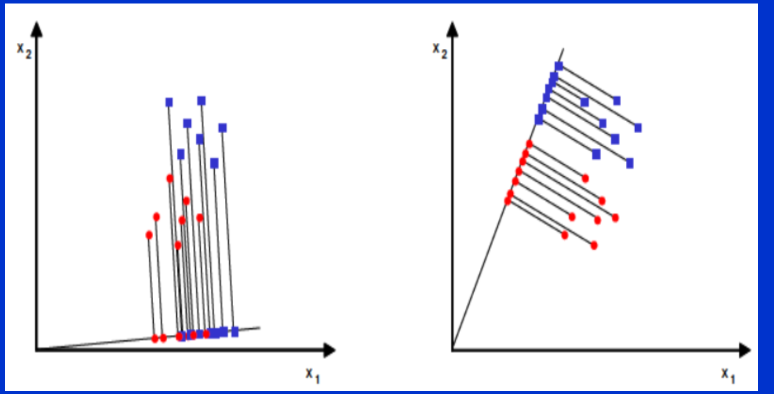


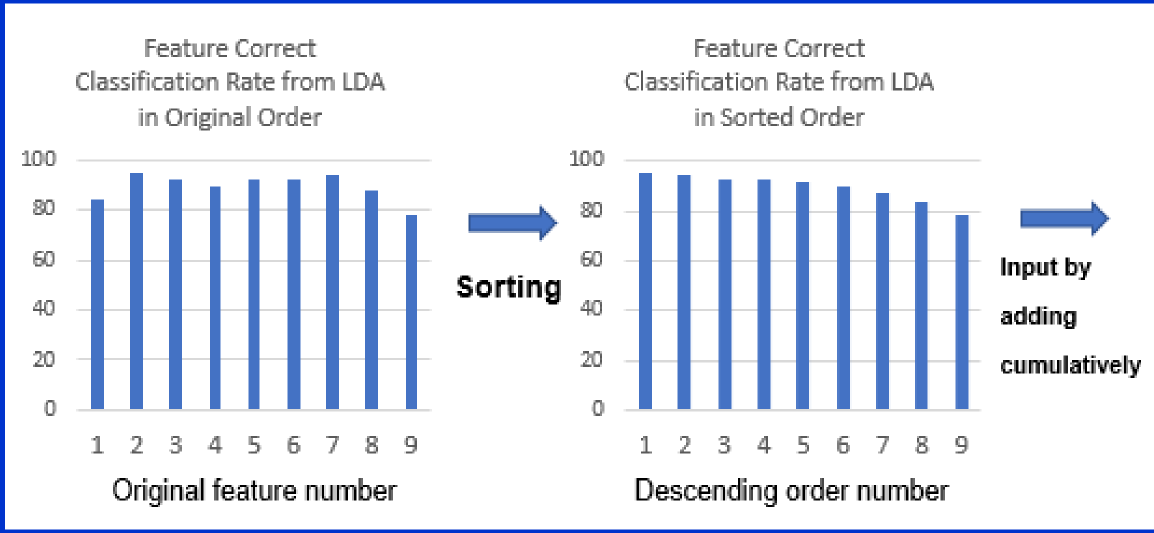


Order of input	Combination of principal components
1	PC1
2	PC1, PC2
3	PC1, PC2, PC3
4	PC1, PC2, PC3, PC4
5	PC1, PC2, PC3, PC4, PC5
6	PC1, PC2, PC3, PC4, PC5, PC6
7	PC1, PC2, PC3, PC4, PC5, PC6, PC7
8	PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8
9	PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9

- Advantage: Effective in reducing high dimensions of data with multicollinearity

Linear Discriminant Analysis (LDA)





Order of input	Combination of features
1	Top1
2	Top1, Top2
3	Top1, Top2, Top3
4	Top1, Top2, Top3, Top4
5	Top1, Top2, Top3, Top4, Top5
6	Top1, Top2, Top3, Top4, Top5, Top6
7	Top1, Top2, Top3, Top4, Top5, Top6, Top7
8	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8
9	Top1, Top2, Top3, Top4, Top5, Top6, Top7, Top8, Top9

- Reduces the dimension by maximizing the ratio of between-class scatter to within-class scatter after supervised learning.

- Advantage : Computation time is short.
- Disadvantage: works well only on data with linear characteristics.

Distance between Classes

Feature selection by using discriminatory power from distance between two classes

$$d_{ij} = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(\mathbf{x}_i^{\,k}, \mathbf{x}_j^{\,m})$$

X_i^k : k^{th} sample in class w_i

$dist(\mathbf{x}_i^{\,k}, \mathbf{x}_j^{\,m})$: Distance between the 2 samples

N_i : The total number of samples in class w_i

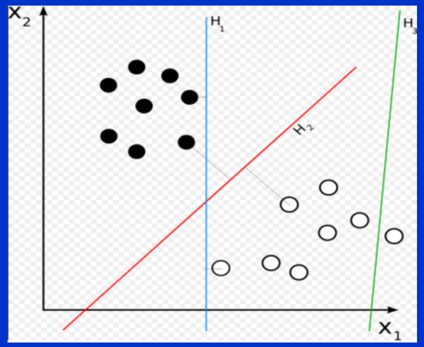
Example of features with long vs. short distance

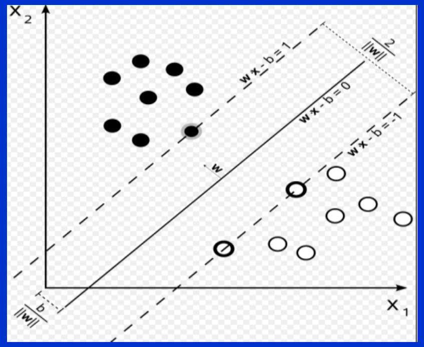
- The distance represents degree of separation of two classes.
- Long distance means high discriminatory power.

- Advantage : CCR can increase further
- Disadvantage: Calculation is computationally expensive if data is large.

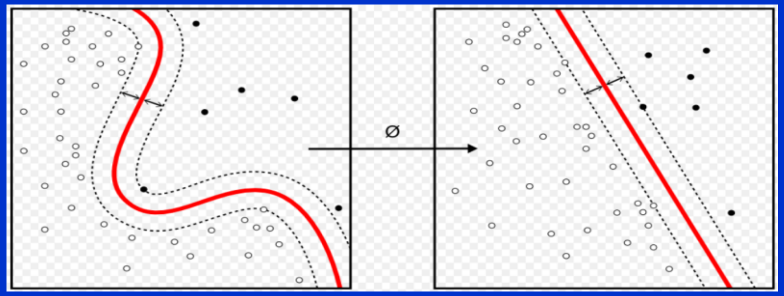
Support Vector Machine (SVM)

- Searches for a hyperplane maximizing margin between the points in the two classes, which is closest to the hyperplane.





- Perform non-linear classification by kernel trick, which maps inputs into high-dimensional feature spaces.



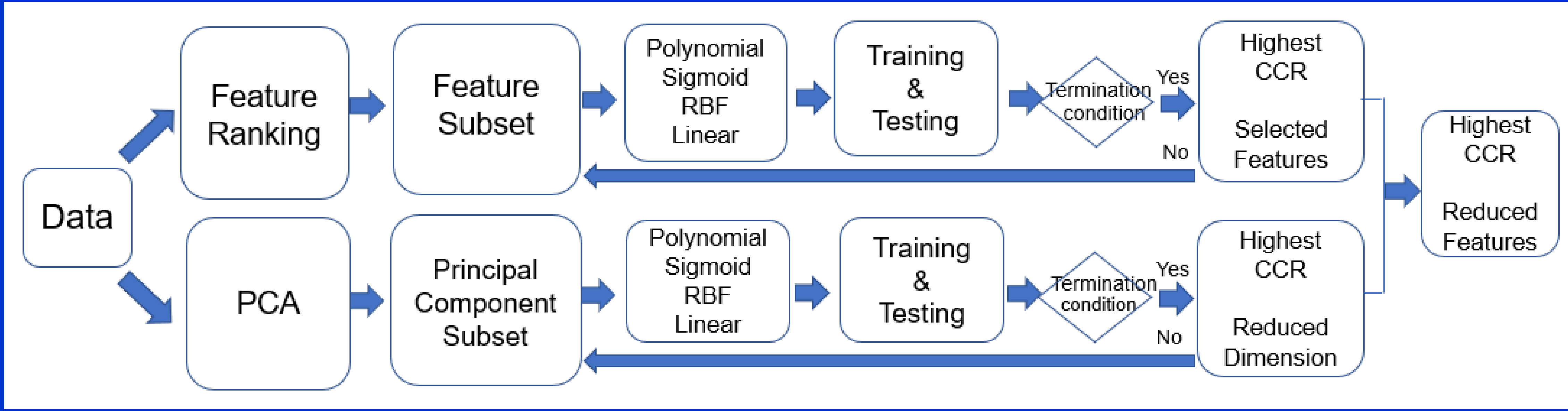
Parameter Optimization of SVM kernels

- Polynomial, sigmoid and RBF require parameter optimization before training.

1. Linear $k(x_{1,} x_2) = x_1^T x_2$ ⟶ No parameter
2. Polynomial $k(x_{1,} x_2) = (\gamma\, x_1^T x_2 + \text{Coef})^d$ ⟶ Search for γ
3. Radial basis $k(x_{1,} x_2) = \exp(-\gamma \| x_1 - x_2 \|^2)$ ⟶ Search for γ and C
4. Sigmoid $k(x_{1,} x_2) = \tanh(\gamma\, x_1^T x_2 + \text{Coef})$ ⟶ Search for γ

- Search for the parameters which produces highest CCR.
- Searching for γ and C in RBF is the most time-consuming due to grid search

Architecture of Rank-PCA based SVM Ensemble Algorithm



- Applied techniques are
  - Various feature ranking criteria (CCR, PCA, LDA, Distance between classes)
  - Various kernels in SVM (Linear SVM, Polynomial, Sigmoid, RBF)
  - Wrapper method
  - Ensemble method

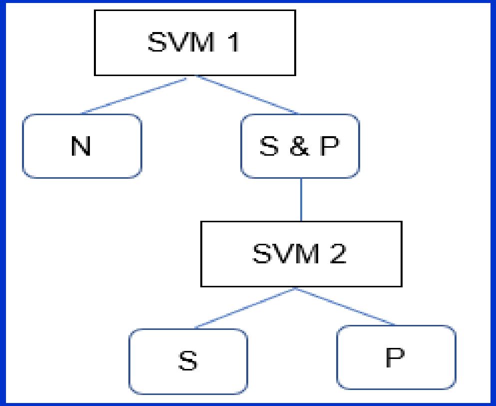
Efficient Algorithm Depending on Feature Type

Criteria of classifying feature type

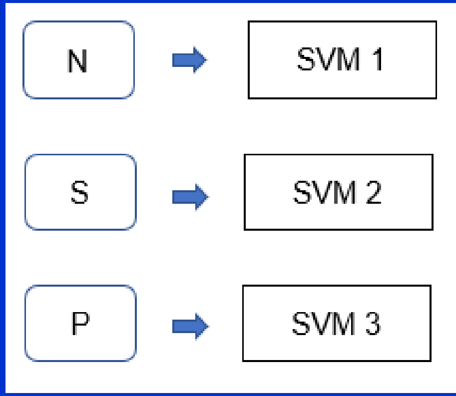
Feature type	Description of type
Metric	(1) Continuous values (e.g. 0.3826)
	(2) Integers representing degree (e.g. 1, 2, 3, 4, ..., 10)
Categorical	Symbols representing category (e.g. 1,2,3, & A, B, C)

- Features in all data is marked as Metric or Categorical.
- The ratio of metric features is calculated per each data.
- To eliminate redundancy, compose efficient ensemble by selecting only the kernels which contributes to highest CCR.

Comparison of Classification Architecture



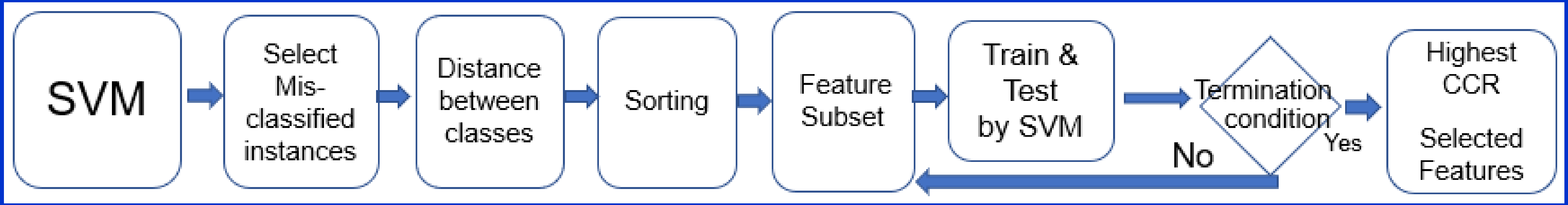
Binary Decision Tree (BDT)



Class-Dedicated SVM

- BDT has been used in literature to extend the binary SVM to multiclass.
- In BDT, misclassification from SVM1 negatively affects SVM2.
- Dedicated SVM improves performance, focusing on increasing CCR of each class.
- In this research, the performances of the two architecture are compared.

