

BOOSTED FEATURE SELECTION FOR CLASS DEDICATED SVM
AND ITS APPLICATION IN FETAL HEALTH PREDICTION

BY

JINPYO LEE

B.S.E. Seoul National University, 2003
M.S.E. University of Michigan, 2011

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Systems Science & Industrial Engineering
in the Graduate School of
Binghamton University
State University of New York
2019

© Copyright by Jinpyo Lee 2019

All Rights Reserved

Accepted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Systems Science & Industrial Engineering
in the Graduate School of
Binghamton University
State University of New York
2019

APRIL 29, 2019

Dr. Mohammad Khasawneh, Chair
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Susan Lu, Faculty Advisor
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Nagendra Nagarur, Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Harold W. Lewis, Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Changqing Cheng, Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Kenneth Chiu, Outside Examiner
Department of Computer Science, Binghamton University

Abstract

This research developed improved classification methodology for Cardiotocography (CTG) data. CTG data has been widely used for diagnosing fetal health until delivery. However, the high dimension and multiclass characteristic of CTG data are barriers in improving the performance further in correct classification rate (CCR) and computational efficiency. This research implemented various experiments on feature selection and extraction methods, kernel selection in SVM and ensemble methods.

Firstly, this research experimented new methodology of classification model by using both feature selection and feature extraction, and SVM with 4 kernels by a wrapper method to obtain the highest CCR with reduced computational complexity. According to experimental results, the algorithm of 4 combination ensemble resulted in the highest CCR with 75% reduced time complexity. The alternative combinations reduced it further with slightly degraded CCR. More efficient ensemble depending on feature types reduced computation time further by 39% for 2-class and 70% for multiclass data.

Secondly, Linear Discriminant Analysis (LDA) and distance between classes were used as alternative feature ranking criteria for feature selection. According to comparison of performance, the CCRs of feature ranking by SVM, LDA, distance between classes and PCA are at almost same level in case of 2-class data. In case of multiclass data, the distance between classes is the most effective ranking criteria. The distance between

classes is effective on multiclass and large-scale data with large instances or high dimension.

Finally, improved classification methodology for CTG data is proposed. As the first step, boosted feature selection was developed. One vs. all multiclass classification architecture with a wrapper composed of feature ranking by the same classifier SVM and distance between classes among misclassified instances resulted in increase of CCR by 3.5% compared to the highest in literature and 49.2% feature reduction compared to the case without feature selection. By applying boosted feature selection and feature extraction by K-means clustering, Class-dedicated SVM increased CCR and sensitivity by 6.9% and 0.131 compared to the highest in literature. This proposed methodology is expected to increase the efficiency of diagnosis based on CTG data by predicting the pathologic state more accurately.

Acknowledgements

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 this dissertation.

Table of Contents

List of Tables.....	ix
List of Figures.....	xi
List of Abbreviations.....	xiv
Introduction.....	1
Chapter 1. Literature Review.....	16
1.1 Feature Selection.....	16
1.2 Feature Extraction.....	23
1.3 Classification.....	28
1.3.1 Support Vector Machine.....	28
1.3.2 Ensemble Method.....	38
1.4 Feature Selection of Cardiotocography Data.....	42
1.5 Summary.....	43
Chapter 2. Methodology.....	46
2.1 Research Framework.....	46
2.1.1 Developing Feature Selection / Extraction Methodology.....	46
2.1.2 Comparing Performance of Various Feature Ranking Criteria.....	50
2.1.3 Developing Boosted Feature Selection Method.....	50
2.1.4 Developing Classification Methodology for Cardiotocography Data.....	51
2.2 Techniques applied to Proposed Model.....	52
2.2.1 Data Preparation.....	52
2.2.2 Outlier Treatment.....	55
2.2.3 Feature Classification Rate Ranking Method.....	57
2.2.4 Principal Component Analysis.....	58
2.2.5 Linear Discriminant Analysis.....	59
2.2.6 Distance between Classes.....	61
2.2.7 Support Vector Machine.....	62
2.2.8 Parameter Optimization.....	65
2.2.9 Feature Selection of Cardiotocography Data.....	67
2.2.10 Improved Classification Methodology for Cardiotocography Data.....	70
2.3 Summary.....	74
Chapter 3. Experimental Result on 2-class and Multiclass Data.....	76
3.1 Summary of Performance of 4 Ensemble Algorithms.....	76
3.2 Comparison of Performance with Approaches in Literature (2-class data).....	80

3.3 Efficient Algorithm Depending on Feature Type.....	82
3.4 Efficient Algorithm for Multiclass Data.....	85
3.5 Comparison of Performance with Approaches in Literature (Multiclass data).....	88
3.6 Comparison of Four Feature Ranking Criteria.....	89
Chapter 4. Experimental Result on Cardiotocography Data.....	92
4.1 Boosted Feature Selection of Cardiotocography Data.....	92
4.2 Validation of Boosted Feature Selection by Applying to Other Classifiers.....	99
4.3 Improved Classification Methodology for Cardiotocography Data.....	100
4.4 Procedure to Search for Highest Performance Model.....	105
4.5 Procedure to Apply to Diagnosis Activity.....	106
4.6 Validation of Methodology by Applying to Other Multiclass Data.....	107
Chapter 5. Conclusions.....	112
References.....	118

List of Tables

Table 1. Gaps in literature and methodologies in this research.....	45
Table 2. The component in the combinations of the 4 kinds of Rank-PCA ensemble algorithms.....	48
Table 3. The criteria for categorizing features depending on feature type.....	49
Table 4. The characteristics of Ten kinds of 2-class data sets.....	53
Table 5. The characteristics of Seven kinds of multiclass data sets.....	53
Table 6. The description of the features in Cardiotocography data.....	55
Table 7. The order of input combinations of principal components.....	59
Table 8. Three kinds of one vs. all classification for 3 class data set.....	68
Table 9. Comparison of performance with other methodologies in literature.....	81
Table 10. The characteristics of features of all 2-class data (Metric/Categorical).....	83
Table 11. The result of preprocessing methods, kernels and efficient algorithm.....	83
Table 12. The comparison of computation time between 4 combination algorithm and efficient algorithm depending on feature type.....	85
Table 13. Characteristics of multiclass data.....	86
Table 14. The characteristics of features of multiclass data (Metric/Categorical).....	86
Table 15. The result of preprocessing methods, kernels, efficient algorithm for multiclass data.....	87
Table 16. The comparison of computation time between 4 combinations and efficient algorithm depending on feature type in case of multiclass data.....	88
Table 17. Summary of ensemble classifiers in improved algorithms depending on number of class and feature types.....	88
Table 18. Comparison of performance with SVM-RFE-Taguchi in literature.....	89
Table 19. Comparison of four feature ranking criteria.....	90
Table 20. The correct classification rates of individual features in Cardiotocography data.....	92
Table 21. The number of instances used for calculating distance between Classes.....	93

Table 22. Comparison of performance between SVM without and with boosted feature selection.....	94
Table 23. The features selected for each one vs. all classification.....	95
Table 24. The Comparison of sensitivity and specificity between the cases without and with boosted feature selection on Cardiotocography data.....	96
Table 25. The Comparison of TPR and FPR between the cases without and with boosted feature.....	97
Table 26. Comparison of performance between SVM without and with boosted feature selection on Contraceptive data set.....	98
Table 27. The features selected for each one vs. all classification (Contraceptive data)...	98
Table 28. Comparison of performance of boosted feature selection with other classifiers on 5 binary class data.....	99
Table 29. Comparison of performance among various methodologies in 2 classification architectures on Cardiotocography Data.....	101
Table 30. Comparison of various models in determining optimum number of clusters (Training data 75%, Testing data 25%).....	104
Table 31. Comparison of performance among various methodologies on 3 multiclass data.....	108

List of Figures

Figure 1. Research framework for developing classification methodology for Cardiotocography data.....	7
Figure 2. The 4 phases of the research flow.....	46
Figure 3. The research flow developing 4 kinds of ensemble algorithms.....	48
Figure 4. The research flow of developing efficient algorithms depending on feature type.....	49
Figure 5. The research flow of the 2 nd phase, Comparing performance of various feature ranking criteria.....	50
Figure 6. The research flow of the 3 rd phase, Developing boosted feature selection method and validating performance.....	51
Figure 7. The research flow of the final phase, Developing classification methodology for Cardiotocography data and validating the effectiveness.....	52
Figure 8. The process of feature correct classification rate ranking method.....	58
Figure 9. The process of using LDA as a feature ranking criteria.....	61
Figure 10. The process of using distance between classes as a feature ranking criteria...62	
Figure 11. The Flow of Algorithm – Preprocessing by SVM & Distance measure + SVM (classifier).....	69
Figure 12. Flow of algorithm – PCA + AdaBoost + Wrapper method for dimension decision.....	70
Figure 13. Flow of algorithm – Boosted feature selection by DT + AdaBoost + Wrapper method.....	70
Figure 14. Flow of algorithm – Boosted feature selection by DT + Random Forest + Wrapper method.....	70
Figure 15. Comparison of classification architecture – Binary Decision Tree (BDT) vs. Class-dedicated SVMs.....	71
Figure 16. The architecture of improved classification methodology for Cardiotocography data.....	71
Figure 17. The confusion matrix of the 3 classes.....	73
Figure 18. The reduction rate of computation time of 3 ensemble algorithms per the number of instances in data.....	77

Figure 19. Average Correct Classification Rate vs. Computation Time (4 ensemble algorithms).....	78
Figure 20. The feature reduction rate of 4 ensemble algorithms.....	79
Figure 21. Comparison of CCR per each 2-class data from 4 feature ranking criteria.....	90
Figure 22. Comparison of average CCR from 4 feature ranking criteria on 2-class data....	91
Figure 23. Comparison of CCR per each multiclass data from 4 feature ranking criteria...	91
Figure 24. Comparison of average CCR from 4 feature ranking criteria on multiclass data.....	91
Figure 25. The CCRs of individual features in Cardiotocography data.....	92
Figure 26. The distance between two classes in original order of features in the 3 one vs. all classifications.....	93
Figure 27. Sorted features in descending order of the 3 one vs. all classifications.....	94
Figure 28. The graphical representation of the change of correct classification rate as the input features cumulatively increase.....	95
Figure 29. The Classification matrix of P vs. N or S Classification.....	96
Figure 30. The TPR and FPR on ROC Curve in case of boosted feature selection.....	97
Figure 31. The graphical representation of the change of correct classification rate as the input features cumulatively increase (Contraceptive data set).....	98
Figure 32. Comparison of CCR(%) among ensembles with different feature selection /extraction methods.....	99
Figure 33. The relation represented by linear regression.....	100
Figure 34. Comparison of performance between BDT and Class-dedicated architecture.	101
Figure 35. Comparison of performance of 4 methodologies with Class-dedicated architecture.....	102
Figure 36. Comparison of performance with the methodology in literature.....	103
Figure 37. Comparison of clustering architecture, Model B vs. Model C.....	104
Figure 38. Comparison of sensitivity and specificity in the models with various numbers of clusters.....	105
Figure 39. Tree diagram to search for the highest performance model.....	106

Figure 40. Tree diagram for diagnosing fetal state.....	107
Figure 41. Comparison of CCR (%) of 3 methodologies on 3 multiclass data.....	108
Figure 42. Comparison of each class CCR/100 in 4 methodologies on Contraceptive data.....	109
Figure 43. Comparison of each class CCR/100 in 4 methodologies on Vehicle data.....	109

List of Abbreviations

ANFIS: Adaptive Neuro Fuzzy Inference System

ANN: Artificial Neural Network

AUC: Area Under Curve

BDT: Binary Decision Tree

CART: Classification And Regression Tree

CCA: Canonical Correlation Analysis

CCR: Correct Classification Rate

CTG: Cardiotocography

GA: Genetic Algorithm

GS: Grid Search

LDA: Linear Discriminant Analysis

LLE: Locally Linear Embedding

MLK: Multiple Kernel Learning

PC: Principal Component

PCA: Principal Component Analysis

PSO: Particle Swarm Optimization

RBF: Radial Basis Function

RF: Random Forest

RFE: Recursive Feature Elimination

SVD: Singular Value Decomposition

SVM: Support Vector Machine

Introduction

The medical decision support system based on classification is getting more widely used in medical diagnosis. Cardiotocography data has been used in monitoring the fetal status until delivery. The decision support system using Cardiotocography data is expected to increase the accuracy of diagnosis and save the medical cost for the diagnosis. In the literature related to the research of classification methodologies on Cardiotocography data, Support Vector Machine (SVM) has been mostly used as a classifier. SVM is one of the high-performance machine learning algorithms with a number of advantages such as capability of high Correct Classification Rate (CCR), the ability to avoid overfitting and the ability of searching for global solution. However, the computational complexity of SVM become higher in case that the data is multiclass or the size of data gets larger, due to its long computation time required in grid search of parameters of kernel.

In order to overcome this limitation, this research set up two research goals, which are achieving both highest CCR and reduced computational complexity. To meet these goals simultaneously, this research experimented four algorithms of different combinations between feature selection / extraction methods and kernels in SVM to search for efficient algorithm while overcoming the phenomenon that the CCR is in trade-off relation with the time complexity.

Firstly, in order to achieve the first goal, feature selection / extraction methods, i.e., feature classification rate ranking method and PCA were used simultaneously and complementarily, composing ensemble with 4 kernels. The feature classification rate

ranking method can utilize the characteristics of original features, rearranging them in descending order in terms of CCR. On the other hand, PCA extracts features from original data and store them in smaller dimensions in order of maximum variability. The feature selection / extraction methods provide the wrapper method with conditions for forward feature elimination, which is computationally inexpensive. As an effort to reduce the time complexity, this research reduced the size of data on which grid search of RBF kernel parameters is implemented. By using this result, this research experimented 4 algorithms with different combinations between feature selection / extraction methods and kernels. This research experimented various combinations of kernels with different characteristics in terms of computation time and CCR to search for efficient algorithm which achieves the two goals regardless of the feature type of data sets.

Secondly, this research expanded the research framework to multiclass data and searched for efficient algorithm for multiclass data, and the efficient algorithms depending on feature types, i.e., whether the feature is metric or categorical, which are more efficient in terms of time complexity. In addition, Linear Discriminant Analysis (LDA) and discriminatory power by distance between classes are experimented to search for feature selection algorithm for Cardiotocography data.

Thirdly, the performance of boosted feature selection is validated by applying to other multiclass data or other ensemble algorithms, i.e., AdaBoost and Random Forest. After initial implementation of boosted feature selection methodology, which sorts and selects features by feature ranking based on misclassified instances from SVM and wrapper method, the methodology is applied to other multiclass data to validate the performance. In addition, the methodology is applied to ensembles methods such as AdaBoost and

Random Forest by using the misclassified instances from decision tree, and the performance is compared to SVM without feature selection. The condition of data on which the boosted feature selection methodology is the most effective, is discovered.

Finally, improved classification methodology for Cardiotocography data is proposed. The class-dedicated classification architecture is used to overcome the disadvantage of binary decision tree architecture which has been mostly used in literature. As a method to increase classification performance and reduce number of features, K-means clustering algorithm and fuzzy membership function is used. The number of clusters are adjusted in either each class or merged classes to search for the clustering structure which results in the highest classification performance. The developed classification methodology is also applied to other multiclass data to validate the performance.

Background

Cardiotocography was introduced into obstetrics' practice in early 1970s. The Cardiotocography is a method of prenatal examination before delivery and during delivery. Cardiotocogram is a recording of 2 distinct signals, which are Fetal Heart Rate (FHR) and Uterine Activity (UA). The use of Cardiotocography data has been relying on the manual reading and interpretation of the patterns in Cardiotocography data by practitioners. However, the variation of the reading and interpretation results in wrong decision to Caesarian section or wrong diagnosis of fetal status. For this reason, more scientific and systematic analysis of the Cardiotocography data by using data mining technique is required.

So far, numerous methodologies have been used to analyze Cardiotocography data. Neural network was used in classifying the 3 classes, normal, suspicious and pathologic. Adaptive neuro-fuzzy inference systems (ANFIS) was used in classifying 1,655 examples with only normal and pathologic classes. SVM and Genetic Algorithm (GA) were used in classifying the same number and composition of examples. LS (Least Squares)-SVM with RBF kernel was used based on Binary Decision Tree (BDT) classification architecture and parameter optimization by Particle Swarm Optimization (PSO). SVM was also used with feature extraction by Hybrid K-means clustering algorithm based on BDT classification architecture. However, the 3-class classification of Cardiotocography data has had a limitation in increasing the performance further in terms of all criteria, i.e., overall correct classification rate (CCR), sensitivity and specificity. A few of the literature developed classification methodologies on 2-class Cardiotocography data. However, the 2-class data with only normal and pathologic is easier condition to achieve higher classification performance due to the lack of the suspect class data on border area between the two classes.

Another factor to be considered in developing improved classification methodology for Cardiotocography data is computational complexity. As the dimension of data gets higher, the training process of classifying algorithm requires larger CPU processing and longer computation time. The high dimensionality negatively affects the classification performance due to noise and redundant data. In addition, the computational complexity is likely to be high because the Cardiotocography data is a multiclass data.

For this reason, improved classification methodology for Cardiotocography data should be also efficient in terms of computational complexity. SVM is a machine learning algorithm which shows a high-performance in terms of CCR. The researchers in machine

learning have been researching how to reduce the computational complexity of SVM further as described as following.

Multiple Use of Kernels in SVM: Bi, J. et al. (2004) proposed SVM model with mixture of kernels for the first time in literature. The authors proved that the multiple use of kernels in SVM, combined with boosting method, produces better CCR and is computationally more efficient. Following the authors, Wang, Z. et al. (2007) developed an improved Multiple Kernel Learning (MKL) Algorithm, which is MultiK-MHKS (Modification of Ho-Kashyap algorithm with Squared approximation of the misclassification error) by using a feature extraction method, Canonical correlation analysis (CCA). CCA can maximally correlates the m views in the transformed coordinates. By using m kernels, the original input data can be mapped into m feature spaces, where each feature space can be taken as one view of the original input data. 2 kinds of combinations, i.e., $CCA + MHKS_1$ and $CCA + MHKS_2$, are used. The former is combined in the feature level and the latter is combined in the decision level. In the 2 combinations, all the candidate kernels (linear, RBF, polynomial) are used. As a result of experiment on various data, the authors proved that MultiK-MHKS algorithm produces a comparable or superior performance in classification rate and has the competitive efficiency in computation. The same author along with other researchers, Wang, Z. et al. (2014) researched a method to reduce both the time and space complexity of Multiple Kernel Learning (MKL) and proposed an efficient MKL classification machine based on the Nystrom approximation. The proposed improved method is Nystrom approximation matrix with Multiple KMHKSS (NMKMJKS). According to the classification performance comparisons on various data sets, combining

several kernels and applying Nystrom approximation matrix technique can bring a better recognition rate.

Analysis of Computational Complexity of SVM: Bhavsar and Ganatra (2012) surveyed and summarized the variations of SVM in terms of their performance including computational complexity. The variations of SVM have been researched in literature to increase speed efficiency, space efficiency and ability to classify multiclass data. The authors categorized the variations of SVM into 3 groups. The first group is decomposition-based algorithms, which were developed to reduce the memory requirement of SVM because it grows with the squares of number of training instances. Decomposition based methods divide a large optimization problem into a series of smaller problems in which each problem only involves a couple of selected variables in order to make the optimization process more efficient. The variations of SVM in the first group is time-consuming because it considers only memory issue. The 2nd group is variant based algorithms. This category of variations of SVM reduces the training time at the price of classification rate. The final group is multiclass-based algorithms. Originally, SVM was developed to perform binary classification. This category of variations of SVM have been researched to solve multiclass problems because most of the classification problems involve multiclass data. Abdiansah and Wardoyo (2015) implemented time complexity analysis of SVM in LibSVM. Basically, the complexity is divided into 2 kinds, i.e., time complexity and space complexity. The authors focused only on time complexity, which deal with how long the algorithm is executed. In addition, the authors selected SVM in their research because SVM provides a global solution for data classification. Lin et al. developed LibSVM to

facilitate the researchers' use of SVM. This research computed the complexity of SVM algorithm by using C++ and JAVA and compared the result of performance on 3 different data sets. The results show that the complexity of SVM (LibSVM) is $O(n^3)$ and that the time complexity of C++ is lower than that from JAVA. The authors found that the growth of data affects and increases the total computation time.

This research implemented various experiments to increase the performance of classification algorithm on Cardiotocography data in terms of CCR and computational complexity. Firstly, the performance of various kernels in SVM is compared. Secondly, as a preprocessing method, the performances of various feature selection method and feature extraction method are compared. Finally, the ensemble methods are applied in the experiments to improve performance. As a result, this research determined the improved classification methodology for Cardiotocography data and validated the performance.

Research Framework

The research framework used in this research is divided into 4 phases as shown in Figure 1.

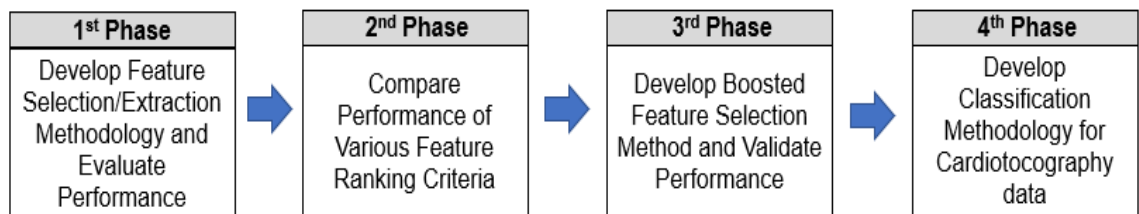


Figure 1. Research framework for developing classification methodology for Cardiotocography data

The 1st phase: In the 1st phase, this research implemented experiments on various combinations between feature selection / extraction methods and kernel in SVM on 10 kinds of 2-class data and 7 multiclass data. At first, the feature selection by sorting according to the CCR from SVM and wrapper method are used. Secondly, features in data are extracted by PCA and the dimension is decided by the same wrapper method. The performances from the feature selection and feature extraction are compared to search for algorithm to achieve highest CCR and lowest computation complexity simultaneously. Finally, the feature types are analyzed, and efficient ensemble algorithm is developed depending on the feature type.

The 2nd phase: In this phase, the performances of different feature ranking criteria, i.e., PCA, LDA, distance between classes on misclassified instances are compared. As the criteria to rank features, the 3 criteria are used in preprocessing stage. The performances are compared to the criteria by SVM with RBF kernel. The experiments are implemented to both 2-class data and multiclass data. Based on the result, the most suitable feature ranking criteria and improved classification architecture for Cardiotocography data are determined.

The 3rd phase: In this phase, the boosted feature selection methodology is developed, and the effectiveness is validated by applying the methodology to other multiclass data and other classifying algorithms. Firstly, the methodology is applied to Contraceptive data which has the same 3-classes as the Cardiotocography data. Secondly, the methodology is applied to ensemble algorithms, AdaBoost and Random Forest by using decision tree as the classifier for misclassification in preprocessing stage.

The 4th phase: In final phase, improved classification methodology for Cardiotocography data is developed. Class-dedicated SVMs are used to overcome the disadvantage and limitation of the BDT classification architecture which has been used in some of the literature related to the classification of Cardiotocography data. The boosted feature selection methodology is used per each binary classification to maximize the CCR of each class. As feature extraction method, K-means clustering algorithm and fuzzy membership function is used to increase the CCR and reduce the extracted number of feature and computational complexity.

Research Problem

Cardiotocography data has been researched in numerous literatures to facilitate medical diagnosis on fetal status before delivery and during delivery of pregnancy. However, the sensitivity has not been improved above 0.852 due to the high dimensionality, multiclass characteristics, and the limitation on classification architecture. Ocak, H. (2012) used SVM and Genetic Algorithm (GA) in the classification of only normal and pathologic class instances of Cardiotocography data. GA is used in parameter optimization of SVM. The proposed methodology by the author outperforms ANFIS-based classifier and ANN-based classifier. The same author proposed ANFIS-based model (2013) and observed the sensitivity 0.966 and specificity 0.972. However, the result is obtained by using only normal and pathologic class instances, which lacks adaptability to real diagnosis activity. There have been literatures which used binary decision tree (BDT) classification architecture on the total 3-class Cardiotocography data. Yilmaz & Kilicier (2013) used LS (Least Squares)-SVM with RBF kernel and parameter optimization by Particle Swarm

Optimization (PSO) on BDT classification architecture. The CCR of this methodology is 91.62%. However, the disadvantage is low sensitivity, 0.767. Chamidah & Wasito (2015) used feature extraction method by Hybrid K-means clustering and SVM on BDT classification architecture. The methodology is effective in reducing number of features to 7, however, its sensitivity is as low as 0.852. The BDT classification architecture has a limitation in increasing sensitivity above 0.852.

