ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

AN EXPERIMENTAL ANALYSIS OF FEATURE SELECTION ALGORITHMS IN HYPERSPECTRAL IMAGE CLASSIFICATION

M.Sc. THESIS

Hamed GHOLAMI VIJOUYEH

Communication Systems Department

Satellite Communication and Remote Sensing Programme

JUNE 2017

ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

AN EXPERIMENTAL ANALYSIS OF FEATURE SELECTION ALGORITHMS IN HYPERSPECTRAL IMAGE CLASSIFICATION

M.Sc. THESIS

Hamed GHOLAMI VIJOUYEH
(705141005)

Communication Systems Department

Satellite Communication and Remote Sensing Programme

Thesis Advisor: Asst. Prof. Dr. Gülşen TAŞKIN KAYA

JUNE 2017

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ BİLİŞİM ENSTİTÜSÜ

HİPERSPEKTRAL GÖRÜNTÜLERIN SINIFLAMASINDA ÖZNİTELİK SEÇİM ALGORİTMALARININ DENEYSEL ANALİZİ

YÜKSEK LİSANS TEZİ

Hamed GHOLAMI VIJOUYEH
(705141005)

İletişim Sistemleri Anabilim Dalı

Uydu Haberleşmesi ve Uzaktan Algılama Programı

Tez Danışmanı: Asst. Prof. Dr. Gülşen TAŞKIN KAYA

HAZIRAN 2017

Hamed GHOLAMI VIJOUYEH, a M.Sc. student of ITU Informatics Institute 705141005 successfully defended the thesis entitled “AN EXPERIMENTAL ANALYSIS OF FEATURE SELECTION ALGORITHMS IN HYPERSPECTRAL IMAGE CLASSIFICATION”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Asst. Prof. Dr. Gülşen TAŞKIN KAYA** ....
Istanbul Technical University

Jury Members : **Assoc. Prof. Dr. Esra ERTEN** ....
Istanbul Technical University

Assoc. Prof. Dr. Füsun BALIK ŞANLI ....
Yıldız Technical University

Date of Submission : 05 May 2017
Date of Defense : 05 June 2017

To my wife, To my parents,

FOREWORD

When I am writing this foreword, I realize that my master education will end very soon. By reviewing my graduate study duration at the Informatics Institute of Istanbul Technical University (ITU), not only ITU has given me a very strong education in Satellite Communication and Remote Sensing, but also it helped me to make my confidence to follow my profession goals in attending my education. However, my experiences in ITU is like a medicine: a good thing for health, but sometimes tastes bitter.

Firstly, I would like to thank my advisor, Dr. Gülşen Taşkın Kaya for her support and patience. I specially appreciate her tough questions, which pushed me to make my work more entire. I feel very lucky to be her student. Without her guidances I can not imagine how I could have finished my master's degree. I also sincerely thank to Dr. Elif Sertel who provides me some facilities in ITU CSCRS center and I used her very useful guidances in my thesis. I also like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) for funding this work with the project number 115Y032.

I would specially like to thanks my lovely wife, Mina. She has supported me throughout my education at ITU and has a high devotion to me get to this point. There is no doubt that without her assists I could not be able to continue my education. My parents, deserve special thanks for their continued encouragement and support. My brother, Ali, who have encourage me in this path. Undoubtedly, without such a team like this, I could not be in this place today.

Finally, I want to thank my late uncle who assisted me during my life, anywhere that I needed help and consultation. I am sad that he will not see my graduation.

June 2017 Hamed GHOLAMI VIJOUYEH

TABLE OF CONTENTS

Page

ABBREVIATIONS

ITU	: Istanbul Technical University
HSI	: Hyperspectral Image
GIS	: Geographical Information System
FE	: Feature Extraction
FS	: Feature Selection
AI	: Artificial Intelligence
MI	: Mutual Information
Chi2	: Chi-Square
CIFE	: Conditional Informative Feature Extraction
CMIM	: Conditional Mutual Information Maximization
DISR	: Double Input Symmetrical Relevance
Fisher	: Fisher Score
Gini	: Gini Index
InfoG	: Information Gain
JMI	: Joint Mutual Information
MIFS	: Mutual Information based Feature Selection
MIM	: Mutual Information Maximisation
MRMR	: Maximum Relevance and Minimum Redundancy
ReliefF	: Relief-F
FSDTree	: Forward Feature Selection using Decision Tree
Single	: Single Feature Selection using Logistic Regression Classifier
RF	: Random Forest
SVMRFEK	: Recursive Feature Elimination using Non-Linear Kernel-Based SVM
SVMRFEL	: Recursive Feature Elimination using Linear Kernel SVM
SVM	: Support Vector Machine
KNN	: K-Nearest Neighbour
AVIRIS	: Airborne Visible/Infrared Imaging Spectrometer
ROSIS	: The Reflective Optics System Imaging Spectrometer

SYMBOLS

nm	**:** Nanometer

LIST OF TABLES

LIST OF FIGURES

AN EXPERIMENTAL ANALYSIS OF FEATURE SELECTION ALGORITHMS IN HYPERSPECTRAL IMAGE CLASSIFICATION

SUMMARY

Recently, hyperspectral images have been an attractive subject for many researches in remote sensing area since they provide abundant information due to their wide range of spectral bands. On the one hand, providing such a huge amount of data by hyperspectral images may lead to complexity and bring some redundancy due to high correlation among the hyperspectral bands. On the other hand, this redundancy often negatively effects the classification of hyperspectral data by imposing extra computational costs without providing any advantageous information to the performance of the classifier. Moreover, the redundancy or using more features may lead to a decrease in the classification accuracy, which is known also as Hughes effect.

In order to reduce the redundancy and increasing the performance of the classification methods, feature selection algorithms have been carried out to remove irrelevant features and highlight the efficient features of dataset to achieve a significant accuracy with minimum costs. The feature selection methods are typically presented in three categories based on how they combine the selection algorithm and the model building: filter-based methods which select suitable features using a search method; wrapper methods that validate the selected features with a classifier; and embedded methods which utilize the profits of two prior methods.

There have been many studies related to feature selection not only in developing novel methods but also in application of the methods to hyperspectral image classification. To our knowledge, there is no any general analysis over hyperspectral remotely sensed datasets, involving a wide range of feature selection methods to compare them in the same experimental environments. In this work, a comprehensive experimental analysis with seventeen mostly used state of art feature selection algorithms is conducted extensively analyzed with two well-known classifiers, that are K-nearest neighbours and support vector machines, on seven common hyperspectral remotely sensed datasets. The contribution of this thesis is to present an extensive benchmark study on using feature selection algorithms with hyperspectral datasets to help researchers to comprehend the behaviour of feature selection methods on different cases. The analysis of feature selection algorithms are carried out by considering different number of training samples and different number of ranked features count. Besides, the methods are assessed based on four evaluation criteria which are classification accuracy, stability of feature selection methods, ability of ranked features to separate the classes of a dataset, and computational cost.

According to the results obtained from the experiments, the filter-based methods, which are improved by mutual information measures, are more profitable than the other filter-based methods, even wrapper and embedded techniques. Although, filter

methods are known as unstable method, they achieve accurate classification results as well as low computational time. Some wrapper and embedded methods also perform significant classification accuracy while filter-based methods also enhance a higher level of generalization.

HİPERSPEKTRAL GÖRÜNTÜLERIN SINIFLAMASINDA ÖZNİTELİK SEÇİM ALGORİTMALARININ DENEYSEL ANALİZİ

ÖZET

Son günlerde, hiperspektral görüntüler geniş spektrum bantlarından dolayı bol miktarda bilgi sağladığı için, uzaktan algılama alanında yapılan birçok araştırma da çekici bir konu olmuştur. Ancak, hiperspektral görüntülerle çok fazla miktarda veri sağlanması, hiperspektral bantlar arasındaki yüksek korelasyona bağlı olarak sınıflandırmada karmaşıklığa neden olabilir ve bilgi fazlalığı yaratabilir. Dolayısyla, bu bilgi fazlalığı, sınıflandırıcının performansına çoğunlukla bir katkı sağlamazken ekstra hesaplama zararı getirerek hiperspektral verilerin sınıflandırılmasını olumsuz yönde etkiler. Fazla öznitelik kullanılması, Hughes efekti olarak da bilinen sınıflandırma doğruluğunda bir azalmaya neden olabilir.

Artıklığı azaltmak ve sınıflandırma yöntemlerinin performansını arttırmak için öznitelik seçim algoritmaları, asgari maliyetlerle belirgin bir doğruluğu sağlamak için fazla öznitelikleri kaldırmak ve veri kümesinin etkin özniteliklerini çıkarmak için kullanılmaktadır. Öznitelik seçimi (ÖS), uzaktan algılamada hiperspektral görüntüler alanında etkili ve avantajlı bir araştırma alanıdır. Çok sayıda ilgisiz ve gereksiz öznitelik içeren bu büyük veride, öznitelik seçimi yaparak, veri fazlalığı, çok fazla bilgi kaybına uğramadan azaltılabilir. Bununla birlikte, dikkate alınan hiperspektral veri kümesi için uygun bir öznitelik seçimi yönteminin belirlenmesi uygulamacılar açısından önemli olabilmektedir.

Öznitelik seçme yöntemleri tipik olarak, seçim algoritmasını ve model oluşturma yöntemini nasıl birleştirdiklerine bağlı olarak üç kategoriye ayrılmaktadır: Bunlar, arama yöntemini kullanarak uygun öznitelikleri seçen filtre temelli yöntemler, seçilen öznitelikleri bir sınıflandırıcıyla doğrulayan sarmalayıcı yöntemler ve iki önceki yöntemin kazançlarını kullanan gömülü yöntemler şeklinde sıralanmaktadır.

Öznitelik seçimi konusu üzerinde yapılan birçok çalışma sadece yeni yöntemlerin geliştirilmesi üzerinde değil, yöntemlerin hiperspektral görüntü sınıflandırmasına uygulanması üzerinde de yapılmaktadır. Bildiğimiz kadarıyla, hiperspektral uzaktan algılanmış veri kümeleri üzerinde, öznitelik seçimi yöntemlerinin aynı deneysel ortamlarda karşılaştırıldığı genel bir analiz çalışması literatürde mevcut değildir. Bu çalışmada, en çok kullanılan en gelişmiş on yedi öznitelik seçimi algoritması ile kapsamlı bir deneysel analiz yapılmıştr. Destek Vektör Makineleri (DVM) ve K-En Yakın Komşuluk (K-EYK) sınıflandırıcıları kullanılarak literatürde bilinen yedi hiperspektral uzaktan algılama veri kümesi üzeride kapsamlı olarak analiz edilmiştir. Bu tezin katkısı, araştırmacıların, farklı tiplerdeki öznitelik seçimi yöntemlerinin davranışını anlamasına yardımcı olmak amacıyla hiperspektral veri kümeleri ile öznitelik seçimi algoritmalarının kullanımı hakkında kapsamlı bir değerlendirme çalışması sunmaktır. Öznitelik seçimi algoritmalarının analizi, farklı sayıda eğitim örneği alınarak da analiz edilmiştir. ÖS yöntemeleri, sınıflandırma

doğruluğu, öznitelik seçimi yöntemlerinin kararlılığı, sınıflandırılmış özniteliklerin bir veri kümesinin sınıflarını ayırabilme kabiliyeti ve hesaplama maliyeti olmak üzere dört ana değerlendirme kriterine göre değerlendirilmiştir. Bu tez, öznitelik seçme yöntemleri ve bunların uzaktan algılama alanındaki hiperspektral veri kümeleri üzerine odaklanmıştır. Bu tez beş bölümden oluşmaktadır.

Birinci bölüm, bu çalışmayı tanıtmakta ve uzaktan algılama alanında kullanılan hiperspektral görüntülerden bazı yönler sunmaktadır. Buna ek olarak, bu bölümde boyut azaltıcı ve sınıflandırma yöntemleri ile ilgili temel tanımlarda verilmektedir. Ayrıca, hiperspektral görüntüleri ve öznitelik seçimi alanında yapılan literatürdeki daha önceki eserlerile ilgili bir literatür özeti verilmektedir. Bu çalışmanın kısa bir amacı ve motivasyonuna da bu bölümde yer verilmiştir.