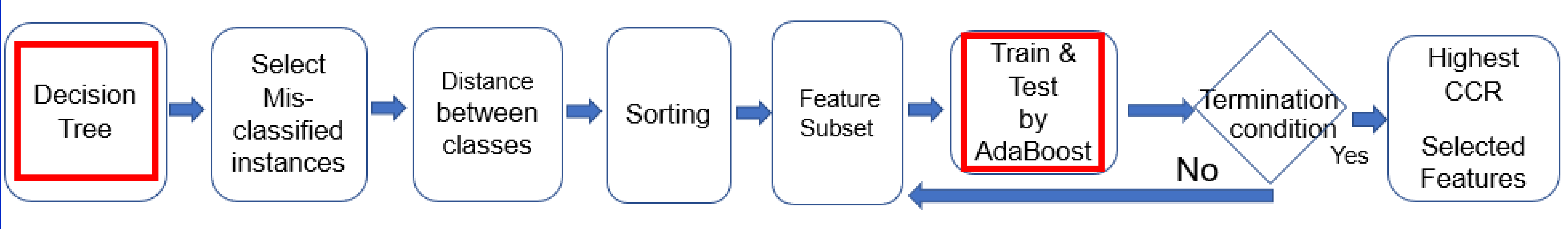
Boosted Feature Selection

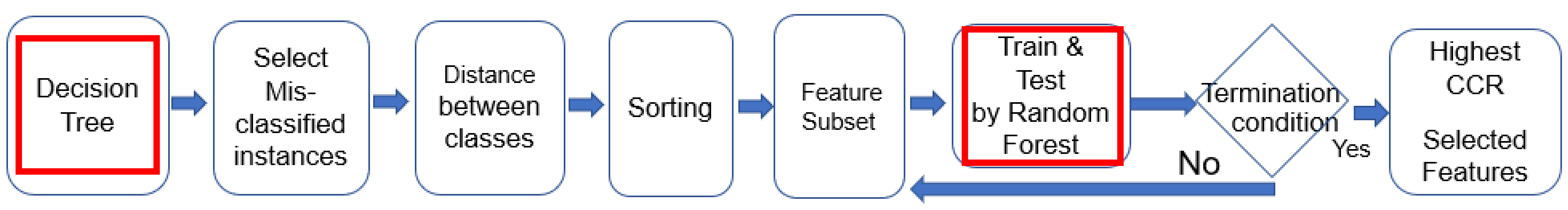


Flow of algorithm – Feature ranking by SVM + Distance between classes + SVM

- Select misclassified instances from SVM, applying boosting concept.
- Calculate the distance between classes among the misclassified instances.
- Sort features according to the distance and use wrapper method to select feature subset for the highest CCR.

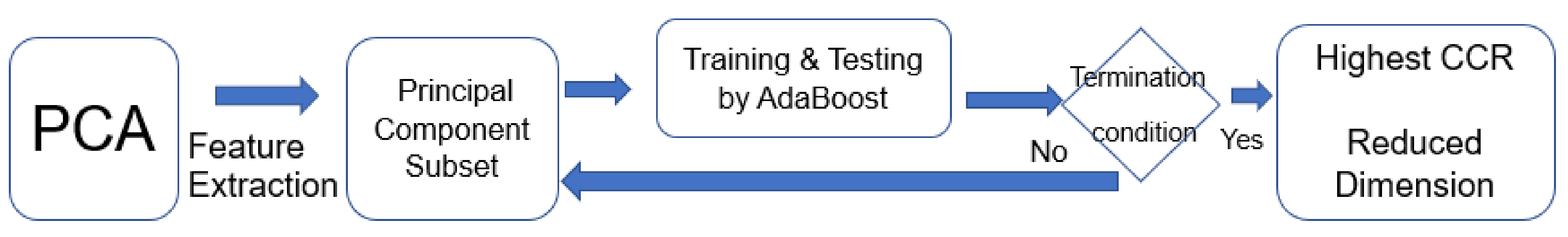
Validation of Boosted Feature Selection by Applying to Other Classifiers





- Applied to AdaBoost and Random Forest to verify effectiveness regardless of classifiers.
- The feature ranking is based on misclassification from decision tree because they use decision tree as basic classifier.

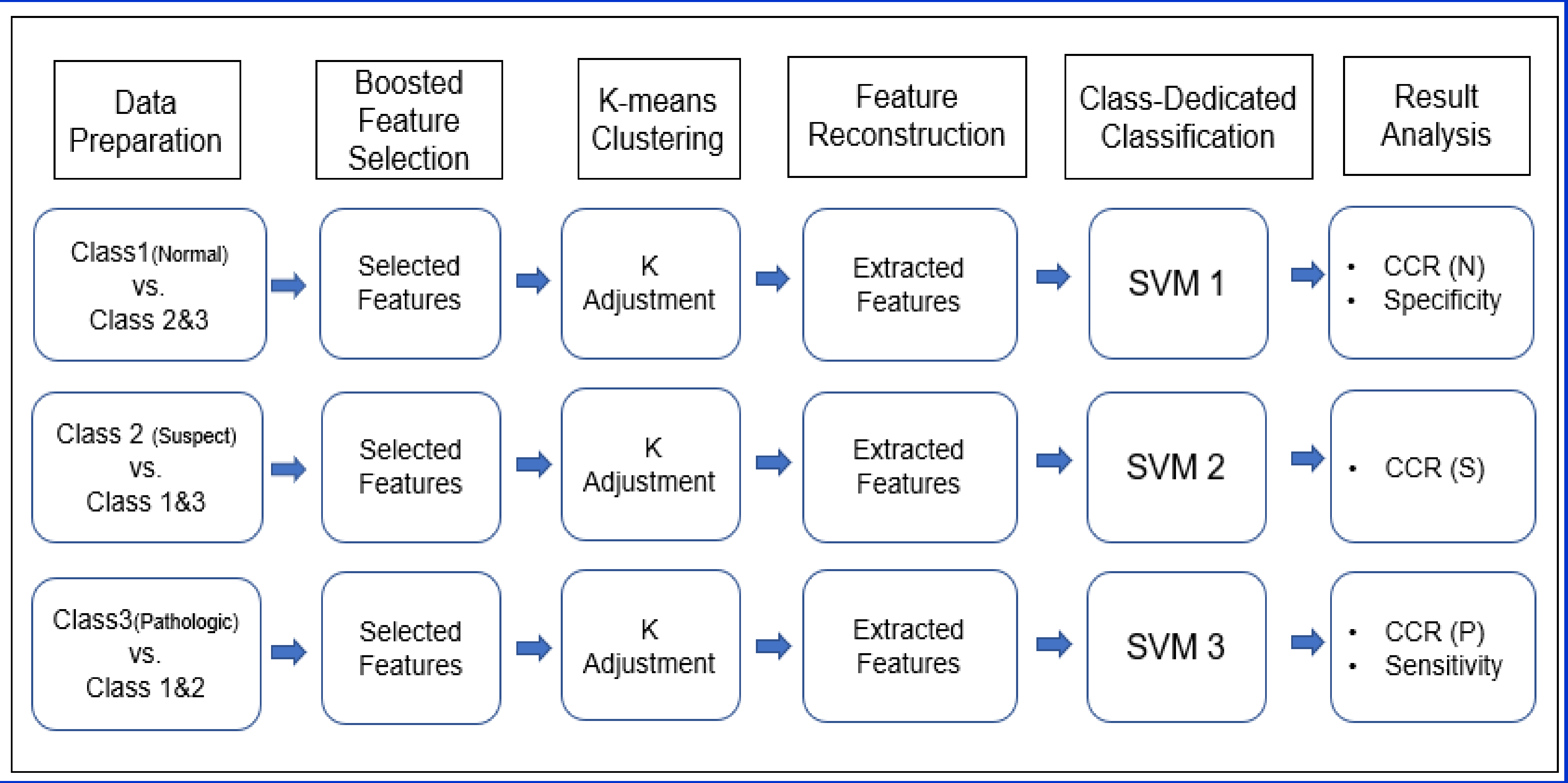
Validation of Boosted Feature Selection by Applying to Other Classifiers



PCA + AdaBoost + Wrapper method for dimension decision

- In case of AdaBoost which is sensitive to noise, PCA is also implemented as a preprocessing because PCA reduces noise in data.

Flow of Algorithm - Improved Classification Methodology



Flow of algorithm

Process of Improved Classification Methodology

- Boosted feature selection selects the features with higher discriminatory power.

$$d_{ij} = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(\mathbf{x}_i^{\,k}, \mathbf{x}_j^{\,m})$$

X_i^k : k^{th} sample in class w_i

$dist(\mathbf{x}_i^{\,k}, \mathbf{x}_j^{\,m})$: Distance between the 2 samples

N_i : The total number of samples in class w_i

- The optimum number of clusters are determined by k-means clustering algorithm.

$$\min_S \sum_{i=1}^{k} \sum_{x \in s_i} || x - \mu_i ||^2$$

- Calculate the values of mean by K-means clustering algorithms.

Process of Improved Classification Methodology

- New Features are extracted by Fuzzy Membership Function

- Fuzzy Membership Function

$$f_{np}(\mathrm{X}_{\mathrm{j}}^{\mathrm{i}}) = 1 - \frac{|X_j^{\mu_{np}} - X_j^i|}{max\,|X_j^{\mu_{np}} - X_j^n|} \quad \text{if } \min(X_j^n) \le X_j^i \le \max(X_j^n), \forall \mathrm{n} \in S_{np}$$

$$f_{np}(\mathrm{X}_{\mathrm{j}}^{\mathrm{i}}) = 0 \quad \text{if otherwise;}$$

- Calculate the output value of fuzzy membership function by using the information on new patterns.

Process of Improved Classification Methodology

- New Feature Extraction

$$EF_{np} = \frac{1}{NSF} \sum_{j=1}^{NSF} f_{np}(X_j^i), \quad 1 \leq np \leq K^N + K^S + K^P \quad \text{(clustering within each class)}$$

$$1 \leq np \leq K^N + K^{S+P} \quad \text{(clustering within N and within S+P)}$$

$$1 \leq np \leq K^S + K^{N+P} \quad \text{(clustering within S and within N+P)}$$

$$1 \leq np \leq K^P + K^{N+S} \quad \text{(clustering within P and within N+S)}$$

- Extract new features by summing the output of all patterns.
- Search for improved model by adjusting the number of clusters.

Process of Improved Classification Methodology

- Confusion Matrix

	Class 1 (Normal) Predicted			Class 2 (Suspect) Predicted			Class 3 (Pathologic) Predicted	
Original	TP_1	FP_1	Original	TP_2	FP_2	Original	TP_3	FP_3
	FN_1	TN_1		FN_2	TN_2		FN_3	TN_3

$$\text{Specificity} = \text{CCR of Class 1} = \frac{TP_1}{TP_1+FP_1}$$

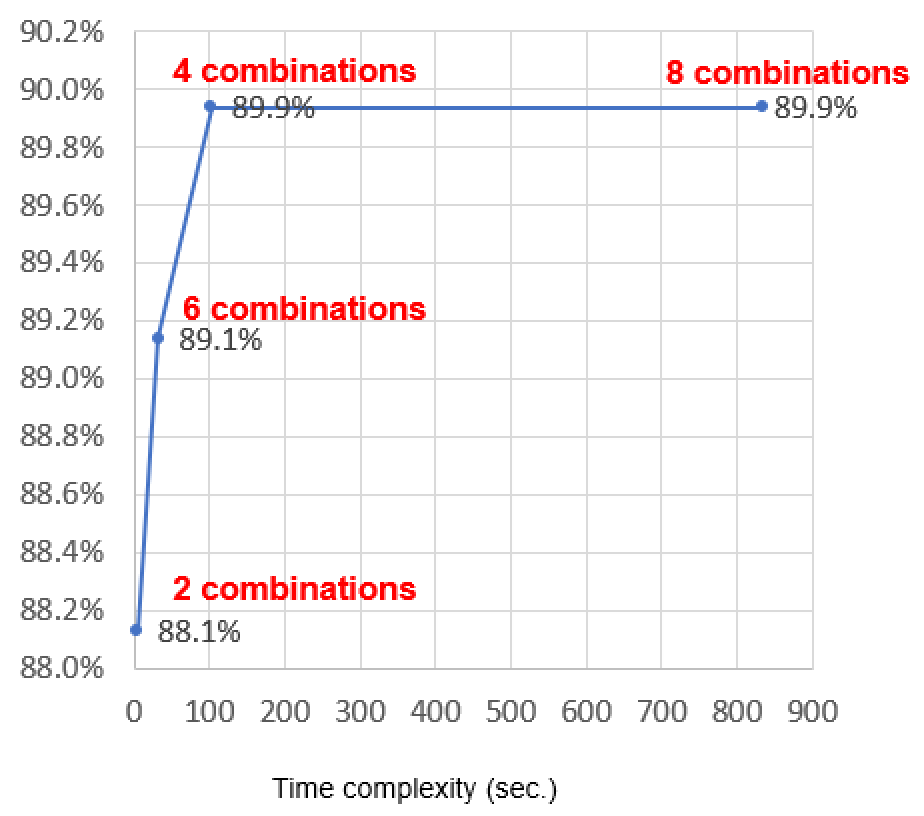
$$\text{CCR of Class 2} = \frac{TP_2}{TP_2+FP_2}$$