SVM has been widely used in numerous application fields because of its advantages over other machine learning algorithms such as its capability of producing high CCR, the ability to avoid overfitting and the ability of searching for global solution. In the recent technological trend that machine learning is getting more applied in embedded devices, the efficiency of computation is getting more emphasized. The research on methodology of reducing computational complexity of SVM while maintaining comparable correct classification rate has been conducted by a number of researchers in literature. Bi, J. et al. (2004) developed multi-kernel learning method for the first time in literature. The authors devised a boosting approach to classification and regression based on column generation by using a mixture of kernels. The authors' proposed model searches for the optimal kernel in SVM automatically and significantly reduces the testing time compared to existing methods. Wang, Z. et al. (2007) developed an improved Multiple Kernel Learning (MKL) algorithm by using Canonical correlation analysis (CCA). As a result of experiment on various data, the authors proved that their proposed algorithm produces a comparable or superior performance in classification rate and has the competitive efficiency in computation. In addition, Wang, Z. et al. (2014) researched a method to reduce both the time and space complexity of Multiple Kernel Learning (MKL) and proposed an efficient

MKL classification algorithm. According to the classification performance comparisons on various data sets, their proposed approach can bring a better recognition rate.

Ensemble method also has been used in literature related to SVM. In Zhang & Yang (2008), an ensemble of classifiers with genetic algorithm (GA) based feature selection, was developed. The GA-ensemble is composed of ANN, Decision tree and SVM. According the experimental results, the proposed GA-ensemble outperforms other algorithms and is an useful method for classification and feature selection. Huang, M.W. et al. (2017) researched SVM ensemble methods in breast cancer prediction with a motivation that there had been very few studies focused on examining the prediction performances of SVM based on different kernel functions. In addition, the authors evaluated the performance of SVM and SVM ensembles over small and large-scale breast cancer data sets. Finally, they found that linear kernel based SVM ensemble with bagging method and RBF kernel based SVM ensemble with boosting method performed better in case of small-scale data set while RBF kernel based SVM with boosting method performed better in case of large-scale data set.

In literature, there have been approaches using multi-kernel and ensemble method. However, the researches have limitations in the aspect that the various combinations of feature subsets are not considered by a wrapper method. In addition, multicollinearity among features are not considered in the combinations of feature subsets. A data-driven approach should be researched in depth in order to increase the performance of classification models. Even if the multicollinearity among features may significantly affects the performance of classifier, the methodologies in literature have not considered the influences of multicollinearity. From this point of view, this research tried to increase the

performance of classifier in terms of CCR and computational complexity by introducing more data-driven approach to SVM algorithm.

Research Objectives

In response to the identified research problem, this research established the following research objectives.

1. Develop feature selection and feature extraction methods which contribute in building SVM classification model achieving highest CCR and least computational complexity.
2. Analyze the advantages and disadvantages of 4 kernels (Polynomial, Sigmoid, Radial basis function & Linear) and develop algorithm to integrate the advantages of each kernel.
3. Analyze the reasons for high computational complexity and develop methodology to reduce the computational complexity.
4. Develop feature selection method, feature extraction method and classification model to achieve the highest CCR and least computational complexity depending on and regardless of the feature types in data sets.
5. Develop improved feature selection / extraction and classification methodology for multiclass Cardiotocography data, overcoming the disadvantage of the classification architecture which has been used in literature.

Research Questions

This research aims to answering the following questions:

1. How to construct classification model for the highest CCR and least computational Complexity?
2. How to use the 4 kernels in SVM in appropriate way to combine the advantage of each kernel?
3. What is the reason for the long computation time of SVM and how to reduce it while maintaining the highest CCR?
4. What are the efficient algorithms depending on and regardless of the feature types in data sets?
5. What is the improved classification methodology for Cardiotocography data in terms of preprocessing, feature selection or extraction, and classification architecture?

Research Contribution and Significance

Firstly, the research developed a methodology to achieve higher CCR with more reduced time complexity compared to the methodology with highest performance on the same data. The performance of the ensemble algorithm with 4 combinations between feature selection / extraction methods and kernel in SVM, is almost equivalent to that from Genetic Algorithm (GA) in literature in terms of CCR and feature reduction rate. However, it is significantly more efficient than GA in terms of time complexity by reducing computation time more than 75%. In addition, the performance of the proposed algorithm

with 4 combinations is better than GA-ensemble which is composed of ANN, decision tree and SVM in literature. The CCR is higher than the method in literature, by 4.5% with equivalent feature reduction rate.

Secondly, this research developed data-driven methodology to reduce time complexity by using the feature types, whether the features are metric or categorical. The efficient ensemble algorithms depending on the feature types are researched to reduce computation time further. As a result, two efficient ensemble algorithms depending on the feature type are developed for 2-class data, which reduces the computation time further by 39% compared to 4 combination algorithm. In case of multiclass data, one efficient algorithm which can be used regardless of the feature type, are developed. The algorithm reduces the computation time further by 70% compared to 4 combination algorithm. The efficient algorithm depending on feature type increases the CCR of multiclass data by 3.1% compared to the highest CCR in literature.

Thirdly, this research developed boosting-based feature selection methodology, which ranks features according to the distance between 2 classes among misclassified instances from the same classifier, SVM. This methodology prioritizes the feature maximizing the discriminant power on misclassified instances, using boosting method. As a result of applying one vs. all multiclass classification architecture, preprocessing method by SVM with RBF kernel, distance between classes and the wrapper method, the CCR on Cardiotocography data increases by 1.0% and the feature reduction rate is 49.2% compared to the case without feature selection. The CCR is 4.0% higher than the highest rate in recent approaches on the same data set in literature. The effectiveness of the boosting-based feature selection is supported by application to other data set.

Finally, the research contributes in the development of reliable and efficient medical decision support system to diagnose fetal health status by using Cardiotocography data. The proposed classification methodology achieves CCR 98.5%, which is 6.9% higher than the highest CCR in literature. The proposed classification methodology also achieves the sensitivity 0.983, which is 0.131 higher than current highest sensitivity in literature. Currently, The highest sensitivity among classification methodologies on 3-class Cardiotocography data in literature, is 0.852, which results in the low reliability of decision support system which diagnoses pathologic status of pregnancy as true pathologic class. In other words, 14.8% of pathologic status of pregnancy is classified as suspect or normal class by the decision support system, and additional medical examination and extra medical cost are required to diagnose the status accurately as pathologic class, which degrades the reliability and efficiency of the decision support system. In addition, this research contributes in the high adaptability of the decision support system for Cardiotocography data because current 3-class data can be used without elimination of suspect class in training of the proposed classifiers, and the Cardiotocography data in real diagnosis activity can be directly used in testing and predicting the fetal status of pregnancy until delivery.

Dissertation Overview

The rest of this dissertation is organized as follows: Chapter 1 includes a comprehensive literature review. Chapter 2 includes Methodology. Chapter 3 includes Experimental Result on 10 kinds of 2-class data and 7 multiclass data. Chapter 4 includes Experimental Results on Cardiotocography data. Finally, Chapter 5 includes Conclusion.

Chapter 1. Literature Review

This chapter provides a literature review on Feature selection, Feature extraction, Classification methodologies which consist of SVM and Ensemble method, and feature selection of Cardiotocography data.

Feature selection is the process to select features which contribute most to the prediction variable or output. Feature extraction is the process to extract new features which are informative and not redundant. Hira and Gillies (2015) reviewed feature selection and feature extraction methods which have been applied on microarray data and summarized the difference between the 2 methods. The advantage of feature selection is that it preserves data characteristics for interpretability while its disadvantages are weak discriminatory power, long training time and possibility of overfitting. The advantage of feature extraction is that it has higher discriminating power and controls overfitting when it is unsupervised while its disadvantage is the loss of data interpretability and the possibility of expensive transformation.

1.1 Feature Selection

The main objective of feature selection is to select a subset of features from the entire dataset that can provide the same information that the entire feature set can provide. There are various objectives of feature selection depending on the perspectives of

researchers as follows. The first objective is for constructing faster and more cost-effective models. Feature selection process selects minimum number of features instead of entire set of features, which reduce the execution time of the model. The second objective is to avoid overfitting and improve performance. Feature selection process can remove noisy, redundant and irrelevant features, which prevents overfitting and increases the accuracy and effectiveness of model. The final objective is to deeply understand the process that generated data. Feature selection process provides the opportunity to better understand the relationships between or among attributes, and the underlying process that generates the data. As the criteria for feature selection, information gain measures, distance measures, dependency measures and consistency measures have been used in literature. The feature selection methodologies are largely divided into 3 categories, i.e., (1) Filter method (2) Wrapper method (3) Embedded method as follows.

(1) Filter methods: Filter method is the simplest strategy for feature selection. In this method, the selection of feature is implemented independent of the learning algorithm without any feedback from the classification algorithm. Filter method use an independent evaluation criterion to measure the usefulness of features to include in a candidate feature subset. The different independent criteria reflect the types of intrinsic properties of features that can be used to evaluate the goodness of a feature set. Because the classification algorithm does not control the feature selection process, the performance and accuracy of algorithm are poor.

Variable Ranking Method: The variable ranking method a method to sort or arrange features of data according to certain criterion. This method has been widely used in

literature. Many variable selection algorithms include variable ranking method as a principal or auxiliary selection mechanism because it has advantages in its simplicity, scalability, and good empirical success. (Guyon and Elisseeff, 2003). The variable ranking method is not necessarily used in building classifiers. It is also used in the microarray analysis to discover a set of drug leads, finding genes that discriminates between healthy and disease patients. According to the classification of Kohavi and John (1997), variable ranking method is a filter method and a preprocessing step and independent of the choice of the classifier. Chang, Y.W. and Lin, C.J. (2008) used 4 kinds of feature ranking methods. Feature ranking is a method useful to gain knowledge of data and identify relevant features. The authors explored the performance of combining linear SVM with various feature selection methods. The 4 kinds of feature ranking methods used in this literature is as follows. Firstly, Fisher score is used in this paper. The authors used a variation of F-score by following Chen and Lin. (2006). Fisher score or F-score is an simple and effective criterion to measure the discrimination between a feature and the label. F-score is independent of the classifiers. Secondly, feature ranking by linear SVM weight is used in this paper. After obtaining a linear SVM model, w in the formula can be used to decide the relevance of each feature. If the w of a specific feature is large, the feature plays an important role in the decision function. SVM solves an unconstrained optimization problem. The third feature ranking method used in this paper is D-AUC (Area under curve). D-AUC is the feature ranking method from checking the change of AUC with/without removing each feature. The final feature ranking method is D-ACC (Accuracy). D-ACC is the feature ranking method from checking the change of accuracy with/without removing each feature. The last 2 performance -based methods can use any classifier. The authors

tried to select the best models among the 4 kinds of model. Although the models showed excellent performance on predictions, the used models could not provide information on the underlying causal relationship between features. Vakharia et al. (2016) used feature ranking methods in bearing fault diagnosis. Feature ranking methods such as Chisquare, ReliefF method are used to select most informative features and reduce the size of feature vector. The authors' experimental results show that the proposed method produces good fault identification accuracy with minimum number of features. Al-Salemi, B. (2018) used feature ranking method for improving performance of boosting-based multi-label text categorization. Most boosting algorithms iteratively examine all training features to generate weak hypothesis, and this is the reason for the increase of learning time. The authors developed RFBoost to manage the problem. The proposed method first ranks the training features and then, filters and used only a subset of the highest-ranked features. The authors investigated 7 feature ranking methods, i.e., information gain, chi-square, GSS-coefficient, mutual information, odds ratio, F1 score and accuracy. Experimental results show that mutual information yields the best performance for RFBoost.

mRMR (Minimum-redundancy-maximum-relevance) feature selection: Peng, H. et al. (2005) proposed a feature selection method that can use either mutual information, correlation, or distance/similarity scores to select features. The purpose of this algorithm is to penalize the feature's redundancy. The relevance of a feature set S for the class c is defined by the average value of all mutual information values between the individual feature f_i and the class c , as shown in equation (1).

$$D(S, c) = \frac{1}{|S|} \sum_{f_i \in S} I(f_i; c). \quad (1)$$

The redundancy of all features in the set S is the average value of all mutual information values between the feature f_i and the feature f_j , as shown in equation (2).

$$R(S) = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \quad (2)$$

The mRMR criterion is the combination of the two above explained measure, represented by equation (3).

$$\text{mRMR} = \max_S \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i; c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \right] \quad (3)$$

(2) Wrapper methods: Because optimal feature selection depends on the heuristics and biases of the classification algorithm, wrapper method selects the feature based on underlying classification algorithm. The feedback from the classification algorithm is used to evaluate the quality of selected features, which results in high performance of classification algorithm. Kohavi and John (1997) showed a wrapper method which searches for an optimal feature subset to a particular classification algorithm. Wrapper methods offer a simple and powerful way to address problem of feature selection regardless of the chosen machine learning algorithm. The disadvantage of wrapper method is that it is often criticized because they require massive amounts of computation, but it is not necessarily so. (Guyon and Elisseeff, 2003). SVM-RFE (Recursive feature elimination, Guyon et al., 2002) is a variation of SVM which computes the ranking weights for all features and sort the features according to weight vectors as the classification basis. SVM-RFE is a wrapper method and an iteration process of the backward feature elimination. (Huang, M.L. et al. 2014). A variation of SVM-RFE was researched by Samb et al. (2012). In their research, 2 kinds of local search methods were experimented in order to improve the performance of SVM-RFE. SVM-RFE + BF(Bit-Flip) local search and SVM-RFE +

AT(Attribute-Flip) local search were experimented. Training (50%), Testing (50%) cross-validation was applied. Wrapper methods have 2 kinds of feature elimination methods, which are forward feature elimination methods and backward feature elimination methods. Forward elimination refers to a search that begins at the empty set of features. The advantage of this method is that it is computationally inexpensive because building classifier when there are a few features in the data is much faster. Backward elimination refers to a search that begins at the full set of features. Backward elimination is computationally expensive compared to forward elimination. SVM-RFE is a kind of backward feature elimination method.

Particle Swarm Optimization (PSO): This is a computational method which optimizes a problem by iteratively trying to improve a candidate solution regarding a given measure of quality. This algorithm has a population of candidate solutions, which are particles. Each particle is affected by locally best-known position but is also guided to the best-known position in the search space. The movement of particle is updated, and the particles move to the best solution eventually. Sakri et al. (2018) used PSO in feature selection for breast cancer recurrence prediction. The authors used PSO for feature selection into 3 classifiers, naïve Bayes, K-nearest neighbor, and fast decision tree learner, with objective of increasing overall prediction accuracy. Their conclusion was that the two latter classifiers performed better with PSO compared to that without PSO. Unler and Murat (2010) developed a modified discrete PSO algorithm for the feature subset selection problem. The authors proposed an adaptive feature selection procedure which dynamically accounts for the relevance and dependence of the features to be admitted into the feature subset. This

methodology proved to be competitive in both the classification accuracy and the computational performance.

Genetic Algorithm (GA): Genetic algorithm is optimum search method is based on the evolution process of biology. This is the method inspired by natural selection process in nature. GA is widely used to generate high quality solutions to optimization and search problems by using operator such as mutation, crossover and selection. GA uses positive region-based dependency measure to calculate the fitness of each chromosome where each chromosome is made of genes which represent the presence of attributes from data set. GA-based feature selection and parameter optimization for SVM was researched by Huang and Wang (2006). The authors in this paper argued that obtaining the optimal feature subset and SVM parameters must occur simultaneously. The authors implemented GA-based approach and Grid algorithm on several real-world data sets and compared the performance in terms of prediction accuracy and computation time.

(3) Embedded methods: Embedded methods improve the disadvantage of both filter method and wrapper method. Filter methods select features regardless of classification algorithm which results in poor performance of classifier while wrapper methods are computationally expensive as the classifier needs to run many times. Embedded methods construct feature subsets as part of building a classifier, combining the advantages of both filter and wrapper methods as a result. Embedded method searches and selects the features which significantly contribute to the performance of model. Rodriguez-Galiano et al. (2018) evaluated filter, wrapper and embedded methods in their use in ranking feature importance. Machine learning algorithms such as CART, Random Forest (RF) and SVM

are used as wrappers considering 4 different sequential search approaches: (1) The sequential backward selection (SBS) (2) The sequential forward selection (SFS) (3) The sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). As a result, the authors found that RF with SFFS had the best performance and good interpretability with 3 selected features and that only wrapper method can reduce the number of features.

Other feature selection methods in literature: Trambaiolli et al (2017) compared 8 feature selection methods in classifying EEG (Electroencephalography) data, i.e., (1) Consistency-Based Filter (CBF) (2) Correlation-Based Feature Selection (CFS) (3) Filtered Subset Evaluator (FSE) (4) Chi Squared (CS) (5) Gain Ratio (GR) (6) Relief-F (7) Symmetrical Uncertainty (SU) (8) Ensemble Feature Selection (EFS), and observed that Filtered Subset Evaluator performed best among the 8 feature selection methods. Bissan Ghaddar and Naoum-Sawaya (2018) used a new approach, SVM and feature selection based on iteratively adjusting a bound on the \mathbf{l}_1 norm of the classifier vector in order to reduce the number of features. The authors applied it to high-dimensional data sets such as (1) on-line review of firms to market feedback information (2) Cancer classification based on gene expression, and found that their approach is computationally tractable in high-dimensional data.

1.2 Feature Extraction

The literatures on feature extraction methods are summarized in this section. The summarized methods are Principal component analysis, Linear discriminant analysis,

Canonical correlation analysis, Singular value decomposition, ISOMAP and Locally linear embedding.

Principal Component Analysis

Zhai, G. et al. (2015) researched new method of material identification of loose particles in sealed electronic devices by utilizing PCA and SVM. The existing method mainly depends on time, frequency and wavelet domain features. The conventional method has a disadvantage in that the selected features are often overlapped and redundant, resulting in unsatisfactory material identification accuracy. This purpose of this paper is to improve the accuracy of material identification by using PCA and SVM. PCA was used to extract less correlated features from time and frequency domain. The experimental result shows that the new method can identify the material more effectively than previous method. Gao, X. et al. (2016) developed an improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastern process. In modern production systems, fault detection and process supervision are very important issue. The authors used PCA in reducing feature dimension on preprocessed data. In addition, the authors used a multi-class SVM by optimizing parameters with grid search (GS) method. The authors argue that GS generates comparable classification accuracy to genetic algorithm (GA) or particle swarm optimization (PSO) while being more efficient than the 2 other methods. Finally, the authors proved the effectiveness of the proposed SVM integrated GS-PCA fault diagnosis approach by comparing it with other related fault diagnosis methods.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method to reduce a data from the viewpoint of optimal classification. LDA reduces the dimension of eigen vector of data by maximizing the ratio of between-class scatter to within-class scatter after supervised learning. Safo, S. and Ahn, J. (2016) used linear discriminant analysis in multiclass data. The authors developed an improved approach to apply linear discriminant analysis to multi-class problems by reducing heavy computational cost. Silva, A et al. (2016) used 2-dimensional linear discriminant analysis for classification of 3-way chemical data. The 2-dimensional linear discriminant analysis (2D-LDA) algorithm was originally proposed to be used in face image processing for the extraction of features with maximal discriminant power. The authors used it for chemical data for the first time. The experimental results show that 2D-LDA produces better classification rate than other methods with no feature extraction. Uncini, A. et al. (2017) optimized electrodiagnostic accuracy by using sparse LDA.

The difference between PCA and LDA is explained as follows. PCA is the method to analyze the trend of all elements in a data regardless of class in case that multiple class exists in a data set. PCA searches for the axis on which all the elements can be spread most broadly. On the other hand, LDA searches for the axis on which the within-class scatter is minimized and the between-class scatter is maximized.

Canonical Correlation Analysis

Canonical-correlation analysis (CCA) is a way of inferring information from cross-covariance matrices. Wang, Z. et al. (2007) developed an improved Multiple Kernel Learning (MKL) Algorithm, which is MultiK-MHKS (Modification of Ho-Kashyap algorithm with Squared approximation of the misclassification error) by using Canonical

correlation analysis (CCA). CCA can maximally correlates the m views in the transformed coordinates. By using m kernels, the original input data can be mapped into m feature spaces, where each feature space can be taken as one view of the original input data. 2 kinds of combinations, i.e., CCA + MHKS₁ and CCA + MHKS₂, are used. The former is combined in the feature level and the latter is combined in the decision level. In the 2 combinations, all the candidate kernels (linear, RBF, polynomial) are used. As a result of experiment on various data, the authors proved that MultiK-MHKS algorithm produces a comparable or superior performance in classification rate and has the competitive efficiency in computation. Shen, C. et al (2014) derived conditions under which Generalized Canonical Correlation Analysis improves classification performance of the projected datasets, compared to standard Canonical Correlation Analysis using only 2 data sets.

Singular Value Decomposition

Singular-value decomposition (SVD) is a factorization of a real or complex matrix. It is the generalization of the eigen decomposition of a positive semidefinite normal matrix to any $m \times n$ matrix via an extension of the polar decomposition. SVD has been widely applied in signal processing and statistics. Tanchotsrinon, W. et al. (2015) used SVD in predicting Human Papillomavirus (HPV) genotypes. HPV is a small double-stranded and most common sexually transmitting DNA virus. The authors of the research used ChaosCentroid, ChaosFrequency and SVD simultaneously in feature extraction of HPV genotypes. The authors transformed HPV genomes with 7,000 – 10,1000 base pairs onto features of 1 -11 dimensions, proving the effectiveness of their proposed methods as a feature extraction technique for predicting the HPV genotypes. Yan, M. (2013) developed

dual adaptive K-SVD algorithm based on a rank symmetrical relationship. The K-SVD algorithm is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data. Experimental results conducted on the ORL and Yale face databases and the results show the effectiveness of the proposed method by the authors.

ISOMAP

Isomap is a nonlinear dimensionality reduction method and is also a few widely used low-dimensional embedding methods. Isomap provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbors on the manifold. This algorithm is highly efficient and generally applicable to a broad range of data sources and dimensionalities. Park, H. (2012) used Isomap induced manifold embedding, which is a more advanced manifold learning technique, in distinguishing Alzheimer disease. The authors felt the necessity of analyzing neuroimaging data, which is high dimensional and cumbersome to analyze. The existing work utilized another learning method, multidimensional scaling (MDS) to shape information for distinguishing Alzheimer's disease from normal. The author's proposed method is effective in dimensional reduction of high dimensional data and can be used in time series data in the situation that the result of MRI scan is time-series data. Bu, Y. et al. (2014) used Isomap and SVM in stellar spectral subclass classification. The authors used Isomap to extract the features within stellar spectra and found that Isomap is more efficient than PCA in extracting features from the original data. The authors compared the performance of Isomap-based SVM with traditional PCA-based SVM with default γ in SVM. The comparison result shows that the performance of Isomap-based SVM is better.

Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) method is based on simple geometric intuitions. If a data is sampled from a smooth manifold, the neighbors of each point remain nearby and are similarly co-located in the low dimensional space. In LLE method, each point in the data set is linearly embedded into a locally linear patch of the manifold. As a result, low-dimensional data is created in which the locally linear relations of the original data are preserved. Liu, X. et al. (2013) used LLE for MRI based Alzheimer's disease classification. The authors used the unsupervised learning algorithm of LLE to transform multivariate MRI data of regional brain volume and cortical thickness to a locally linear space with fewer dimensions, while also utilizing the global nonlinear data structure. After experiment, the authors found that classification using embedded MRI features outperformed the classifications using the original features directly. In addition, they noticed that the improvements from LLE were obtained from all tested classifiers, i.e., regularized logistic regressions, SVM and linear discriminant analysis. Coy, B. (2012) used LLE in dimensional reduction for analysis of unstable periodic orbits (UPO) in physics. The author analyzed a 4-dimensional dynamic system that generates a strange attractor. The analysis was dependent on the UPOs that were found by the method of close returns in the 4-dimensional phase space. Surrogate UPOs were found in the 4-dimensional phase space and pairs of these orbits were embedded in 3-dimensions using LLE. As a result, a table of linking numbers was computed for a range of control parameters values which shows that the organization of the UPOs is consistent with that of a Lorenz-type branched manifold with rotation symmetry.