İkinci bölüm, araştırmada ele alınan yöntemlerin genel olarak tanımlarını içermektedir. Tüm öznitelik seçimi ve sınıflandırma yöntemleri kısaca özetleri verilmektedir.

Üçüncü bölüm, üç alt bölümden oluşmaktadır. İlk bölüm, çalışmanın içerdiği hiperspektral veri kümelerininin genel tanımlarını içermektedir. Kullanılan hiperspektral veri kümeleri, uzaktan algılama alanındaki kullanımı çok yaygındır ve bu alanda yapılan çalışmalarda geniş ölçüde kullanılmaktadır. İkinci bölümde, öznitelik seçim yöntemlerinin analizinde kullanılan değerlendirme kriterlerinin neler olduğundan söz edilmektedir. Çalışmada, dört farklı değerlendirme kriteri dikkate alınmıştır. Bunlar, sınıflandırma doğruluğu, öznitelik seçimi yöntemlerinin kararlılığı, seçilen özniteliklerin sınıfları ayırma becerisi ve ÖS yöntemlerinin hesaplama şeklindedir. Son bölümde ise, deneylere geçmeden önce yapılan gerekli parametre seçimlerinden söz edilmektedir.

Dördüncü bölüm, hiperspektral veri kümeleri ile yapılan deneylerin sonuçlarını göstermekte ve elde edilen bulgular hakkında kapsamlı bir tartışma imkanı sunmaktadır. Deney sonuçları, öznitelik seçme yöntemleri ve sınıflandırıcıların hiperspektral veri kümeleri üzerindeki davranışlarına göre ayrı ayrı incelenmiştir.

Son olarak, beşinci bölümde, tez kapsamında yapılan çalışmanın elde edilen genel sonuçları özetlenmektedir. Elde edilen sonuçlara göre, filtre temelli yöntemler, hiperspektral veri kümeleri üzerinde yapılan farklı değerlendirme kriterleri çerçevesinde sarmalayıcı veya katıştırılmış tipteki yöntemlere göre daha avantajlı yöntemlerdir.

Bu tez kapsamında elde edilen sonuçları ışığında, uzaktan algılama alanında hiperspektral görüntülerin sınıflandırma problemleri ve analizleri için "mutual information" tabanlı filtre yöntemlerinin kullanılması önerilmektedir. Ayrıca "Random Forest (RF)" ve "FSTree" yöntemleri de başarılı sonuçlar vermektedir. "Mutual information" yöntemlerinin avantajları aşağıda sıralanmaktadır:

- Bu yöntemler herhangi bir sınıflandırıcıdan bağımsızdır ve herhangi bir parametre ayarı gerektirmezler. Dolayısıyla, bu yöntemlerin uygulanması oldukça kolaydır.

- Bu yöntemlerin hesaplama zamanı sarmalayıcı ve katıştırılmış yöntemlerden daha düşüktür.

- "Mutual information" tabanlı yöntemler farklı örnekler ve veriler için iyi bir genellştirme yeteniğine sahiptir.

"RF" yöntemi, hiperspektral veri kümelerinin boyut azaltma sorunlarıyla başa çıkmanın başka bir alternatifi olabilir. Bu yöntem, düşük hesaplama süresinin yanı sıra yüksek düzeyde bir sınıflandırma doğruluğu vermektedir. Ancak, "RF" yönteminin uygulanması filtre tabanlı yöntemlere göre daha zordur.

"FSDTree" yöntemi, sınıflandırma doğruluğu açısından oldukça iyi bir yöntemdir ancak hesaplama maliyeti yüksektir. Hesaplama zamanının önemli olmadığı uygulamalar için, bu yöntem ile başarılı sonuçlar elde edilmesi mümkündür.

1. INTRODUCTION

1.1 Introduction

Recent promotes in remote sensing and geographical information system (GIS) is directed to find new ways to extend the hyperspectral images (HSI) and its sensors. In case of remote sensing, HSI is a recently developed technology that allocates the scientists and researchers in their investigations. The main subject of hyperspectral remote sensing has begun in the mid-80's to use for mapping minerals by the geologists.

Hyperspectral image is an imaging technique that collects information from objects, based on their electromagnetic spectrum [2]. Using advantages of thousands of sensors, HSI spectrometer can measure about 100 to 200 spectral bands with 5 to 10 nm through an extensive wavelength, mostly in the range of 400 to 2500 nm, whereas, multispectral imaging are usually composed of about 5 to 10 bands with a large bandwidth (70-400 nm).

Since the HSI has the ability to provide a detailed information about the objects, a variety of application such as object discovering, material identification, and target detection have been reported in the literature [3] [4] [5] [6]. The HSI imagery is generally collected as a data cube with spacial information in the X-Y plane, and spectral data in the Z-direction as shown in Figure 1.1.

Classification algorithms are quite effective tools to extract the information from the HSI. However, hyperspectral data always come with a huge number of redundant and correlated bands that might cause a poor classification accuracy. Besides, the redundancy in features also brings an extra computational cost without contributing any useful information to classification performance [7]. Therefore, processing such a huge volume of data might become a quite difficult task especially when supervised classification methods are used. Another problem, often reported in the context of classification of hyperspectral images in the literature is the Hughes effect or

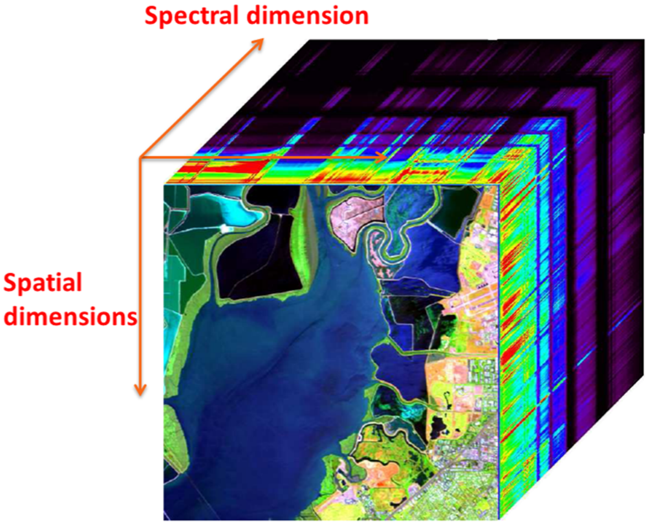


Figure 1.1 : Data cube of a hyperspectral image [1].

phenomenon [8] that can have a major unfavourable impact on the classification accuracy. In classification analysis and with assuming a fixed training set, classification accuracy increases with the addition of new features. The rate of increase, declines and eventually, the accuracy will begin to decrease with adding more features. Hence, feeding more features to a classification method may cause a decrease in the classification accuracy [9]. To tackle with this issues, dimensionality reduction techniques have been carried out as very useful tools to effectively use the classification methods, to reduce the computational time and to optimally use data storage requirements [10]. Moreover, reducing the number of features may lead to increase the classification accuracy in some cases [11].

The dimensionality reduction methods can be categorized into two main sections: feature extraction, and feature selection. Feature extraction (FE) methods map the original features into a low dimensional space. These methods provide more separable features in the low dimensional space, but FE modifies the physical properties of the features. In contrast, feature selection (FS) methods rank the features or pick a subset of features with respect to their ability to generate an accurate classification performance or ignore some features that are considered as redundancy [12]. The ambition of the FS is to select a batch of features which brings as much information as possible. Through keeping the physical explanation of the features, the FS might be a better choice than the FE, especially when dealing with the analysis of real datasets.

There exists three different approaches for the FS: filter, wrapper, and embedded methods [13]. Filter methods clarify unsuitable features usually by using a search method based on statistical criterion without considering any classifier [11]. On the contrary, wrapper methods utilize a classification technique to validate the selected features [14]. The wrappers, often generate better results for final classification accuracy than filter methods while they are optimized for a particular learning method. However, while they have an inside evaluator and each feature set is separately considered, the wrapper methods are more expensive to run. The wrapper methods also can be uncompromising in case of having the big data including many features and classes like HSI datasets. Moreover, in case of the wrapper methods, the methods need to be re-run from a problem to another. Hence, the filter methods provide more general solution than the wrappers. Besides, the embedded methods use the advantages of both above strategies.

The different FS methods result various subsets of the original feature set. Moreover, for a specified classification purpose, a particular FS method should be chosen. An FS method may obtain significant results in a classification problem while in another problem it can not show the same outcomes and using another FS method is preferable. For all these reasons and challenges, feature selection persists a hot topic for researchers in the area of machine learning and data analysis [15].

1.2 Related Works

In the context of hyperspectral remote sensing literature, the studies related to the feature selection have been generally focused on methodology itself. Accordingly the analyses have been conducted to provide the effectiveness of their proposed approach in comparison to several existing algorithms based on a few hyperspectral datasets. Pal and Foody analysed the impact of the features on SVM classification method in context of hyperspectral images. They achieved that the SVM classification accuracy declines with adding more features especially when the size of training set is small. They also found that the feature selection is a useful method to increase the classification accuracy [16]. Ghamisi and Benediktsson presented a new FS approach based on a binary optimization method using fractional-order Darwinian particle swarm optimization. They proposed their method via its impact on SVM, and on attribute profile (AP) vectors [17]. Patra et al. introduced a supervised method to select efficient features in hyperspectral images by using Rough set theory. They evaluated their method in comparison to three other methods that are fuzzy C-mean, divergence, and mutual information [18]. Hossain et al. denoted a one-class oriented method to select the suitable features. They used mutual information (MI) as a FS criterion and applied to cluster space. Each class was classified in a sequence. They used hyperspectral and Lidar datasets for their experiments [19]. Persello and Bruzzone presented a kernel-based feature selection method. Their method chooses a subset of original feature set which are relevant and invariant. The selected features should be discriminant in the considered classification problem and stable along different domains such as source and the target domains [20].

Furthermore, some studies have optimized the common feature selection methods. Zabalza et al. proposed an optimized PCA approach called Folded-PCA which resolved PCA drawbacks. They said that although PCA had been broadly used in feature selection and feature extraction, it tolerated three main drawbacks which are high computational cost, large memory requirement, and low ability in processing large dimensional data like HSI [21].

To our knowledge, for a researcher who looking for a feature selection method as a tool to solve a specific problem in the area of remote sensing, there is no comprehensive analysis covering the most of the FS methods with using hyperspectral remotely sensed datasets, and comparing them in the same environmental conditions meaning that conducting the experiments on the same datasets in the area of remote sensing. Moreover, while HSI analysis is very costly due to their high dimensionality, there is no significant assessment including a large collection of commonly used remotely sensed hyperspectral datasets. For instance, Bolon-Canedo et al. have investigated important feature selection methods on some synthetic Artificial Intelligence (AI) datasets with aiming to review the performance of feature selection method with taking into account of irrelevant features, noise in the data, redundancy and interaction between attributes, as well as a small ratio between number of samples and number of features [22]. Besides, Pohjalainen et al. developed supervised and unsupervised feature selection methods by focusing on paralinguistic analysis using standard K-Nearest Neighbours (KNN) as a classifier. They showed that the classification of paralinguistic dataset using FS methods with KNN classifier, leads to achieve equivalent or even better performance than using support vector machine (SVM) or random forest as a classifier [23]. Ang et al. provided a review on the supervised, unsupervised, and semi-supervised feature selection methods in gene selection analysis. They also discussed the challenges and problems faced in order to obtain better diseases prediction or fining new diseases. The paper implied that there are still many open opportunities for further improvements. The authors utilized most commonly used gene micro-array expression datasets [24]. Vergara and Estévez presented a review of the feature selection methods without considering any dataset to analyse the methods. They showed that modern feature selection techniques must go beyond the concept of relevance and redundancy to include complementarity. They developed a framework based on mutual information which is able to optimize the FS method [25]. Brown et al. demonstrated an unifying framework for feature selection methods by optimizing the conditional likelihood. They are not pursued an exhaustive analysis but, displayed a valuable comparison between information-based feature selection techniques using 15 common machine learning datasets [26].

1.3 Purpose of Thesis

The aim of this research is to present a benchmark study covering an experimental analysis of mostly used state of art feature selection methods on variety of hyperspectral remotely sensed datasets [27]. The motivation of this thesis is to exploit a comprehensive survey for using a remote sensing researcher to understand the performance of the FS algorithms on a specified hyperspectral dataset.