$$\text{Sensitivity} = \text{CCR of Class 3} = \frac{TP_3}{TP_3+FP_3}$$

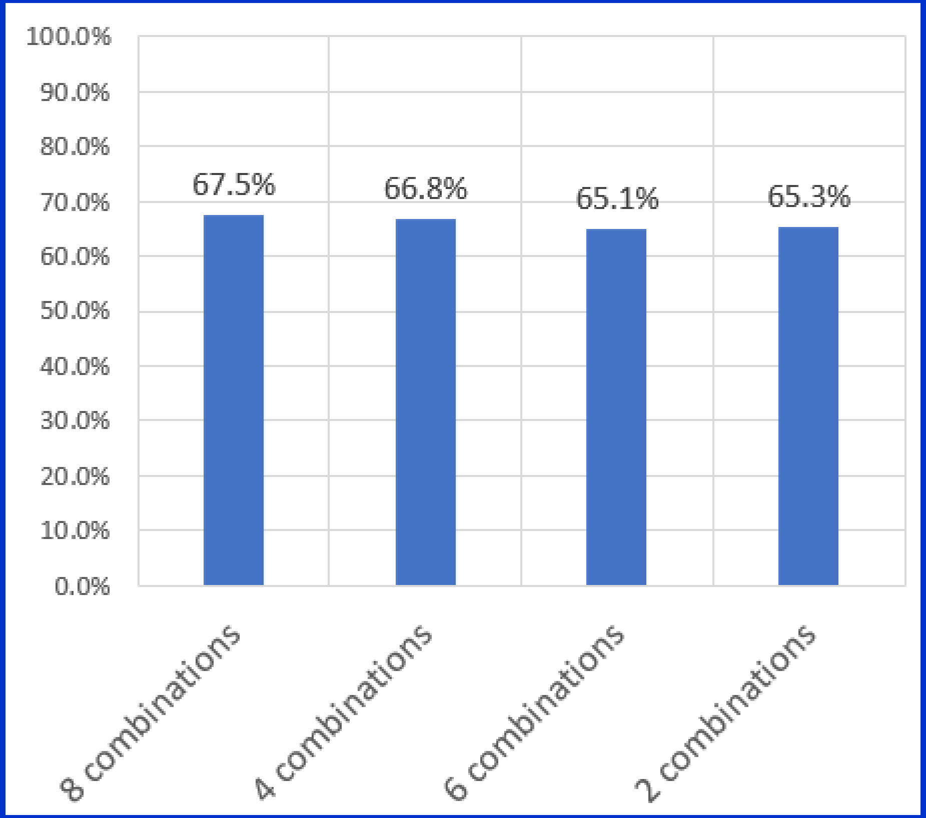
- Each of the binary classifications is class-dedicated SVM.
- In each, Normal, Suspect, Pathologic is regarded as positive, respectively.

4. Experimental Result on 2-Class and Multiclass Data

Performance of 4 Kinds of Rank-PCA Ensemble Algorithms



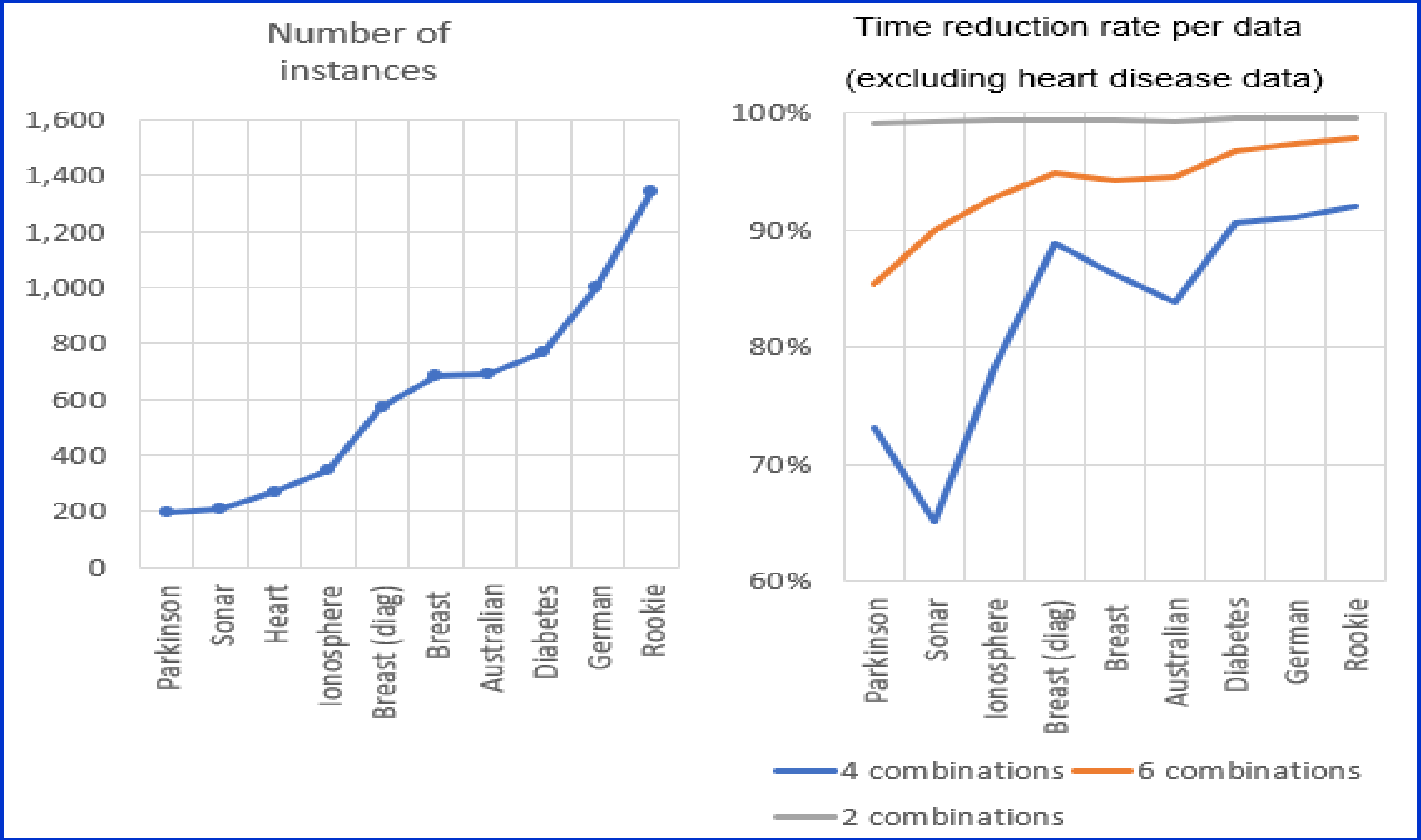
CCR vs. Time



Feature reduction rate

- Various combinations show different performance of CCR and time complexity.
- Reduced data is applied to GS of RBF (Rank - $\frac{1}{12}$, PCA - $\frac{1}{3}$) in 4 combinations.
- Feature reduction rates are almost at the same level.

Summary of Performance of 3 Kinds of Algorithms



- Time reduction rates increase significantly as instances increase.

Performance of Efficient Algorithm Depending on Feature Type – 2 class data

No.	Data	Number of instances	Number of features	Number of Metric features	Ratio of Metric feature	Highest CCR		Efficient Algorithm
						Feature ranking or PCA	Kernel	
1	Parkinson	195	22	22	100.0%	PCA	Radial	Feature ranking (L) + PCA (S/R/L)
2	Sonar	208	60	60	100.0%	PCA	Radial	
3	Breast (diag)	569	30	30	100.0%	PCA	Linear	
4	Breast	683	9	9	100.0%	PCA	Sigmoid/Radial/Linear	
5	Rookie	1,340	19	19	100.0%	Feature ranking	Linear	
6	Diabetes	768	8	8	100.0%	Feature ranking	Linear	
7	Ionosphere	351	33	32	97.0%	PCA	Radial	
8	Australian	690	14	8	57.1%	Feature ranking	Polynomial	Feature ranking (P/S/R/L)
9	Heart	270	14	5	35.7%	Feature ranking	Polynomial/Sigmoid Radial/Linear	
10	German	1,000	20	4	20.0%	Feature ranking	Radial	

- Only the kernels contributing to highest CCR, are selected.
- Feature ranking is more effective on data with lower metric ratio.
- Further time reduction is 39%, compared to 4 combination algorithm.

Performance of Efficient Algorithm depending on Feature Type – Multiclass data

No.	Data	Number of classes	Number of instances	Number of features	Number of Metric features	Ratio of Metric feature	Highest CCR		Efficient Algorithm
							Feature ranking or PCA	Kernel	
1	Iris	3	150	4	4	100.0%	Feature ranking	Sigmoid	Feature ranking (S/R)
2	Vehicle	4	846	18	18	100.0%	Feature ranking PCA	Radial	
3	Soybean	15	266	35	35	100.0%	Feature ranking	Radial	
4	Contraceptive	3	1473	9	2	22.2%	Feature ranking	Radial	
5	Dermatology	6	358	34	1	2.9%	Feature ranking PCA	Sigmoid/Radial/Linear	
6	Zoo	7	101	16	0	0.0%	Feature ranking PCA	Sigmoid/Radial/Linear	
7	Flare	6	1389	12	0	0.0%	Feature ranking	Radial	

- PCA is not effective in producing higher CCR on multiclass data.
- Further time reduction is 70%, compared to 4 combination algorithm.

Performance of 4 combinations of Rank-PCA Ensemble on 2-class data

NF = Number of used Features

No.	Data	Proposed					GA		Proposed		GA-ensemble		Proposed		SVM RFE+AT	
		Sensitivity	Specificity	CCR		NF	CCR	NF	CCR	NF	CCR	NF	CCR	NF	CCR	NF
				Training	Testing											
1	Parkinson	1.000	0.750	91.7	94.9	6										
2	Breast (diag)	1.000	1.000	100.0	100.0	4										
3	Rookie	0.838	0.564	83.0	73.9	5										
4	German	0.397	0.939	90.2	78.0	13	85.6	13								
5	Australian	0.855	0.916	97.6	97.1	1	88.1	3								
6	Diabetes	0.579	0.885	89.9	80.5	7	81.5	3.7								
7	Heart	1.000	1.000	100.0	100.0	1	94.8	5.4								
8	Breast	0.981	0.989	97.5	100.0	2	96.2	1								
9	Sonar	0.798	0.929	97.1	90.5	11	98	15	87.8	10.0	84.0	12				
10	Ionosphere	1.000	0.938	98.6	97.1	10	98.6	6	98.6	11.0	93.5	10	97.1	15	86.3	5
Average		**0.845**	**0.891**	**94.6**	**91.2**	**6.0**	**91.8**	**6.7**	**93.2**	**10.5**	**88.7**	**11.0**	**97.1**	**15.0**	**86.3**	**5.0**
Cross-validation		Training 90%, Testing 10%							Training 80%, Testing 20%				Training 50%, Testing 50%			

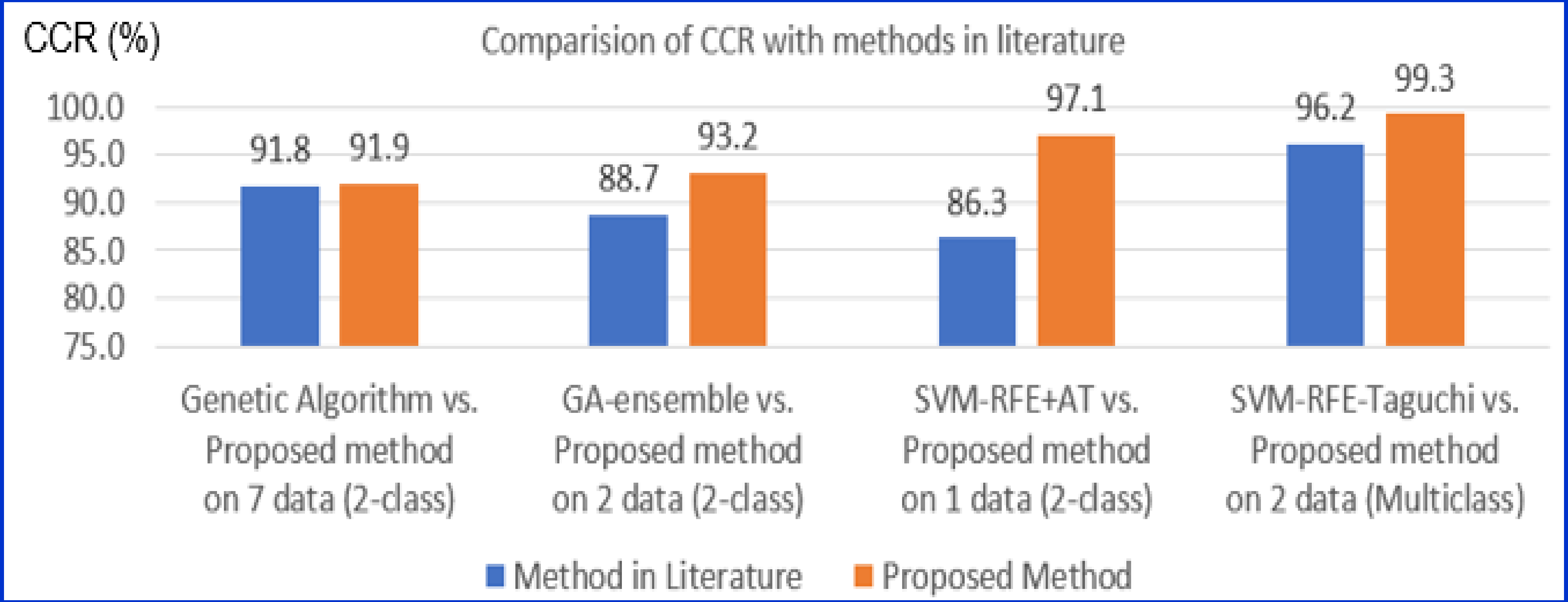
- Performances are compared to literature with the same Cross-validation.
- The comparison results are graphically represented in upcoming slides.