1.3 Classification

1.3.1 Support Vector Machine

Staelin, C. (2003) developed an algorithm for selecting support vector machine (SVM) meta-parameter values which is based on ideas from design of experiments (DOE). The author compared the performance of the proposed method with the standard grid search method using LIBSVM and the RBF kernel, both in terms of the quality of the final result and the work required to obtain the result. The grid search was done with 20 samples per parameter with uniform resolution in \log_2 – space. The proposed method utilized five iterations and less than 50 samples to find the C and γ values that gave the best cross-validation error while the grid search required 20 samples per dimension, for 400 samples overall, an eight-fold difference in computational cost. The comparison of performance is as follows in Table 10. The accuracies between the 2 methods are almost equivalent, however, the significantly less samples were used for the proposed method. The authors developed an algorithm based on techniques from DOE that can reliably find very good parameter settings for SVMs with RBF kernel with relatively little efforts across a wide range of machine learning problems. Lebrun, G et al. (2004) developed a new method for training SVM on large data sets by applying Vector Quantization (VQ). VQ is method to construct example prototypes and then generate smaller training data set. The authors showed that their proposed method finds the optimal solution faster than the classical grid search. In addition, VQ finds a decision function with reduced complexity when the data set includes noisy or error examples. Table 8 shows the classification rates and training times using the author’s method and a classic grid search. The proposed method in this paper significantly reduces the training time on huge databases while the classification rate

is almost the same. The authors evidenced that the training time is reduced by 108 times in case of shuttle data set whose instances are 58,000, however, the training time is reduced by 30 times in case of Satimage data set whose instances are 6,435. Huang and Wang (2006) researched GA-based feature selection and parameter optimization for SVM. The authors in this paper argues that obtaining the optimal feature subset and SVM parameters must occur simultaneously and that GA has the potential to generate both the optimal feature subset and SVM parameters at the same time. The goal of the authors' research is to realize it without degrading SVM classification accuracy. The authors implemented GA-based approach and Grid algorithm on several real-world data sets and compared the performance in terms of prediction accuracy and computation time. The authors proved that their proposed GA-based approach significantly improves the classification accuracy and has fewer input features for SVM compared to Grid algorithm. In all of the data sets used in the experiment, the prediction accuracy of GA-based approach is higher than that from Grid algorithm. The authors argue that the proposed GA-based approach significantly improves the classification accuracy and has fewer input features for SVM even if the average running time of GA-based approach is slightly inferior to that of the Grid algorithm. Wang, Z. et al. (2007) developed an improved Multiple Kernel Learning (MKL) Algorithm, which is MultiK-MHKS (Modification of Ho-Kashyap algorithm with Squared approximation of the misclassification error) by using Canonical correlation analysis (CCA). CCA can maximally correlates the m views in the transformed coordinates. By using m kernels, the original input data can be mapped into m feature spaces, where each feature space can be taken as one view of the original input data. 2 kinds of combinations, i.e., CCA + MHKS₁ and CCA + MHKS₂, are used. The former is combined in the feature

level and the latter is combined in the decision level. In the 2 combinations, all the candidate kernels (linear, RBF, polynomial) are used. As a result of experiment on various data, the authors proved that MultiK-MHKS algorithm produces a comparable or superior performance in classification rate and has the competitive efficiency in computation. Chang, Y.W. et al. (2008) used feature ranking by using linear SVM. Feature ranking is a method useful to gain knowledge of data and identify relevant features. The authors explored the performance of combining linear SVM with various feature selection methods. The 4 kinds of feature ranking methods used in this literature is as follows. Firstly, Fisher score is used in this paper. Fisher score or F-score is a simple and effective criterion to measure the discrimination between a feature and the label. F-score is independent of the classifiers. Secondly, feature ranking by linear SVM weight is used in this paper. After obtaining a linear SVM model, w in the below formula can be used to decide the relevance of each feature. If the w of a specific feature is large, the feature plays an important role in the decision function. SVM solves the following unconstrained optimization problem in equation (4).

$$\min_{w,b} \quad \frac{1}{2}w^T w + C \sum_{i=1}^l \xi(w, b; x_i, y_i) \quad (4)$$

The third feature ranking method used in this paper is D-AUC (Area under curve). D-AUC is the feature ranking method from checking the change of AUC with/without removing each feature. The final feature ranking method is D-ACC (Accuracy). D-ACC is the feature ranking method from checking the change of accuracy with/without removing each feature. The last 2 performance -based methods can use any classifier. The authors tried to select the best models among the 4 kinds of model. Although the models showed excellent performance on predictions, the used models could not provide information on

the underlying causal relationship between features. Maldonado, S. et al. (2009) developed HO (Hold-out)-SVM, a wrapper method for feature selection using SVM and a sequential backward selection. The authors compared the result with filter method or SVM-RFE. The proposed method outperformed other filter and wrapper methods in terms of the classification rate, the ability to adjust better to a data and the ability to avoid overfitting. The method can be used with any kernel function and can be easily generalized to variations of SVM, such as Support Vector Regression and multi-class SVM. Li, S. et al. (2010) came up with a strategy to combine a comprehensive learning particle swarm optimizer (CLPSO) with Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for effectively tuning the SVM parameters. The proposed method identifies multiple local optima of generalization bounds rather than locating a single local optimum, which can significantly improve the stability of parameter decision. Grid search is a time-consuming method when dealing with multiple parameters while the numerical methods are very sensitive to the initial value of parameters. The proposed method in this paper overcome the disadvantages of the 2 methods. The experimental results in Table 18 show that the proposed method can tune the parameters of 2 kinds of experimented SVM, i.e., L1-SVM and L2-SVM effectively and produces competitive performance compared to other optimized classifiers. In addition, the proposed method in this paper produces more stable performance than the gradient methods and is less computationally expensive than the grid search method. Li and Sun (2011) researched a straightforward wrapper method to predict business failure by using SVM. The authors utilized a straightforward wrapper approach to make the model produce more accurate prediction. The wrapper approach employed a forward feature selection method, which is composed of feature ranking method and

feature selection. The authors also used linear SVM to select features for all SVMs in the wrapper because non-linear SVMs have the possibility of overfitting. In addition, the authors used a robust re-sampling approach to evaluate model performances for the task of business failure prediction in China. As the experimental result, the non-linear SVM with radial basis function kernel and features selected by linear SVM are significantly better than all the other SVMs. Samb, M.L. et al. (2012) developed a novel SVM-RFE-based feature selection method for classification problems. The authors overcame the limitation of SVM-RFE and proposed improved SVM-RFE algorithm. Even if SVM-RFE is one of the most effective methods, it is a greedy search method that only hopes to find the best possible combination for classification. The authors combined the SVM-RFE algorithm with local search operators based on operation research and artificial intelligence. 2 kinds of local search methods, i.e., SVM-RFE+BF (Bit-Flip) and SVM-RFE+AT(Attribute-Flip) were researched and experimented in the paper. Out of the 3 data used in the experiment, SVM-RFE+AT shows the highest prediction accuracy in 2 data and SVM-RFE+BF shows the highest accuracy in 1 data. The authors concluded that the reuse of features previously removed during SVM-RFE process improves the quality of the final classifier. Bhavsar, H. et al. (2012) surveyed and summarized the variations of SVM in terms of their performance including computational complexity. The variations of SVM have been researched in literature to increase speed efficiency, space efficiency and ability to classify multiclass data. The authors categorized the variations of SVM into 3 groups. The first group is decomposition-based algorithms. The memory requirement of SVM grows with the squares of number of training algorithms. Decomposition based methods divide a large optimization problem into a series of smaller problems in which each problem only

involves a couple of selected variables in order to make the optimization process more efficient. The variations of SVM in the first group is time-consuming because it considers only memory issue. The 2nd group is variant based algorithms. This category of variations of SVM reduces the training time at the price of accuracy. The final group is multiclass-based algorithms. Originally, SVM was developed to perform binary classification. The variations of SVM have been researched to solve multiclass problems because most of the classification problems involve multiclass data. Amami, R. et al. (2013) proposed a method to find the suitable kernel with which SVM may achieve good generalization performance. In addition, the authors analyzed the performance of SVM classifier when the parameters take very small or very large values. The authors used 2 kinds of feature extraction methods which are MFCC (Mel-frequency cepstral coefficients) and PLP (Perceptual Linear Prediction). The authors also used 2 different methods to research the suitable number of frames to utilize, which are Middle frames and Fuzzy c-means clustering (FCM). Finally, the authors implemented a comparative study to analyze the impact of the choice of the parameters, kernel tricks and feature representations on the performance of the SVM classifier. The authors found that SVM with RBF kernel and 36-dimensional MFSS features perform best among other combinations of the different kernels and features. As the kernel width parameters and the penalty parameters tends to be smaller, the accuracy and the runtime improve. However, the authors could not conclude on which kernel is optimal for a given learning task, leaving it as a future research issue. Wang, Z. et al. (2014) researched a method to reduce both the time and space complexity of Multiple Kernel Learning (MKL) and proposed an efficient MKL classification machine based on the Nystrom approximation. The proposed improved method is Nystrom approximation matrix

with Multiple KMHKSs (NMKMJKS). According to the classification performance comparisons on various data sets, combining several kernels and applying Nystrom approximation matrix technique can bring a better recognition rate. Huang, M.L. et al. (2014) researched SVM-RFE based feature selection and Taguchi parameters optimization for multiclass classifier. This research combined feature selection and SVM-RFE to investigate the classification accuracy of multiclass problems. In addition, Taguchi method was combined with SVM classifier to optimize C and γ to increase classification accuracy for multiclass classification problems. The experiments result shows that the prediction accuracy can be more than 95% for the 2 multiclass data. Singla, A. et al. (2014) researched a novel classification methodology based on progressive transductive SVM (PTSVM) learning. PTSVM is a semisupervised version of SVM, which takes in account both unlabeled data and labeled data in building classifier. The existing PTSVM techniques select the transductive samples by exploiting only the properties of the SVM classifier while it does not consider the low-density region of the feature space as well as poor initial training set in the definition of the criterion for selecting wrong transductive samples. For this reason, the PTSVM's probability of selecting wrong transductive patterns is high and the classification performance can be poor. The proposed technique in this paper exploits a k-nn technique and the cluster assumption for selecting accurate transductive samples as well as the properties of the SVM classifier. The authors evaluated the effectiveness of the proposed method by comparing the performance with other PTSVM based approaches in the literature and observed that the proposed method provides better accuracy compared to the existing techniques on the same data sets. Chen, G. and Chen, J. (2015) introduced a wrapper method, cosine similarity measure support vector machines (CSMSVM). The

proposed method eliminates irrelevant or redundant features during classifier construction by introducing cosine distance into SVM. The cosine distance has the advantage over the Euler distance or other distance on increasing the probability of correct classification. In addition, the proposed feature selection method in this paper is a kind of wrapper method. The authors compared the novel method with well-known feature selection techniques with experiments and proved that CSMSVM outperformed the other methodologies in improving the pattern recognition accuracy with fewer features. Zhai, G. et al. (2015) researched new method of material identification of loose particles in sealed electronic devices by utilizing PCA and SVM. The existing method mainly depends on time, frequency and wavelet domain features. The conventional method has a disadvantage in that the selected features are often overlapped and redundant, resulting in unsatisfactory material identification accuracy. This purpose of this paper is to improve the accuracy of material identification by using PCA and SVM. PCA was used to extract less correlated features from time and frequency domain. The experimental result shows that the new method can identify the material more effectively than previous method. Abdiansah, A. et al. (2015) implemented time complexity analysis of SVM in LibSVM. Basically, the complexity is divided into 2 kinds, i.e., time complexity and space complexity. The authors focused only on time complexity, which deal with how long the algorithm is executed. In addition, the authors selected SVM in their research because SVM provides a global solution for data classification. Lin et al. developed LibSVM to facilitate the researchers' use of SVM. This research computed the complexity of SVM algorithm by using C++ and JAVA and compared the result of performance on 3 different data sets. The results show that the complexity of SVM (LibSVM) is $O(n^3)$ and that the time complexity of C++ is

lower than that from JAVA. The growth of data affected and increased the total computation time. Gao, X. et al. (2016) developed an improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastern process. In modern production systems, fault detection and process supervision are very important issue. The authors used PCA in reducing feature dimension on preprocessed data. In addition, the authors used a multi-class SVM by optimizing parameters with grid search (GS) method. The authors argue that GS generates comparable classification accuracy to genetic algorithm (GA) or particle swarm optimization (PSO) while being more efficient than the 2 other methods. Finally, the authors proved the effectiveness of the proposed SVM integrated GS-PCA fault diagnosis approach by comparing it with other related fault diagnosis methods. Martins, S. et al. (2016) researched the optimal parameterization of SVM for change detection mapping in Funil hydroelectric reservoir in Brazil. The main goal of the research was to test the performance of polynomial function and RBF, and to identify which input parameters combinations are the best to use in SVM algorithm for the reservoir. Finally, the authors found that the RBF kernel is the best SVM's kernel function to be used in classifying the time-series images. Huang, M.W. et al. (2017) researched SVM ensemble methods in breast cancer prediction with a motivation that there had been very few studies focused on examining the prediction performances of SVM based on different kernel functions. In addition, the authors evaluated the performance of SVM and SVM ensembles over small and large-scale breast cancer data sets. Finally, they found that linear kernel based SVM ensemble with bagging method and RBF kernel based SVM ensemble with boosting method performed better in case of small-scale data set while RBF kernel based SVM with boosting method performed better in case of large-scale data set. Lee, S.B. et

al. (2017) researched parameter search methodology of SVM for improving performance. The authors in this paper proposed a search method that explores parameters C and σ of SVM to reduce computation time while maintaining high prediction accuracy. A traditional grid search method requires significant computation time because it searches for all available combinations of C and σ values to find optimal combinations of parameters which produces the best performance. This paper proposes a deep search method that reduces computation time. The experimental results show that the proposed deep search algorithm outperforms the conventional algorithms in terms of performance and search time.

Ghaddar and Naoum-Sawaya (2018) researched a methodology for feature selection based on iteratively adjusting bound in large size of data. The large-scale data is costly for collection, storage, and processing. To overcome this limitation, the authors searched for minimal number of features for SVM binary classifier by using 2 data sets, i.e., the medical diagnosis of tumor on microarray data and sentiment classification of on-line reviews from Amazon, Yelp, and IMDb. The used classifiers are Joint feature selection SVM, SVM-C2, SVM-CR1 and SVM-RFE. The experimental results show that the proposed classification and feature selection approach is simple, computationally less expensive and produces high classification accuracy.

Lin, X. et al. (2018) proposed a method, SVM-RFE-OA, which combines the classification accuracy rate and the average overlapping ratio of the samples to decide the number of features to be selected from the feature rank of SVM-RFE. This paper proposes 2 techniques of selecting discriminative feature subsets based on SVM-RFE. The proposed method eliminates at most one-third of the samples in each class to make sure enough samples kept for the training and protect heavy overlapping area in the features to make the calculation of the feature weights more stable and accurate. The

authors' experimental results on 8 biological data sets show the validation of these techniques.

1.3.2 Ensemble Method

In literature, ensemble methods have been used in classification (Huang et al. 2017, Pujari & Gupta 2012, Zhang & Yang 2008). The advantage of ensemble method is that it can result in prediction accuracy with higher reliability compared to the method which uses a single model. There are several kinds of ensemble models, e.g., bagging, boosting and stacking. Bagging is abbreviation of Bootstrap Aggregation, which is an algorithm of training identical model by restored random sampling data and aggregating the predicted variables. The purpose of bagging is to increase the stability and accuracy of algorithm. The typical errors coming from most of training is either underfitting due to high bias or overfitting due to high variance. Ensemble method minimizes these kinds of errors. Especially, bagging avoid overfitting by selecting the median value from the results obtained from each model. Generally, bagging method aggregates the results by voting in case of categorical data, and aggregates by averaging the results in case of continuous data. A typical bagging algorithm is Random Forest. Originally, the boundary of decision tree model is in discrete-shape, however, Random Forest could overcome this by combining multiple decision tree models. Boosting focuses on solving difficult problems while bagging focuses on building general model. Booting method assign weighted value on the model which solved difficult problem and select the model as the final model. Boosting also implements restored random sampling, which is the same with Bagging. However, the difference from bagging method is that boosting assign weighted value on the model which

solved the difficult problem. Bagging implement training in parallel while boosting implement training sequentially. After training, boosting distributes the weighed value on models again according to the results. Boosting focuses on predicting difficult problem because more weighted value is added on difficult problem while less weighted value is added on easy problem. Boosting produces high correct classification rate, however, has a disadvantage that it is weak at outlier. There are various kinds of boosting model such as AdaBoost, XGBoost, GradientBoost. Especially XGBoost shows highest performance. Stacking or Meta modelling is a little different from the above two methods. Stacking creates model which produces the best performance by combining different models. The models to be used in stacking can be various algorithms such as SVM, Random Forest, KNN etc. Stacking method can take the advantages of each models and complement disadvantages of each models. The disadvantage of stacking method is that it is computationally expensive. In literature, there are papers which used ensemble model in classification. An ensemble of classifiers (ANN, Decision tree & SVM) with genetic algorithm (GA) in feature selection was researched (Zhang and Yang, 2008). In the authors' research, each of the classifier assesses the data and features with their own strategies. Multi-objective GA is employed to balance their assessments and facilitate their diversity. They showed that the GA ensemble model outperformed other algorithms in comparison and found to be the best method for classification. Pujari & Gupta (2012) used ensemble model by combining 3 classifiers, i.e., CART (Classification and Regression Tree), CHAID (Chi-squared Automatic Interaction Detection) and QUEST (Quick, Unbiased, Efficient Statistical Tree) by confidential-weighted voting scheme. The authors evaluated the performance of the 3 different classifiers and its ensemble model by applying them on

Ionosphere data set. The performance of all classifier was investigated by using statistical measures like accuracy, specificity and sensitivity. Gain chart and R.O.C. (Receiver operating characteristics chart) are also used for measuring performances. As a result, the ensemble model with feature selection has achieved a remarkable performance with highest accuracy of 93.84% on the data set. Huang et al. (2017) used SVM in breast cancer prediction in literature. Breast cancer prediction has long been regarded as an important research problem in medical and healthcare fields. However, there have been very few researches focused on evaluating the prediction performances of SVM based on different kernel functions. In addition, there have been also few researches on whether SVM ensemble outperform SVM with single kernel. The authors in this paper examined the prediction performance of SVM based on different kernel functions and different size of data. The classification accuracy, ROC, F-measure, and computation times of training SVM and SVM ensembles are measured and evaluated by comparison. After implementing experiment, they found that linear kernel based SVM ensemble with bagging method and RBF kernel based SVM ensemble with boosting method performed better in case of small-scale data set while RBF kernel based SVM with boosting method performed better in case of large-scale data set. Wang, Z. et al. (2007) developed an improved Multiple Kernel Learning (MKL) Algorithm, which is MultiK-MHKS (Modification of Ho-Kashyap algorithm with Squared approximation of the misclassification error) by using Canonical correlation analysis (CCA). CCA can maximally correlates the m views in the transformed coordinates. By using m kernels, the original input data can be mapped into m feature spaces, where each feature space can be taken as one view of the original input data. 2 kinds of combinations, i.e., $CCA + MHKS_1$ and $CCA + MHKS_2$, are used. The former is

combined in the feature level and the latter is combined in the decision level. In the 2 combinations, all the candidate kernels (linear, RBF, polynomial) are used. As a result of experiment on various data, the authors proved that MultiK-MHKS algorithm produces a comparable or superior performance in classification rate and has the competitive efficiency in computation. The same author and other researchers, Wang, Z. et al. (2014) researched a method to reduce both the time and space complexity of Multiple Kernel Learning (MKL) and proposed an efficient MKL classification machine based on the Nystrom approximation. The proposed improved method is Nystrom approximation matrix with Multiple KMHKSs (NMKMJKS). According to the classification performance comparisons on various data sets, combining several kernels and applying Nystrom approximation matrix technique can bring a better recognition rate.

1.4 Feature Selection of Cardiotocography Data

Several papers in literature used SVM in classification problem on Cardiotocography data as follows. First of all, Yilmaz and Kilikcier (2013) used Least Squares (LS)-SVM with Particle Swarm Optimization (PSO) and Binary Decision Tree (BDT) in determining fetal state from the data. The authors used PSO in parameter optimization of LS-SVM. LS-SVM is a modified version of SVM and its computational load is much less than SVM. LS-SVM solves the problem by use of a set of linear equations while SVM uses a quadratic programming in solving problem. Because SVM is basically a binary classifier, the authors used BDT to extend the binary SVMs to multiclass classification. The authors' experimental results demonstrated that their proposed method resulted in the correct classification rate of 91.62%.

Secondly, Chamidah and Wasito (2015) used hybrid K-means clustering-based feature extraction and SVM in classifying fetal state from the data. The author obtained the correct classification rate of 90.64%. The author extracted features from the 3 classes, which are normal, suspect and pathologic, by using K-means clustering and found out hidden patterns for each class. Afterward, the author reconstructed features by using membership function. By using the resulting pattern from the membership function, the author obtained 7 reduced features. Finally, the author trained the SVM by using training data and tested by using testing data, based on the reduced features. The methodology of hybrid K-means clustering-based feature extraction and SVM was also used in classifying breast cancer data by Zheng, B et al, 2014. The authors reduced computation time while maintaining the highest accuracy in literature.

1.5 Summary

The literature related to the Feature selection, Feature extraction, Support vector machine, Ensemble method and feature selection methodology for Cardiotocography data were reviewed in Chapter 1. Various kinds of feature selection methods were reviewed. The variations of SVM in literature were summarized. The literature related to the detail methodologies or subject in the proposed algorithms in this research, i.e., Variable ranking method, Ensemble method and Computational complexity, were also summarized.

As a result of the literature review, the gaps in the literature were detected and summarized in Table 1 with the implemented methodologies in this research. Firstly, in literature, Feature accuracy ranking method and PCA has not been used simultaneously

and complementarily. In this research, the 2 preprocessing methods were used complementarily and the advantages of the 2 methods were merged. Secondly, no literature has used any method to reduce instances of training data to reduce computation time for grid searching parameters of SVM. This research reduced the instances of training data from the conviction that doing so is critical in reducing total computation time and searched for the reduced size of training data. Thirdly, searching for efficient algorithms depending on the features types was not used in literature. This research developed feature-oriented algorithms to reduce the computation time further. Fourthly, various options regarding the choice between the correct classification rate and computation time were not provided in the literature. The correct classification rate and computation time are in trade-off condition in most of machine learning algorithm. This research provided various algorithms with different correct classification rate and computation time to satisfy the different requirements of the users in various application fields in the future.

As a result of the literature review related to the classification methodology of Cardiotocography data, the gaps in the literature were added. Fifthly, boosted feature selection methodology by using wrapper method based on feature ranking according to the largest distance between classes among misclassified instances, has not been used in literature. This research used boosted feature selection methodology and proved the effectiveness by applying to other data and classifiers. Finally, Class-dedicated SVMs have not been applied to the classification of 3-class Cardiotocography data. This research proposed class-dedicated classification architecture to increase the performance of classification methodology for Cardiotocography data.

Table 1. Gaps in literature and methodologies in this research

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.
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
5	Chang, Y.W. et al. (2008) Maldonado, S. et al. (2009) Ocak, H (2012) Ocak, H (2013) Chamidah & Wasito (2015) Yilmaz & Kilicier (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 & Kilicier (2013)	Class-dedicated SVMs have not been applied to the classification of 3-class Cardiotocography data.	This research developed class-dedicated architecture to increase the performance of classification methodology for Cardiotocography data.

Chapter 2. Methodology

2.1 Research Framework

This chapter provides a detailed explanation on the methodologies employed in this research. This research largely consists of 4 phases as shown in Figure 2. In the 1st phase, this research developed feature selection and extraction methodology and evaluated the performance. In the 2nd phase, this research compared the performance of various feature ranking criteria. In the 3rd phase, this research developed the boosted feature selection method and validated the performance. In the final phase, this research developed classification methodology for 3-class Cardiotocography data set.

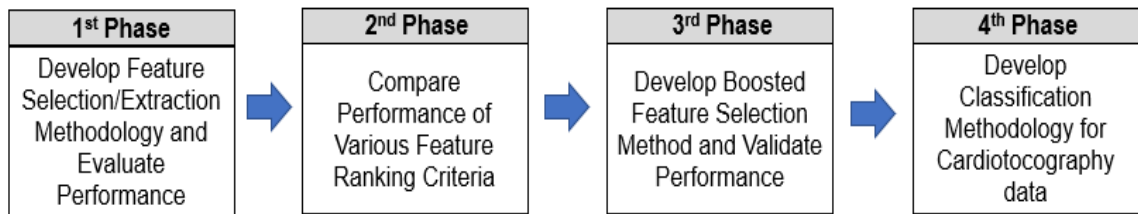


Figure 2. The 4 phases of the research flow

The detail research flow per each phase is described from 2.1.1.

2.1.1 Developing Feature Selection / Extraction Methodology

The 4 kinds of Rank-PCA ensemble algorithms

This research used ensemble method by combining 2 methods, i.e., feature ranking method and PCA, and 4 kinds of kernel in SVM, i.e., polynomial, sigmoid, radial basis function (RBF) and linear SVM. Firstly, the feature selection and extraction methods are combined by ensemble to utilize the characteristics of original features and newly extracted features from PCA simultaneously. The feature ranking method sorts individual feature in descending order, based on the CCR from SVM. Multicollinearity is the correlation between or among features, which prevents classification model from increasing CCR further. PCA is effective in increasing CCR since it reduces dimension and eliminates the multicollinearity simultaneously. In summary, the ensemble method which consists of the feature selection and extraction method results in achieving highest CCR because the highest CCR is expected to come from either feature ranking method or PCA. Secondly, 4 kernels in SVM are combined by ensemble to use the advantage of each kernel. Each kernel used in SVM has different characteristics in terms of the performance criteria, i.e., CCR, computation time and feature (dimensional) reduction rate. If the 4 kernels are combined with the feature ranking method and PCA, totally 8 kinds of combination are expected to produce different performance in terms of the 3 performance criteria.