For the experimental analysis, 17 number of FS methods are be tested with two well-known classification methods over 7 remotely sensed hyperspectral datasets in terms of classification accuracy, stability of feature selection methods, ability of the selected features to separate classes, and computational time of FS methods.

1.4 Thesis Overview

This thesis is focused on feature selection methods and their impact on hyperspectral datasets in the area of remote sensing. The outline of this thesis is organized in 5 chapters.

The first chapter introduces this work and gives some aspects from hyperspectral images used in the area of remote sensing. In addition, this chapter contains fundamental definitions about dimensionality reduction and classification methods. It also contains a look to previous works that were done in the area of feature selection of remotely sensed hyperspectral images. A brief purpose of this work and the motivation also is given in this chapter.

The chapter 2 describes the methodologies considered in this research. All feature selection and classification methods are briefly given. The feature selection methods are separately explained in three categories.

The chapter 3 is constructed in three sections. First section, illustrates the hyperspectral datasets that are included in this work. All these datasets are well-known in the area of remote sensing and are broadly used in the studies conducted in this area. In second section, the assessment criteria are described to evaluate the feature selection methods. Four evaluation criteria are considered in this work: classification accuracy, stability of feature selection methods, ability of the selected features to separate classes, and computational time of FS techniques. The experiments and their settings are demonstrated in the last section of this chapter.

The chapter 4 demonstrates the results of experiments conducted with hyperspectral datasets and gives a comprehensive discussion about the findings. The experimental outcomes are investigated with respect to the behaviour of all feature selection methods and classifiers over all hyperspectral datasets.

Finally, the chapter 5 gives a conclusion of this thesis. According to the obtained results, the filter-based methods are more profitable than the wrapper or embedded techniques in terms of different evaluation criteria for hyperspectral datasets.

2. METHODOLOGY

This chapter explores several well-known feature selection methods based on SVM and KNN classifiers. For each method, a reference is given to the reader to be able to get more detail analysis about a specified method.

2.1 Feature Selection Methods

Feature selection methods are categorized as three parts depending on their evaluation capability of individual feature or feature sets. The FS methods used in this study are briefly explained below:

2.1.1 Filter methods

2.1.1.1 Chi square (Chi2)

This method uses the *Chi-Square* distribution that is a special case of the *Gamma* distribution and is one of the most broadly-used probability distributions. Chi2 utilizes the *Chi-Square* (X^2) statistic to discretize numeric attributes of the features repeatedly until some discrepancies are found in the data. X^2 is calculated by this formula:

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2.1}$$

where k is the number of classes, A_{ij} is the number of patterns in the i^{th} interval of j^{th} class, and E_{ij} is the expected frequency of A_{ij} that is determined from this equation:

$$E_{ij} = R_i \times \frac{C_j}{N} \tag{2.2}$$

where R_i is the number of patterns in the i^{th} interval, C_j is the number of patterns in the j^{th} class, and N is the total number of patterns [28]. Chi2 is a quite useful method on numeric and ordinal type of data.

2.1.1.2 Conditional informative feature extraction (CIFE)

This method is based on information theory and uses the mutual information to reduce the redundancies between the features based on their relevant classes that cause a maximized joint class-relevant information [29]. CIFE is a special case of MI based feature selection methods which the coefficient of MI is equal to 1 [25].

2.1.1.3 Conditional mutual information maximization (CMIM)

The CMIM is a very fast FS method that uses conditional mutual information (MI). This method maximizes the MI of features that picked up individually or weakly dependant pairs. The CMIM is a forward selection method and ranks the features by comparing each feature with the selected one to determine the good features. The good features mean that if $\hat{I}(Y;\acute{X}|X)$, which is the estimation of the quantity of information shared between X (feature vector) and Y (related class), is large for each selected X. In other words, $\acute{X}$ is suitable if it has information about Y and these information have not been caught by any of the X already picked [30].

2.1.1.4 Double input symmetrical relevance (DISR)

DISR uses two major characteristic of feature selection: a combination of variables that returns more information about related class rather than the information obtained from individual variables [31], or a combination of best performing subset when there is no information how to combine the variables [32].

2.1.1.5 Fisher score (Fisher)

This method uses discriminative methods, and generative statistical models to determine the most relevant features and selects the features in such a way that the distances between them and the other features with different classes are as large as possible while the distances between features with the same class labels are as small as possible [33]. The Fisher method scores i^{th} using this formula:

$$\text{Fisher Score} = \frac{\sum n_j(\mu_{ij} - \mu_i)^2}{\sum n_j \rho_{ij}^2} \tag{2.3}$$

where μ_{ij} is the mean and ρ_{ij} is the variance of the i^{th} feature in the j^{th} class, μ_i is the mean of the i^{th} feature and n_j is the number of samples in the j^{th} class [34].

2.1.1.6 Gini index (Gini)

This filter method uses *Gini coefficient* to measure a feature capability to detect class separability. The Gini Index is determined independently for each feature. The more important feature has the smaller Gini index. After the whole features are weighted, the top N features having smallest Gini index are selected [35]. The Gini Index for a feature is calculated by this formula:

$$\text{GiniIndex} = 1 - \sum_{j=1}^{n} p_j^2 \tag{2.4}$$

where p_j is the relative frequency (probability of occurrence) of class j.

2.1.1.7 Information gain (InfoG)

InfoG is an easy method to implement that counts the number of obtained bit information with a corresponding class [36]. To select the valuable features, the entropy of the data both for whole classes and each class is calculated, and the features with highest discrimination are selected [37]. This method values each term by:

$$\text{InfoG}(S_x,x_i) = H(S_x) - \sum_{v=values(x_i)}^{|S_{x_i}=v|/|S_x|} H(S_{x_i}=v) \tag{2.5}$$

where S_X is the set of training examples, x_i is the vector of i^{th} variables and $|S_{x_i}=v| \big/ |S_x|$ is the fraction of examples of the i^{th} variable having value v and:

$$H(S) = -p_+(S)\log_2 p_+(S) - p_-(S)\log_2 p_-(S) \tag{2.6}$$

where $p_\pm(S)$ is the probability of a training sample in the set S to be in the positive/negative class [38].

2.1.1.8 Joint mutual information (JMI)

This method is a model-independent technique and uses mutual information to detect the relevant features. The MI checks each feature pair individually while the relevance of a set of these pairs is described by JMI [39]. JMI is selected the most relevant features to the class rather than the MI. It takes into consideration the relevance of feature and class when the subset of features were selected.

2.1.1.9 Mutual information based feature selection (MIFS)

MIFS technique uses mutual information to select the features. Firstly, it computes MIs for each feature from an initialized feature set with an output class. Secondly, it selects the first feature and finds the feature that maximizes the MI. This method uses greedy method to select the features. The steps of MIFS can be given as follows [40]:

1. Set the first set of n features as F and define S as an empty set (initialization).

2. Compute $I(C;f)$ for each $f \in F$ (MI computation).

3. Find the feature f that maximizes MI and set $F = \frac{F}{\{f\}}$ and $S = \{f\}$.

4. Repeat levels 2 and 3 until the $|S|$ meets the needed feature rank (greedy selection).

5. Proceed S as the output.

2.1.1.10 Mutual information maximisation (MIM)

The MIM method weights the features with considering mutual information [41]. To evade the disadvantages of random sampling such as ignoring informative features, the MIM selects the features that maximize the MI individually with the class prediction. This technique does not assure minimal dependency between the features and may lead to redundancy [42] [43].

2.1.1.11 Maximum relevance and minimum redundancy (MRMR)

The MRMR is a common filter technique to select the features that have the strongest correlation with a classification variable. The MRMR selects the features that are mutually different from each other while still having a high correlation [44]. To show the dependency among the feature pairs, it uses the mutual information. In other words, this strategy consists of selecting a feature f_i among the features f_s that are not selected and maximizes $(u_i - r_i)$, where u_i is relevance of f_i to the class c and r_i is the mean redundancy in these two selected features and can be defined as:

$$r_i = \frac{1}{|f|} \sum_{f_i \in f} I(f_i;c) \tag{2.7}$$

$$u_i = \frac{1}{|f|^2} \sum_{f_i \in f} I(f_i,f_j) \tag{2.8}$$

where $I(f_i;c)$ is the mutual information between f and c that are random variables [16].

2.1.1.12 Relief-F (ReliefF)

ReliefF is an updated version for Relief feature selection technique [45]. It uses the differences of average distance between the nearest point in the same class (near-hit) with the nearest point in a different class (near-miss) [46] [47]. The original Relief uses the single nearest hit and miss but the ReliefF take an average among K nearest hits and misses [48]. The ReliefF is one of the best and successful strategies in the feature selection.

2.1.2 Wrapper methods

2.1.2.1 Forward feature selection using decision tree (FSDTree)

This method uses forward feature selection to select worthy features then evaluate them with decision tree classifier as a validator. The aim of decision trees is to find a model that predicts the target value using decision rules derived from data features [49].

2.1.2.2 Single feature selection using logistic regression classifier (Single)

The Single algorithm is a wrapper method and uses a simple feature selection technique. This method selects each feature individually and evaluate its prediction accuracy using logistic regression classification method. Logistic regression is a very common method that maximizes the sum of the likelihood logarithm and penalizes the regression coefficients using L1 norm [50].

2.1.3 Embedded methods

2.1.3.1 Random forest (RF)

This embedded method selects features by repetitively training a random forest classifier [51] by using ongoing feature set and eliminating the least important feature.

The random forest classifier is a classification method based on random decision trees [52]. This classifier fits classification trees by obtaining a bootstrap sample from the data. A random subset of variables is selected independently from all possible variables at each node of decision trees. The each tree weights the variables by finding the best partition on the selected instances. Random forest is one of the best-known machine learning classification methods and very stable when the training data have small changes [53].

2.1.3.2 Recursive feature elimination using non-linear kernel-based SVM (SVMRFEK)

SVM-RFE is a sequential backward feature elimination method that uses the binary SVM classifier as its evaluator [54]. The SVM-RFE begins with all the features. At each step, feature weights are acquire by comparing th training data with the existing features. Then, the feature with the minimum cost function is removed. This procedure continues until all features are ranked according to the removed order [55]. The cost function is calculated by this formula:

$$| \parallel w \parallel^2 - \parallel w^{(-f)} \parallel^2 | \tag{2.9}$$

where $|w|^2$ is the norm of feature weight vector. The notation $-f$ means that the feature f has been removed [54].

The SVMRFEK is a kernel version of this method. It uses an RBF kernel and can handle non-linear SVM models, but is slower than the original technique. A correlation bias reduction (CBR) [56] strategy is designed to deal with the highly correlated features.

2.1.3.3 Recursive feature elimination using linear kernel SVM (SVMRFEL)

This method is a version of original linear SVM-RFE as in [54] therefore it uses SVM with a linear kernel as its evaluator. This method is faster than the method that uses RBF kernel. Similarly to deal with highly correlated features, a CBR [56] method is used.

2.2 Classification Methods

In machine learning, classification is a supervised method that assigns an input feature vector to one of the existing classes, based on specific classification measures. A linear classification method classify the samples base on the value of a linear combination of features. In this thesis, two well-known classification methods which are SVM with linear kernel as a linear classifier and KNN as a non-linear method are used.

2.2.1 Support vector machine (SVM)

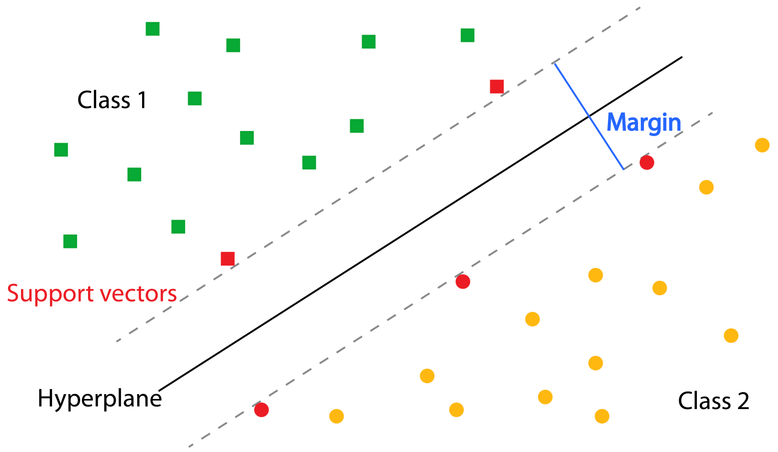


Figure 2.1 : Optimal separating hyperplane in SVM for a linear kernel. The support vectors are indicated by red shapes.