Performance of Efficient Ensemble Algorithm for Multiclass Data

No.	Data	Proposed			SVM-RFE-Taguchi			Difference	
		CCR	Original Number of Features	Number of Selected Features	CCR	Original Number of Features	Number of Selected Features	CCR	Feature Reduction Rate (%)
1	Dermatology	98.6	34	34	95.4	34	23	3.2	-47.8%
2	Zoo	100.0	12	8	97	12	12	3.0	33.3%
Average		**99.3**	**23.0**	**21.0**	**96.2**	**23.0**	**17.5**	**3.1**	-20.0%
Cross-validation		Training 80%, Testing 20%							

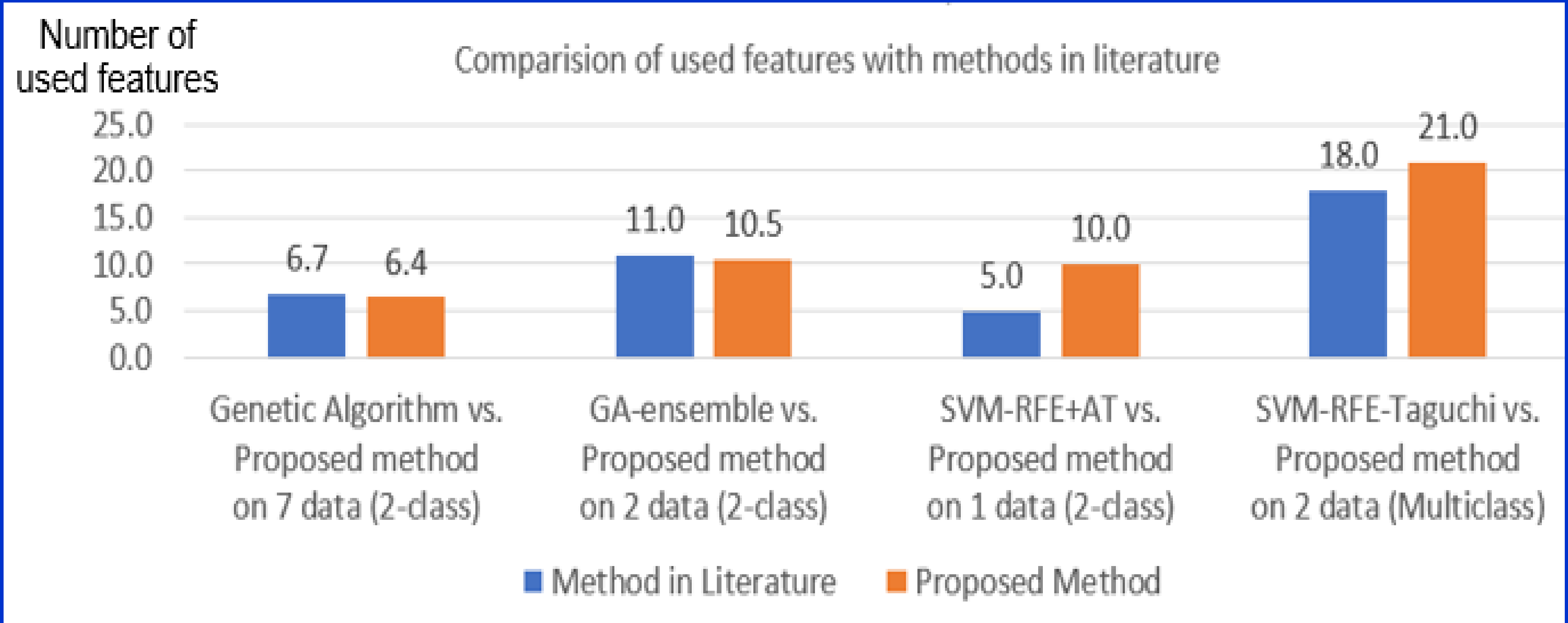
- Compared to literature, proposed efficient algorithm shows higher CCR by 3.1%
- Compared to literature, the feature reduction rate is not always lower.

Comparison with Approaches in Literature



- Compared to GA, proposed method shows equivalent CCR.
- Compared to GA-ensemble (SVM, DT, ANN) & variations of SVM, proposed method shows higher CCR.

Comparison with Approaches in Literature



- However, the number of used features are not always less, compared to literature.

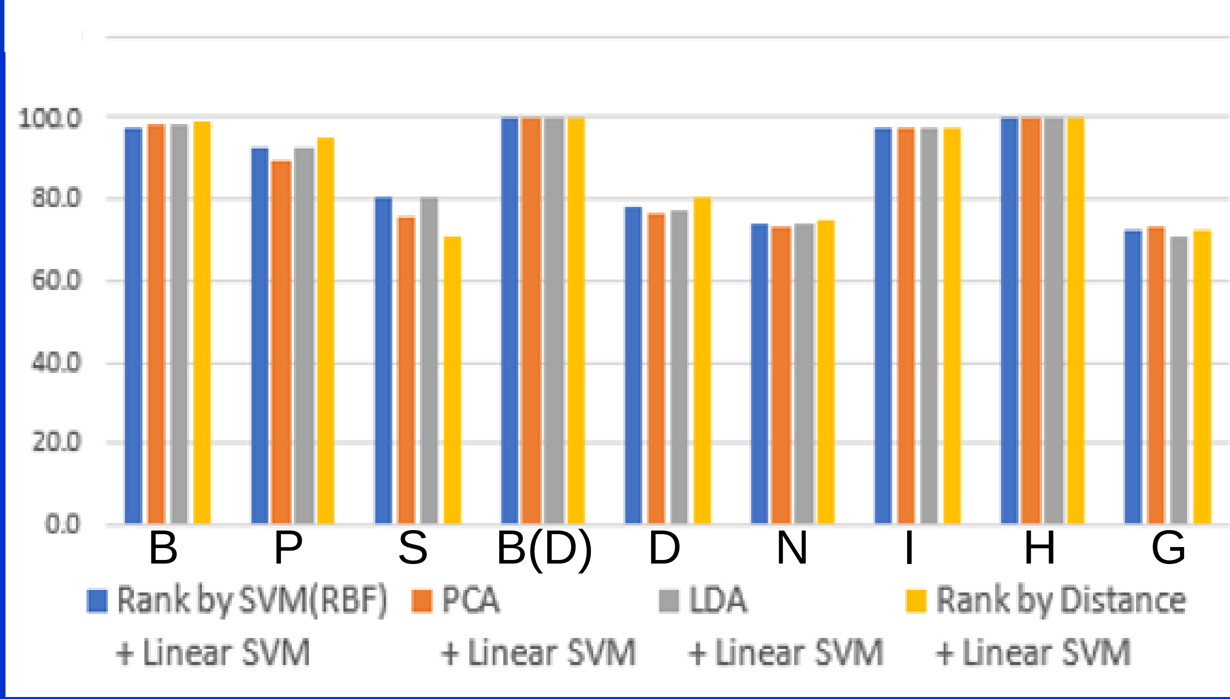
Comparison of 4 Different Feature Ranking Criteria

Class	Data	Number of Classes	Number of Instances	Number of Features	Ratio of Numerical feature	Correct Classification Rate (%)			
						Rank by SVM(RBF) + Linear SVM	PCA + Linear SVM	LDA + Linear SVM	Rank by Distance + Linear SVM
2	Breast	2	683	9	100.0%	97.8	98.5	98.5	99.3
	Parkinson	2	195	22	100.0%	92.3	89.7	92.3	94.9
	Sonar	2	208	60	100.0%	80.5	75.6	80.5	70.7
	Breast(diag)	2	569	30	100.0%	100.0	100.0	100.0	100.0
	Diabetes	2	768	8	100.0%	78.4	76.5	77.1	80.4
	NBA rookie	2	1,340	19	100.0%	73.9	73.5	73.9	74.6
	Ionosphere	2	351	33	97.0%	97.1	97.1	97.1	97.1
	Heart	2	270	14	35.7%	100.0	100.0	100.0	100.0
	German	2	1,000	20	20.0%	72.0	73.0	71.0	72.5
	Average	**2.0**	**598.2**	**23.9**		**88.0**	**87.1**	**87.8**	**87.7**
Multi	Iris	3	150	4	100.0%	96.7	96.7	93.3	89.0
	Soybean	15	266	35	100.0%	90.6	83.0	88.7	98.2
	Vehicle	4	846	18	100.0%	78.1	79.3	78.1	88.6
	Contraceptive	3	1,473	9	22.2%	50.7	50.7	53.9	100.0
	Dermatology	6	358	34	2.9%	97.2	98.6	95.8	99.1
	Flare	6	1,389	12	0.0%	75.9	75.9	75.5	89.9
	Zoo	7	101	16	0.0%	100.0	100.0	100.0	100.0
	Average	**6.3**	**654.7**	**18.3**		**84.2**	**83.5**	**83.6**	**95.0**
Total	**Average**	**3.9**	**622.9**	**21.4**		**86.3**	**85.5**	**86.0**	**90.9**

- 4 feature ranking criteria are compared on 2-class and multiclass data
- The highest CCRs are red-colored.

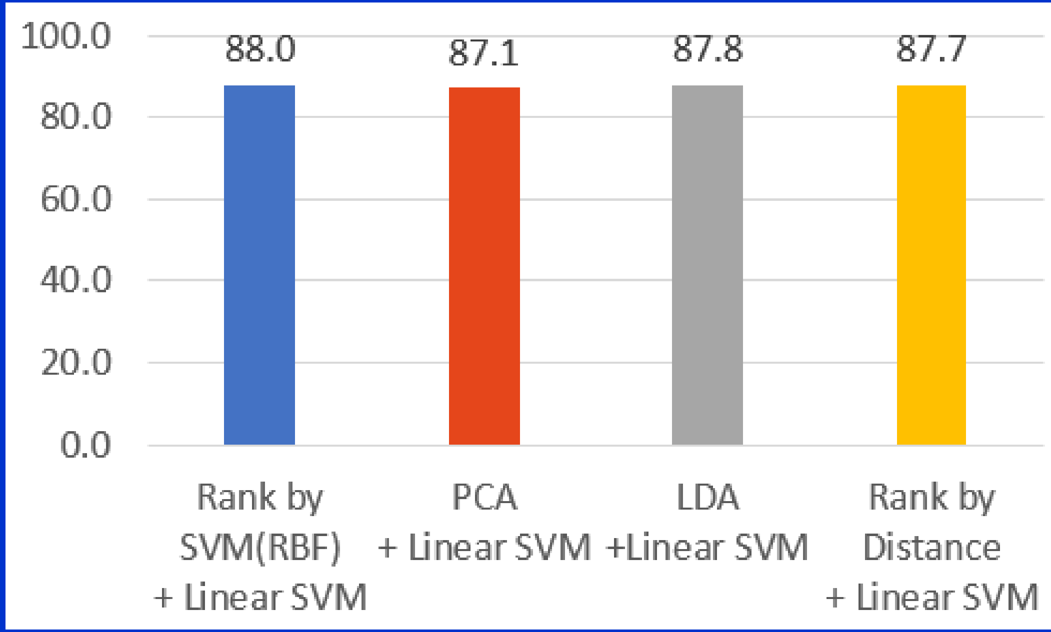
Comparison of 4 Different Feature Ranking Criteria on 2-class Data

CCR (%)



CCR of each data

CCR (%)