At first, 8-combination ensemble algorithm was implemented. The performance is analyzed and used in developing more efficient algorithms. The average CCR and average computation time of all experimented data are summarized. In case of RBF kernel, which is the most time-consuming in grid search of parameters among all kernels, the reduced size of training data for grid search is determined and applied to develop more efficient algorithms. Figure 3 is the research flow of developing 4 kinds of ensemble algorithms.

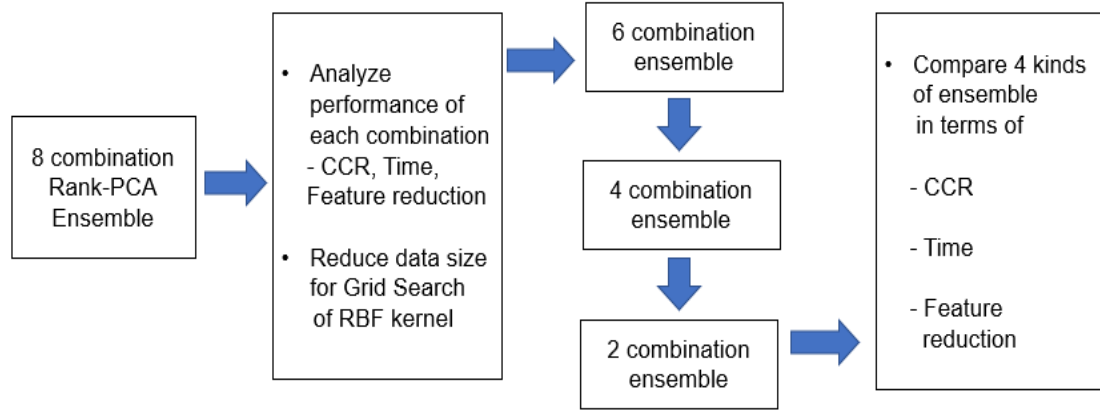


Figure 3. The research flow developing 4 kinds of ensemble algorithms

Totally 4 kinds of ensemble method were experimented as shown in Table 28.

Table 2. The component in the combinations of the 4 kinds of Rank-PCA ensemble algorithms

Feature selection or extraction	Feature ranking				PCA			
	Polynomial	Sigmoid	Radial	Linear	Polynomial	Sigmoid	Radial	Linear
8 Combinations	○	○	○	○	○	○	○	○
6 Combinations	○	○		○	○	○		○
4 Combinations	○		○	○			○	
2 Combinations				○				○

Efficient algorithms depending on feature type

For both 2 class data sets and multiclass data sets, the types of feature of all data sets are investigated, and the features are divided into either metric or categorial to search for efficient algorithm depending on the ratio of metric features to the total features. The criteria for dividing features into the 2 categories are as summarized in Table 3. If a feature is a continuous value such as 0.3862, the feature is a metric value definitely. If a feature is

symbols such as 1, 2, 3, or A, B, C representing a certain category, the feature is a categorical value. If a feature is integers representing a degree of a measure, the feature is also a metric value because the integers have the property of metric values and can be regarded as discretized values of continuous value.

Table 3. The criteria for categorizing features depending on 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)

After the analysis of features, efficient algorithms depending on feature type are developed. The most efficient algorithms with the same CCR as the 4-combination ensemble algorithm, are searched for both 2-class data and multiclass data. Finally, the performance of the efficient algorithms depending on feature type is compared to other recent approaches on the same data in literature to compare the performance in terms of CCR and computation time. Figure 4. is the research flow of developing efficient algorithms depending on feature type.

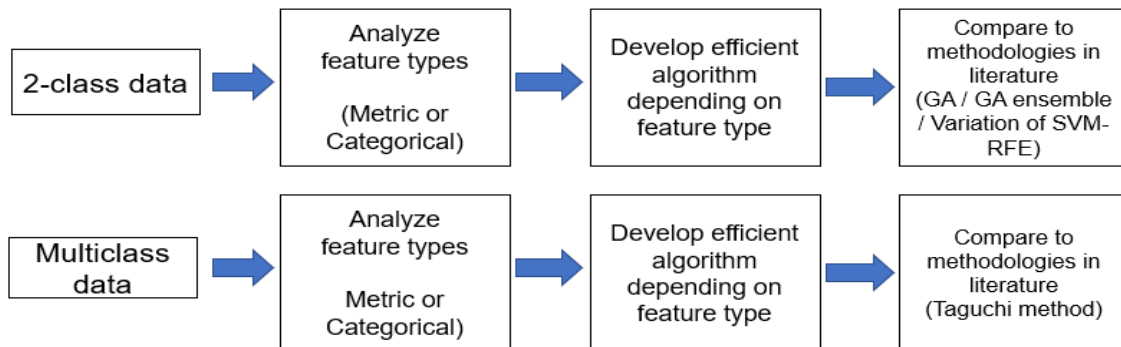


Figure 4. The research flow of developing efficient algorithms depending on feature type

2.1.2 Comparing Performance of Various Feature Ranking Criteria

In the 2nd phase, the performances of Linear discriminant analysis (LDA) and the distance between classes as a feature ranking method, are evaluated. Firstly, LDA is performed on all the individual features of the prepared data, i.e., 10 kinds of 2-class data sets and 7 kinds of multiclass data by use of LDA function in R. The prediction by use of LDA function in R finally shows classification matrix in which the number of instances of correct classification and wrong classification per class are shown. Based on the results, the CCRs of all individual features are calculated. Then, the features are rearranged in the order of high CCR from LDA. Secondly, the distance between classes is calculated on all the individual features of the prepared data. Then, the features are rearranged in the order of long distance. Finally, the conditions of characteristics of data on which each feature ranking criteria are the most effective, is analyzed. Figure 5. is the research flow of the 2nd phase, Comparing performance of various feature ranking criteria.

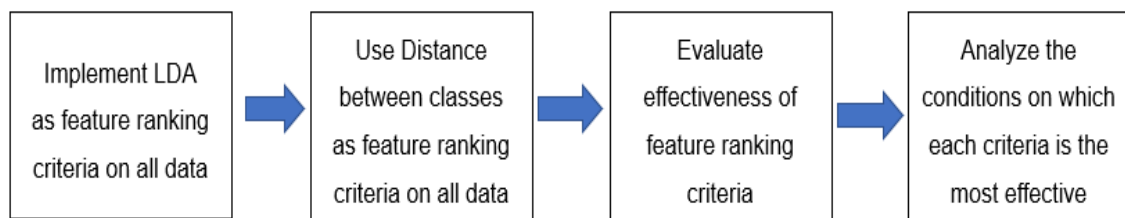


Figure 5. The research flow of the 2nd phase, Comparing performance of various feature ranking criteria.

2.1.3 Developing Boosted Feature Selection Method

In the 3rd phase, this research developed the boosted feature selection method and validated its performance.

Firstly, this research applied this method to the 3-class Cardiocography data. Secondly, the method is applied to the 3-class Contraceptive data, which has different number of instances and features. Then, this research validated the effectiveness of the boosted feature selection method regardless of classifiers. The method is applied to ensemble algorithms, AdaBoost and Random Forest by using decision tree as the classifier for misclassified instances in preprocessing stage. Finally, the experimental results from this method are analyzed and evaluated. Figure 6. is the research flow of the 3rd phase, Developing boosted feature selection method and validating performance.



Figure 6. The research flow of the 3rd phase, Developing boosted feature selection method and validating performance.

2.1.4 Developing Classification Methodology for Cardiocography Data

In the final phase, this research developed classification methodology for Cardiocography data and validated the effectiveness. Firstly, the 3-class Cardiocography data is converted into 3 binary class problems, and boosted feature selection is implemented per each of binary classification problems. Secondly, K-means clustering algorithm is implemented on the selected features by adjusting the number of clusters, and new features are extracted by fuzzy membership function. Thirdly, the performance of the proposed methodology is compared to the highest performance in

literature in terms of all criteria, CCR, sensitivity and specificity. Finally, the methodology is applied to other multiclass data to validate the effectiveness. The Figure 7 is the research flow of the final phase, Developing classification methodology for Cardiotocography data and validating the effectiveness.

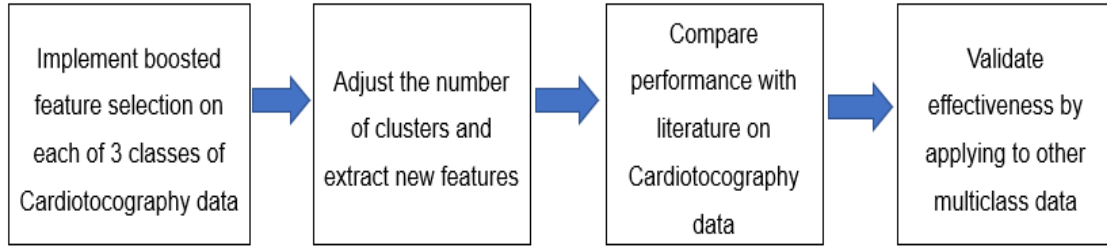


Figure 7. The research flow of the final phase, Developing classification methodology for Cardiotocography data and validating the effectiveness.

2.2 Techniques Applied to Proposed Model

The detailed information on the techniques to the proposed model in this research, i.e., Data preparation, Outlier treatment, Feature classification rate ranking method, Principal component analysis, Linear discriminant analysis, Distance between classes, Support vector machine, Parameter optimization, Feature selection of Cardiotocography data, Improved classification methodology for Cardiotocography data, are presented as follows.

2.2.1 Data Preparation

This research experimented the proposed algorithms on 17 data sets, i.e., 10 kinds of 2-class data and 7 kinds of multiclass data. At first, 10 kinds of 2-class data sets of

various number of instances and features were prepared. Omissions in data sets were deleted to prevent any malfunction of algorithm in R. In addition, all features were normalized to make computation load even and eliminate unnecessary computation load. The detail characteristics of the data sets are summarized in Table 4.

Table 4. The characteristics of Ten kinds of 2-class data sets

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

In addition, 7 kinds of multiclass data sets of various number of instances and features were prepared. In the same way as the 2-class data, any omissions in data sets were deleted and all features were normalized. The detail characteristics of the data sets are summarized in Table 5.

Table 5. The characteristics of Seven kinds of multiclass data sets

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

In Table 4 and Table 5, the data sets are sorted in the ascending order of the number of instances. In both 2-class data and multiclass data, the number of instances and features are diversely distributed so that the change of experimental result depending on the size of data can be observed and analyzed. As the number of instances increase, the computational complexity increases. Because this research intends to reduce the computation time of data, the experimental results from large size of data is more meaningful.

The Cardiotocography Data - This data set consists of measurements of Fetal Heart Rate (FHR) and Uterine Contraction (UC) features on cardiotocograms (CTGs) classified by expert obstetricians. This data set has 2,126 features, 21 features and 3 classes. The 3 classes are fetal state class, i.e., Normal / Suspect / Pathological. The detail description and the characteristic (Metric or Categorical) of the features are summarized in Table 6.

Table 6. The description of the features in Cardiotocography data

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 histogram	M
14	Max	Maximum of FHR histogram	M
15	Nmax	Number of histogram peaks	M
16	Nzeros	Number of histogram 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

The implementation of experiment in this research was carried out on the R.3.5.0 and desktop with the environment, 6th generation Intel Core i5-6400 CPU and 64-bit 12GB DDR3L system memory.

2.2.2 Outlier Treatment

This research implements outlier treatment on the data sets. All the features of a data are normalized according to equation (5).

$$z = \frac{x - \mu}{\sigma} \quad (5)$$

μ = Mean

σ = Standard Deviation

The outlier can be defined by the criteria of standard deviation of a feature, σ , multiplied by a coefficient. The coefficient can be any number, i.e., 2,3,..., or any continuous values, i.e., 2.3, 2.5, depending on the outlier level. $X \leq |3 \sigma|$ include 99.7% of the data samples in normal distribution. $X \leq |2 \sigma|$ include 95.4% of the data samples. $X \leq |1 \sigma|$ include 68.2% of the data samples. The outlier level can be adjusted to obtain better performance per data.

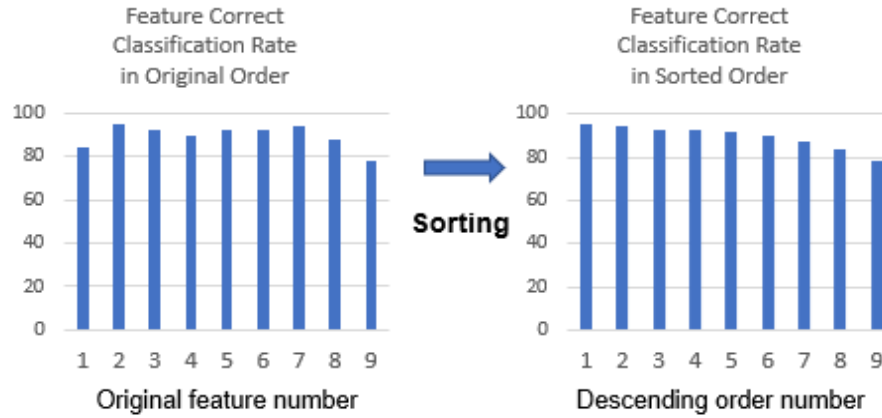
Base of this outlier level determination, two types of outlier treatment methods are reviewed and considered. The first type method is trimming, which is the outlier elimination method. This method is that the instance is totally eliminated if outlier is found even in one dimension of the features. The disadvantage of this method is that the available instances are reduced if the number of outliers increases, consequently degrading the obtainable CCR. The second type method is winsorizing, which is the outlier adjustment method. This method adjusts the outlier values into the upper or lower limit, i.e., 3σ . The advantage of this method is that the values in other dimensions of the outlier feature in a specific instance can be used in training and testing process, consequently minimizing the effects of outlier treatment on degrading CCR.

This research decides to use the second method as outlier treatment because the number of instances are various depending on the data and the instances of a specific data is as less as 195, which is not sufficient in producing improved performance if the instances of outlier are completely eliminated.

2.2.3 Feature Classification Rate Ranking Method

Feature ranking method is one of the 2 methods used in preprocessing phase of this research. This method is referred as ‘Rank’ method in remaining part of this dissertation. The procedure of feature ranking method is explained as following. At first, the CCR of each feature is calculated by using SVM with RBF kernel. The parameter C & γ in RBF kernel is set as default parameters to apply same parameter conditions in calculating the CCR of individual feature. Secondly, the features are sorted in descending order of the calculated CCR. Thirdly, grid searching parameters C & γ is implemented by adding features cumulatively in case of the 3 kinds of kernel, i.e., polynomial, sigmoid and RBF kernel. Linear SVM does not need this process because it has no parameter. All the processes so far are described in Figure 8. Finally, training and testing are implemented by using the optimized parameters to obtain predicted CCR. All the explained processes will be repeated by adding features cumulatively until stopping criteria is met. The iteration loop terminates if the prediction accuracy 100% with only one feature is produced. This algorithm searched for the subset realizing the highest CCR. If multiple subsets realizing the same highest CCR exist, the subset with fewer number of selected features is chosen.

The advantage of the Rank method is that the characteristics of original feature can be used in training and classification steps while the disadvantage is that the multicollinearity among features remains, which is a factor decreasing the CCR of classifier. The multicollinearity among features negatively influences the performance of classifier.



Input by adding cumulatively

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

Figure 8. The process of feature correct classification rate ranking method

2.2.4 Principal Component Analysis

Principal component analysis (PCA) is the 2nd method used in preprocessing phase of this research. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PC). It extracts a set of uncorrelated

variables and store them in features of smaller dimensions. For example, the PC1 is the first principal component which is on the axis maximizing the variability of data. The PC2 is the second principal component which is orthogonal to PC1 and uncorrelated to PC1. This process of searching for principal components are repeated by the number of features.

Table 7. The order of input combinations of principal components

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

The advantage of PCA is that it is effective in reducing number of dimensions of data with multicollinearity. The multicollinearity means that input features in a data are correlated each other. PCA is also effective in dimensional reduction of data with large number of features.

2.2.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method to reduce a data from the viewpoint of optimal classification. LDA reduces the dimension of eigen vector of data by maximizing the ratio of between-class scatter to within-class scatter after supervised learning.

Within-class scatter S_w is generalized as equation (6) and (7).

$$S_w = \sum_{i=1}^c S_i \quad (6)$$

$$S_i = \sum_{\mathbf{x} \in \omega_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T \quad \mu_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{x} \quad (7)$$

Between-class scatter S_B is generalized as equation (8) and (9).

$$S_B = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (8)$$

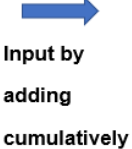
$$\mu = \frac{1}{N} \sum_{\mathbf{x} \in \omega} \mathbf{x} = \frac{1}{N} \sum_{i=1}^c N_i \mu_i \quad (9)$$

Total scatter matrix S_T is the summation of S_B and S_w as equation (10).

$$S_T = S_B + S_w \quad (10)$$

This research used LDA as a feature ranking criteria as following procedure.





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

Figure 9. The process of using LDA as a feature ranking criteria

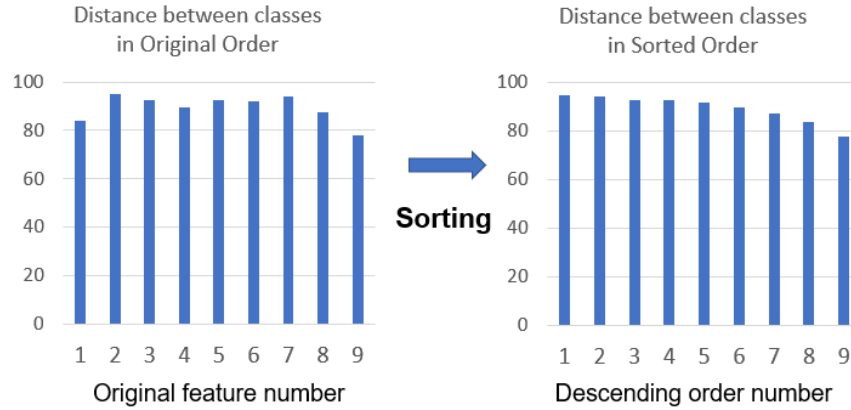
2.2.6 Distance between Classes

The distance between 2 classes can be calculated by equation (11), which is discriminatory power between the class w_i and w_j .

$$d_{ij} = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(\mathbf{x}_i^k, \mathbf{x}_j^m) \quad (11)$$

where X_i^k is k^{th} sample in class w_i . $dist(\mathbf{x}_i^k, \mathbf{x}_j^m)$ is the distance between the 2 samples. N_i is the total number of samples in class w_i .

This distance represents the degree of separation of the classes. As the distance gets longer, the discriminatory power increases. This research uses the distance as a criteria of rearranging features in input subset. The feature with the highest distance is the top-ranked feature. A wrapper method searches for the highest CCR and the corresponding number of features. If this method is combined with outlier adjustment, the discriminatory power increases because the outliers negatively affect the correct calculation of the distance between the classes.



Input by adding cumulatively

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

Figure 10. The process of using distance between classes as a feature ranking criteria

2.2.7 Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm which is used for classification and regression analysis. SVM is an effective classification method with significant advantages. SVM separates the training patterns by searching for optimal hyperplane. In SVM, algorithm searches for a hyperplane realizing maximum margin between the points in the two classes, which is closest to the hyperplane.

If such a hyperplane exists, the hyperplane is maximum-margin hyperplane and the linear classifier is called as the maximum margin classifier. If it is assumed that the set of

D is the data with N points to be used for training as shown in equation (12): Each \mathbf{x}_i is real number vector in p dimension and \mathbf{y}_i is a value representing the class \mathbf{x}_i belongs to. \mathbf{y}_i is either 1 or -1.

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n \quad (12)$$

In case that the above set of training data can be linearly separable depending on \mathbf{y}_i , what divides the data set is referred as hyperplane and it can be represented as the set of X which satisfying equation (13). \cdot is inner product operator and w is normal vector of hyperplane.

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \quad (13)$$

The support vectors ($\mathbf{X}^+, \mathbf{X}^-$) in the given hyperplane is defined as follows.

\mathbf{X}^+ : data which is closest to the hyperplane **among** $\mathbf{y}_i = +1$ data

\mathbf{X}^- : data which is closest to the hyperplane **among** $\mathbf{y}_i = -1$ data

The margin, i.e., the distance between the two hyperplanes which crosses the support vectors, is $\frac{2}{\|\mathbf{w}\|}$. The constraints in equation (14) means that each data point must be located on the correct side of the margin.

$$\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \text{ for all } 1 \leq i \leq n. \quad (14)$$

The SVM problem which meets the hyperplane condition and searches for maximum margin can be represented as the optimization problem in equation (15).

$$\arg \min_{(\mathbf{w}, b)} \|\mathbf{w}\| \quad (15)$$

in which $\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \text{ for all } 1 \leq i \leq n$

However, the above explained optimization problem is difficult to solve because $\|\mathbf{w}\|$ includes square root. In the above optimization problem, w and b do not change in

meeting the optimization problem even if $\|\mathbf{w}\|$ is replaced by $\frac{1}{2} \|\mathbf{w}\|^2$. If $\|\mathbf{w}\|$ is replaced by $\frac{1}{2} \|\mathbf{w}\|^2$, the above problem becomes optimization problem in quadratic programming.

The replaced optimization problem can be represented as equation (16), in which

$$\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \text{ for all } 1 \leq i \leq n \quad (16)$$

If Lagrange multiplier method is used, the above problem can be represented in equation (17) which searches for the saddle point.

$$\arg \min_{\mathbf{w}, b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1] \right\} \quad (17)$$

In the process of searching for saddle point, all the data points that can be classified as $\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1 > 0$ do not have any influence because the α_i of these points are 0. Then, this problem can be solved as technique in quadratic programming. According to Karush-Kuhn-Tucker condition, the solution can be represented as the linear combination of training vector in equation (18).

$$\mathbf{w} = \sum_{i=1}^n \alpha_i \mathbf{y}_i \mathbf{x}_i \quad (18)$$

Only a few α_i have a value larger than 0, the \mathbf{x}_i s are supporting vectors satisfying $\mathbf{y}_i (\mathbf{w} \cdot \mathbf{x}_i - b) = 1$. The offset b can be defined from this formula, as equation (19).

$$\mathbf{w} \cdot \mathbf{x}_i - b = \frac{1}{\mathbf{y}_i} = \mathbf{y}_i, \text{ so } b = \mathbf{w} \cdot \mathbf{x}_i - \mathbf{y}_i \quad (19)$$

If $\mathbf{w} = \sum_{i=1}^n \alpha_i \mathbf{y}_i \mathbf{x}_i$ is used from $\|\mathbf{w}\|^2 = \mathbf{w}^T \cdot \mathbf{w}$

Dual form can be changed to the optimization problem in equation (20).

$$\text{Maximize } \tilde{L}(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \mathbf{y}_i \mathbf{y}_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (20)$$

$$\text{Where } \alpha_i \geq 0, \sum_{i=1}^n \alpha_i \mathbf{y}_i = 0, \text{ for all } 1 \leq i \leq n.$$

$\sum_{i=1}^n \alpha_i \mathbf{y}_i = 0$ is constraint resulting from minimizing b , and kernel is defined by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$$

w can be calculated in terms of α , as equation (21).

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \quad (21)$$

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification by use of kernel trick, which maps inputs into high-dimensional feature spaces. In this research, 3 kinds of kernel, i.e., polynomial, sigmoid and RBF kernel, and linear SVM are used in the proposed classification models. In case that a data is not labeled, an unsupervised learning approach is possible by support vector clustering. SVM can be used to solve various problems in real world such as text categorization, classification of images, hand-written character recognition and biological sciences, etc.

2.2.8 Parameter Optimization

In case of polynomial, sigmoid, RBF kernels, parameter optimization is required before training. The kernel functions and parameters for each kernel are as follows.

1. **Linear** $k(x_1, x_2) = x_1^T x_2 \quad (22)$

In case of Linear SVM, there is no parameter to be optimized.

2. **Polynomial** $k(x_1, x_2) = (\gamma x_1^T x_2 + \text{Coef})^d \quad (23)$

γ needs to be searched by setting Coefficient as default 0 and degree d as default 3.

3. **Radial basis** $k(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2) \quad (24)$

γ and C (cost) need to be searched by grid search.

4. **Sigmoid** $k(x_1, x_2) = \tanh(\gamma x_1^T x_2 + \text{Coef}) \quad (25)$

γ needs to be searched by setting Coefficient as default 0.