Recently, in context of classification of remotely sensed hyperspectral images, a special attention to SVM has been denoted. SVM has often a higher classification accuracy in counter to another common pattern recognition techniques [57] [58]. This method uses support vectors to classify given data. The aim of SVM is to find an optimal hyperplane between classes by maximizing perpendicular distance (the margin).

As shown in figure 2.1, a basic and simple implementation of SVM is to find an optimized linear hyperplane between the samples of two classes that are linearly separable. This means that it is possible to find a linear hyperplane with function $f(x)$ that can separate the two classes. $f(x)$ is the discriminant function and can be defined as:

$$f(x) \; = \; w \, . \, x \, + \, b \tag{2.10}$$

where w is the normal (weight vector) to the line, x is the training data, and b the bias. The optimal hyperplane can be represented in an infinite number of different ways by scaling of w and b. As a matter of convention [59], among all the possible representations of the hyperplane, the one chosen is:

$$| \, w \, . \, x \, + \, b \, | \; = \; 1 \tag{2.11}$$

From the geometry, the distance between a point x and a hyperplane (w, b) can be calculated:

$$\text{Distance} = \frac{| \, w \, . \, x \, + \, b \, |}{\| \, w \, \|} = \frac{1}{\| \, w \, \|} \tag{2.12}$$

Recall that the margin that already defined, is twice the distance to the closest samples:

$$\text{Margin} = \frac{2}{\| \, w \, \|} \tag{2.13}$$

In order to find an optimal hyperplane, the Margin should be maximized. In other word,

$$\text{Max}\frac{2}{\| \, w \, \|} \Rightarrow \text{Max}\frac{1}{\| \, w \, \|} \Rightarrow \min \| \, w \, \| \Rightarrow \min\frac{1}{2} \| \, w \, \|^2 \tag{2.14}$$

To sum up, the constraints model the requirement for the hyperplane to classify correctly all the training examples x_i. Formally,

$$\min\frac{1}{2} \| \, w \, \|^2 \quad \text{subject to} \quad y_i(w.x_i + b) \geqslant 1, \qquad \forall \quad i = 1, 2, ..., N \tag{2.15}$$

where y_i represents each of the labels of the training data. SVM is also independent from Hughes effect [60].

2.2.2 K-nearest neighbours (KNN)

The KNN is one of the simplest methods used in the classification that collects all the available instances and then classifies new instances with respect to their distance based similarity. This method determines the class of an unknown data depending on the class of the nearest neighbours whose classes are already known [61]. It has a parameter *K* (integer and usually small number) that refers to the number of nearest neighbours in the current feature set. The output is a class label that has maximum iteration in *K* nearest neighbours classes. For instance, if *K* is equal to 1, then the class label of the unknown data is clearly allocated as the class of that first nearest neighbour.

3. DATASETS AND EXPERIMENTAL SETTING

3.1 Hyperspectral Datasets

In order to analyse the FS methods, seven hyperspectral remotely sensed datasets were considered in this study. All datasets are very common in the literature of hyperspectral images. The datasets and their properties are shown briefly in Table 3.1.

Table 3.1 : The hyperspectral datasets used in the experiments.

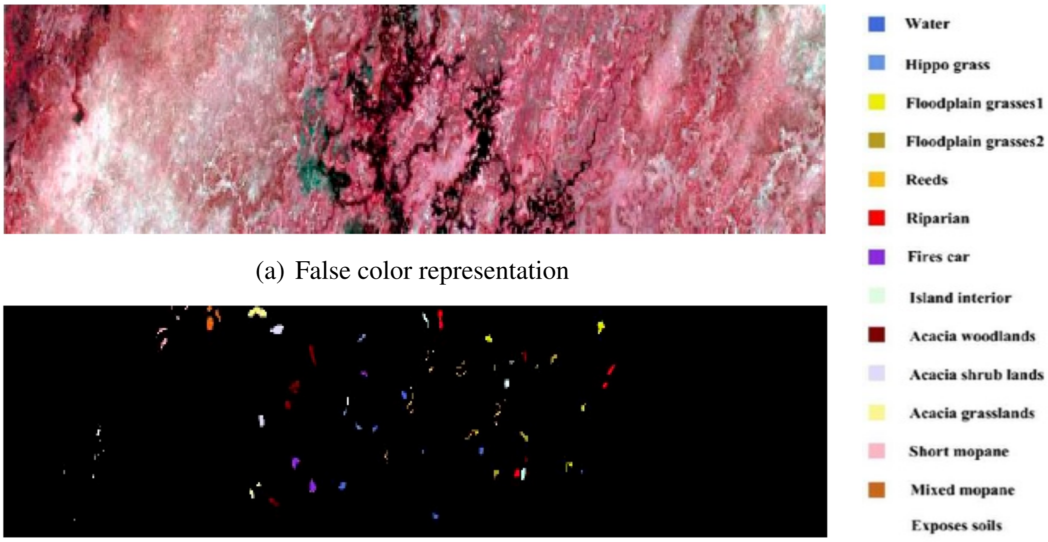
Sensor	Dataset	Measures		
		Bands	Classes	Non-Zero Samples
EO-1	Botswana	145	14	3248
AVIRIS	Indian Pines	200	16	10249
	KSC	176	13	5211
	Salinas	204	16	54129
	SalinasA	204	6	5348
ROSIS	Pavia Center	102	9	148152
	University of Pavia	103	9	42776

3.1.1 Botswana

The Botswana dataset is captured by *Hyperion NASA EO-1* at 30 m pixel resolution over a 7.7 km altitude from Okavango Delta, Botswana, in 242 spectral bands with 400-2500 nm portion of the spectrum. Uncalibrated and noisy bands that cover water absorption are removed, and the 145 spectral bands are remained. The data are analysed in 14 identified classes displaying the land cover types in seasonal swamps, occasional swamps, and drier woodlands [62]. The class 3 (Riparian) with 237 data samples and the class 6 (Woodlands) with 199 data samples are the most complicated classes for this dataset. The true color representation along with ground truth for Botswana dataset is shown in Figure 3.1.

3.1.2 Indian pines

The Indian Pines dataset is acquired by *AVIRIS (Airborne Visible/Infrared Imaging Spectrometer)* sensor over the Indian Pines site, Northwest of Indiana, US. It is



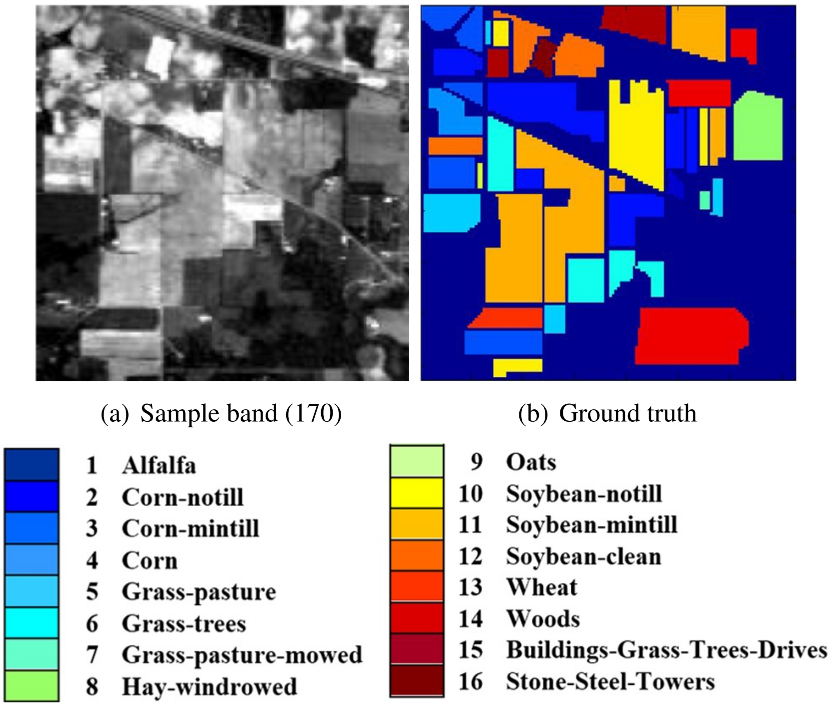
(a) False color representation

(b) Ground truth

(c) Class labels

Figure 3.1 : Botswana dataset

captured with 145 × 145 pixels and 224 spectral bands in the wavelength range 400-2500 nm. This scene (Figure 3.2(a)) is a subset of a larger dataset. The Indian Pines scene contains two-thirds agriculture, and one-third forest or other natural perennial vegetation. There are two major dual lane highways, a rail line, as well as some low density housing, other built structures, and smaller roads. Since the scene is taken in June some of the crops present, corn, soy beans, are in early stages of growth with less than 5% coverage. The ground truth is nominated in 16 classes that



(a) Sample band (170)

(b) Ground truth

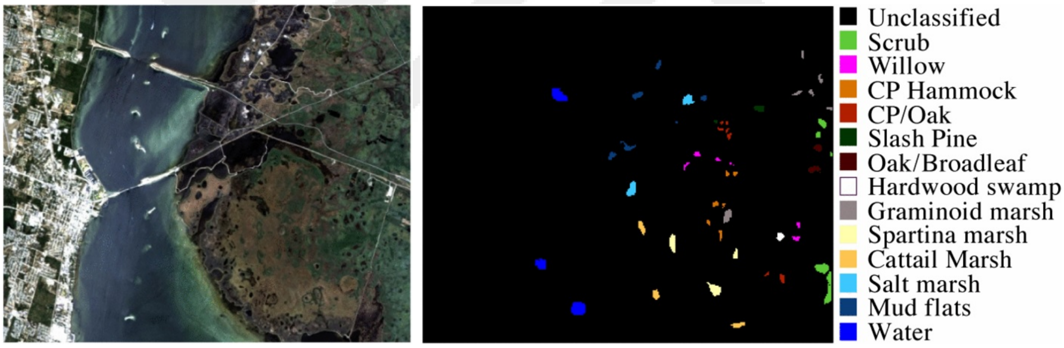
(c) Class labels

Figure 3.2 : Indian Pines dataset

are not all mutually exclusive. The most complicated classes for this dataset are the class 3 (Corn-Min Till) with 834 samples and the class 12 (Soybeans-Heavy Till) that contains 614 data samples. The number of bands are reduced to 200 by removing water absorption and other noisy bands ([104-108], [150-163], and 220) [63]. A sample band, the ground truth and related classes for Indian Pines are shown in Figure 3.2.

3.1.3 Kennedy space center (KSC)

This scene is collected by *AVIRIS* sensor over the Kennedy Space Center, Florida on March 23, 1996, US. It is collected in 224 bands of 10 nm width with center wavelengths from 400-2500 nm from an altitude of approximately 20 km and with a spatial resolution of 18 m. After removing water absorption and low SNR bands, 176 bands are used for the analysis. Training data are selected using land cover maps derived from color infrared photography provided by the Kennedy Space Center and Landsat Thematic Mapper (TM) imagery. For classification purposes, 13 classes representing the various land cover types are defined for this dataset [64]. The KSC dataset have four most complicated classes which are the class 3 (Cabbage palm hammock) with 256 data samples, the class 4 (Cabbage palm/oak) with 252 data samples, the class 5 (Slash pine) with 161 data samples, and the class6 (Oak/broadleaf hammock) with 229 data samples. The true color map of KSC and the ground truth are shown in Figure 3.3.