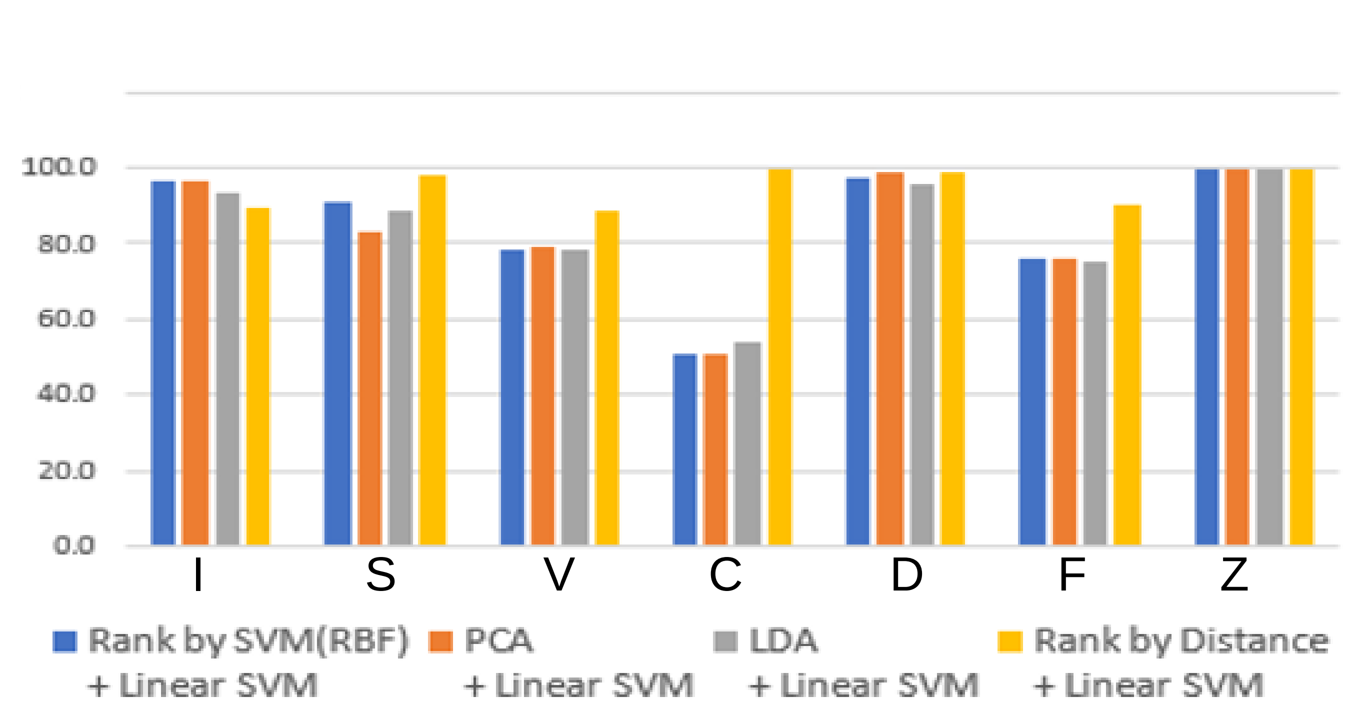


Average CCR

- CCR of the 4 methods are almost same (87.55 ± 0.45%)

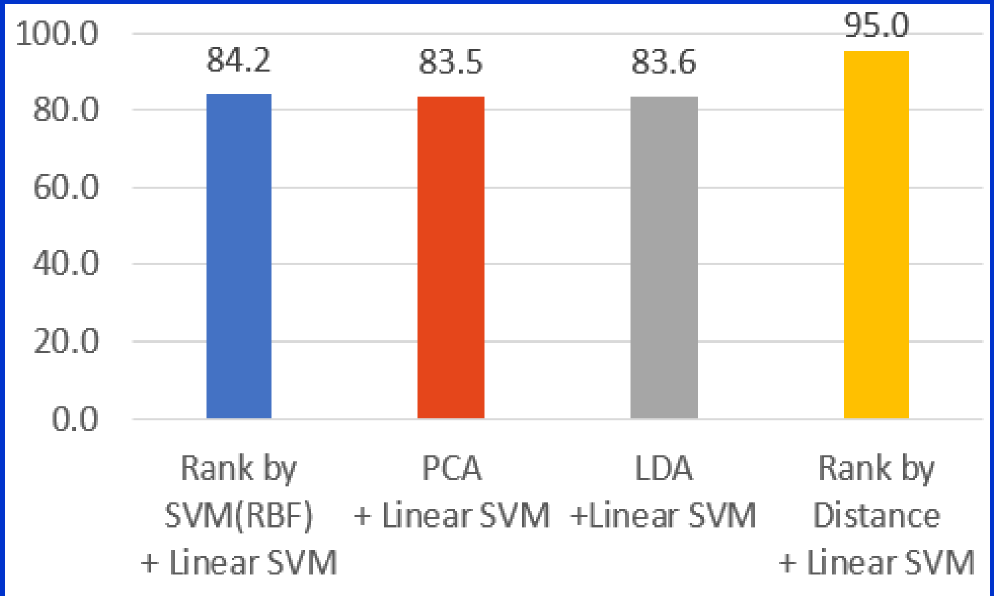
Comparison of 4 Different Feature Ranking Criteria on Multiclass Data

CCR (%)



CCR of each data

CCR (%)



Average CCR

- Distance between classes is the most effective. (95.0%)
- Effective on multiclass & large data with large instances or high dimension.
- The effectiveness is also strengthened by one. vs. all multiclass classification.

5. Experimental Result on Cardiotocography Data

Boosted Feature Selection of Cardiotocography Data

Comparison of performance – SVM without FS vs. Boosted FS

Class	Number of Instances used for Calculating Distance between Classes	SVM (RBF) with No Feature Selection (A)		Rank by Distance using missclassification by SVM + SVM (RBF) (B)		Improvement (B-A)	
		Correct Classification Rate (%)	Number of Selected Feature	Correct Classification Rate (%)	Number of Selected Feature	Correct Classification Rate (%)	Feature Reduction Rate (%)
1 vs. 2&3	151	92.6	21	94.0	11	1.4	47.6
2 vs. 1&3	179	92.2	21	93.5	10	1.3	52.4
3 vs. 1&2	61	97.4	21	97.7	11	0.3	47.6
Average	**130.3**	**94.1**	**21.0**	**95.1**	**10.7**	**1.0**	**49.2**

- CCR increases by 1.0%, features reduced by 49.2%, sensitivity & specificity increases by 2.8% (0.739) and 0.2% (0.997), compared to none feature selection.
- CCR increases by 3.5 % compared to literature (BFS + One vs. all architecture).

Validation by Applying to Other Data (Contraceptive Data)

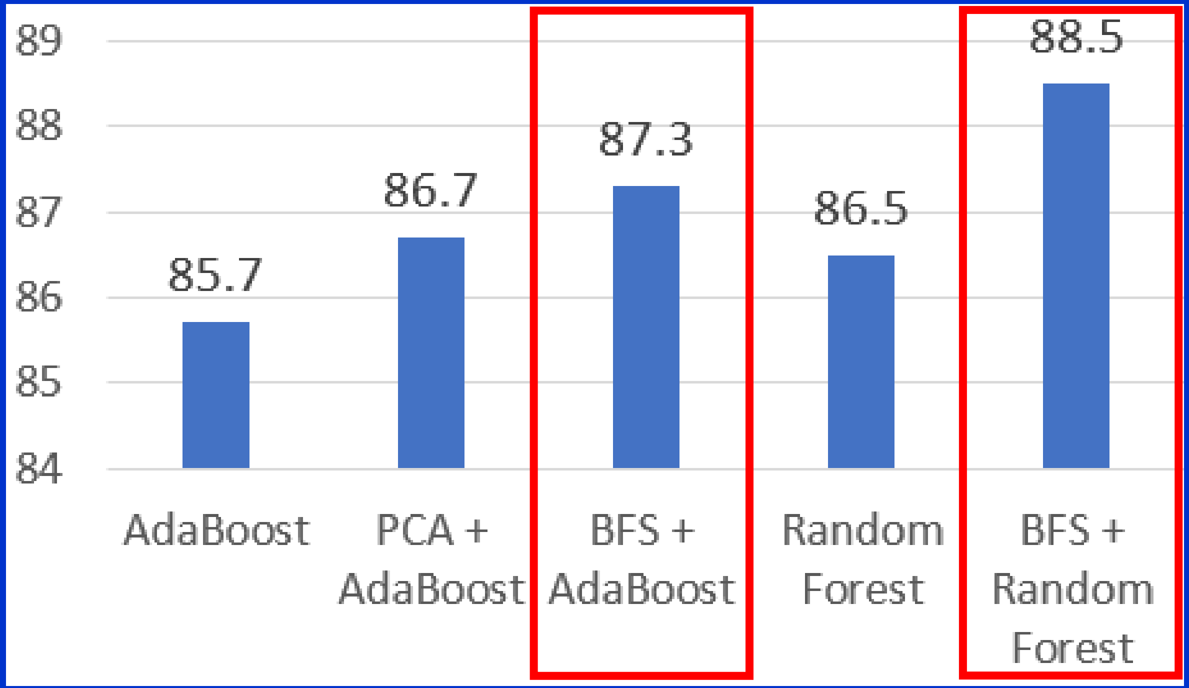
Comparison of performance – SVM without FS vs. Boosted FS

Class	Number of Instances used for Calculating Distance between Classes	SVM (RBF) with No Feature Selection (A)		Rank by Distance using missclassification by SVM + SVM (RBF) (B)		Improvement (B-A)	
		Correct Classification Rate (%)	Number of Selected Feature	Correct Classification Rate (%)	Number of Selected Feature	Correct Classification Rate (%)	Feature Reduction Rate (%)
1 vs. 2&3	482	67.2	9	69.5	3	2.3	66.7
2 vs. 1&3	339	76.9	9	78.7	5	1.8	44.4
3 vs. 1&2	361	66.3	9	67.6	8	1.3	11.1
Average	**394.0**	**70.1**	**9.0**	**71.9**	**5.3**	**1.8**	**40.7**

- Contraceptive data has 3 classes, 1,473 instances and 9 features (2 M & 7 C).
- The CCR increases by 1.8%, and the features are reduced by 40.7%.

Validation of Boosted Feature Selection by Applying to Other Classifiers

Average
CCR (%)
on 5 data



- Boosted Feature Selection increases the CCR of AdaBoost & Random Forest by 1.6% and 2.0% respectively.
- In case of AdaBoost, boosted feature selection is more effective than PCA.

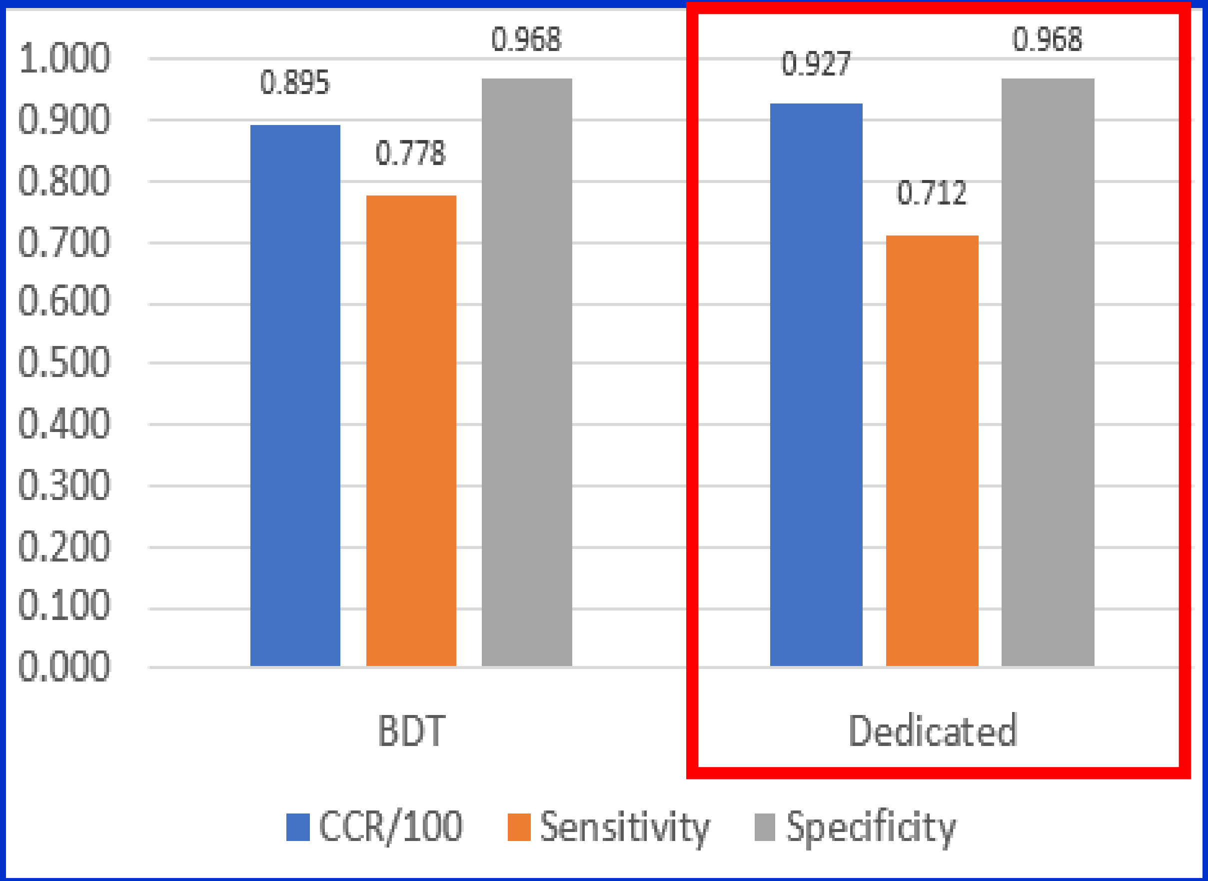
Result of Improved Classification Methodology for Cardiotocography Data