This research searched for the parameters of the 3 kinds of kernel which produces highest CCR by using tuning function in R. The tuning function searches for the optimal parameters. Among the 4 kernels, searching for parameters, γ and C, in RBF is most time-consuming because grid search is needed for the searching. Grid search method calculates all the combinations between γ and C value, which is the reason for long computation time. The optimal parameters and CCR are calculated per each input subset unless stopping criterion is met. In case of γ and C values, the values are different between the Rank and PCA methods. The optimal parameters are different depending on kernel and the combinations of input features. Consequently, the resulting CCRs are different depending on the combinations of input features and parameter setting. The iteration loop by wrapper method continues unless the stopping criterion is met. In the wrapper method, the feature selection and parameter optimization are implemented simultaneously. This research conducted grid search to find parameter pairs (C, γ) which produce highest CCR. Grid searching parameters for RBF Kernel is the most time-consuming compared to that of 3 kinds of kernel. As the 2 parameters get larger, the classification model gets more complex. The parameter γ defines how far the influence of a single training example reaches. Low γ value means the influence of a single training example reaches far. In contrast, high γ value the influence of a single training example reaches only close area. The parameter C (cost) is the penalty parameter which controls the overfitting. C trades off misclassification of training examples against simplicity of the decision surface. In case of RBF, tuning function in R calculates the CCR of every pair of combinations between C and γ , compared them and output the C, γ which produces the highest CCR. Each combination of features

or principal components (PC) results in different optimal parameters, consequently different CCRs. The iteration loop by wrapper method selects the feature or PC subsets which produces the highest CCR with fewer number of selected features or PCs.

2.2.9 Feature Selection on Cardiotocography Data

Selection of Classification Method for Multiclass

Basically, SVM is a binary classifier. For this reason, the following 2 options are reviewed for multiclass classification on Cardiotocography data.

One vs. all – This is the multiclass classification method by M binary SVMs, which has a disadvantage that the samples in one class and the samples in the rest of the classes can be unbalanced. For example, 10% of the samples are in one class and 90% of the samples are in the rest of the classes for 10 class problem.

One vs. one – This is the multiclass classification method by $M(M-1)/2$ binary SVMs and voting, which has a disadvantage that the computation time can be much longer than one vs. all classification method.

Between the 2 options, one vs. all is more widely used even if it has a disadvantage. This research selected one vs. all method in feature selection of Cardiotocography data. The classifications are implemented 3 times as shown in Table 8 and the average of the 3 correct classification rates is regarded as the final correct classification rate.

Table 8. Three kinds of one vs. all classification for 3 class data set.

Sequence	One vs. all
1	1 vs. 2&3
2	2 vs. 1&3
3	3 vs. 1&2

First of all, SVM is trained and tested on Cardiotocography data to detect the misclassified instances. In increasing correct classification rate and selecting features for the highest CCR, increasing the classification rate of misclassified instances is critical. For this reason, this research prioritizes the features whose distance between classes calculated based on the misclassified instances is longest, in cumulatively inserting to classifier by wrapper method. This is the most effective method to determine selected features and maximize the CCR simultaneously.

The RBF kernel used in the SVM for preprocessing is the same kernel used in classifier SVM to maximize the advantage from instance selection for distance calculation. The number of 2 types of misclassified instances, the instances originally labelled as class 1 and classified as class 2, and the instances originally labelled as class 2 and classified as class 1, are summed per 3 kinds of one vs. all classifications for multiclass. The distance between classes are calculated per all features, by using only the detected instances according to equation (26).

$$d_{ij} = \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{m=1}^{N_j} dist(\mathbf{x}_i^k, \mathbf{x}_j^m) \quad (26)$$

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

Not all the instances, but the only detected instances are used in calculating distance because the distance calculation based on all instances cannot guarantee the increase of or at least the equivalent correct classification rates according to the experiment results on other various data sets. At next step, by using a wrapper method inserting cumulatively added features to SVM, the highest correct classification rate and selected features are detected. Figure 11. illustrates the flow of this algorithm.

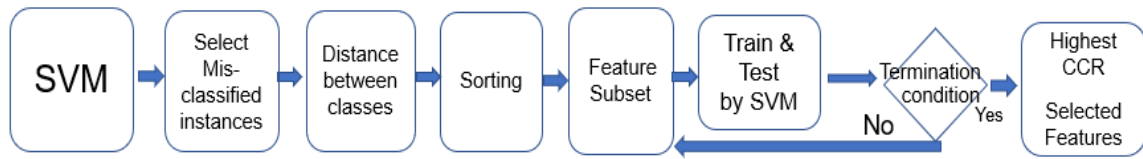


Figure 11. The Flow of Algorithm – Preprocessing by SVM & Distance measure + SVM (classifier)

By using this method, as the cumulative input features increase, the correct classification rate is expected to increase faster compared to the case without preprocessing because the discriminatory power is used effectively used only on the misclassified instances, which are critical in increasing CCR further.

The boosted feature selection methodology is applied to AdaBoost and Random Forest to verify the effectiveness of the methodology. The feature ranking is based on decision tree because AdaBoost and Random Forest use decision tree as a classifier. In case of AdaBoost, which is sensitive to outliers, PCA is also implemented as a preprocessing method because PCA has the effect of noise reduction in data. The result is also compared with the boosted feature selection. The flow of algorithm of the 3 methodologies are represented in Figure. 12, 13 and 14, respectively.

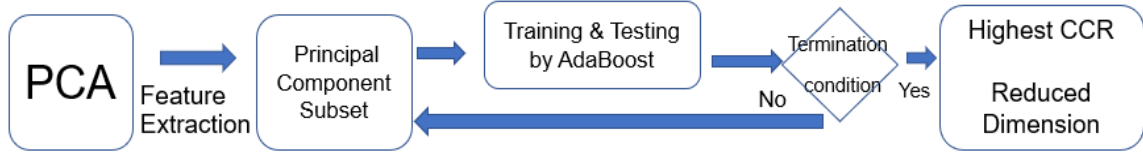


Figure 12. Flow of algorithm – PCA + AdaBoost + Wrapper method for dimension decision

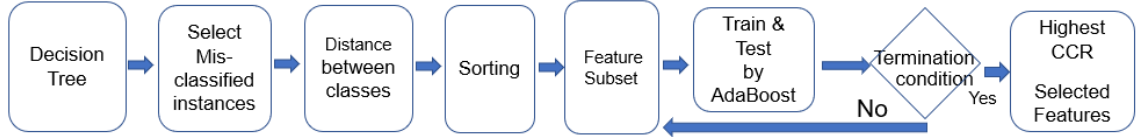


Figure 13. Flow of algorithm – Boosted feature selection by DT + AdaBoost + Wrapper method

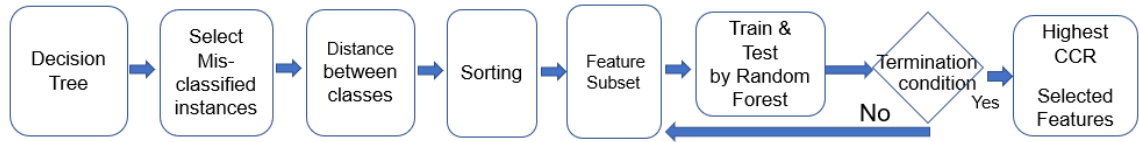


Figure 14. Flow of algorithm – Boosted feature selection by DT + Random Forest + Wrapper method

2.2.10 Improved Classification Methodology for Cardiotocography Data

This research developed class-dedicated SVMs for improved classification methodology for Cardiotocography data. SVM is originally developed for binary classification. For this reason, Binary Decision Tree (BDT) has been used in several literature related to classification of Cardiotocography data to extend the binary SVMs to multi-class problems. However, BDT has a limitation in improving the performance further because the misclassified instances from SVM1 negatively affects the classification performance of SVM 2. In contrast, class-dedicated SVM focuses on increasing CCR of each class, consequently improving the overall performance. In this research, the

performances of the two architecture are compared by the same classification condition and criteria.

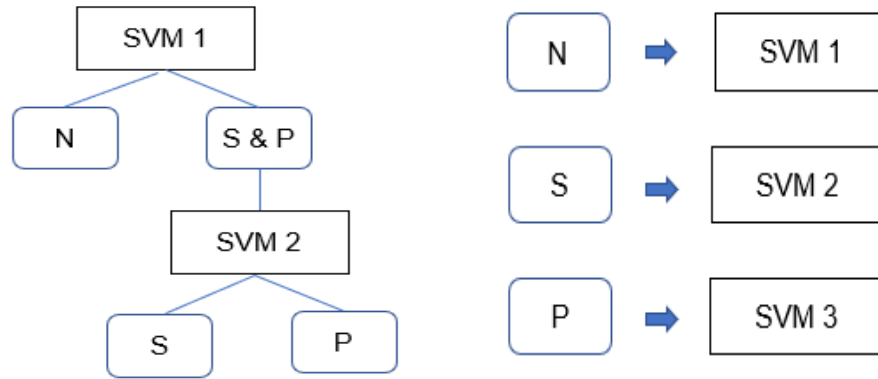


Figure 15. Comparison of classification architecture – Binary Decision Tree (BDT) vs. Class-dedicated SVMs.

The architecture of improved classification methodology for Cardiotocography data is represented in Figure 16.

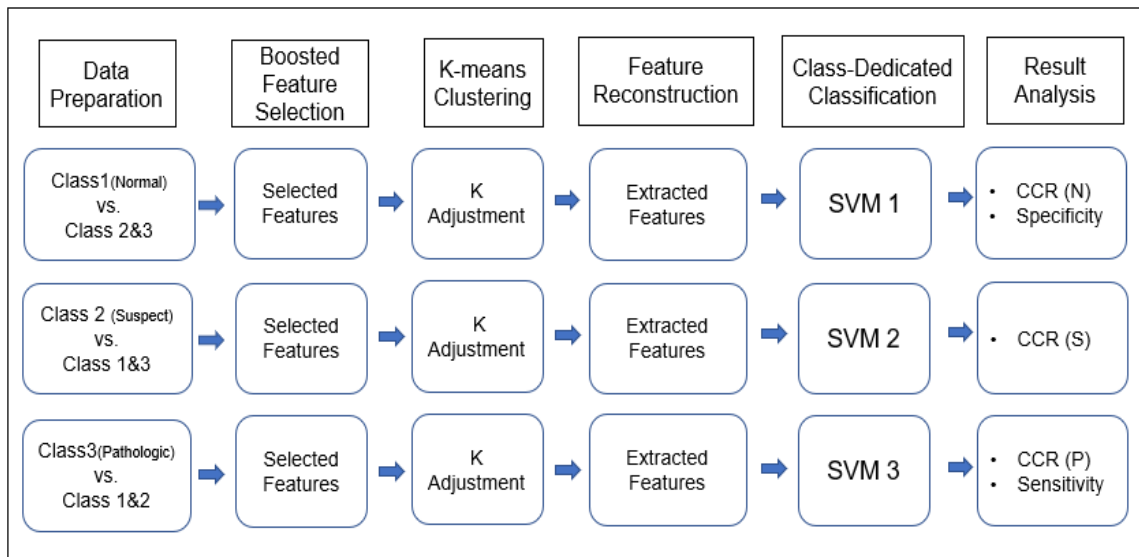


Figure 16. The architecture of improved classification methodology for Cardiotocography data

At first, the 3-class problem is converted into 3-binary class problems so that each of 3 dedicated SVMs can be the dedicated classifier for each of the 3 classes of Cardiotocography data, i.e., normal, suspect and pathologic. The 2nd step is boosted feature selection for each of binary classification problems. The feature selection methodology in 3.2.9 is used for the feature selection of the 3-binary classification problems. The original features are sorted according to the highest discriminant powers, i.e., the largest distance between the 2 classes among the misclassified instances from SVM. The 3rd step is K-means clustering with adjusting the number of clusters, k. K-means clustering partitions the n instances into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ to minimize the within-cluster sum of squares, as shown in equation (27). k is the number of cluster and μ_i is the center of the i th cluster, mean of all points in S_i .

$$\min_S \sum_{i=1}^k \sum_{x \in S_i} ||x - \mu_i||^2 \quad (27)$$

Basically, the suitable number of k is determined in the target class of each of the 3 binary classifications, i.e., within class 1, within class 2, within class 3. Then, the clusters in the other binary class can be determined in either method (1) separately within each class, or (2) within merged classes, i.e., class 2 + class 3, class 3 + class 1, class 1 + class 2. The 4th step is feature extraction by fuzzy membership functions as shown in equation (28). The fuzzy membership function converts the result of clustering into extracted features. If there exist k clusters in input features, these are converted into k newly extracted features, consequently reducing dimension and computational complexity.

$$f_{np} (X_j^i) = 1 - \frac{|X_j^{\mu np} - X_j^i|}{\max |X_j^{\mu np} - X_j^n|} \quad \text{if } \min(X_j^n) \leq X_j^i \leq \max(X_j^n), \forall n \in S_{np} \quad (28)$$

$$f_{np} (X_j^i) = 0 \quad \text{if otherwise;}$$

The newly extracted features are defined by the function (29). NSF is the number of selected features for each class by boosted feature selection. EF_{np} is the extracted feature of the new pattern np by K-means clustering algorithm.

$$EF_{np} = \frac{1}{NSF} \sum_{j=1}^{NSF} f_{np} (X_j^i) , 1 \leq np \leq K^N + K^S + K^P \text{ (clustering within each class)} \quad (29)$$

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

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

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

The 5th step is the classification of each class by using class-dedicated SVMs. Each of the class-dedicated SVM is dedicated to the classification of each class to maximize the CCR of each class and overall CCR, sensitivity and specificity. The confusion matrices of the 3-class and the equations for specificity, sensitivity and CCR of each class are as shown in Figure 17. and equations (30), (31) and (32). In each of the 3 binary classifications of the 3 classes of Cardiotocography data, i.e., normal, suspect and pathologic, each class in target is regarded as positive class, and the other remaining classes are considered as negative.

Normal	Suspect	Pathologic												
<table><tr><td>TP₁</td><td>FP₁</td></tr><tr><td>FN₁</td><td>TN₁</td></tr></table>	TP ₁	FP ₁	FN ₁	TN ₁	<table><tr><td>TP₂</td><td>FP₂</td></tr><tr><td>FN₂</td><td>TN₂</td></tr></table>	TP ₂	FP ₂	FN ₂	TN ₂	<table><tr><td>TP₃</td><td>FP₃</td></tr><tr><td>FN₃</td><td>TN₃</td></tr></table>	TP ₃	FP ₃	FN ₃	TN ₃
TP ₁	FP ₁													
FN ₁	TN ₁													
TP ₂	FP ₂													
FN ₂	TN ₂													
TP ₃	FP ₃													
FN ₃	TN ₃													

Figure 17. The confusion matrix of the 3 classes

$$\text{Specificity} = \text{CCR of Class 1} = \frac{TP_1}{TP_1 + FP_1} \quad (30)$$

$$\text{CCR of Class 2} = \frac{TP_2}{TP_2 + FP_2} \quad (31)$$

$$\text{Sensitivity} = \text{CCR of Class 3} = \frac{TP_3}{TP_3 + FP_3} \quad (32)$$

2.3 Summary

This chapter presents the summary of various methodologies employed in the Rank-PCA-LDA ensemble algorithm. At first, overall Research Framework was explained by 4 phases in 2.1. Afterward, the details of the Techniques applied to Proposed Model are introduced in 2.2. In 2.2.1, Data Preparation, the characteristics of the 17 data sets, necessary measures taken in advance of experiments, are summarized. 10 kinds of 2-class data and 7 kinds of multiclass data with various number of instances and features are prepared. In 2.2.2, applied Outlier Treatment methods are explained in detail. In 2.2.3 Feature Classification Rate Ranking Method and 2.2.4 Principal Component Analysis, the 2 preprocessing methods, i.e., feature correct classification rate ranking method and principal component analysis, are introduced with their advantages and disadvantages. By the Ranking method, the features are sorted in descending order of the calculated correct classification rate from SVM with RBF kernel. The advantage of the Rank method is that the characteristics of original feature can be used in training and classification steps while the advantage of PCA is that it is effective in reducing number of dimensions of data with multicollinearity. In 2.2.5 Linear Discriminant Analysis and 2.2.6 Distance between Classes, 2 other alternative feature ranking criteria are explained in detail. In 2.2.5 Linear

Discriminant Analysis, the theoretical background of LDA is presented. LDA reduces the dimension of eigen vector of data by maximizing the ratio of between-class scatter to within-class scatter after supervised learning. In 2.2.7 Support Vector Machine, the theoretical background of SVM is presented. SVM searches for a hyperplane realizing maximum margin between the points in 2 classes, which are closest to the hyperplane. In addition to linear classification, SVMs can efficiently perform a non-linear classification by use of kernel trick, which maps inputs into high-dimensional feature spaces. In this research, 3 kinds of kernel, i.e., polynomial, sigmoid and RBF kernel, and linear SVM are used in the classification models. In 2.2.8, Parameter Optimization, the process and importance of parameter optimization of kernels are explained. Grid search method calculates all the combinations between γ and C value per all feature input subsets, which is the reason for the long computation time. SVM with RBF kernel shows the longest computation time among all kernels. Wrapper method selects the feature input subsets and parameters which produces the highest prediction accuracy. In 2.2.9, the methodology for feature selection for Cardiotocography is explained. Finally, in 2.2.10, the improved classification methodology for Cardiotocography data is provided. The implementation results of the listed methodologies will be discussed in chapter 3 and chapter 4.

Chapter 3. Experimental Result on 2-class and Multiclass Data

This section presents the experimental results of 4 kinds of ensemble algorithms, i.e., 8 combinations, 6 combinations, 4 combinations and 2 combination ensemble algorithms, followed by the efficient algorithms depending on the characteristics of data such as the number of class, i.e., whether the dependent variable is 2-class or multiclass, and the ratio metric features.

The 8 combinations include all kernels to maximize the CCR regardless of time complexity. 6 combinations exclude the most time-consuming RBF even if it degrades CCR. The 4 combinations include RBF and exclude other inefficient kernels which are not contributing to the highest CCR. Reduced data is applied to grid search of RBF kernel in 4 combinations. In case of feature ranking, 1/12 of training data is applied to grid search. In case of PCA, 1/3 of training data is applied to grid search. Finally, the 2 combinations include only linear SVMs, pursuing the least time complexity even if it degrades CCR.

3.1 Summary of Performance of 4 Ensemble Algorithms

In Figure 18, the data are arranged in the ascending order of the number of instances. The change of time reduction rate of 3 ensemble algorithms per the number of instances in data, is represented in Figure 18.

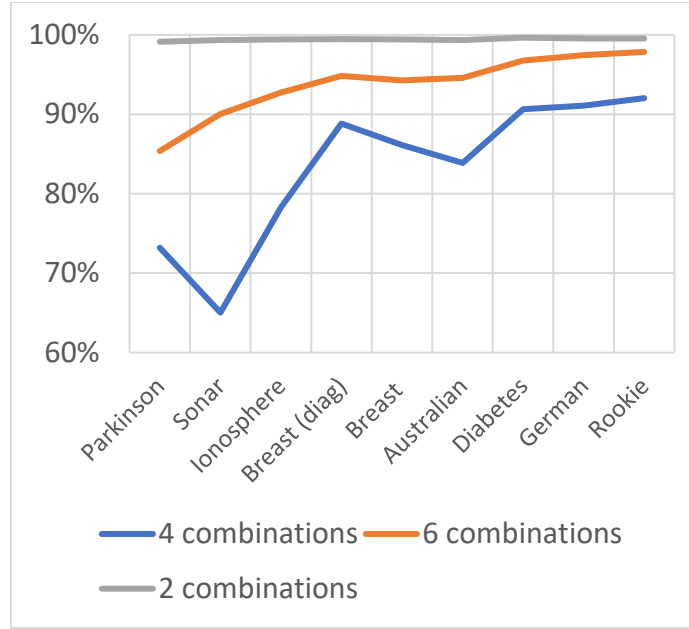


Figure 18. The reduction rate of computation time of 3 ensemble algorithms per the number of instances in data

According to the result, the computation time reduction rates of both 4 combination and 6 combination ensemble algorithms increases significantly as the number of instances of data increases. The reduction rate of 6 combinations is higher than that of 4 combinations. The number of instances increases exponentially as the time reduction rate approaches to 100%, which can be interpreted as such that the effect of time reduction gets greater as the number of instances of data gets larger. However, the number of features has no distinct relation with the time reduction rate but has a weak reversely propositional relation with the time reduction rate, which can be interpreted as such that the effect of time reduction gets smaller as the number of features of data gets larger. Finally, the relation between the multiplication of instances and features, and the time reduction rate is analyzed. According to the analysis, the multiplication of instances and features has no clear relation with the time reduction rate but has a weak proportional relation with time reduction rate. This can

be interpreted as such that the large size data tends to have the more benefit of time reduction from the 4-combination ensemble algorithm by assuming that the size of data is represented by the multiplication of instances and features of data. According to the result, the 8 combination and 4 combination ensemble algorithms produces the highest CCRs in all 10 data sets. The CCR of 6 combinations degrades in 3 data (30%) out of 10 data while the correct classification rate of 2 combinations degrades in 5 data (50%) of 10 data.

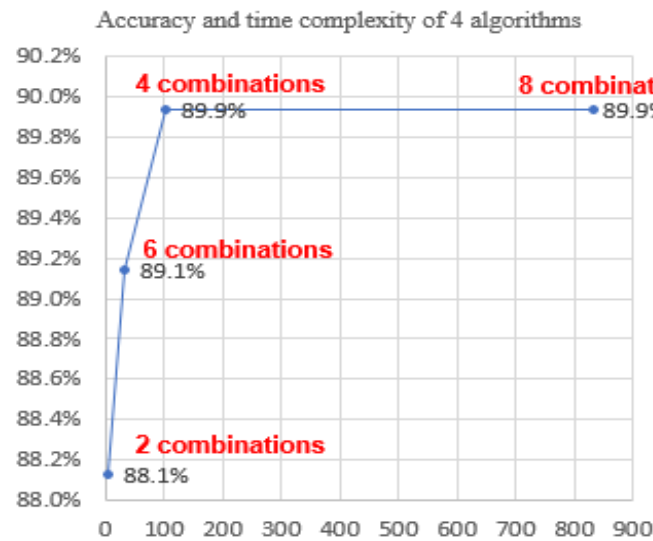


Figure 19. Average Correct Classification Rate vs. Computation Time (4 ensemble algorithms)

Figure 19 represents the performance of the 4 ensemble algorithms in terms of the average CCRs and average computation time. The 4-combination ensemble algorithm reduces the computation time significantly while maintaining the same CCR as the 8-combination algorithm. The 6 and 2 combination algorithms reduce the computation time further at the expense of degrading the CCR by 0.8% and 1.8 % respectively. The time reduction rates of 4, 6 and 2 combinations are 75%, 84% and 97% respectively compared to 8 combinations.

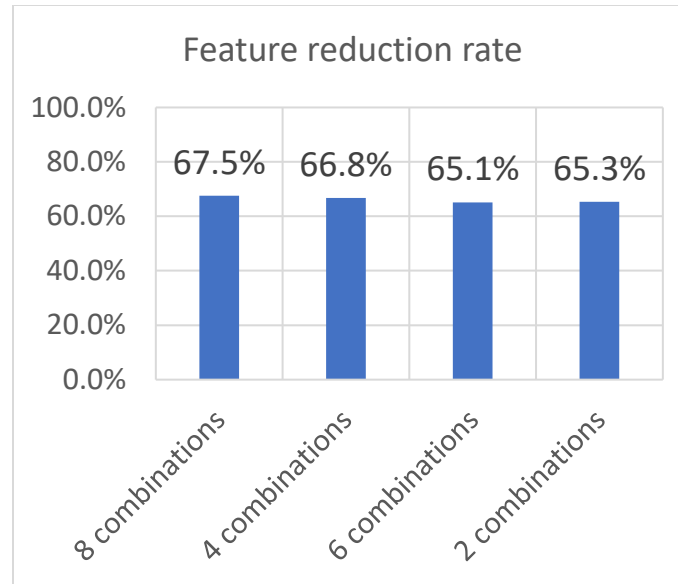


Figure 20. The feature reduction rate of 4 ensemble algorithms

The feature reduction rates of the 4 algorithms are compared as represented in Figure 20. According to the result, the feature reduction rates of the 4 algorithms are almost at the same level. The maximum rate is 67.5% while the minimum rate is 65.1%, the difference being 2.2%. The feature reduction contributes in reducing memory requirement in implementing algorithm. However, the feature reduction does not significantly influence on reducing computation time. This research placed the priority of research on reducing overall computation time while not degrading CCR.

3.2 Comparison of Performance with Approaches in Literature. (2-class data)

Comparison with Genetic Algorithm (GA)

GA-based feature selection and parameter optimization for SVM was researched by Huang and Wang (2006). 7 data sets used in this research were also used in Huang and Wang's research. The number of selected features and CCR in this paper were compared. To compare the performance from the same experimental condition, training (90%) and testing (10%) cross-validation was applied.