(a) True color map

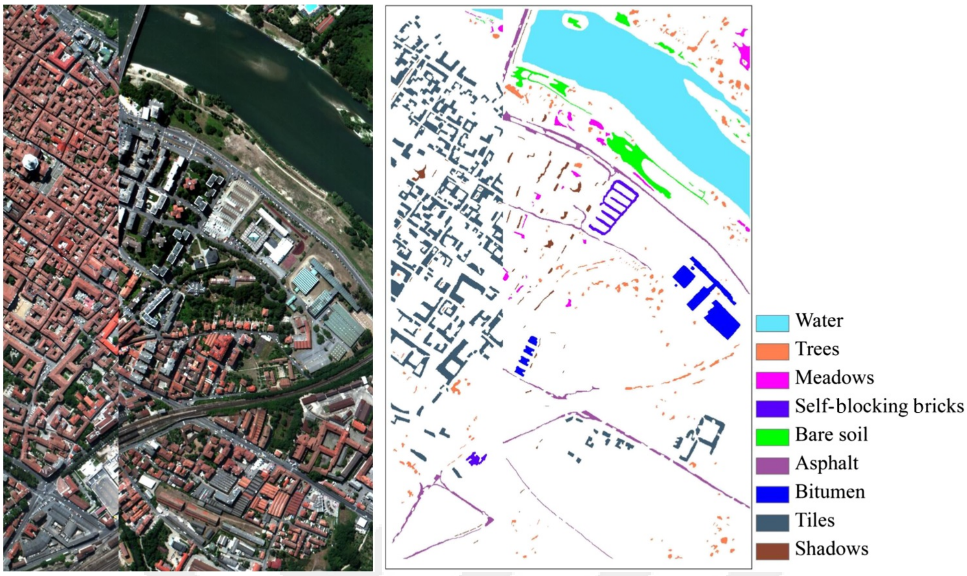
(b) Ground truth

Figure 3.3 : Kennedy Space Center (KSC) dataset

3.1.4 Pavia center and university of pavia

Pavia Center and University of Pavia scenes are captured by *ROSIS (The Reflective Optics System Imaging Spectrometer)* hyperspectral sensor over Pavia, northern Italy.

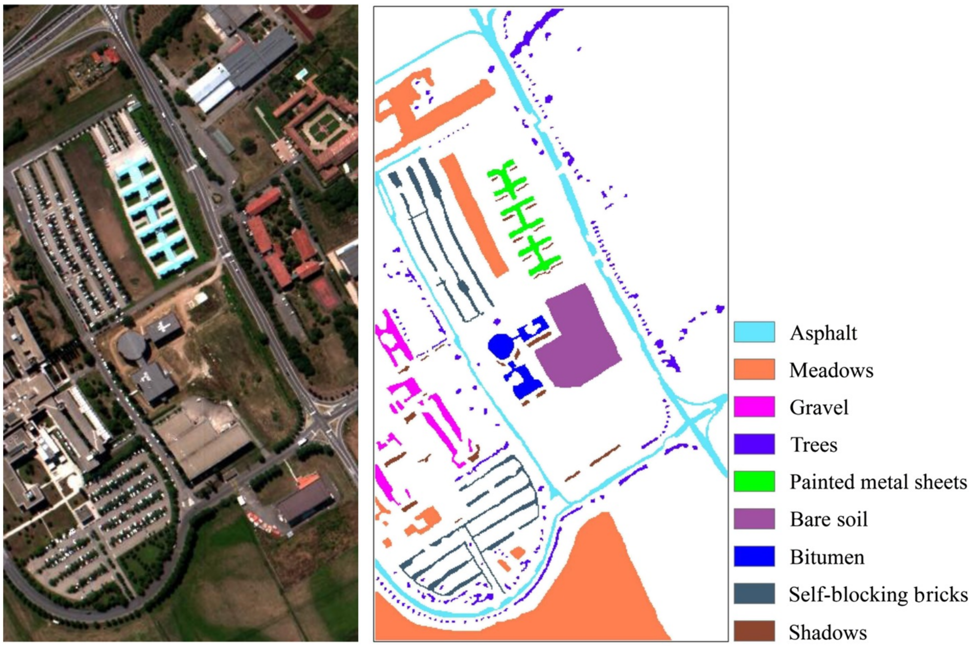
Pavia Center has 102 spectral bands while University of Pavia has 103 spectral bands. The spatial resolution for both datasets is 1.3 meters. The ground truth datasets for both images are provided for 9 classes [65].



(a) True color map (b) Ground truth

Figure 3.4 : Pavia Center dataset

Figure 3.4 shows the original view and the ground truth for Pavia Center hyperspectral dataset. The true color map of University of Pavia and its ground truth dataset are also demonstrated in Figure 3.5.

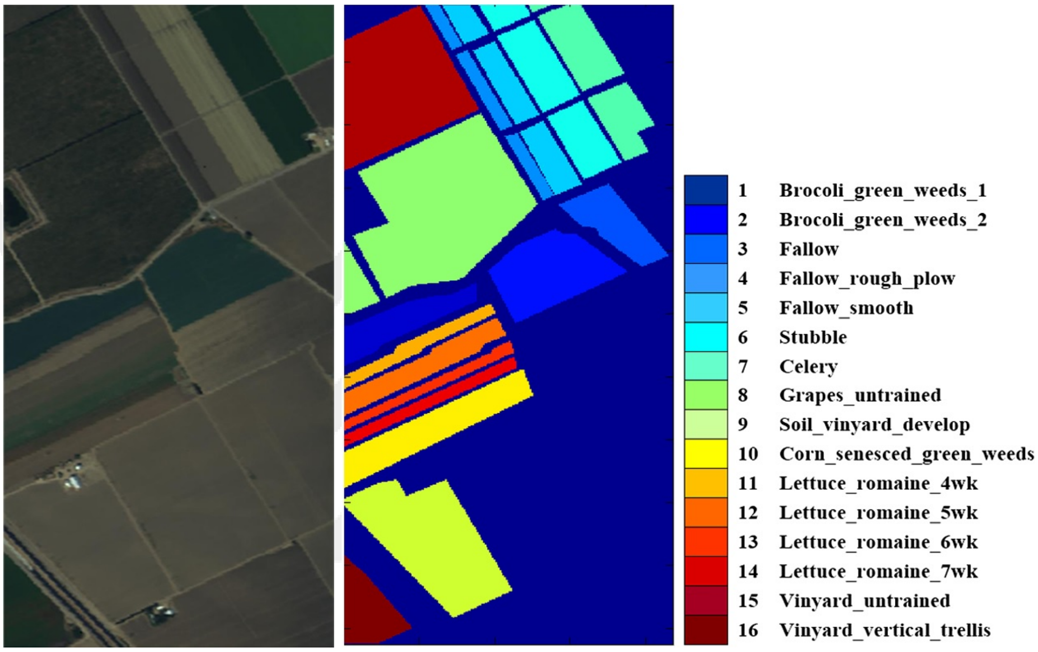


(a) True color map (b) Ground truth

Figure 3.5 : University of Pavia dataset

3.1.5 Salinas datasets

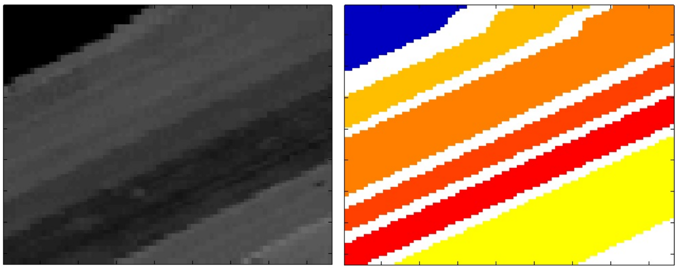
Salinas scene is captured over Salinas Valley, California, US by *AVIRIS* hyperspectral sensor and is characterized by 3.7 meter as spatial resolution. This dataset is a 512 × 217 pixels image and includes 224 spectral bands, and twenty of these ([108-112], [154-167], 224) are discarded because of water absorption effects. The Salinas include 16 land cover classes and mostly contain vegetables, bare soils, and vineyard fields [66]. Figure 3.6 displays the true color map and ground truth dataset for Salinas scene.



(a) True color (b) Ground truth

Figure 3.6 : Salinas dataset

The last dataset called as SalinasA is an small sub-scene of original Salinas dataset with 86 × 83 pixels. As Salinas scene, this sub-scene, has 204 bands. The ground truth for this dataset contains 6 land cover classes [66]. The image with its ground truth is shown in Figure 3.7.



(a) Band 170 (b) Ground truth

Figure 3.7 : SalinasA dataset

3.2 Performance Evaluation Criteria

With the purpose of assessment of the performance of feature selection methods on HSI datasets, four different evaluation criteria were used in this study, that are classification accuracy, stability of feature selection methods, ability of the selected features to separate classes, and computational time of FS methods.

3.2.1 Classification accuracy

The first criterion is classification accuracy as to measure the effect of progressively feeding selected features to a classifier. The aim of this analysis is to show how the accuracy results change with adding more features to a classification method. The classification accuracy is evaluated in terms of different aspects such as the size of training data, dependency of FS methods to a classifiers or datasets, determining the optimal number of features, and finally obtaining the best FS method.

3.2.2 Stability

The stability of FS methods is a measure of robustness of the selected features to small variations in the training dataset and is shown by plotting of top ten selected features for ten different random realization. Obviously, an stable FS method is the one that gives the same or very close feature ranking in each trial with different training datasets.

The aim of investigating the stability of feature selection methods is to find an application domain experts with quantified evidence that the selected features by an stable method are relatively robust to variations of training samples. If a FS method is stable and selects same features for different training data, the captured rank can be used for different range of training data without utilizing the feature selection method again.

3.2.3 Ability of the selected features to separate classes

The aim of feature selection methods is to rank the important features of a dataset. In the ranked feature set, the first features are more important than the other features. Therefore, obtaining the importance of selected features and investigating their ability to distinguish the classes can be another criterion to evaluate the feature selection methods. Evidently, the features that can be separate classes are more important than the other features and play a significant role in analysis of data.

3.2.4 Computational time

The computational time of a FS method is another measure to evaluate the feature selection methods. In order to show, the computationally effectiveness of each FS strategy, the CPU time is recorded on a computer with *Intel Corei7-4710HQ* CPU and *8 GB DDR3* memory.

3.3 Experimental Setting

Inasmuch as all FS algorithms are supervised, the datasets were partitioned randomly into two sections: training data and test data. The training data were included 10, 25, and 50 samples per each class, and the test data were contained the rest of samples of classes for each hyperspectral remotely sensed dataset. To reduce processing time and prevent the reactions of large-value features, training and test data were scaled to the range of $[0,1]$. To obtain the robust results, the experiments were conducted with ten different randomly created training and test datasets, and the average results were reported.

Support vector machine (SVM), and K-nearest neighbour (KNN) were used to evaluate the performance of FS techniques in terms of classification accuracy. In the case of KNN, K neighborhood number was optimized with respect to the leave-one-out error. For SVM case, a linear kernel was used with the penalty parameter $C \in [2^{-5}, 2^{15}]$ which was obtained with a 5-fold cross-validation algorithm. Since the goal of this work is to provide fair results for each feature selection method, all classifiers were used with their default parameters in the implementation of each FS method [67] [26].

In the implementation of the methods, the SVM classification methods were accomplished by using the *LibSVM*[1], which is an integrated software for support vector classification and supports multi-class classification, and the KNN classifier were carried out by using *PRTools*[2], that is a Matlab toolbox for pattern recognition. Moreover, all the feature selection methods were implemented by combining the

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[2] `http://prtools.org/`

feature selection toolbox progressed by *Arizona State University*[1] and *scikit-learn*[2] *Python* machine learning library.

[1] http://featureselection.asu.edu/index.php

[2] http://scikit-learn.org/stable/

4. EXPERIMENTAL RESULTS

The results of the feature selection methods over hyperspectral datasets are demonstrated in this section. The behaviours of the FS methods are assessed in terms of classification accuracy based on two common classifiers, stability of feature selection methods, ability of the selected features to separate classes, and computational time for the learning phase of the FS methods to rank the relevant features. It is necessary to note that all the feature selection methods are evaluated with ten different randomly selected training and test samples for each HSI dataset, and average results are reported.