Methodologies in BDT vs. Class-dedicated architectures

Classification Architecture			Binary Decision Tree			3 Class-dedicated SVMs			
CV	Criteria for Performance Evaluation		Literature		SVM (RBF)	SVM (RBF)	Boosted Feature Selection + Clustering + SVM	Ada-Boost	Random Forest
			C&W (2015)	Y&K (2013)					
Train: 90% Test: 10%	CCR (%)	Training	N/A	N/A	91.2	94.6	98.6	97.3	94.2
		Testing	90.6	91.6	89.5	92.7	98.5	90.6	92.9
	Sensitivity		0.852	0.767	0.778	0.712	0.983	0.824	0.882
	Specificity		0.912	0.969	0.968	0.968	0.995	0.975	0.988
Train: 75% Test: 25%	CCR (%)	Training				93.8	98.7	96.4	94.1
		Testing				90.6	96.3	91.9	93.6
	Sensitivity					0.718	0.983	0.881	0.881
	Specificity					0.978	0.996	0.973	0.988

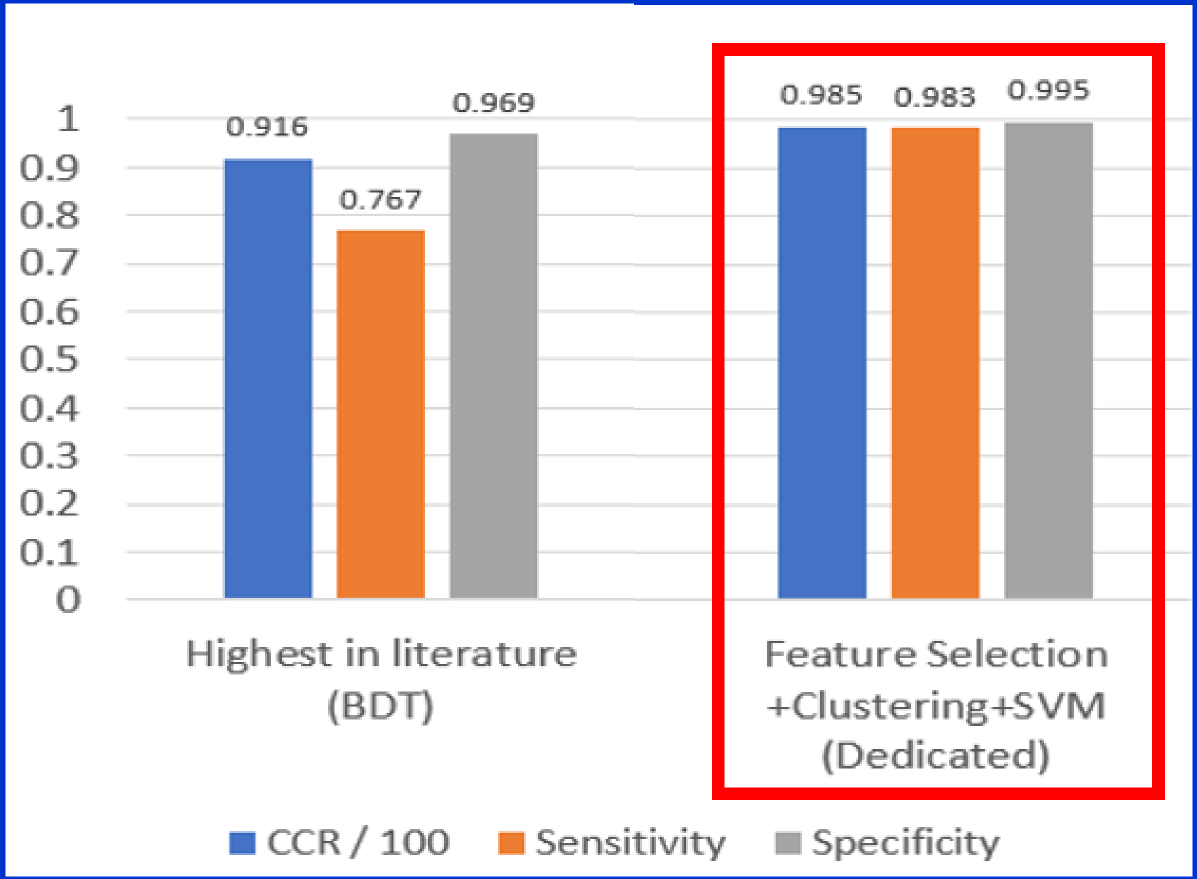
- Class-dedicated SVM is implemented on both 10-fold CV and 4-fold CV.

Comparison of BDT and Class Dedicated Architecture



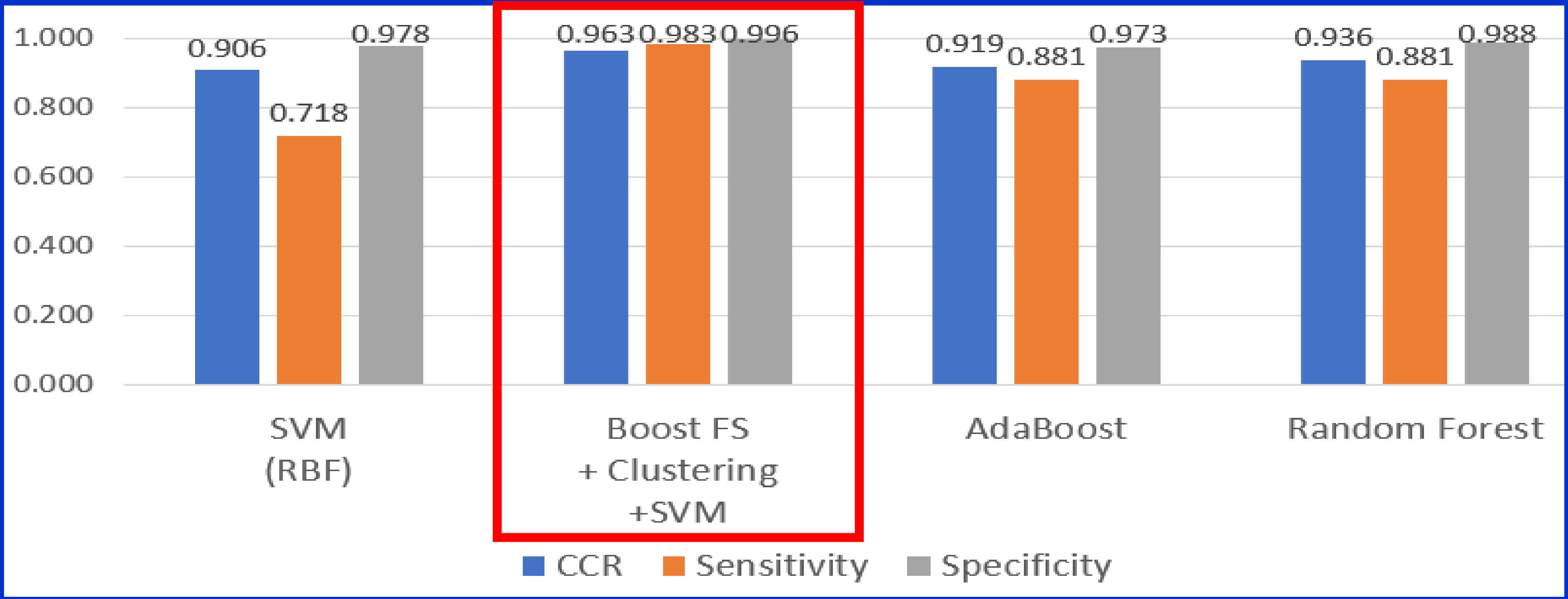
- Two architectures are compared on same condition, SVM with RBF kernel.
- Class-dedicated architecture shows 3.2% higher CCR but 0.066 lower sensitivity.

Comparison of Proposed Methodology with Literature



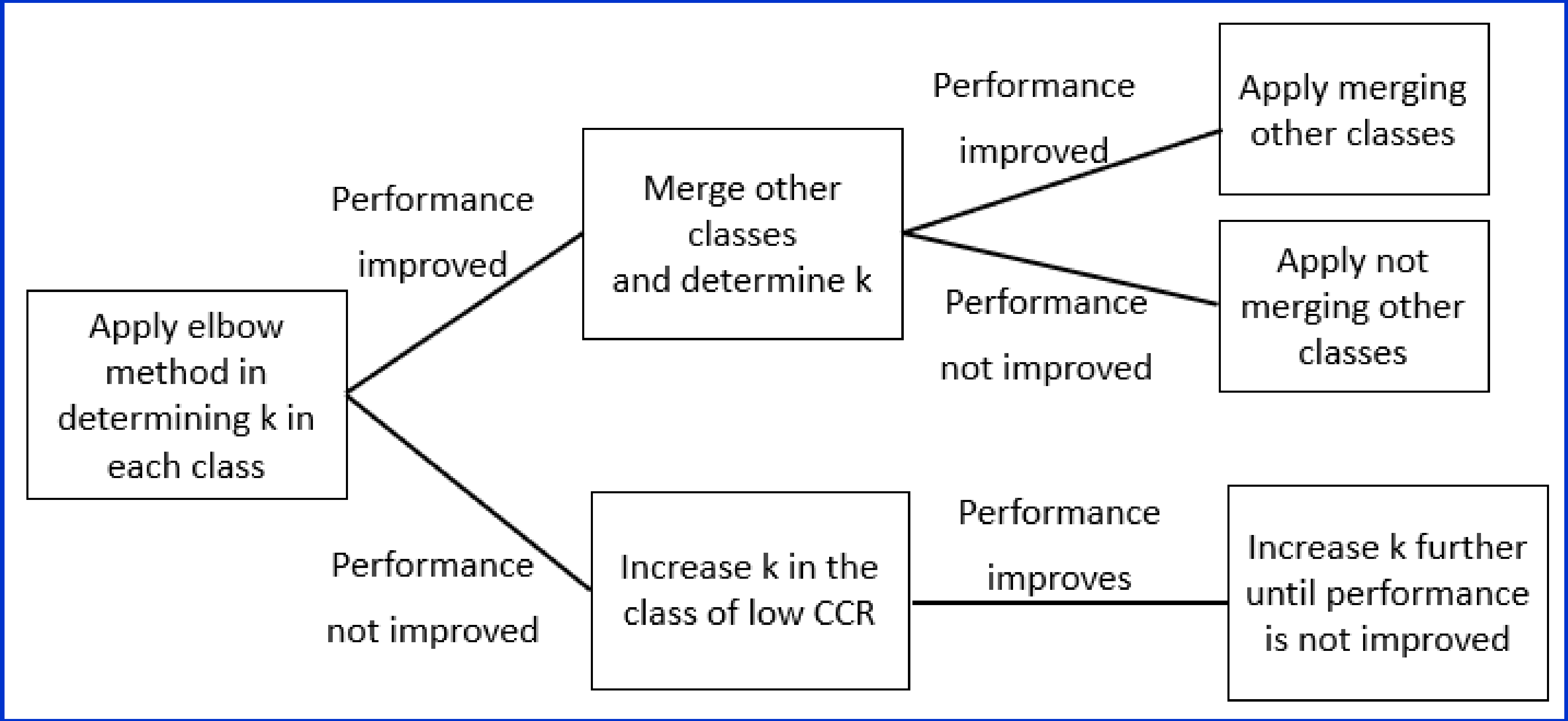
- The proposed methodology overcomes disadvantage of class-dedicated architecture.
- CCR, sensitivity, specificity are increased by 6.9%, 0. 216, 0.026 respectively.

Comparison of 4 Methodologies in Class Dedicated Architecture



- The methodology outperforms SVM (RBF kernel), AdaBoost and RF in all criteria.
- The CCR is higher than RF by 2.7%. The sensitivity is higher than RF by 0.102.