Comparison with GA-Ensemble Method

The GA-ensemble composed of ANN, Decision tree and SVM was researched by Zhang and Yang (2008). 2 data sets used in this research, i.e., Sonar, Ionosphere, were also used in the research of Zhang and Yang. The number of selected features and CCR in this paper were compared. To compare the performance from the same experimental condition, training (80%) and testing (20%) cross-validation was applied.

Comparison with other Feature Selection Method (SVM-RFE+AT)

A variation of SVM-RFE was researched by Samb et al. (2012) with an aim to improve the performance of SVM-RFE. 2 kinds of local search methods, i.e., SVM-RFE+BF (Bit-Flip) and SVM-RFE+AT(Attribute-Flip) were developed and experimented in the paper. 1 data set in this research, i.e., Ionosphere was also used in the research of Samb et al. The number of selected features and CCR in this paper were compared. To compare the performance from the same experimental condition, training (50%) and testing (50%) cross-validation was applied. The highest CCR, 86.3% obtained by SVM-RFE+AT, is compared to that from this research.

Table 9. Comparison of performance with other methodologies in literature (NF means the 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%			

The performance of Rank-PCA ensemble is slightly better than that of GA approach with significantly reduced time complexity. The average CCR of Rank-PCA ensemble is higher than that of GA approach by 0.06%. The average number of selected features of Rank-PCA is less than that of GA approach by 0.3. Overall, the correct classification rates of Rank-PCA ensemble are higher than that of GA approach in 3 data sets out of 7 data sets. Regarding time complexity, the authors in the research mentioned that the average learning time for GA-based approach is slightly inferior to that of the Grid algorithm and that GA-based approach significantly improves the CCR and has fewer input features for SVM. Considering the authors implemented Grid algorithm on not reduced training data sets, the time reduction rate of GA-based approach to 4 combination Rank-PCA ensemble algorithm is assumed to be higher than 75%.

The performance of Rank-PCA ensemble is also better than GA-ensemble. The average CCR of Rank-PCA ensemble is higher than that of GA approach by 4.45%. The average number of selected features of Rank-PCA is less than that of GA approach by 0.5. The GA-ensemble consists of multi-objective GA and ensemble classifiers, i.e., ANN,

Decision tree and SVM. Even if there is no clear mention regarding the time complexity of the proposed algorithm in this paper, it is assumed that the computation time is at least as long as the general time required to GA, which is longer than that of Grid search in case of SVM. The authors used SVM with polynomial kernel and default parameters in this paper, which does not contribute in reducing the total computation time of GA-ensemble.

The CCR of Rank-PCA is higher than SVM-RFE+AT by 11%. However, SVM-RFE+AT used more features than Rank-PCA in realizing the higher CCR. The authors didn't mention time complexity of the proposed algorithm. Originally, SVM-RFE is a kind of wrapper which is computationally expensive than filter method. However, it is uncertain whether the authors used Grid searching parameters or default parameters.

3.3 Efficient Algorithm Depending on Feature Type

This research conducted an experiment to search for efficient algorithm depending on the feature types of data sets. First of all, the characteristics of feature of all data sets were investigated and divided into either metric (M) or categorical (C). The metric characteristic means that the feature is discrete or continuous variables. The categorical characteristic means the feature consists of a few categories. The result is represented in Table 10.

Table 10. The characteristics of features of all 2-class data (Metric/Categorical)

No.	Data	Feature Type																													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	Parkinson	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M									
2	Sonar	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
3	Breast (diag)	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
4	Breast	M	M	M	M	M	M	M	M	M																					
5	Rookie	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M												
6	Diabetes	M	M	M	M	M	M	M	M																						
7	Ionosphere	C	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
8	Australian	C	M	M	C	M	M	M	C	C	M	C	C	M	M																
9	Heart	M	C	C	M	M	C	C	M	C	M	C	C	C	C																
10	German	C	M	C	M	M	C	C	C	C	C	C	M	C	C	C	C	C	C	C											

No.	Data	Feature Type																													
		31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
2	Sonar	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M

Secondly, the ratio of number of metric features to total number of features is calculated. Then, the preprocessing methods (Rank or PCA) and kernel which produce the highest CCR per data were identified. The results are summarized as follows in Table 11.

Table 11. The result of preprocessing methods, kernels and efficient algorithm

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	

In case that the metric feature ratio is higher than 90%: According to the analysis from the above result, in case of metric features, PCA performs better than Rank method in producing the highest CCR. It results in the highest CCR in 5 data out of 7 data sets. However, Rank method with linear SVM also produces highest CCRs in 2 data of all metric features. Consequently, in case that the metric feature ratio is higher than 90%, the combination of PCA with Sigmoid kernel/RBF kernel/Linear SVM and Rank method with Linear SVM is the most efficient algorithm which produce the highest correct classification rate.

In case that the metric feature ratio is lower than 90%: The Rank method with all 4 kinds of kernel can produce the highest CCRs. This can be interpreted as such that Rank method performs better than PCA in producing highest CCRs as the ratio of metric feature gets lower.

Reduction of time complexity by efficient algorithms depending on feature type:

According to the computation time complexity analysis, the computation time of RBF kernel is significantly longer than that of other kernels even after the reduced size of training data for grid search is applied. In case of the efficient algorithms depending on feature type, the average computation time of all data sets from the 4-combination algorithm, 103 seconds, is reduced to 60 seconds with the same CCRs as the 4-combination algorithm, as represented in Table 12. The proposed efficient algorithms depending on feature type reduce the computation time by 39% compared to that of 4-combination algorithm. This reduction is possible by using the RBF kernel only 1 time with either Rank or PCA method, instead of using it for both preprocessing methods in 4-combination algorithm.

Table 12. The comparison of computation time between 4 combination algorithm and efficient algorithm depending on feature type

No.	Data	4 combinations		Feature-optimized Algorithm		Time reduction rate
		Highest accuracy	Time complexity (sec.)	Highest accuracy	Time complexity (sec.)	
1	Parkinson	94.9%	40	94.9%	30	26.3%
2	Sonar	87.8%	229	87.8%	120	47.7%
3	Heart	100.0%	1.8	100.0%	1.8	0.0%
4	Ionosphere	98.6%	113	98.6%	68	39.7%
5	Breast (diag)	100.0%	53	100.0%	1.2	97.7%
6	Breast	98.5%	35	98.5%	25	30.4%
7	Australian	91.3%	57	91.3%	28	50.3%
8	Diabetes	78.4%	42	78.4%	34	21.0%
9	German	76.0%	176	76.0%	66	62.6%
10	Rookie	73.9%	279	73.9%	230	17.4%
Average		89.9%	103	89.9%	60	39.3%

3.4 Efficient Algorithm for Multiclass Data

This research also conducted an experiment to search for efficient algorithm for multiclass data. As same as the 2 class data sets, the characteristics of feature of all data sets were investigated and divided into metric or categorial. The result is represented in Table 13 and Table 14.

Table 13. Characteristics of multiclass data

No.	Data	Number of class	Number of instances	Number of feature	Number of Numerical feature	Ratio of Numerical feature
1	Iris	4	150	4	4	100.0%
2	Vehicle	4	846	18	18	100.0%
3	Soybean	15	266	35	35	100.0%
4	Contraceptive	3	1473	9	2	22.2%
5	Dermatology	6	358	34	1	2.9%
6	Zoo	7	101	16	0	0.0%
7	Flare	6	1389	12	0	0.0%

Table 14. The characteristics of features of multiclass data (Metric/Categorical)

No	Data	Feature Type																																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	Iris	M	M	M	M																															
2	Vehicle	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M																		
3	Soybean	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
4	Contraceptive	M	C	C	M	C	C	C	C	C																										
5	Dermatology	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	M	
6	Zoo	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C																				
7	Flare	C	C	C	C	C	C	C	C	C	C	C																								

In the same way, the ratio of the number of metric features to the number of total features is calculated, and the preprocessing methods (Rank or PCA) and kernels which produce the highest CCR per data were identified. The results are summarized as follows in Table 15.

Table 15. The result of preprocessing methods, kernels, efficient algorithm for 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	

According to the analysis from the above result, the dedicated algorithms for each feature type (metric or categorical) are not identified in case of multiclass data. Rank method with Sigmoid / RBF kernel can produce the highest CCRs of all multiclass data. As same as the case of 2-class data, the efficient algorithm reduces the computation time by 70% compared to that of 4-combination algorithm by using the RBF kernel only 1 time with Rank method, instead of using it for both preprocessing methods in 4-combination algorithm. The average computation time of all data sets, 99 seconds, is reduced to 30 seconds with slightly improved CCRs compared to that from 4-combination algorithm as represented in Table 16. In case of Iris data and Zoo data, the computation time is increased by 171% and 167% respectively. However, the increased computation time of efficient algorithm is less than 4 seconds, which is far less than the computation time of efficient algorithm of other multiclass data.

Table 16. The comparison of computation time between 4 combinations and efficient algorithm depending on feature type in case of multiclass data

No.	Data	4 combinations		Feature-optimized Algorithm		Time reduction rate
		Highest accuracy	Time complexity (sec.)	Highest accuracy	Time complexity (sec.)	
1	Iris	96.7%	1.4	100.0%	3.8	-171.4%
2	Vehicle	81.7%	135	81.7%	36	73.1%
3	Soybean	94.3%	112	94.3%	52	53.5%
4	Contraceptive	54.1%	161	54.1%	36	77.9%
5	Dermatology	98.6%	117	98.6%	50	57.5%
6	Zoo	100.0%	0.6	100.0%	1.6	-166.7%
7	Flare	76.3%	164	76.3%	30	81.5%
Average		86.0%	98.7	86.4%	29.9	69.7%

The selection of ensemble classifiers from the preprocessing methods and kernels in SVM in the efficient algorithms depending on number of class of data and the ratio of metric feature, are summarized in Table 17.

Table 17. Summary of ensemble classifiers in improved algorithms depending on number of class and feature types

Class	Ratio of numerical feature	Rank				PCA			
		Polynomial	Sigmoid	Radial	Linear	Polynomial	Sigmoid	Radial	Linear
2 class	Higher than 90%				O		O	O	O
	Lower than 90%	O	O	O	O				
Multiclass	Higher than 90%		O	O					
	Lower than 90%		O	O					

3.5 Comparison of Performance with Approaches in Literature (Multiclass data)

This research compared the performance of efficient algorithm with other feature selection method (SVM-RFE) and parameters optimization method (Taguchi method,

Huang et al. 2014) for multiclass data. The performance of the efficient ensemble algorithm is better than that of SVM-RFE-Taguchi. The average correct classification rate of efficient ensemble algorithm is higher than that of SVM-RFE-Taguchi by 3.1%. However, the average feature reduction rate of the efficient algorithm is less than that of SVM-RFE-Taguchi by 20%. The result of comparison of performance is summarized in Table 18.

Table 18. Comparison of performance with SVM-RFE-Taguchi in literature

No.	Data	SVM-RFE-Taguchi			Optimized Algorithm			Difference	
		Accuracy (%)	Original number of feature	Number of selected feature	Accuracy (%)	Original number of feature	Number of selected feature	Accuracy (%)	Feature reduction rate
1	Dermatology	95.4	34	23	98.6	34	34	3.2	-47.8%
2	Zoo	97.0	12	12	100.0	12	8	3.0	33.3%
Average		96.2	23	18	99.3	23	21	3.1	-20.0%
Cross-validation		Training 80%, Testing 20%							

3.6 Comparison of Four Feature Ranking Criteria

This research also used LDA as a feature ranking criteria. The CCRs of individual features can be calculated by using LDA. As same as the feature classification rate ranking method and PCA method, the features can be rearranged in input subset according to the high CCR from LDA. This research implemented the same methodology on all data sets to obtain increase of generality of the experimental result. This research also employed the distance between classes as criteria of rearranging features in input subset to SVM, and then applied this methodology on all data to validate its effectiveness. The comparison result of CCR among the 4 kinds of feature ranking or feature extraction methods is summarized in Table 19 and graphically represented in Figure 21 through 24. In case of 2-class data, the CCR of the 4 methods are almost at the same level. The average and variation

of CCR is $87.55 \pm 0.45\%$. In case of multiclass data, the distance between classes is the most effective. The average CCR on 7 multiclass data is 95.0%, more than 10% higher than other feature ranking criteria, i.e., CCR from SVM, PCA and LDA. In case of the linear SVM with feature ranking by distance between classes, one vs. all multiclass classification architecture is applied. The distance between classes is effective feature ranking criteria on multiclass and large-scale data with large number of instances or high dimension. The effectiveness is also strengthened by one vs. all multiclass classification architecture.

Table 19. Comparison of four 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

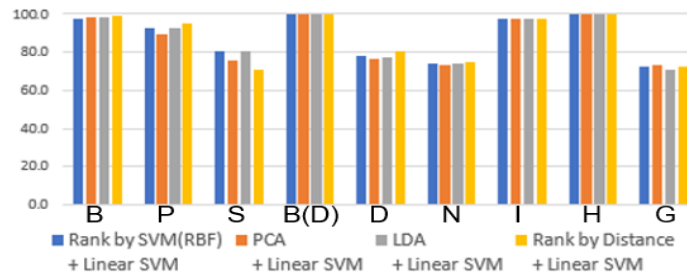


Figure 21. Comparison of CCR per each 2-class data from 4 feature ranking criteria

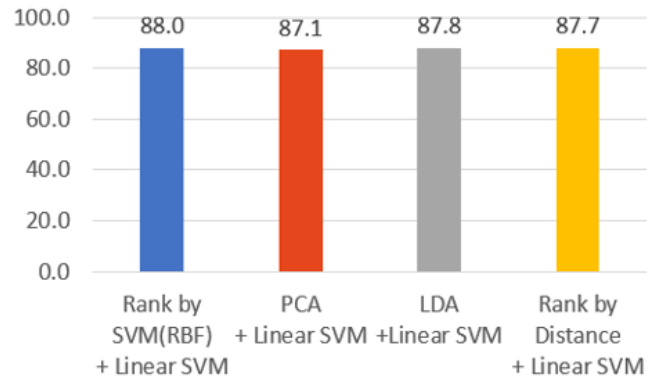


Figure 22. Comparison of average CCR from 4 feature ranking criteria on 2-class data

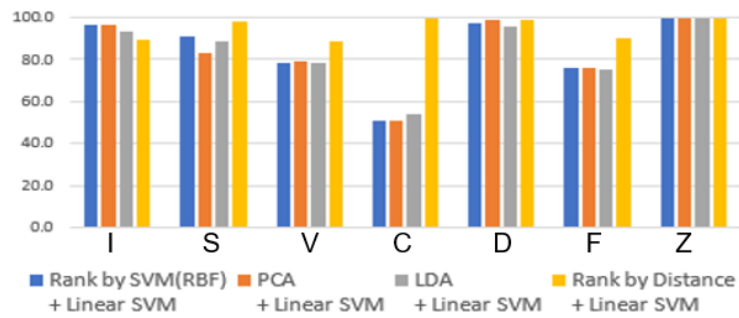


Figure 23. Comparison of CCR per each multiclass data from 4 feature ranking criteria

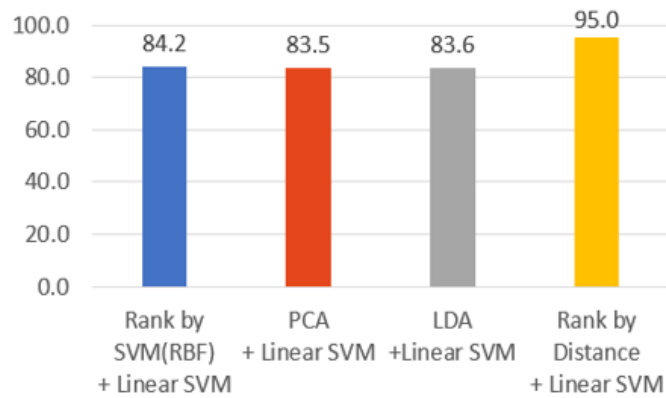


Figure 24. Comparison of average CCR from 4 feature ranking criteria on multiclass data

Chapter 4. Experimental Result on Cardiotocography Data

4.1 Boosted Feature Selection of Cardiotocography Data

First of all, the CCRs of individual features in the data are checked by using SVM with RBF kernel. The results are represented in Table 20 and Figure 25.

Table 20. The CCRs of individual features in Cardiotocography data

One Vs. All Classification	CCR of Individual features																					The highest CCR
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
C1 vs. C2&C3	79.0	77.8	77.8	78.8	77.7	78.1	81.5	83.2	85.2	81.7	76.8	80.6	78.4	77.5	77.7	77.8	81.1	80.8	78.8	80.8	77.8	85.2
C2 vs. C1&C3	86.5	86.1	86.0	86.1	86.1	86.1	86.1	86.4	88.0	86.7	86.0	85.6	86.4	86.1	86.1	86.1	86.4	86.1	85.9	85.9	86.1	88.0
C3 vs. C1&C2	91.7	91.7	91.7	91.6	91.5	92.0	94.4	92.7	91.5	93.2	91.6	92.5	91.4	91.7	91.6	91.7	94.5	94.4	93.2	91.3	91.7	94.5

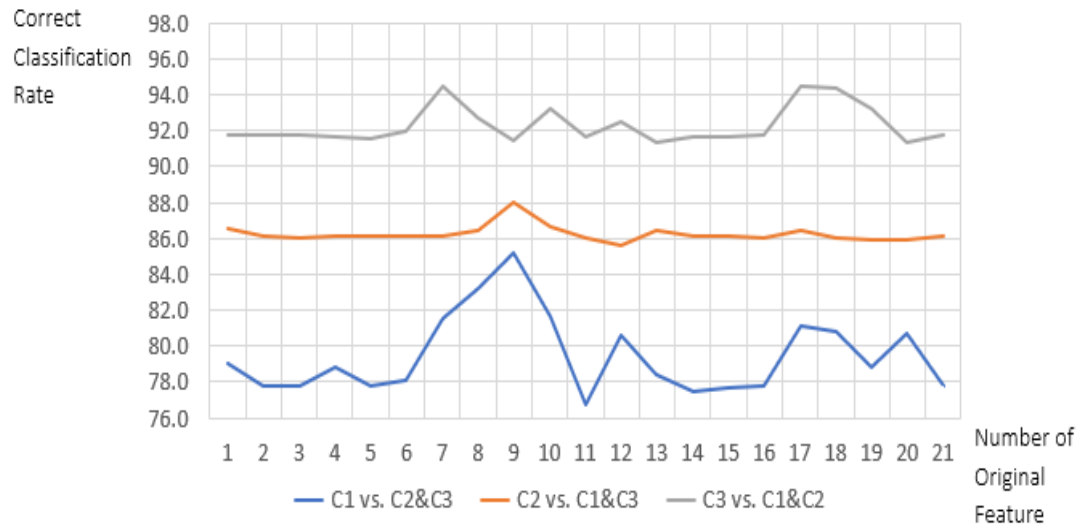


Figure 25. The CCRs of individual features in Cardiotocography data

According to the result, the CCR of C3 vs. C1&C2 is the highest (94.5%) and the CCR of C1 vs. C2&C3 is the lowest among the 3 one vs. all classifications (85.2%). C1 vs. C2&C3 shows the highest variation across the features. As a next step, the distances between two classes are calculated by using the misclassified instances per each one vs. all classification. The numbers of used instances are 151, 179 and 61 respectively per each one vs. all classification as summarized in Table 21.

Table 21. The number of instances used for calculating distance between classes

Class	Number of Instances used for Calculating Distance between Classes
1 vs. 2&3	151
2 vs. 1&3	179
3 vs. 1&2	61
Average	130.3

Then, the distances between 2 classes are calculated per each feature. The results are graphically represented in Figure 26. The distances of features at C3 vs. (C1&C2) is different from that from the other 2 classifications with high degree.

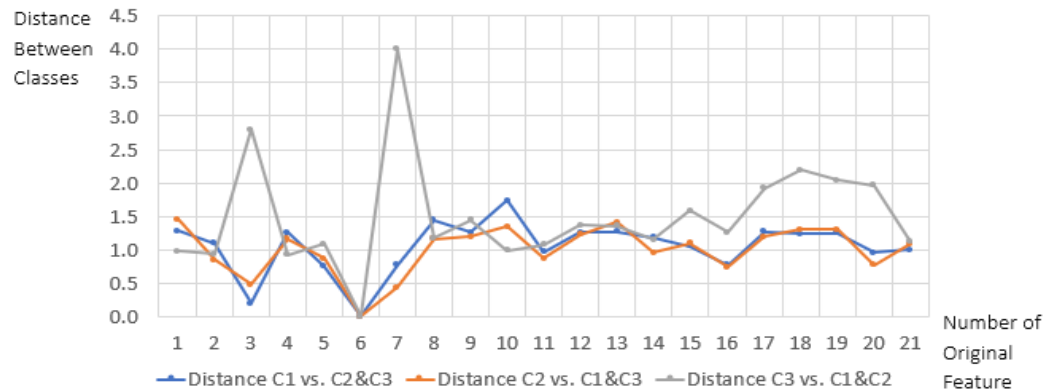


Figure 26. The distance between two classes in original order of features in the 3 one vs. all classifications.

The features are rearranged according to descending order of the distance as shown in Figure 27.

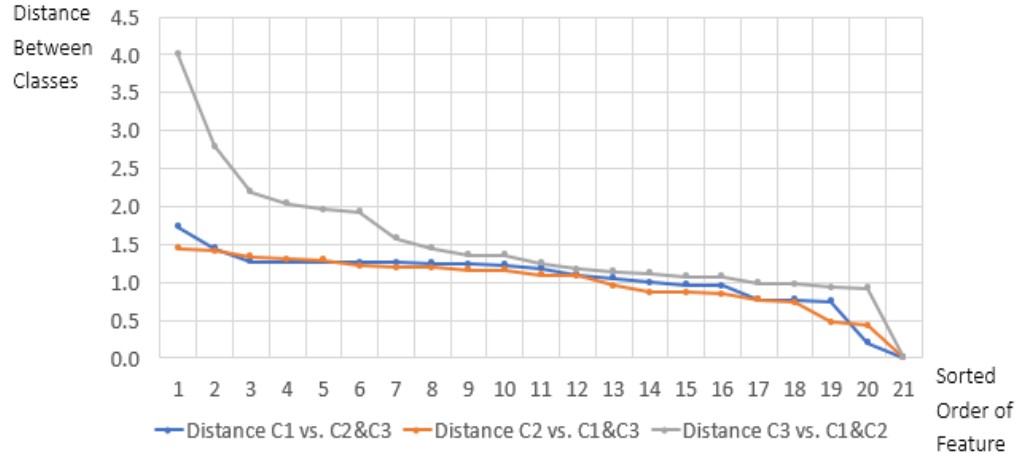


Figure 27. Sorted features in descending order of the 3 one vs. all classifications

The performance of boosted feature selection methodology for Cardiocography data is compared with SVM without feature selection as summarized in Table 22.

Table 22. Comparison of performance between SVM without and with boosted feature selection

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

The average CCR improved by 1.0% compared to the case without feature selection. In each of all the 3 classifications of one vs. all for multiclass, i.e., 1 vs. 2&3, 2 vs. 1&3, 3 vs. 1&2, the CCRs improved by 1.4%, 1.3% and 0.3% respectively. The average feature

reduction rate is 49.2% compared to the case without feature selection. The combination of selected features for each classification are different. The graph in Figure 28 is the representation of the change of CCRs per each classification as the input features cumulatively increase. The number of features for each classification are 11,10 and 11 respectively. The highest CCRs per each of the one vs. all classification are indicated by red circles.

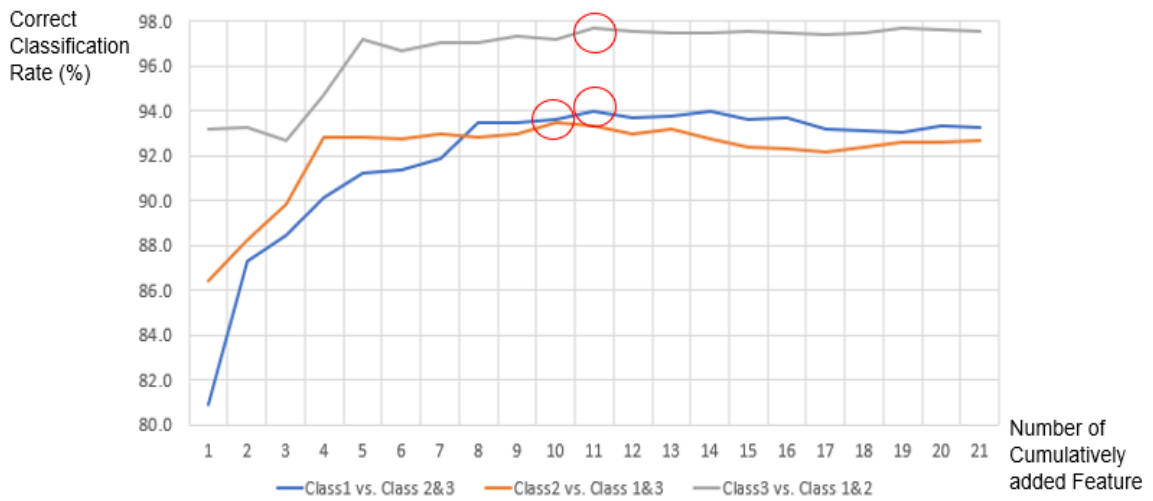


Figure 28. The graphical representation of the change of CCR as the input features cumulatively increase

The features selected for each of one vs. all classification is represented in Table 23.