4.1 Classification Accuracy

The classification accuracy, obtained from applying the classifiers on the ranked feature sets, is considered as the first evaluation criteria. The procedure has two main schemes. In the first part as the learning phase, each training data is used to obtain the relevant features for once, and in the second part as the classification phase, the sets of ranked features are given to each classifier. Since the intention of this thesis is to investigate the FS methods, not classifiers, the parameters of each classifier are optimized based on each training data to achieve the highest classification accuracy. Moreover, because the obtained results are too many due to using dozens of FS methods and hyperspectral datasets, only the important results are considered in this section. For researchers, who may want to see the whole results, the classification accuracy tables as well as their plots are demonstrated in appendices A.1 and A.2.

Based on the outcomes achieved from the experiments, it is observed that the higher number of features makes the classification accuracy higher, except in one case: SVM with using 10 training samples per class over the KSC dataset. As shown in Figure 4.1, adding more than a specific number of features (after 50th features approximately) leads to a slight decrease in the learning curves of classification accuracy in case of some FS methods. This means that by using small size of training samples in the

KSC dataset, these feature selection methods may select some irrelevant features which play a negative role in the classification accuracy. Therefore, according to the results obtained, linear SVM may not be an appropriate classifier for KSC dataset when the size of training data is small.

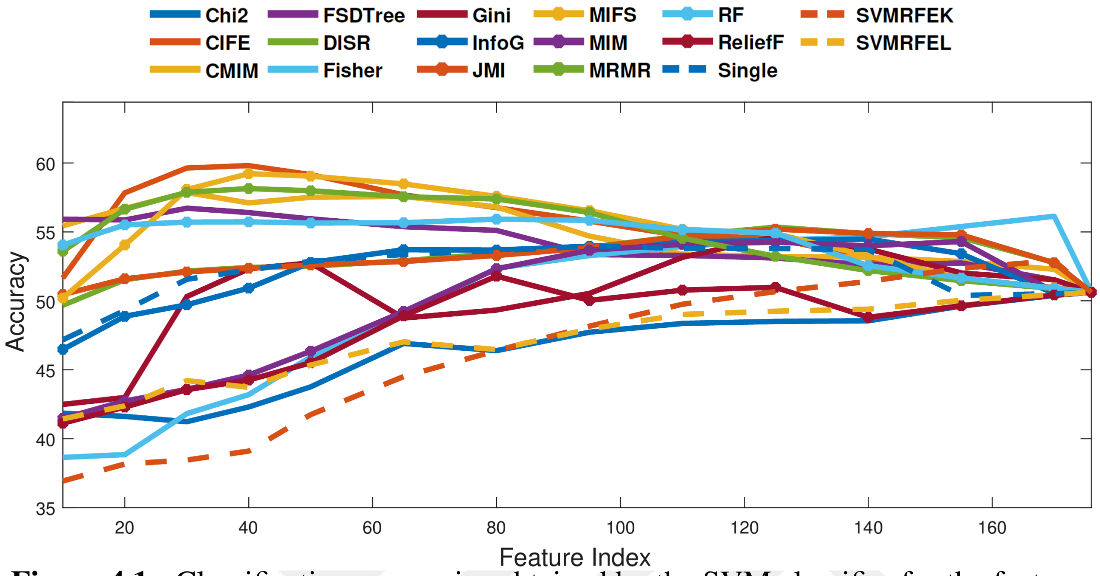


Figure 4.1 : Classification accuracies obtained by the SVM classifier for the features ranked by different feature selection methods for KSC dataset with 10 training samples per class.

In order to study the effects of feature selection methods on the classification accuracy, the behaviour of methods are further analysed in term of different aspects such as the size of training samples, classifiers and datasets, and the number of features used in classification.

4.1.1 Correlation between size of training data and classification accuracy

This section shows the effects of training data size on the classification accuracy for each feature selection method. To achieve robust results that are independent from using different classification method, one classifier is considered and while the SVM results is higher than the KNN, the SVM classification method is chosen. The aim is to compare the classification accuracies obtained from different size of training data by using fixed number of ranked features. Because the first 50 ranked features show the more stable results, first 50 features are considered for each FS method. The maximum classification accuracy obtained from whole available features is used as a base for this comparison.

The achieved results denote that increasing the size of training data leads to an increase in the classification accuracy. Table 4.1 shows the classification accuracy results in Botswana dataset for different size of training data. The last row shows the classification accuracy with using whole feature set. As can be seen use of larger training sample makes higher accuracy.

Table 4.1 : The classification accuracy (%) of first 50 features for Botswana dataset using SVM classifier with different training data size."10SpC", "25SpC", and "50SpC" represent the size of training data.

Method	**10 SpC**	**25 SpC**	**50 SpC**
Chi2	85.1	90	92.3
CIFE	82.9	87.5	89.9
CMIM	**86.4**	90.7	**92.9**
FSDTree	86.3	90.6	92.4
DISR	**86.9**	**90.8**	**92.5**
Fisher	81.1	85.9	87.9
Gini	78.6	84.4	87.3
InfoG	84.5	87	88.3
JMI	86.3	**90.9**	**92.9**
MIFS	82.9	87.3	89.2
MIM	**86.4**	**90.8**	**92.5**
MRMR	83.3	86.1	88.3
RF	86.3	90.4	92.3
ReliefF	84.6	89.6	91.1
Single	84.8	88.9	90.1
SVMRFEK	79.2	83.9	86
SVMRFEL	85	89.5	90.8
Maximum Acc.	87.5	91.9	94.1

The purpose of this section is to find methods that capture higher accuracy with a small size of training data. The methods CMIM, DISR, JMI, and MIM achieve higher classification accuracy than the other methods in Botswana dataset, and the classification accuracies are closer to the classification accuracy obtained by whole features. The CMIM shows accurate results with using 10 and 50 training samples per class, meaning that this method works also fine with a small size of training data. In contrast, JMI obtains higher classification accuracy than the other methods when the size of training data is large enough. Furthermore, the DISR and MIM show the better results for all size of training data. This means that these two methods can work fine with a small training data.

Table 4.2 : The classification accuracy (%) of first 50 features for Indian Pines dataset using SVM classifier with different training data size."10SpC", "25SpC", and "50SpC" represent the size of training data.

Method	**10 SpC**	**25 SpC**	**50 SpC**
Chi2	54.6	60.2	62.7
CIFE	50.1	57	60.8
CMIM	52.2	61.1	65.8
FSDTree	**54.8**	**63**	**67.5**
DISR	52.8	61.4	64.5
Fisher	54.2	60.4	62
Gini	51	56.1	58.6
InfoG	52.1	58.1	62.9
JMI	53.4	61.1	63.9
MIFS	50.7	57.6	60.5
MIM	**55.4**	59	59
MRMR	51	55.1	58.2
RF	52.2	**63.8**	**67.6**
ReliefF	51.2	55.7	56.9
Single	51.7	56.2	62.7
SVMRFEK	51.9	53.8	56.3
SVMRFEL	52.6	55	57.1
Maximum Acc.	53.8	63.1	68.4

The classification accuracy obtained by ranked features is not necessarily always smaller than the maximum classification accuracy. For example, in Table 4.2 which shows the accuracy results for the Indian Pines dataset, the MIM method demonstrates higher accuracy than the maximum classification accuracy obtained by the whole features. It is worthy to say that, the MIM shows the higher accuracy than the other methods when the size of training data is small. The FSDTree is another successful method which achieves accurate results for different size of training data meaning that this method not only shows the higher classification accuracy than the maximum accuracy, but also a small training sample is enough. In contrast, the RF is a method which needs sufficient number of training samples.

Table 4.3 presents the best feature selection methods for each dataset. The methods are selected where the classification accuracy is higher than the other methods in case of first 50 features. Overall, the methods which uses mutual information measures such

Table 4.3 : The best FS methods using SVM classifier with considering top 50 ranked features for different size of training data over all datasets. "10SpC", "25SpC", and "50SpC" demonstrate the size of training samples.

Dataset	**10 SpC**	**25 SpC**	**50 SpC**
Botswana	DISR	JMI	CMIM, JMI
Indian Pines	MIM	RF	RF
KSC	CIFE	DISR	FSDTree
Pavia Center	MIM, RF	MIM	FSDTree, JMI
University of Pavia	MRMR	MIM	DISR, MRMR
Salinas	DISR	DISR	FSDTree
SalinasA	DISR	DISR	DISR

as CMIM, MIM, JMI, and MRMR as well as FSDTree, RF, and DISR are the methods that show the accurate performance for different size of training data.

4.1.2 Dependency of FS methods to classifiers

The purpose of investigating FS methods in this section is to find an independent method from any classifier. More precisely, the aim is to determine the methods that achieve a higher accuracy in all classification methods. In order to get a clear view of classifiers effects on feature selection methods, the classification accuracies obtained by first 50 features of each ranked feature set (because 50 features give more stable accuracy than first 10 and 30 features) are considered.

Table 4.4 : The classification accuracy (%) of first 50 features for Botswana dataset."10SpC", "25SpC", and "50SpC" represent the size of training data.

Method	**KNN**			**SVM**		
	10 SpC	**25 SpC**	**50 SpC**	**10 SpC**	**25 SpC**	**50 SpC**
Chi2	77.6	80.8	85.3	85.1	90	92.3
CIFE	75.5	80	84.2	82.9	87.5	89.9
CMIM	79.3	84	86.9	**86.4**	90.7	**92.9**
FSDTree	**80.6**	83.8	**87.5**	86.3	90.6	92.4
DISR	**80.9**	83.8	86.6	**86.9**	**90.8**	**92.5**
Fisher	75	78.2	80.7	81.1	85.9	87.9
Gini	69.9	74.8	76.6	78.6	84.4	87.3
InfoG	77.3	79.9	80.1	84.5	87	88.3
JMI	80	**84.3**	**87.6**	86.3	**90.9**	**92.9**
MIFS	76.2	80.4	83.3	82.9	87.3	89.2
MIM	79.1	81.5	85.5	**86.4**	**90.8**	**92.5**
MRMR	76.8	78.2	80.4	83.3	86.1	88.3
RF	79	**84.5**	86.6	86.3	90.4	92.3
ReliefF	75.8	80.7	84.7	84.6	89.6	91.1
Single	79.3	82.1	84.2	84.8	88.9	90.1
SVMRFEK	73.3	77.2	79.7	79.2	83.9	86
SVMRFEL	77	82.2	85.4	85	89.5	90.8

Table 4.4, shows the classification accuracies of the considered first 50 features in Botswana dataset with regard to three different size of training samples. As can be seen, the DISR method achieves accurate performance with both classifiers when the training data is small. This means that the DISR is a classifier-independent method which works fine with just a small training samples.

Based on the achieved results in Botswana dataset, some of the feature selection methods are dependent to the considered classifier. For example, the MIM works with SVM better than the KNN or the RF is a method that provides higher accuracy in KNN classifier. However, this does not mean that these two methods are dependent to the classifiers in all HSI datasets. For instance, the RF is a classifier-independent method in Indian Pines dataset. In case of Botswana dataset, the FSDTree and JMI are two methods which show the better performance than the other FS methods for both SVM and KNN classifiers.

The methods which obtain higher classification accuracy than the other methods are illustrated in Table 4.5. The results are demonstrated for each classifier. If a method is the winner for both classifiers, it means that the method is a classifier-independent method for the related dataset. For instance, the CMIM is an independent method for the Salinas and SalinasA datasets and shows a higher performance for both KNN and SVM. However, in case of KSC dataset, this method works fine with the SVM only.

Table 4.5 : The first two best FS method for two classifiers with considering classification accuracy of 50 first ranked features over all datasets. Some methods present almost identical performance that are come together.

Dataset	**Best FS Methods**			
	KNN		**SVM**	
	1st	**2nd**	**1st**	**2nd**
Botswana	FSDTree	JMI	JMI	FSDTree
Indian Pines	FSDTree	RF	FSDTree	RF
KSC	RF	CIFE	CMIM, JMI	FSDTree, DISR
Pavia Center	JMI	FSDTree, RF, MRMR	FSDTree	RF, MIM
University of Pavia	MRMR	FSDTree	MRMR	MIM
Salinas	FSDTree, CMIM	DISR	FSDTree	CMIM, DISR
SalinasA	CMIM	JMI, RF	DISR, JMI	CMIM, Gini

A noticeable point that worth to say is about the SalinasA dataset. In this dataset, the Gini is one of the successful feature selection methods when using the SVM as classifier. The Gini is a mediocre method for almost all the datasets and does not show a significant classification accuracy. However, this method achieves accurate performance in the SalinasA dataset. This is possibly due to the small number of classes in this dataset.