Tree Diagram in Searching for the High Performance Model



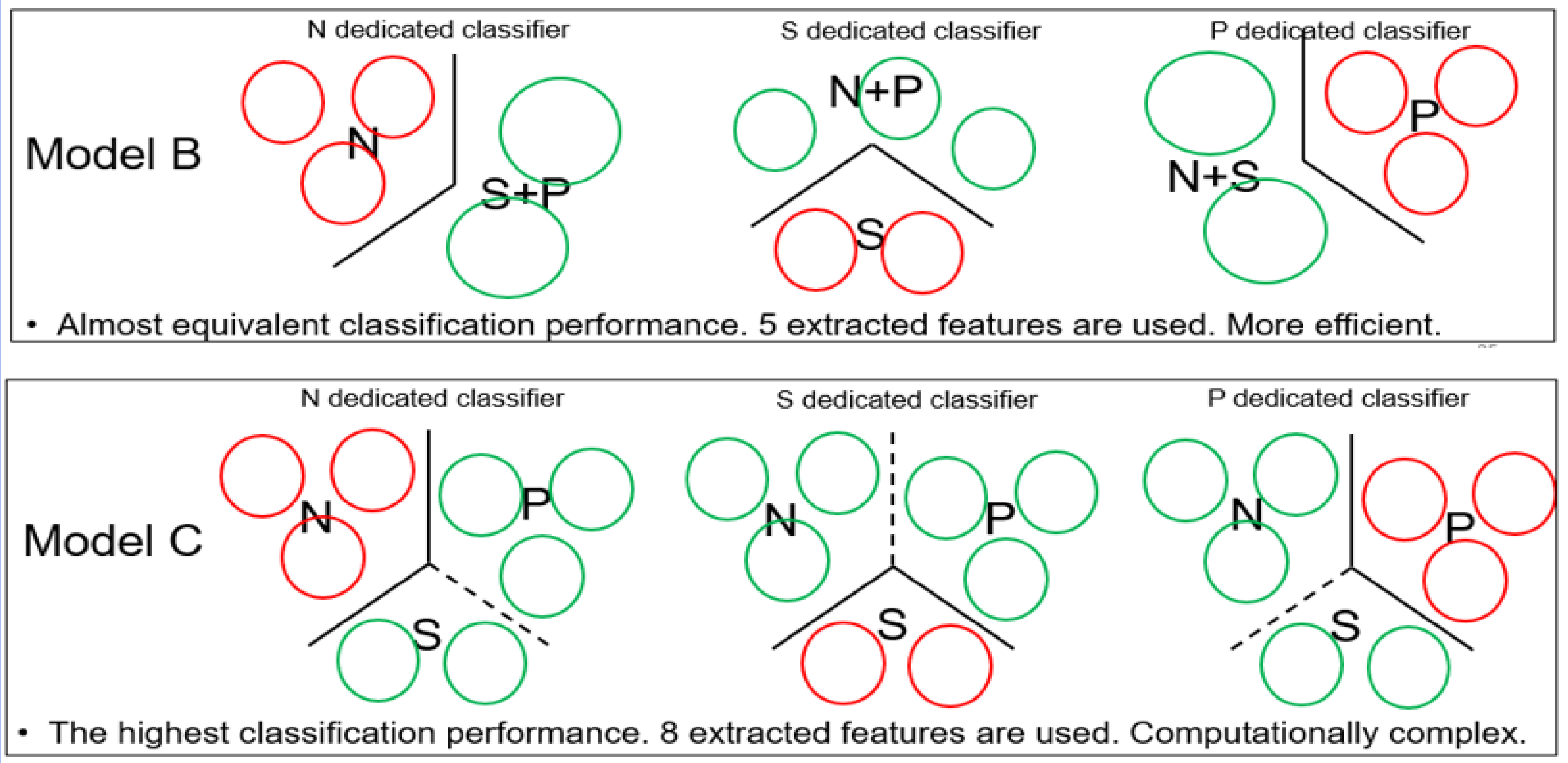
- The number of extracted features can be reduced by merging other classes.
- The performance can be improved by increasing k in the low CCR class.

Determining Optimal Number of Clusters (k) for Improved Model

Model		A	B	C	D	E
Number of Clusters	Class N	k=1	k=3 (N) vs. k=2 (S&P)	k=3	k=3	k=3
	Class S	k=1	k=2 (S) vs. k=3 (N&P)	k=2	k=3	k=4
	Class P	k=1	k=3 (P) vs. k=2 (N&S)	k=3	k=3	k=3
Number of Reduced Features		3	5	8	9	10
CCR (%)	Training	90.0	97.3	98.7	97.7	98.2
	Testing	82.6	94.8	96.3	96.3	96.9
Sensitivity		0.721	0.978	0.983	0.920	0.898
Specificity		0.978	0.987	0.996	0.996	0.998

- The model B and C show the higher performances compared to other models.
- Model B reduced features from 21 to 5. (The least is 7, Chamidah and Wasito, 2015)
- The number of clusters in Class S is critical in increasing sensitivity.

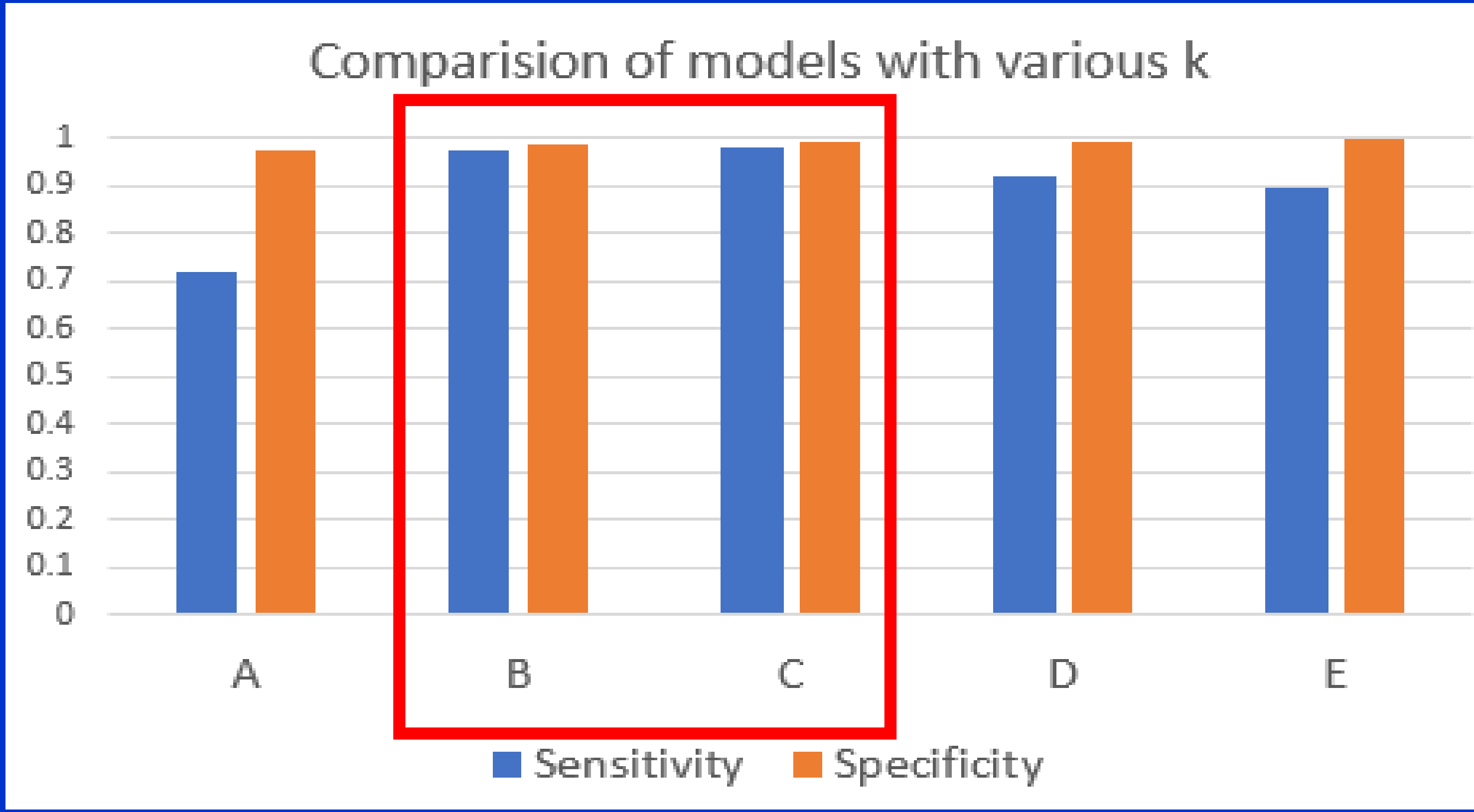
Comparison of Clustering Structure between Model B and Model C



- In model B, k is determined within target class and within merged other classes.
- In model C, k is determined in each of the 3 classes.

Performance Evaluation by Sensitivity & Specificity

The performance of 5 models depending on different number of clusters (k)



- Both the model B (0.983 / 0.996) and C (0.978 / 0.987) show the higher performances in terms of sensitivity / specificity compared to other models.

Procedure to Apply to Diagnosis Activity

- Positive predictive value:

  Probability that subjects with a positive screening test truly have the disease

- Negative predictive value:

  Probability that subjects with a negative screening test truly don’t have the disease

- Apply the concept to 3 classes of Cardiotocography data.

Procedure to Apply to Diagnosis Activity

- Calculate (1) Normal (Negative) (2) Suspect (3) Pathologic (Positive) Predictive Values, respectively.

- Confusion Matrix

Class 1 (Normal) — Predicted; Original

TP_1	FP_1
FN_1	TN_1

Class 2 (Suspect) — Predicted; Original

TP_2	FP_2
FN_2	TN_2

Class 3 (Pathologic) — Predicted; Original

TP_3	FP_3
FN_3	TN_3

$$\text{Normal Predictive Value} = \frac{TP_1}{TP_1 + FN_1}$$

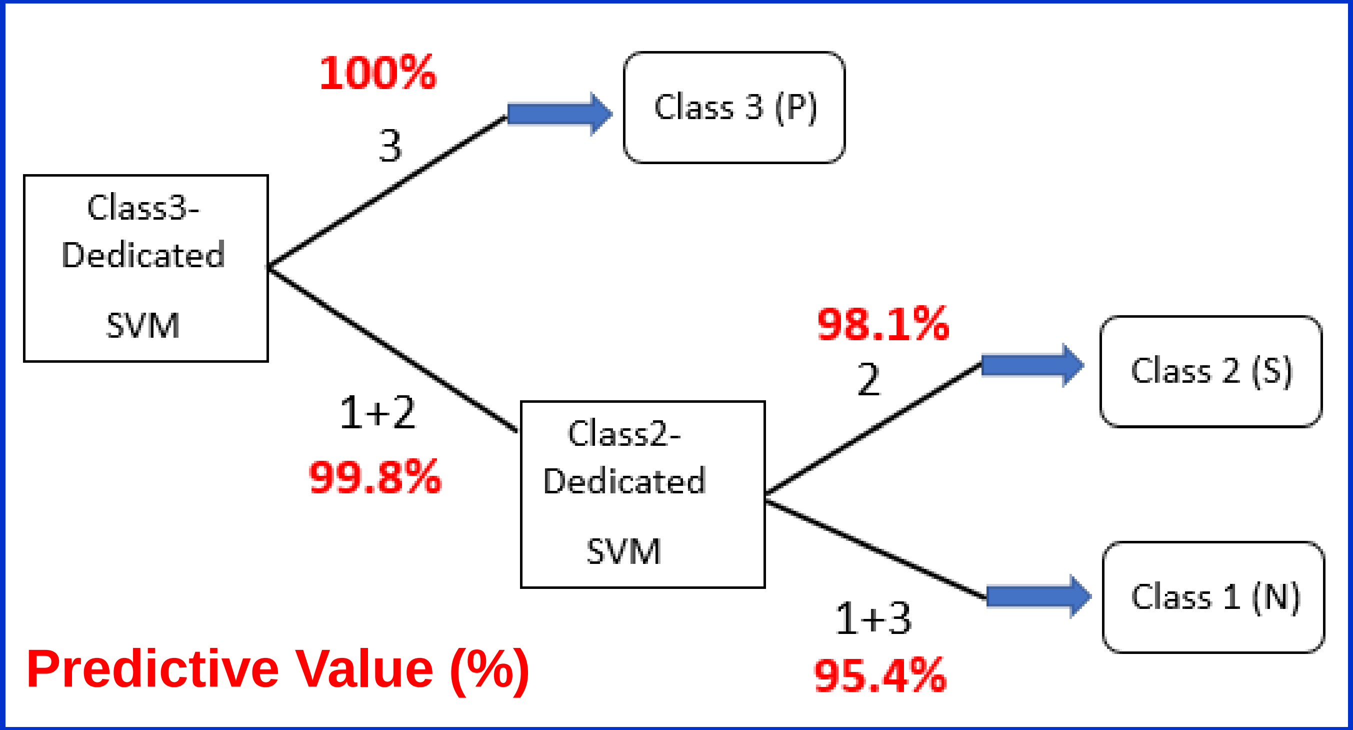
$$\text{Suspect Predictive Value} = \frac{TP_2}{TP_2 + FN_2}$$

$$\text{Pathologic Predictive Value} = \frac{TP_3}{TP_3 + FN_3}$$

In Model B, Predictive values are 95.1%(Normal), 98.1%(Suspect), 100.0%(Pathologic)

Tree Diagram for Diagnosing Fetal State

- Prioritize use of class-dedicated SVM of higher predictive value: Class3 > Class 2
- Class 1-dedicated SVM is not needed because the predictive value is the lowest.