Table 23. The features selected for each one vs. all classification

(C1:Normal, C2: Suspect, C3:Pathologic)

One vs. all Classification	The Number of Selected Features																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
C1 vs. C2&C3	0			0				0	0	0		0	0	0			0	0	0		
C2 vs. C1&C3	0			0				0	0	0		0	0				0	0	0		
C3 vs. C1&C2			0				0		0			0	0		0	0	0	0	0	0	

Among the 6 misclassification types of the 3 kinds of one vs. all classifications, reducing the misclassification type from original class P (Pathologic) to predicted class N

(Normal) or S (Suspect) is the most meaningful because the purpose of Cardiotocography is to diagnose the fetal condition of pregnant women and detect any disease in advance of delivery of baby. According to the above experimental result, the proposed methodology reduces the number of misclassifications of this type from 51 to 46, which is 9.8% reduction rate, which is represented in Figure 29. Reducing it is the same as increasing sensitivity or True Positive, which increased by 2.8% from 71.0% in the case without feature selection to 73.9% to the case with boosted feature selection. In addition, the specificity, True Negative, increased by 0.2% from 99.5% to 99.7%. This improvement was shown in Table 24.

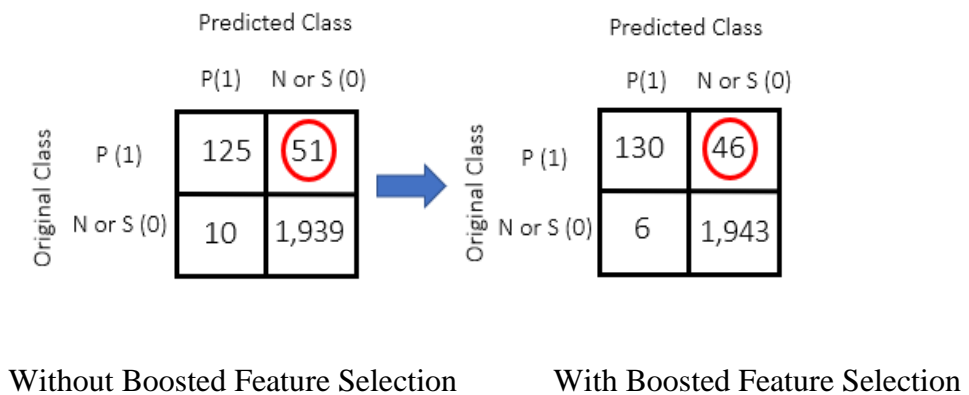


Figure 29. The Classification matrix of P vs. N or S Classification

Table 24. The Comparison of sensitivity and specificity between the cases without and with boosted feature selection on Cardiotocography data.

	Without Feature Selection	With Boosted Feature Selection	Increase
Sensitivity	71.0%	73.9%	2.9%
Specificity	99.5%	99.7%	0.2%

According to the result of comparison of True Positive Rate (TPR) and False Positive Rate (FPR) without feature selection and boosted feature selection, the latter case is located in the point indicated by red point in Figure 30. The area under ROC curve (AUC) gets closer to 1 by the boosted feature selection, which means that the performance of the proposed methodology improves compared to the case without feature selection.

Table 25. The Comparison of TPR and FPR between the cases without and with boosted feature selection on Cardiotocography data.

	Without Feature Selection	With Boosted Feature Selection	Increase
True Positive Rate	0.710	0.739	0.029
False Positive Rate	0.005	0.003	-0.002

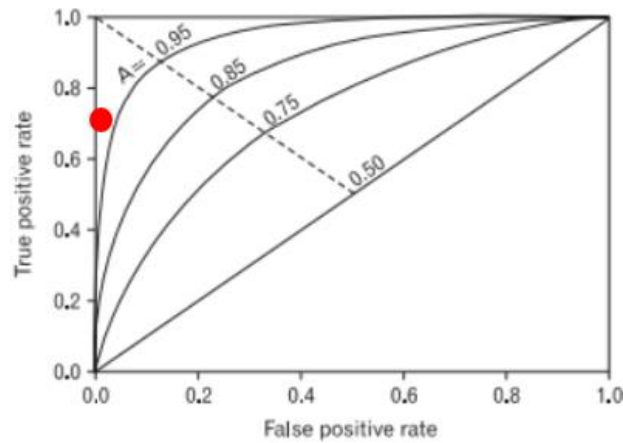


Figure 30. The TPR and FPR on ROC Curve in case of boosted feature selection

To verify the effectiveness of the proposed methodology, it is applied to other data set, which is Contraceptive data set. This data has 1,473 instances, 9 features and 3 classes. The features consist of 2 metric and 7 categorical features. The experimental result is shown in Table 26. The correct classification rate improves by 1.8% and the average feature reduction rate is 40.7%. The improvement of correct classification rate is 0.8 % higher than

that from Cardiotocography data and the average reduction rate is 8.5% lower than that from Cardiotocography. However, the improvements are significant and finally it is proved that the proposed methodology is effective feature selection methodology.

Table 26. Comparison of performance between SVM without and with boosted feature selection on Contraceptive data set

Class	Number of Instances used for Calculating Distance between Classes	SVM (RBF) with No Feature Selection (A)		Rank by Distance using misclassification 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

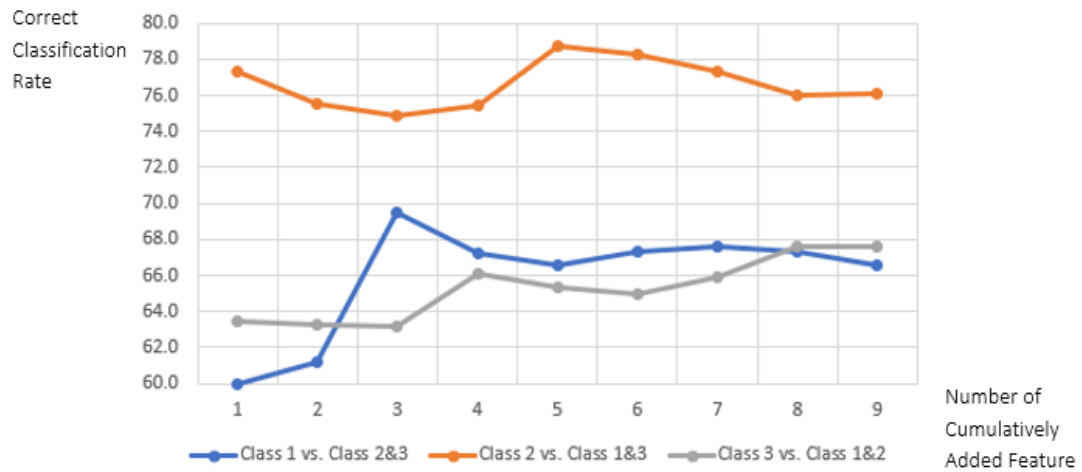


Figure 31. The graphical representation of the change of correct classification rate as the input features cumulatively increase (Contraceptive data set)

The features selected for each of one vs. all classification is represented in Table 27.

Table 27. The features selected for each one vs. all classification (Contraceptive data)

One vs. all Classification	The Number of Selected Features								
	1	2	3	4	5	6	7	8	9
C1 vs. C2&C3	O						O		O
C2 vs. C1&C3	O			O			O	O	O
C3 vs. C1&C2	O	O	O	O		O	O	O	O

4.2 Validation of Boosted Feature Selection by Applying to Other Classifiers

The experimental results on 5 binary-class data is shown in Table 28 and Figure 32. In case of Cardiotocography data and Contraceptive data, which are 3-class data, 2 classes out of 3 classes are selected to experiment the performance of 2-class data. In addition, another large size of data, Bank Marketing data, is included in the experiment to verify the effect of size of data more accurately.

Table 28. Comparison of performance of boosted feature selection with other classifiers on 5 binary class data

Data	Class	Number of Instances	Number of Features	Number of Instances X Number of Features	AdaBoost					Random Forest		
					No Feature Selection	PCA	Increase	Mis by DT + Distance + Wrapper	Increase	No Feature Selection	Mis by DT + Distance + Wrapper	Increase
Cardiotocography (S vs. P)	2	471	21	9,891	94.7	94.7	0.0	94.7	0.0	96.8	96.8	0.0
Contraceptive (1 vs. 3)	2	1,140	9	10,260	64.9	63.2	-1.7	66.7	1.8	66.7	69.3	2.6
German	2	1,000	20	20,000	71.5	77.0	5.5	76.0	4.5	71.0	77.5	6.5
Cardiotocography (N vs. P)	2	1,830	21	38,430	97.8	98.9	1.1	99.5	1.7	98.4	98.9	0.5
Bank Marketing	2	45,211	16	723,376	99.74	99.84	0.10	99.79	0.05	99.82	99.85	0.03
Average	2	9,930	17	160,391	85.7	86.7	1.0	87.3	1.6	86.5	88.5	1.9

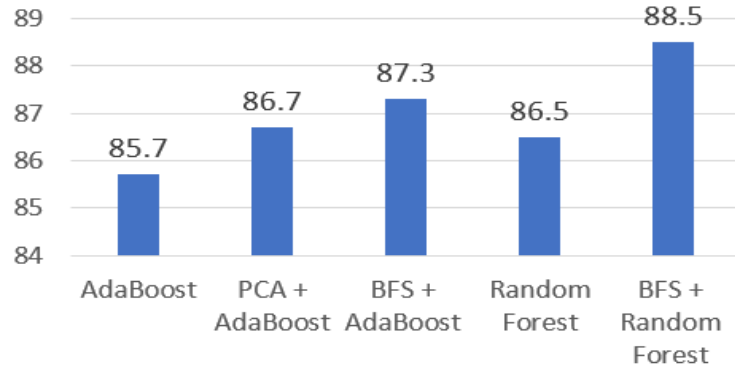


Figure 32. Comparison of CCR (%) among ensembles with different feature selection /extraction methods

According to the experimental result, the boosted feature selection increases the CCR of AdaBoost and RF by 1.6% and 1.9% respectively. In case of AdaBoost, the boosted feature selection is more effective than PCA. After implementing RF with boosted feature selection methodology on totally 13 binary class data, the condition of data which can maximize the increase of CCR, is discovered. The increase of CCR is proportional to the number of instance and the number of features, and reversely proportional to the CCR of RF with no feature selection. As the size of data gets larger, the increase of CCR increases even if the change is not strictly proportional. Other unique and inherent characteristics in each data affects the increase.

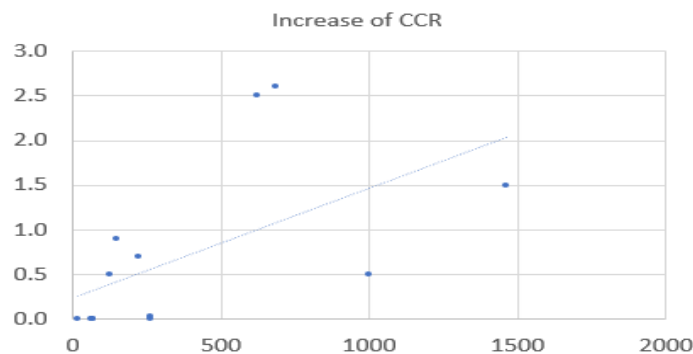


Figure 33. The relation represented by linear regression

$X: (100\text{-CCR of Random Forest with no feature selection}) \times 0.01 \times \text{number of instances} \times 0.25 \times \text{number of features}.$

$Y: \text{Increase of CCR (\%)} \text{ compared to RF with no feature selection}$

4.3 Improved Classification Methodology for Cardiotocography Data

The experimental results are summarized in Table 29. The experiments are implemented by using 2 kinds of classification architectures, i.e., BDT and 3-class dedicated SVMs. The 2 architectures are compared based on the same methodology, SVM with RBF kernel and no feature selection to measure the difference of performance due to the difference of the classification architecture.

Table 29. Comparison of performance among various methodologies in 2 classification architectures on Cardiotocography Data

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

The comparison result is graphically represented in Figure 34. Compared to BDT, CCR of 3-class dedicated architecture increases by 3.2% while the sensitivity decreases by

0.066. and the specificity is the same. The low sensitivity is the disadvantage of the 3-class dedicated architecture compared to BDT.

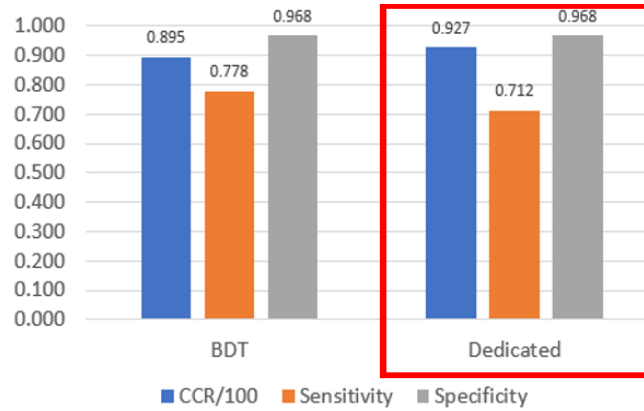


Figure 34. Comparison of performance between BDT and Class-dedicated classification architecture

The performance of the proposed methodology, SVM with boosted feature selection and clustering, is compared to that of (1) SVM with RBF kernel and no feature selection, (2) AdaBoost (3) RF, based on the Class-dedicated architecture. The comparison results are graphically represented in Figure 35. According to the result, the CCR of the proposed methodology is 96.3%, which is 5.7% higher than that of (1). The sensitivity of the proposed methodology increases by 0.265 and specificity increases by 0.018, compared to (1), resulting the values close to 1. The sensitivity of the proposed methodology is even higher than that of AdaBoost or RF by 0.102.

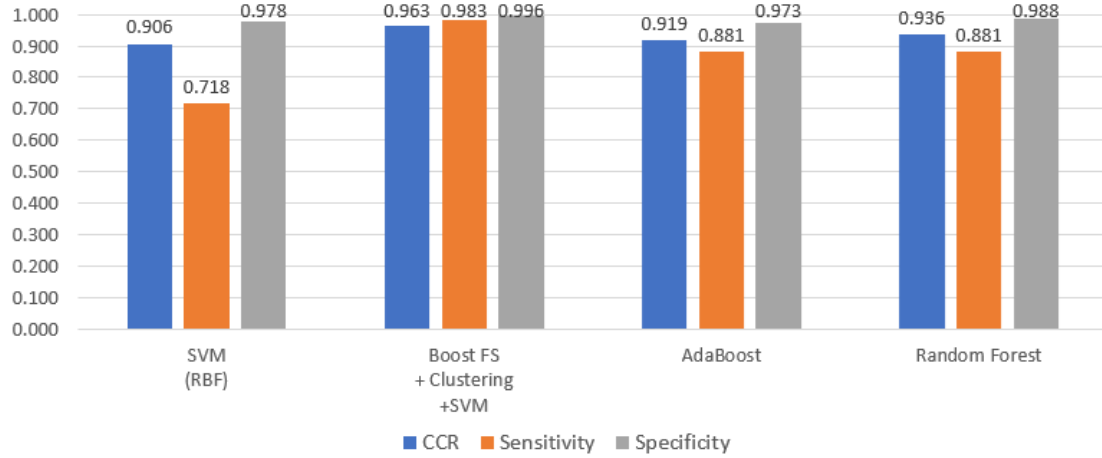


Figure 35. Comparison of performance of 4 methodologies with Class-dedicated architecture

The performance of the proposed methodology is compared to the highest performance in literature related to Cardiotocography data, Yilmaz & Kilikcier, 2013. In addition, SVM (RBF kernel) with all features is implemented by using (A) BDT and (B) Dedicated SVM architecture respectively to compare the difference of performance due to the difference of classification architecture. The experimental result shows that the CCR of (B) increases by 3.2% while the sensitivity decreases by 0.07, compared to (A).

In case of AdaBoost and RF, the sensitivity increased by 0.05 and 0.1 respectively compared to (A). In case of the proposed methodology, CCR increases by 9.0%, and the sensitivity and specificity increase by 0.205 and 0.03 respectively, compared to (A). In summary, the performance of proposed classification methodology outperforms all the other methodologies, i.e., SVM without preprocessing, AdaBoost and RF.

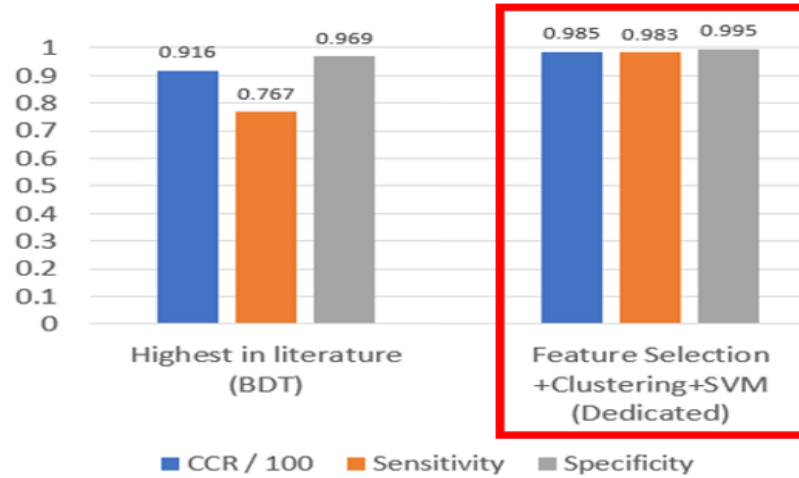


Figure 36. Comparison of performance with the methodology in literature

The proposed methodology outperforms the highest performance in literature in terms of all criteria by overcoming the disadvantage of the class-dedicated architecture, which is the low sensitivity.

Table 30 shows the experimented various models with different number of clusters to search for appropriate number of clusters. In model B, the k is determined within target class and within merged other classes while it is determined within each class in other models, i.e., A, C, D & E. The model B and C show the highest performances in terms of sensitivity and specificity.

Table 30. Comparison of various models in determining optimum number of clusters (Training data 75%, Testing data 25%)

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

Figure 37 shows the difference of the clustering structure between model B and model C. Even if the model C shows a slightly higher performance in terms of sensitivity and specificity compared to model B, model B is computationally more efficient because the number of extracted features of model B is 5 while it is 8 in model C. In model B, the other 2 classes are merged in clustering while the clustering is implemented in each class in model C.

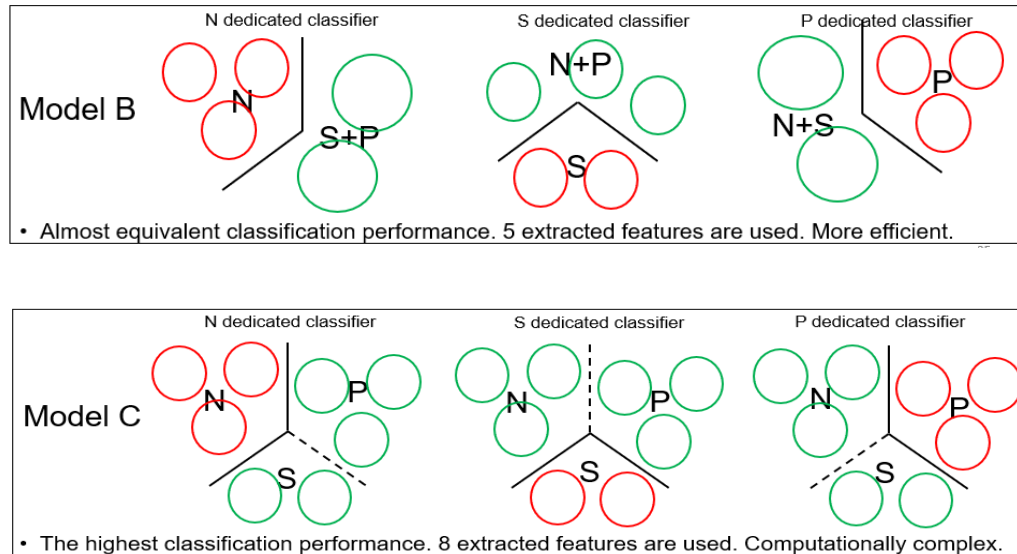


Figure 37. Comparison of clustering architecture, Model B vs. Model C.

The sensitivity and specificity of the 5 models are graphically represented in Figure 38. The sensitivity and specificity of the model B and C are at the highest level among the 5 models.

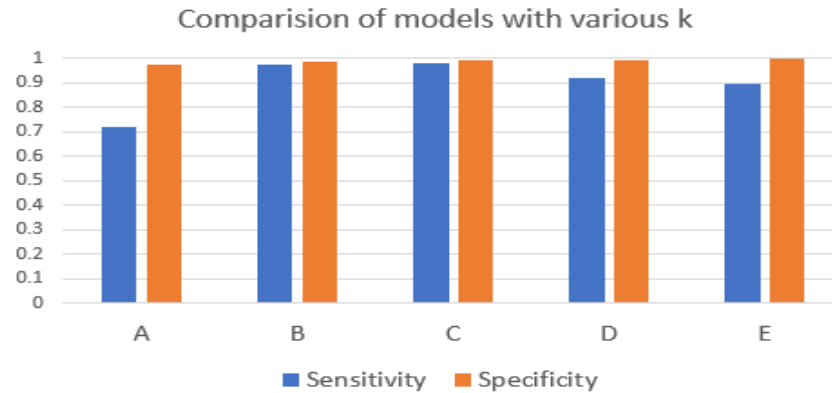


Figure 38. Comparison of sensitivity and specificity in the models with various numbers of clusters

4.4 Procedure to Search for the Highest Performance Model

Figure 39 is the tree diagram to search for the highest performance model. Firstly, the number of clusters is calculated by elbow method. The model with the obtained k needs to be checked to see whether the classification performance of the model is improved compared to literature or not. If the performance is improved, merging other classes and determining k is recommended to reduce features further. If merging other classes and determining k results in performance improvement, the obtained k can be used as the final number of clusters. If the merging other classes and determining k decreases performance, the other classes should not be merged in determining k. If classification performance is not improved at the first step after applying elbow method, the k in the class of low CCR should be increased until the performance is not improved anymore.

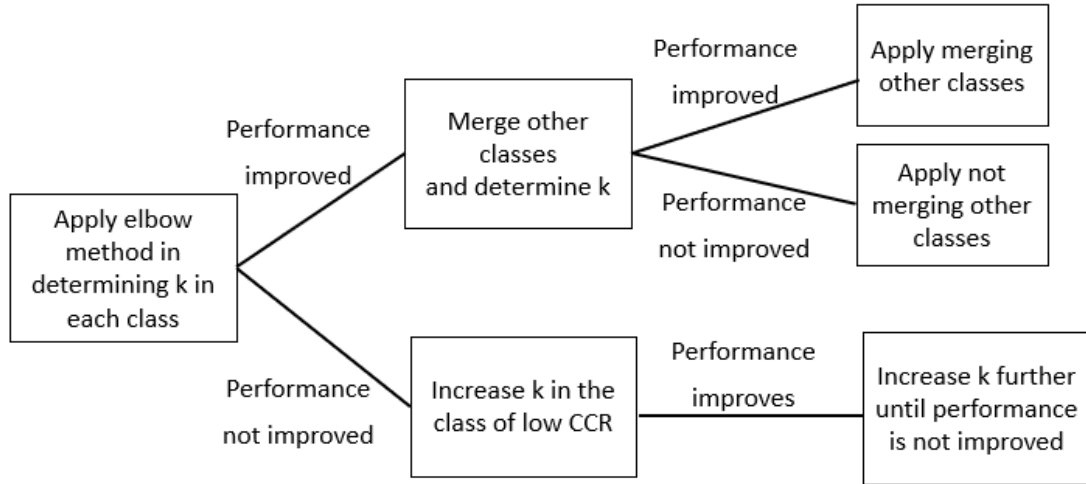


Figure 39. Tree diagram to search for the highest performance model

4.5 Procedure to Apply to Diagnosis Activity

Figure 40 shows the tree diagram for diagnosing fetal state by using the proposed classification methodology. Firstly, predictive values of each of the 3 classes need to be calculated by using the equation (33), (34) and (35).

$$\text{Normal Predictive Value} = \frac{TP_1}{TP_1 + FN_1} \quad (33)$$

$$\text{Suspect Predictive Value} = \frac{TP_2}{TP_2 + FN_2} \quad (34)$$

$$\text{Pathologic Predictive Value} = \frac{TP_3}{TP_3 + FN_3} \quad (35)$$

In case of model B in Table 73, Normal predictive value, Suspect predictive value and Pathologic predictive value are 95.1%, 98.1% and 100.0% respectively. The tree diagram uses the class-dedicated SVM by prioritizing the class with higher predictive value.