It is easy to see that the method FSDTree is the most successful methods with both SVM and KNN classifiers in all the datasets, which means that this method is an independent method from any classification technique for a wide range of HSI datasets. However, in the University of Pavia dataset, FSDTree does not obtain a highest accuracy with SVM, but achieves high accuracy while it is lower than the MRMR and MIM classification accuracy. In this dataset, the MRMR is an accurate method for both SVM and KNN classifiers. In addition, as the FSDTree, the DISR is another independent FS method in Salinas dataset.

The methods FSDTree, RF, and MI based feature selection methods such as JMI, MRMR, and CMIM are the methods with the highest classification accuracy using both KNN and SVM classification methods which make them independent and powerful methods for hyperspectral datasets.

4.1.3 Dependency of FS methods to datasets

This section determines the best FS methods for each dataset separately. For each dataset, the best methods are the methods that achieve the highest classification accuracy for different size of training data and different number of considered features using both classifiers.

Table 4.6 : The best FS methods for each hyperspectral dataset. The best method is the method with highest achieves in classification accuracies.

Sensor	Dataset	Measures		Best Methods	
		Bands	Classes	1st Rank	2nd Rank
EO-1	Botswana	145	14	FSDTree	JMI
AVIRIS	Indian Pines	200	16	FSDTree	RF
	KSC	176	13	CIFE	RF
	Salinas	204	16	FSDTree	CMIM, JMI
	SalinasA	204	6	FSDTree	CMIM, JMI
ROSIS	Pavia Center	102	9	FSDTree	JMI, RF
	University of Pavia	103	9	FSDTree	MRMR

Table 4.6 illustrates the best feature selection methods for each dataset in terms of their classification accuracies. In order to have an explicit view, this table also shows the number of classes, feature, and the type of sensors for each dataset. The best

methods are the methods having the largest number of maximums in classification accuracy. For instance, in Pavia Center dataset the FSDTree, RF, and JMI perform the highest classification accuracy for different size of training data and different number of features than the other methods.

Obviously, the best methods that demonstrate in the Table 4.6 are categorized into two main groups: the methods which use decision trees such as FSDTree and RF, and the methods which use MI measures like JMI, CMIM, and MRMR. The FSDTree is the method that frequently repeated in the table. This method is an appropriate FS method for all HSI datasets except in KSC dataset where the CMIM and RF are found as the best methods.

From a point of view, it can be said that the embedded and wrapper methods like FSDTree and RF gain the more significant results than the filter-based methods. However, filter-based methods like JMI, CMIM, and MRMR are also achieve noticeable results with lower computational costs than the wrapper and embedded methods. Next, the computational cost of feature selection methods will be investigated in Section 4.4.

4.1.4 The optimal number of features

The target of feature selection methods is to find the important features in the original feature set that can increase the classification accuracy. Accordingly, the features that are ranked in the beginning of the feature set is more important than the other features and lead to rise the classification accuracy. In order to select the number of effective features and ignore the less compatible ones from the ranked feature set, a threshold is needed [22]. Finding this threshold is always one of the questions of researchers that is not very easy to solve. In order to assess a feature selection method and determining the optimal number of the features, standard deviation of ten different realization of classification is a criteria that might be considered. Obviously, the low standard deviation means that the classification accuracies of different feature selection methods provides more stable learning curve. Hence, a method can be thought as a best method if it achieves a higher classification accuracy with a lower standard deviation by using minimum number of ranked features.

Table 4.7 : Classification accuracy (%) for Botswana dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	64.4	73.7	77.6	69.9	79.9	80.8	75.9	83.5	85.3	75.1	83.4	85.1	80.2	87.9	90	83.5	89.6	92.3
CIFE	56.7	64.3	75.5	63.3	70.5	80	70.7	76.4	84.2	67.8	74.4	82.9	73.7	80.3	87.5	79	84.2	89.9
CMIM	71.4	76.6	79.3	77.8	**84**	84	79.5	**86.9**	86.9	79.1	84.5	**86.4**	84.1	89.7	90.7	86.3	**92.3**	**92.9**
FSDTree	**77.6**	**78.9**	**80.6**	**83.7**	**83.6**	83.8	**85.5**	**87**	**87.5**	**83.8**	**85.6**	86.3	**89**	**90.4**	90.6	**90.4**	92.1	92.4
DISR	71.6	77.5	**80.9**	78.3	82.9	83.8	79.7	84.2	86.6	78.9	85.2	**86.9**	85.1	89.5	**90.8**	85.5	90.1	**92.5**
Fisher	52	70.1	75	51.5	76.3	78.2	51.2	78.9	80.7	62.2	77.8	81.1	62.3	83.6	85.9	62.9	85.1	87.9
Gini	54	69.7	69.9	58.3	74.2	74.8	59.5	76	76.6	61.1	77.5	78.6	65.1	83	84.4	67	85.9	87.3
InfoG	63.6	72.8	77.3	69.3	75.9	79.9	70.8	77.4	80.1	70.6	80	84.5	76.5	83.6	87	78.5	85.6	88.3
JMI	72.6	77.9	80	78.8	83.2	**84.3**	**84.3**	**86.9**	**87.6**	81	**85.8**	86.3	**86.1**	**90.3**	**90.9**	**89.7**	**92.2**	**92.9**
MIFS	61.2	68.4	76.2	64.8	71.8	80.4	68.2	75.7	83.3	72.4	76.8	82.9	74.9	80.9	87.3	77.8	83.3	89.2
MIM	**75.8**	78.6	79.1	78	81.4	81.5	78.7	84.1	85.5	**82.4**	85.2	**86.4**	85.3	89.3	**90.8**	84.5	89.8	**92.5**
MRMR	57.1	68.5	76.8	61	68.4	78.2	63.3	70.3	80.4	68.7	76.9	83.3	70.5	78.4	86.1	71.3	80.2	88.3
RF	75.6	**79.4**	79	**78.9**	82.6	**84.5**	79.3	84.1	86.6	81.1	85.2	86.3	84.6	88.2	90.4	84.7	89.6	92.3
ReliefF	55.5	73.8	75.8	68.8	79.5	80.7	72.1	83	84.7	69.7	83.2	84.6	77.9	87.6	89.6	79.7	89.3	91.1
Single	64.3	75.4	79.3	67.4	75.5	82.1	63.6	75.8	84.2	71.3	81.6	84.8	74.2	83.3	88.9	71.7	84.2	90.1
SVMRFEK	46.6	55.5	73.3	49.8	58.9	77.2	49.1	61.9	79.7	50.3	63	79.2	53	67	83.9	52.1	69.4	86
SVMRFEL	67.9	76.2	77	73.8	80.8	82.2	76.3	83.9	85.4	75.4	83.5	85	80.8	87.5	89.5	82.2	89.3	90.8
All feature		80.8			85.2			88.1			87.5			91.9			94.1	

Table 4.7 demonstrates the classification accuracies while Table 4.8 shows the standard deviations for Botswana dataset. In this dataset, with using 10 training samples per class, the FSDTree is the method that achieves higher accuracy than the other methods in case of using first 10, 30 and 50 features. This means that this method obtains proper accuracy with a low number of features. In addition, FSDTree shows an acceptable standard deviation. This method shows the same performance in almost all the datasets which gives it a priority when selection an advantageous FS method is considered. The whole results of standard deviations for all datasets are demonstrated in Appendix B.

Table 4.8 : Standard deviation (%) for Botswana dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	7.7	3	**1.6**	5.5	2	1.9	2.8	0.9	0.7	3.9	2	1.9	3.6	1	0.9	2.1	**0.6**	**0.4**
CIFE	4.9	4.5	4.7	6.7	4.7	**1**	6.1	4.9	1.9	3.6	3.1	3	4.1	2.4	0.9	5.1	2.6	1
CMIM	6.8	3.4	1.8	2.6	**1.2**	1.7	4.1	0.8	**0.5**	5.3	2.5	2.1	2.5	0.9	1	3.7	**0.7**	0.8
FSDTree	2.6	2.3	2	**2**	1.7	1.8	**1**	1.4	0.7	2.6	2.5	2.1	**0.8**	**0.7**	0.6	**0.7**	**0.6**	1
DISR	5.4	2.3	**1.5**	3.8	1.3	1.3	2.3	1.1	0.6	3.8	1.8	1.9	3.2	0.8	0.7	1.8	1.1	0.7
Fisher	9.9	5.1	1.7	7.5	1.5	**1.1**	**1**	1.4	0.8	8.1	2.5	**1.3**	5.4	**0.6**	**0.4**	**0.6**	1.2	0.7
Gini	**2.2**	**1.6**	1.9	**0.8**	1	1.7	**1**	1	1	**1.9**	**1.4**	**1.3**	1.2	0.8	0.9	**0.7**	**0.7**	**0.6**
InfoG	7.8	4	3.9	7	1.5	2.8	7.5	0.9	1.9	5.3	2.5	2.6	4.8	1.1	1.8	5.4	0.9	1
JMI	5.7	2.9	2.6	2.8	1.9	1.9	1.7	**0.7**	**0.4**	3.1	1.7	**1.8**	**1.1**	**0.7**	0.7	1.2	**0.6**	**0.6**
MIFS	5.8	4.5	2.6	7.1	4.6	2.1	6.7	5.4	1.4	5.2	3.2	2.5	5.4	3	0.9	5.4	2.9	0.9
MIM	3.2	2.4	3	2.5	2	2.2	2.4	1.2	**0.4**	2.6	2	1.9	2.1	**0.7**	1	2.2	1.3	1
MRMR	4.8	3.3	3.2	6.7	5	1.7	7.6	6.2	2.6	5.2	2.9	2.4	5.6	3	1.4	6.1	3.7	1.3
RF	**2**	**2.1**	2.5	4.1	3	2.7	2.1	1.9	1.8	2	2.5	2	2.9	2.2	1.4	2.3	1.9	1.3
ReliefF	6.3	3.1	2.7	7.6	1.7	2.1	7.3	**0.6**	0.8	3.1	1.6	1.9	4.5	1.2	0.9	4.5	**0.6**	0.7
Single	10	5.2	2.6	7.9	6.3	2.8	9.3	5.4	3	9.7	3.5	1.9	6	5	1.6	7.5	3.8	1.5
SVMRFEK	3.2	2.7	1.8	3.5	**0.9**	1.5	**1**	1.3	1.2	4.5	**1.5**	**1.8**	4.8	1.2	0.7	0.9	1	0.7
SVMRFEL	2.3	**2.1**	2.1	**0.8**	1.4	1.4	**0.7**	0.8	**0.5**	**1.7**	2	2	**0.8**	1	**0.5**	0.9	0.8	0.7
All feature		2.1			1.8			0.8			2			0.5			0.3	

As mentioned before, the DISR method shows a higher accuracy than the other methods with small size of training data in Botswana dataset in SVM classifier. This method achieves this performance exactly with using 50 first features which indicates that this method requires a sufficient number of features. The standard deviation of the DISR is also very low. The CMIM is another method that performs an accurate result with 30 and 50 features, meaning that this method requires a sufficient number of features. In this dataset, the JMI shows a different behaviour. This method works fine with small amount of features only when the size of training data is large enough.

In the KSC dataset, the FSDTree and RF result in higher accuracy than the other methods by using 10 first ranked features. It can be said that for the datasets like KSC which have high correlation between data, these two FS methods show better classification accuracy than the other methods with using small number of ranked features. Whereas, in datasets with low data dependencies, filter-based methods achieve the highest accuracy with using a few ranked features. For instance, in the Salinas dataset, the JMI, which is a filter-based method, shows the higher classification accuracy than the other methods when the number of features considered in the classification is low for all size of training data. The MIM is another filter-based method which shows the same behaviour in SalinasA dataset. Other methods that not described in this section have almost same performances for all three considered feature count and size of training data.

4.1.5 The best FS methods

According to the parameters that are discussed before, to researchers who are looking for an appropriate FS methods in case of classification accuracy for different HSI datasets, the best feature selection methods are shown in Table 4.9. To select the best methods, all the results from investigating factors that impact the classification accuracy such as classification method, size of training samples, and number of ranked features are considered.