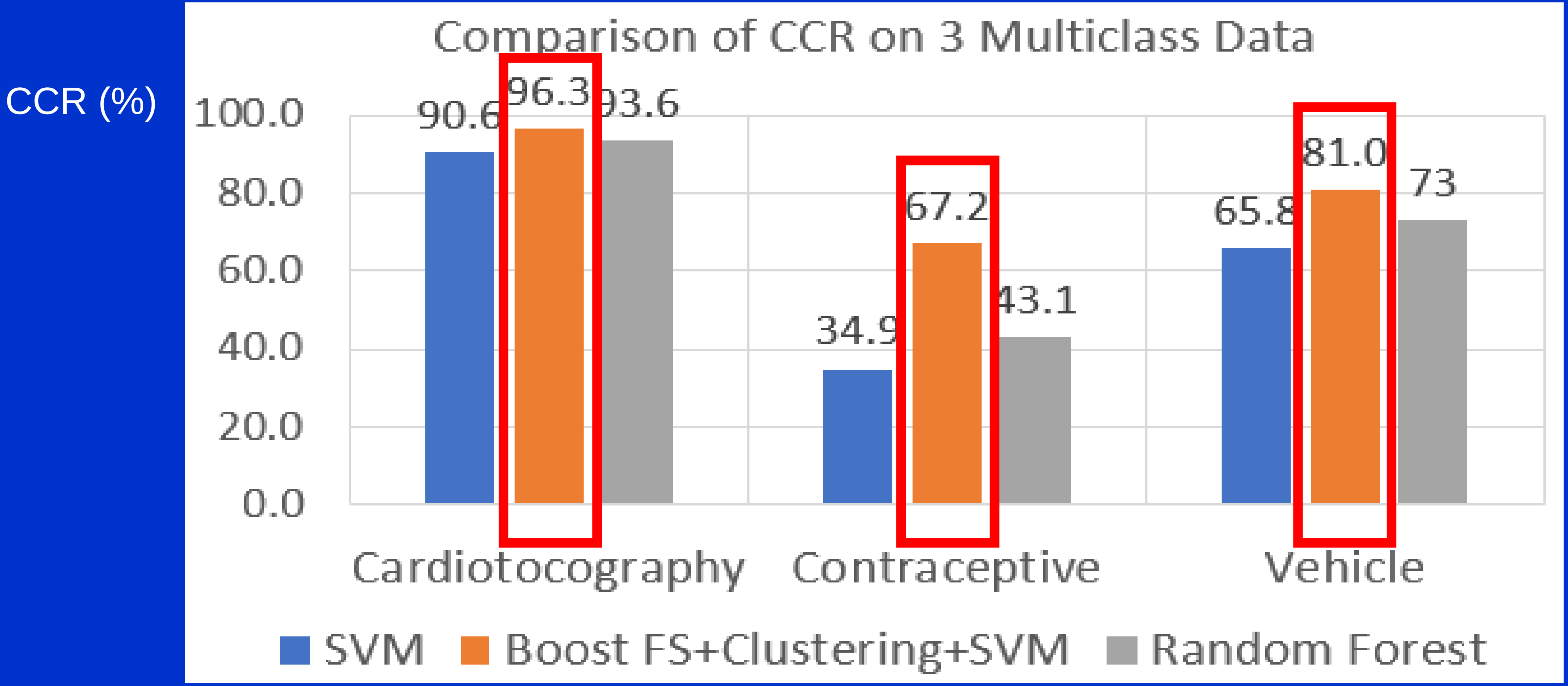
- If prediction from class 3 dedicated SVM is class 3, the outcome is Pathologic.
- If prediction is class 1 or 2, the outcome depends on class 2 dedicated SVM.

Result of Validation of Methodology by Applying to Other Multiclass Data

No.	Data	Number of Classes	Number of Instances	Number of Original Features	Number of Extracted Features	Criteria for Performance Evaluation		Methodologies				Increase (B-A)
								SVM (RBF) (A)	Boosted Feature Selection + Clustering + SVM (B)	AdaBoost	Random Forest	
1	Cardio-tocography	3	2,126	21	8	CCR (%)	Training	93.8	98.7	96.4	94.1	4.9
							Testing	90.6	96.3	91.9	93.6	5.7
						Sensitivity		0.718	0.983	0.881	0.881	0.265
						Specificity		0.978	0.996	0.973	0.988	0.018
2	Contra-ceptive	3	1,473	9	6	CCR (%)	Training	62.0	87.7	53.5	43.2	25.7
							Testing	34.9	67.2	41.4	43.1	32.3
						Class 1 CCR		0.473	0.811	0.497	0.529	0.338
						Class 2 CCR		0.183	0.681	0.345	0.299	0.498
						Class 3 CCR		0.277	0.498	0.360	0.400	0.221
3	Vehicle	4	846	18	13	CCR (%)	Training	83.4	87.9	78.4	67.4	4.5
							Testing	65.8	81.0	65.4	73.0	15.2
						Class 1 CCR		0.933	0.955	0.883	1.000	0.022
						Class 2 CCR		0.458	0.676	0.311	0.333	0.218
						Class 3 CCR		0.429	0.727	0.464	0.518	0.298
						Class 4 CCR		0.844	0.860	0.900	1.000	0.016

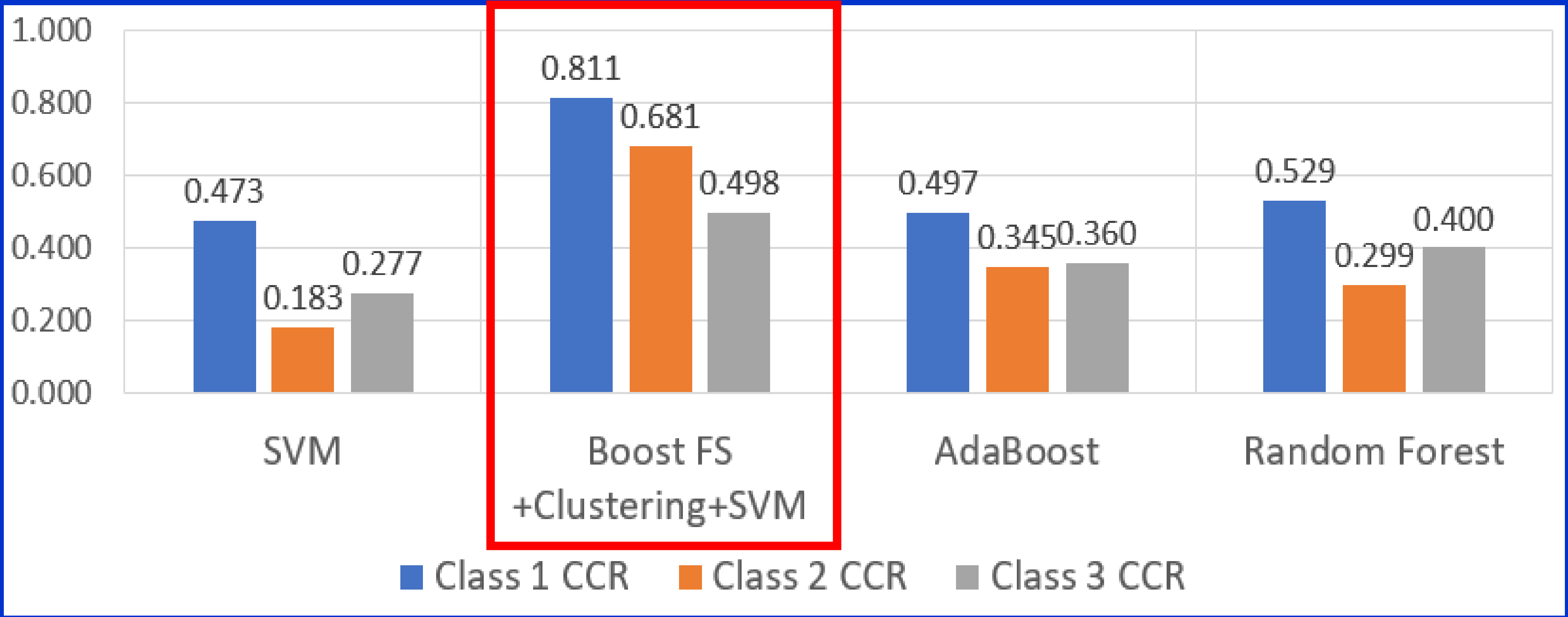
- The methodology is applied to data with different number of instances and features.

Comparison of CCR on 3 Multiclass Data



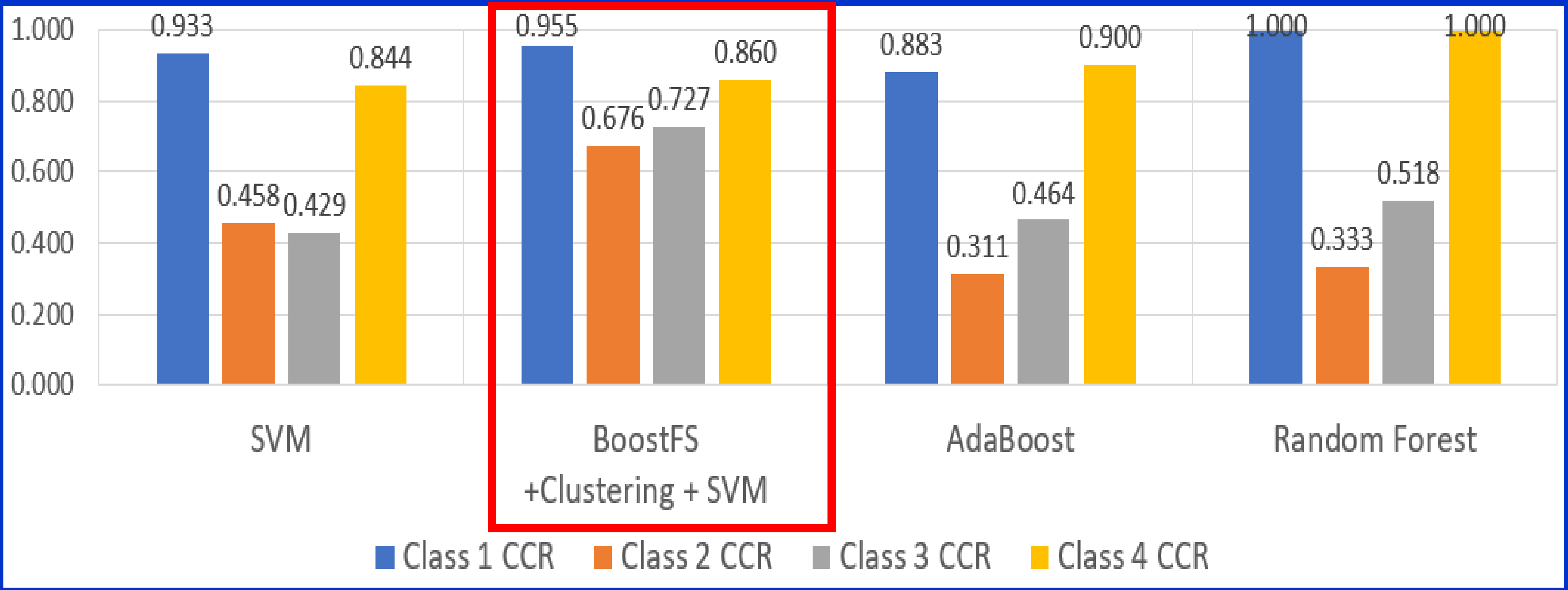
- Proposed methodology outperforms Random Forest and SVM on all 3 data.
- CCR is higher than RF by 11.6 %, and higher than SVM by 17.7% on average.

Comparison of CCR of Each Class on Contraceptive Data



- The proposed methodology shows balanced higher sensitivity, specificity, CCR of class compared to SVM without FS, AdaBoost and Random Forest.

Comparison of CCR of Each Class on Vehicle Data



- The proposed methodology shows balanced higher sensitivity, specificity, CCRs of class compared to SVM without FS, AdaBoost and Random Forest.

6. Conclusion

1. Implemented various methodologies in terms of feature selection / extraction methods, kernel selection in SVM and composition of ensemble methods to improve the performance of SVM.

- The CCR of proposed feature ranking-PCA ensemble outperforms the method on the same 2-class and multiclass data in literature.
- Proposed efficient methodology for large-scale data by reducing data size for Grid Search of RBF kernel. Time complexity is significantly reduced.
- Proposed efficient methodology depending on feature type.
- Compared feature ranking criteria and selected suitable one for multiclass data.

2. Proposed Boosted feature selection methodology on Cardiotocography data, prioritizing the features with maximum discriminatory power.

- Features reduced by 49.2% compared to SVM without feature selection.
- CCR increases by 3.5% compared to literature.
- CCR increases by 1.0% compared to SVM without feature selection.
- Effective on other multiclass data. (feature reduction rate 40.7%)
- The methodology is more effective on data with larger error rate.
- The methodology is effective on AdaBoost and Random Forest.

3. This research proposed improved classification methodology on Cardiotocography data for more accurate diagnosis on fetal state.

- Used the boosted feature selection, feature extraction by K-means clustering and class-dedicated SVM for 3-class Cardiotocography data.
- Overcame the disadvantage of BDT classification architecture.
- Increase: CCR by 6.9%, sensitivity by 0.131 compared to literature.
- The pathologic class is predicted 13.1% more accurately compared to literature.
- The features are reduced from 21 to 5, reducing computational complexity.
- Contribute in building more reliable and efficient decision support system.

Titles of Two Paper Drafts

1. Boosted Feature Selection Methodology for Class Dedicated SVM and Its Application in Fetal Health Prediction. (focused on Methodology)

2. Boosted Feature Selection for Class Dedicated SVM and Its Application in Fetal Health Prediction. (focused on Application)

Acknowledgement

I deeply thank faculty advisor, Professor Susan Lu and Committee members, Professor Nagendra Nagarur, Professor Harold W. Lewis, Professor Changqing Cheng for a lot of advices for my dissertation research and writing of this dissertation.