The class 3-dedicated SVM is used at first because the class of the highest predictive value is class 3. If the prediction result from class 3-dedicated SVM is class 3, the outcome is Pathologic. If the prediction result is class 1 or class 2, the outcome depends on the result from class 2-dedicated SVM. The class 1-dedicated SVM is not used because the predictive value is the lowest.

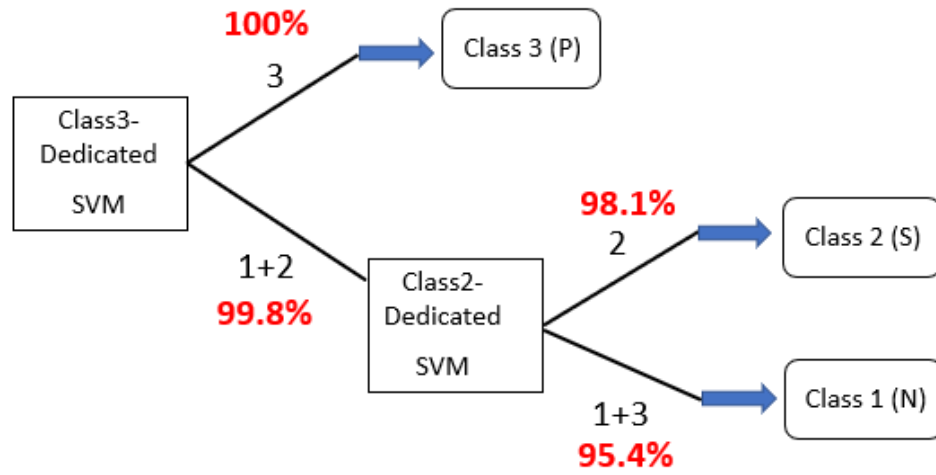


Figure 40. Tree diagram for diagnosing fetal state

4.6 Verification of Methodology by Applying to Other Multiclass Data

To verify the effectiveness of the proposed methodology, it is applied to 2 other multiclass data sets, i.e., Contraceptive data and Vehicle data. In addition to CCR, sensitivity and specificity are measured. In case of non-disease data, which don't have positive or negative classes, the equivalent criteria, i.e., the CCR of each class, is measured alternatively. The CCR of each class is defined in equation (36).

$$\text{CCR of each class} = \frac{\text{Correctly predicted number of instances in the class}}{\text{Original number of instances in the class}} \quad (36)$$

Table 31. Comparison of performance among various methodologies on 3 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 CCR of the proposed methodology increases by 32.3% on Contraceptive data, and by 15.2% on Vehicle data, compared to that of SVM with RBF kernel and no feature selection, as shown in Figure 41. The improvement of performance is more significant in the data with lower CCR.

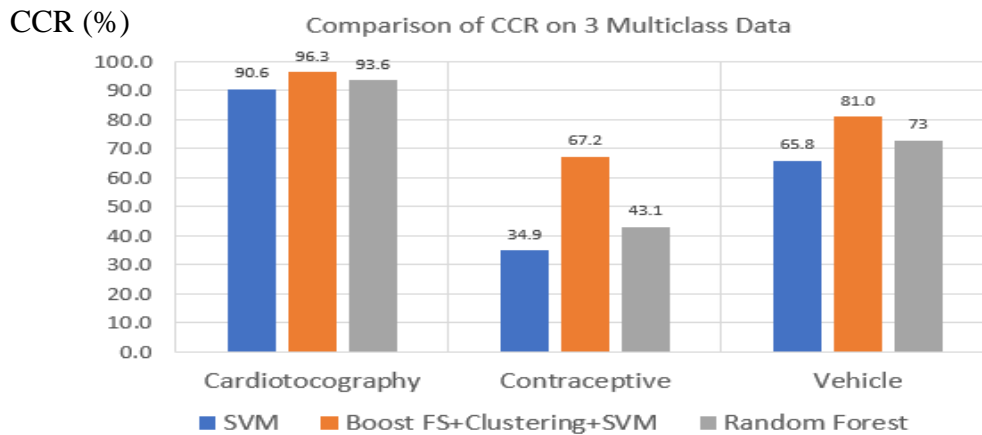


Figure 41. Comparison of CCR (%) of 3 methodologies on 3 multiclass data

The CCR of each class increases by 0.352 (Average/Class) on Contraceptive data, and by 0.139 (Average/Class) on Vehicle data. The comparison result of the 4 methodologies on the 2 data is represented in Figure 42 and Figure 43.

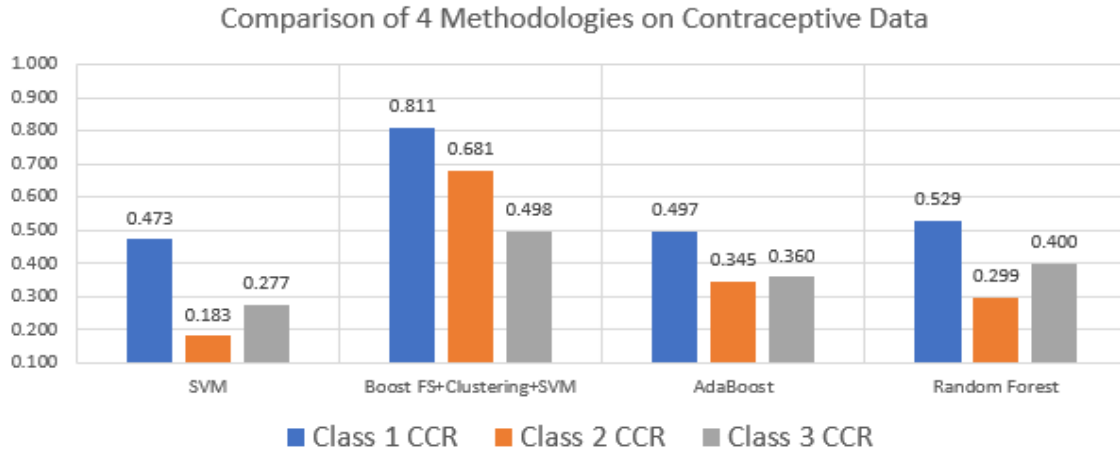


Figure 42. Comparison of each class CCR/100 in 4 methodologies on Contraceptive data

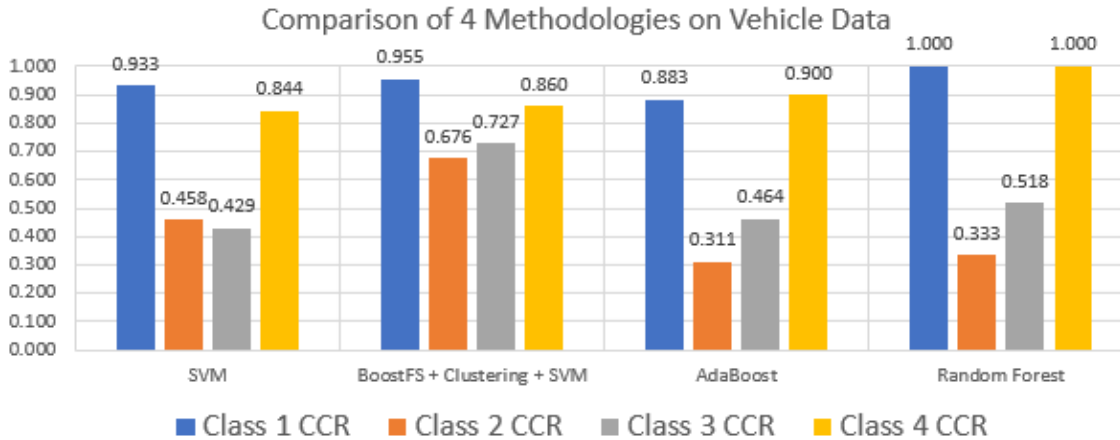


Figure 43. Comparison of each class CCR/100 in 4 methodologies on Vehicle data

According to the experimental result, the proposed methodology is effective in increasing performance of multiclass data in terms of all criteria, i.e., CCR, sensitivity, specificity. The proposed methodology especially increases low sensitivity and specificity,

or equivalent CCR of each class of multiclass data, compared to other classification methodologies. The number of clusters, k can be determined by adjusting k in tuning process. The optimal number of k obtained from elbow method can be considered as 1st candidate k numbers. However, it does not always result in improved model. Even in that case, the overall CCR can still be lower than those from other methodologies. In this case, increasing k in the class of low CCR is effective in increasing overall and the class CCR. The performance of proposed methodologies in Vehicle data in Figure 47, is obtained by k tuning process. Merging classes in clustering is an effective method in reducing number of extracted features, consequently reducing computational complexity for training. However, this method does not always result in the highest performance model. In conclusion, the proposed methodology is expected to significantly increase classification performance of multiclass data in various application fields including medical diagnosis with Cardiotocography data.

Chapter 5. Conclusion

The result and finding of this research are summarized as follows.

- The proposed ensemble algorithms result in highest CCR by using both individual feature correct classification rate-based ranking method and PCA in preprocessing step complementarily.
- The ensemble composed of 2 preprocessing methods, 3 kernels and linear SVM is effective in achieving highest CCR regardless of the characteristics of data.
- The time required in grid search of parameters of RBF kernel can be significantly reduced by using reduced training data while maintaining equivalent CCR.
- As a result of experimenting various combinations of preprocessing method and kernels, 4-combination algorithm was turned out to reduce overall computation time by 75% while maintaining the same CCR.
- In order to decrease the time complexity further, this research experimented 6 and 2 combination algorithms, respectively. 6-combinations reduces the time further by 84% at the expense of degraded CCR by 0.8%. 2-combination algorithm, which consists of only linear SVMs and does not need the time for parameter optimization, reduces computation time by 97%, but degrades CCR by 1.8%.
- The feature reduction rate of the 4 algorithms are almost same level. The primary goal of the proposed ensemble algorithms is to obtain highest CCR.

- The most critical factor in determining the time for parameter optimization is the number of instances. The 2nd influencing factor is the number of features. As the scale of data gets larger, the proposed ensemble algorithms get more effective in reducing time complexity.
- The performance of the proposed algorithm with 4 combinations almost equivalent to that from genetic algorithm (GA) in literature in terms of CCR and feature reduction rate. However, it is significantly more cost-effective than GA in terms of time complexity by reducing computation time more than 75%.
- The performance of the proposed algorithm with 4 combinations is better than GA-ensemble composed of ANN, Decision tree & SVM in other research. The CCRs are higher by 4.5% with equivalent feature reduction rate.
- The efficient ensemble algorithms depending on the feature type were explored to reduce computation time further. As a result, 2 efficient algorithms depending on the feature type were developed for 2-class data, which reduces the computation time further by 39% compared to 4 combination algorithms. In case of multiclass data, 1 efficient algorithm regardless of the characteristics of feature, were developed, which reduces the computation time further by 70% compared to 4 combination algorithms.
- The proposed algorithm significantly reduced the time complexity of SVM while achieving slightly higher or equivalent CCR compared to other recent approaches. The outcome of this research can be utilized in the various application fields where the reduced time complexity of classification is critical.

- According to the comparison result of 4 feature ranking criteria, in case of 2-class data, the CCR of the 4 methods are almost at the same level. In case of multiclass data, the distance between classes is the most effective. Distance between classes is effective on large-scale and multiclass data.
- As a result of applying one vs. all multiclass classification method, preprocessing method by SVM with RBF kernel, distance between classes and the wrapper method, the CCR on Cardiotocography data increased by 1.0% and the feature reduction rate was 49.2% compared to the case without feature selection. The CCR is 3.5% higher than the highest CCR in recent approaches on the same data set in literature.
- The effectiveness of the above proposed feature selection methodology is supported by the result from other multiclass data set.
- The boosted feature selection method which prioritizes the features with the highest discriminant power, K-means clustering and Class-dedicated SVM results in the increase of CCR by 6.9%, and the increase of sensitivity by 0.131 compared to the highest performance in literature related to the classification methodology research on Cardiotocography data.
- The effectiveness of the proposed classification methodology is validated by applying it to other multiclass data sets.

First of all, this research implemented various methodologies in terms of feature selection and extraction, kernel selection in SVM and ensemble methods to improve the performance of SVM. The CCR of proposed methodology outperforms the method in the

same 2-class and multiclass data in literature. The feature reduction rate of the proposed methodology is not always less than literature, however, time complexity is significantly reduced. To maximize the advantage of SVM, which is high correct classification rate, this research searched for the methodology to reduce the computational complexity of SVM especially in case that the data size gets larger. This research proposed efficient methodology for large-scale data by reducing data size for Grid Search of RBF kernel. In addition, this research proposed efficient methodology depending on feature type.

Secondly, boosted feature selection methodology by using feature ranking and wrapper method based on the distance between classes among misclassified instances from SVM is effectively applied to Cardiotocography data. The features are reduced by 49.2% compared to SVM without feature selection. CCR increases by 3.5% compared to literature and increases by 1.0% compared to SVM without feature selection. The result is significant and is also supported by the result of application to other multiclass data set which shows the feature reduction rate 40.7%. The methodology is more effective on data with higher error rate, and is also effective on other classifiers, i.e., AdaBoost and Random Forest.

Finally, this research proposed the classification methodology on Cardiotocography data by using the boosted feature selection, K-means clustering and class-dedicated SVMs. The research contributes in the development of more reliable and efficient decision support system for diagnosing fetal status using Cardiotocography data. The proposed classification methodology achieves CCR 98.5%, which is 6.9% higher than the highest CCR in literature. In addition, the proposed methodology shows the specificity 0.983, which is 0.131 higher than the currently highest sensitivity in literature. The highest sensitivity among classification methodologies on 3-class Cardiotocography data in

literature, is 0.852. The decision support system based on the highest performance in literature will diagnose only 85.2 % of pathologic status of pregnancy as true pathologic class. In other words, 14.8% of pathologic status of pregnancy is classified as suspicious or normal class by the decision support system, and additional medical examination and medical cost are required to diagnose the status as pathologic class, which will degrade the efficiency and reliability of the medical decision support system. In addition, this research contributes in realizing high adaptability of the decision support system for Cardiotocography data. The currently available 3-class data can be used without elimination of suspect class in training process of the classifiers of the proposed methodology. The Cardiotocography data in real diagnosis activity can be used in testing and predicting the fetal status of pregnancy without any separation of data according to the classes. This improvement in performance is resulted from the use of class-dedicated classification architecture instead of BDT classification architecture which is mostly used in literature. The number of features is also reduced to 5 from the least number of features in literature, 7, decreasing the computational complexity further in training process.

The proposed methodology with boosted feature selection, and feature extraction by K-means clustering and fuzzy membership function shows the importance of preprocessing step as an efficient tool to improve classification performance and reduce computational complexity. Basically, the number of clusters by elbow method is used in clustering of each class. However, it does not guarantee the highest performance of classification. Clustering is unsupervised machine learning algorithm which searches for the optimum number of clusters within a data. However, if there are multiple classes in which optimum clustering is independently implemented, the maximizing classification

performance will require different setting of the number of clusters in consideration of the relation between or among classes. It is necessary to increase the number of clusters in the class of low sensitivity, specificity or low CCR to increase the classification performance of the class.

References

Books:

Joseph F. Hair Jr et al. Multivariate Data Analysis, 7th Edition, Prentice Hall, 2010

Hastie, Trevor, The Element of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, 2017

Laurene Fausett, Fundamentals of Neural Networks, Prentice Hall, 1994

J.-S.R. et al. Neuro-Fuzzy and Soft Computing, Prentice Hall, 1997

Leo Breiman, Machine Learning, 45,5-32, Kluwer Academic Publishers, 2001

Raghav Bali & Dipanjan Sarkar, R Machine Learning by Example, Packt Publishing, 2016

Richard S. Sutton and Andrew G. Barto, Reinforcement Learning, The MIT Press, 2017

Bishop, C.M. Pattern Recognition and Machine Learning, Springer. 2006

Papers:

Wang, Z. et al. (2007) MultiK-MHKS: A Novel Multiple Kernel Learning Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-8.

Bhavsar, H. and Ganatra, A. (2012) Variations of Support Vector Machine Classification Technique: A survey. *International Journal of Advanced Computer Research*, 2(4), 223-227.

Wang, Z. et al. (2014) Multi-kernel classification machine with reduced complexity, *Knowledge-Based Systems*, 65, 83-95.

Abdiansah, A. & Wardoyo, R. (2015) Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM, *International Journal of Computer Applications*, 128(3), 28-34.

Bi, J. et al. (2004) Column-Generation Boosting Methods for Mixture of Kernels, *KDD*, August 22-25.

Huang, M.W. et al. (2017) SVM and SVM Ensembles in Breast Cancer Prediction, *PloS one*, 12(1), 1-14.

- Zhai, G. et al. (2015) Material identification of loose particles in electronic devices using PCA and SVM. *Neurocomputing*, 148, 222-228.
- Gao, X., & Hou, J. (2016) An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process. *Neurocomputing*, 174, 906-911.
- Chang, Y.W. & Lin, C.J. (2008) Feature Ranking Using Linear SVM. *JMLR Workshop and conference proceedings*, 3,53-64.
- Huang, C.L., & Wang, C.J. (2006) A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, 31, 231-240.
- Agrawal, R., & Srikant, R. (1993) Fast Algorithms for Mining Association Rules, *IBM Almaden Research Center*, San Jose.
- Everingham, Y. et al. (2016). Accuracy prediction of sugarcane yield using a random forest algorithm, *Agronomy for Sustainable Development*, 36, 1-9.
- Masarat, S. et al. (2016). Modified parallel random forest for intrusion detection systems, *The Journal of Supercomputing*, 72(6), 2235-2258.
- Lei, C. et al. (2017) A random forest approach for predicting coal spontaneous combustion. *Fuel*. 223, 63-73.
- Zheng, H. et al. (2009). Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions. *Agriculture, Ecosystems and Environment*, 132, 98-105.
- Liu, C. et al. (2017) Forecasting copper prices by decision tree learning, *Resources Policy*, 52, 427-434.
- Tayefi, M. et al. (2018) Evaluating of associated risk factors of metabolic syndrome by using decision tree, *Comparative Clinical Pathology*, 27. 215-223.
- Kohavi and John (1997) Wrapper for Feature Subset Selection. *Artificial Intelligence special issue on relevance*, 97, 273-324.
- Guyon, I. & Elisseeff, A. (2003) An Introduction to variable and feature selection. *Journal of machine learning research*, 3, 1157-1182.
- Guyon, I. et al. (2002) Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46, 389-422.

- Huang, M.L. et al. (2014) SVM-RFE based feature selection and Taguchi Parameters Optimization for Multiclass SVM classifier. *The Science World Journal*, Volume 2014, 1-10.
- Samb. M. L. et al. (2012) A novel RFE-SVM-based feature selection approach for classification, *International Journal of Advanced Science and Technology*. 43, 27-36.
- Rodriguez-Galiano, V.F. et al. (2018) Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods, *Science of the total environment*, 624, 661-672.
- Peng, H. et al. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8),1226–1238.
- Sakri, S. et al. (2018) Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction, *IEEE Special Section on Big Data Learning and Discovery*, 6, 29637-29647.
- Unler, A. & Murat, A. (2010) A discrete particle swarm optimization method for feature selection in binary classification problems, *European Journal of Operational Research*, 206, 528-539.
- Trambaiolli, L. et al. (2017) Feature selection before EEG classification supports the diagnosis of Alzheimer's disease, *Clinical Neurophysiology*, 128, 2058-2067.
- Staelin, C. (2002) Parameter selection for support vector machine. *HP Laboratories Israel*, HPL-2002-352 (R.1) 1-4.
- Lebrun, G. et al. (2004) SVM training time reduction using vector quantization, *IEEE*, 0-7695-2128-2/04.
- Maldonado, S. & Weber, R. (2009) A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, 179, 2208-2217.
- Li, S. & Tan, M. (2010) Tuning SVM parameters by using a hybrid CLPSO-BFGS algorithm, *Neurocomputing*, 73, 2089-2096.
- Li, H., & Sun, J. (2011) Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study, *Expert Systems with Applications*, 38, 12747-12756.
- Amami, R. et al. (2013) Practical selection of SVM supervised parameters with different feature representations for vowel recognition. *International Journal of Digital Content Technology and its application*. 7-9, 418-424.

- Singla, A. et al. (2014) A novel classification technique based on progressive transductive SVM, *Pattern Recognition Letters*, 42, 101-106.
- Chen, G. & Chen, J. (2015) A novel wrapper method for feature selection and its applications, *Neurocomputing*, 159, 219-226.
- Martins, S. et al. (2016) Support Vector Machine algorithm optimal parameterization for change detection mapping in Funil Hydroelectric Reservoir, *Model, Earth System, Environment*, 2016. 2-138.
- Lee, S.B. et al. (2017) Parameter Search methodology of support vector machines for improving performance, *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, 7(3), 329-337.
- Ghaddar, B., & Naoum-Sawaya, J. (2018) High dimensional data classification and feature selection using support vector machines, *European Journal of Operational Research*, 265, 993-1004.
- Lin, X. et al. (2018) Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics, *Molecules*, 23(52),1-10.
- Chen, Y., & Lin, C. (2006) Combining SVMs with various feature selection strategies, *Feature extraction, foundations and applications*, Springer, 2006.
- Vakharia et al. (2016) Bearing Fault Diagnosis Using Feature Ranking Methods and Fault Identification Algorithms, *Procedia Engineering*, 144, 343-350.
- Al-Salemi, B. et al. (2018) Feature ranking for enhancing boosting-based multi-label text categorization. *Expert Systems with Applications*, 113, 531-543.
- Jing, C., & Hou, J. (2015) SVM and PCA based fault classification approaches for complicated industrial process. *Neurocomputing*, 167, 636-642.
- Safo, S. & Ahn, J. (2016) General sparse multi-class linear discriminant analysis, *Computational Statistics and Data Analysis*, 99, 81-90.
- Silva, A et al. (2016) Two-dimensional linear discriminant analysis for classification of three-way chemical data, *Analytica Chimica Acta*, 938, 53-62.
- Uncini, A. et al. (2017) Optimizing the electrodiagnostic accuracy in Guillain-Barre syndrome subtypes: Criteria sets and sparse linear discriminant analysis, *Clinical Neurophysiology*, 128, 1176-1183.
- Pujari, P. and Gupta, J.B. (2012) Improving classification accuracy by using feature selection and ensemble model, *International Journal of Soft Computing and Engineering (IJSCE)*. 2-2, 380- 385.

- Zhang, Z. & Yang, P. (2008) An ensemble of classifiers with genetic algorithm based feature selection, *IEEE Intelligent Informatics Bulletin*, 9-1, 18-24.
- Hira, Z.M. & Gillies, D.F. (2015) A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, *Advances in Bioinformatics*, 1-13.
- Coy, Benjamin (2012) Dimension Reduction for Analysis of Unstable Periodic Orbits Using Locally Linear Embedding, *International Journal of Bifurcation and Chaos*, 1-10.
- Yan, Miniun (2013) Dual Adaptive K-SVD Algorithm Based on a Rank Symmetrical Relationship, *Journal of Software*, 8-7, 1550-1555.
- Shen, C. et al. (2014) Generalized canonical correlation analysis for classification, *Journal of Multivariate Analysis*, 130, 310-322.
- Park, H & ADNI (2012) Isomap induced manifold embedding and its application to Alzheimer's disease and mild cognitive impairment, *Neuroscience Letters*, 513, 141-145.
- Liu, X. et al. (2013) Locally linear embedding (LLE) for MRI based Alzheimer's disease classification, *NeuroImage*, 83,148-157.
- Alter, O., & Golub, G.H. (2006) Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening, *PNAS*, 103(32), 11828-11833.
- Bu, Y., et al. (2014) Stellar spectral subclasses classification based on Isomap and SVM, *New Astronomy*, 28, 35-43.
- Feng, S.X, et al. (2008) Support Vector Machine Integrated CCA for Classification of Complex Chemical Patterns, *Second International Conference on Future Generation Communication and Networking*, 21- 24.
- Yilmaz, E & Kilicier, Caglar. (2013) Determination of Fetal State from Cardiotocogram Using LS-SVM with Particle Swarm Optimization and Binary Decision Tree, *Computational and Mathematical Methods in Medicine*, 1-10.
- Chamidah, N. & Wasito, Ito (2015) Fetal State Classification from Cardiotocography Based on Feature Extraction Using Hybrid K-Means and Support Vector Machine, *ICACSI, 2015 IEEE*, 37-41.
- Zheng, B. et al. (2014) Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert Systems with Applications*, 41, 1476-1482.

Krupa, N et al. (2011) Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine, *BioMedical Engineering Online*, 2011, 10/1/6.

Wang, C. & You, W. (2013) Boosting-SVM: effective learning with reduced data dimension, *Appl Intell*, 2013, 39,465-474.