Obviously, the demonstrated methods are assorted in three context: FSDTree, RF, and MI based methods (such as JMI, CIFE, CMIM, and MRMR). The FSDTree is one of the most successful methods for classification accuracy in case of different aspects that mentioned before for almost all datasets. The RF is a successful method for Indian

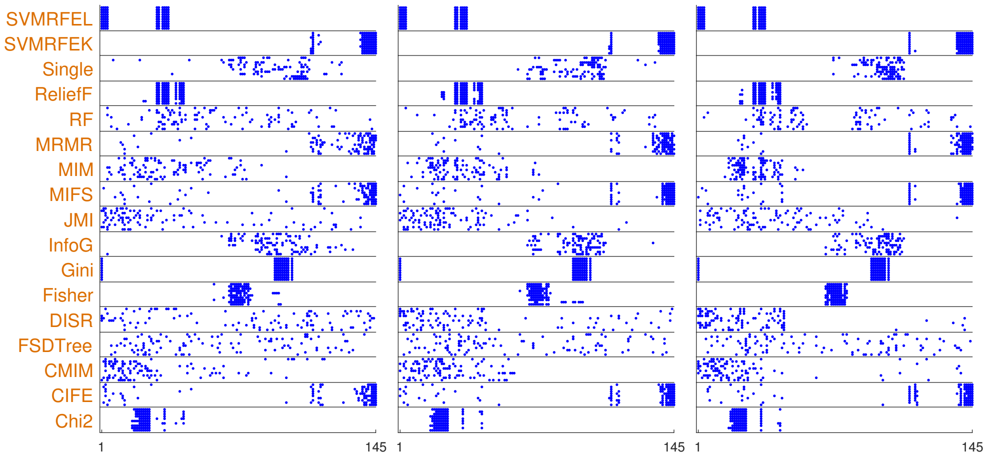
Table 4.9 : The best FS method for HSI datasets based on their classification accuracy and standard deviation results.

Dataset	**Best FS Methods**	
	1st	**2nd**
Botswana	FSDTree	JMI
Indian Pines	FSDTree	RF
KSC	CIFE	FSDTree
Pavia Center	JMI	FSDTree
University of Pavia	MRMR	FSDTree
Salinas	FSDTree	CMIM
SalinasA	CMIM	JMI

Pines dataset. Besides, the methods JMI, CIFE, CMIM, and MRMR also show the highest classification accuracy in the most of hyperspectral datasets.

4.2 Stability of Feature Selection Methods

In this section, the reaction of feature selection methods in terms of their stability without considering any classification method is presented. The aim is to determine the stability of feature selection methods by measuring their robustness to changing the training data. Put differently, when a stable method is utilized to rank the features of a hyperspectral dataset, the one does not need to use this method for each time when the training data is slightly changed. In order to achieve a robust results, the stability of FS methods is examined for 10 different realization of training data. Evidently, a stable method shows the same or very close ranked features in each realization.



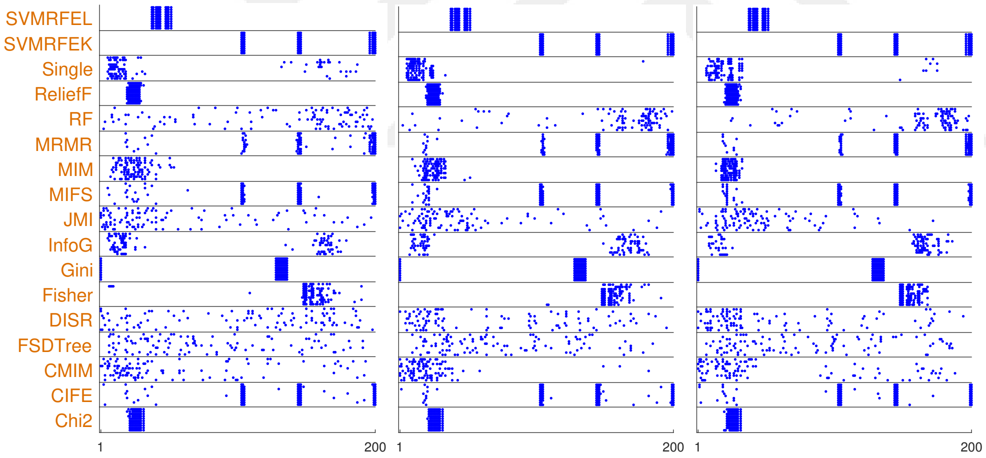
(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

Figure 4.2 : Distribution of the top 10 features provided by FS methods on ten different realizations for Botswana dataset.

Figure 4.2, shows the results of stability for three different-size of training data over the Botswana dataset. The blue dots in plots indicate the index of top 10 selected features that are ranked by each FS method.

It is obvious that the dense plot for each feature selection method, which means the method selects same or very close features in each realization, shows a behaviour of stable method. For example, in Botswana dataset, the methods Chi2, Fisher, Gini, ReliefF, SVMRFEL, and SVMRFEK are very stable methods in all size of training data. The mentioned methods are known as stable methods in almost all datasets meaning the stability of these methods are not depend on the datasets.

Moreover, with increasing the size of training data in the Botswana dataset, the stability of almost FS methods is increased or at least remains constant, expect in two case: JMI and FSDTree. For these two methods, increasing the size of training samples does not effect the stability of methods, however the JMI and FSDTree are totally unstable methods for the Botswana dataset.

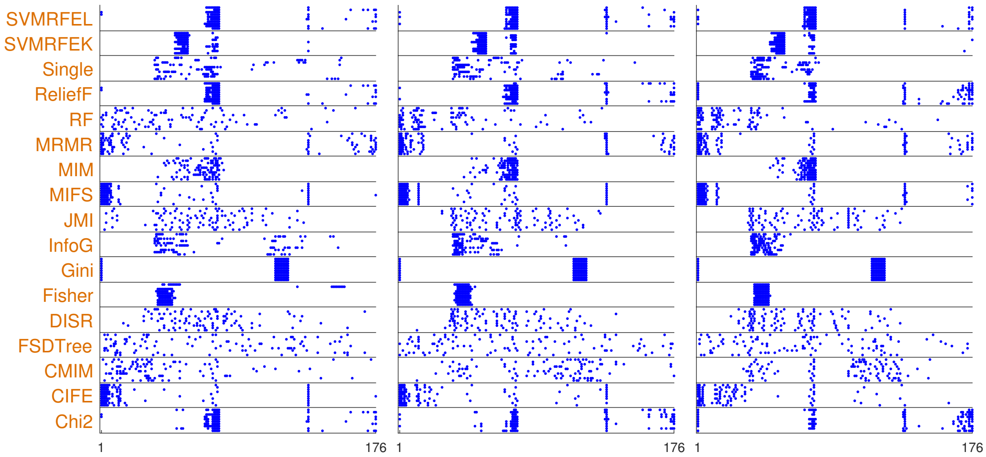


(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

Figure 4.3 : Distribution of the top 10 features provided by FS methods on ten different realizations for Indian Pines dataset.

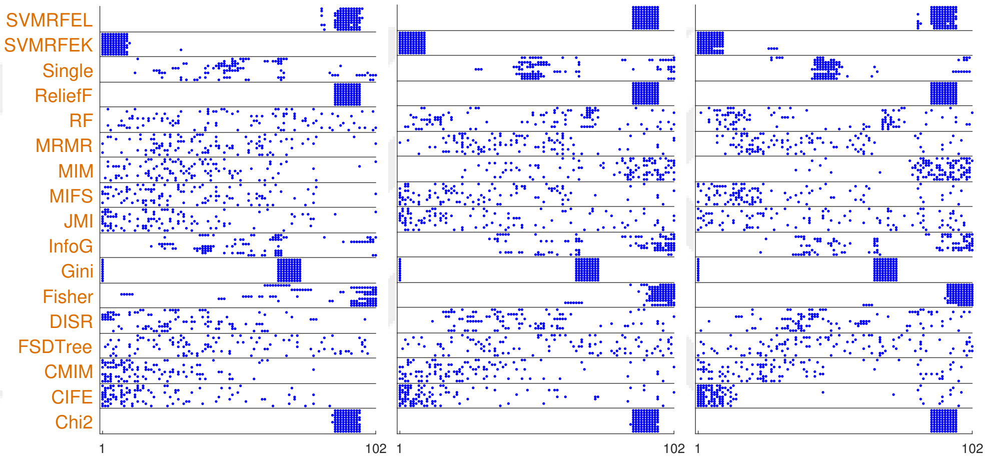
The stability of feature selection methods for the Indian Pines dataset is shown in Figure 4.3. As can be seen, FS methods may not demonstrate similar stability results in all dataset. For example, the "Single" method is a stable technique in Indian Pines dataset. However, in the Botswana, this method is an unstable method.

The stability of feature selection methods for other hyperspectral datasets is illustrated in Figures 4.4 to 4.8. It is worth to say that, in all datasets, almost filter-based methods like JMI, MIM, MIFS, and MRMR that use mutual information measures are not very



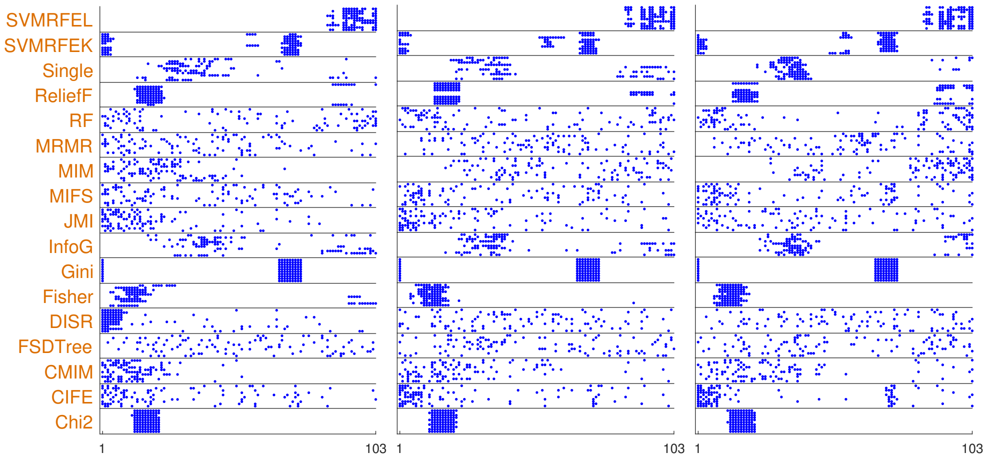
(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

Figure 4.4 : Distribution of the top 10 features provided by FS methods on ten different realizations for KSC dataset.



(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

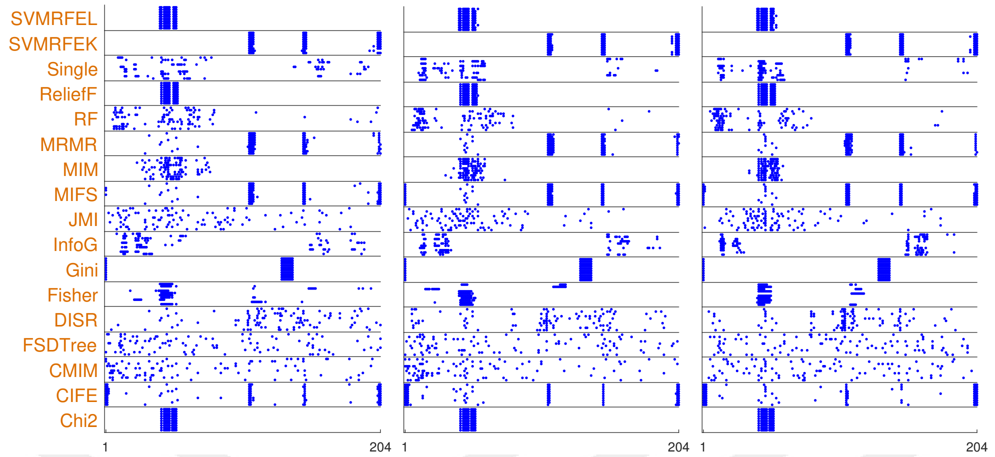
Figure 4.5 : Distribution of the top 10 features provided by FS methods on ten different realizations for Pavia Center dataset.



(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

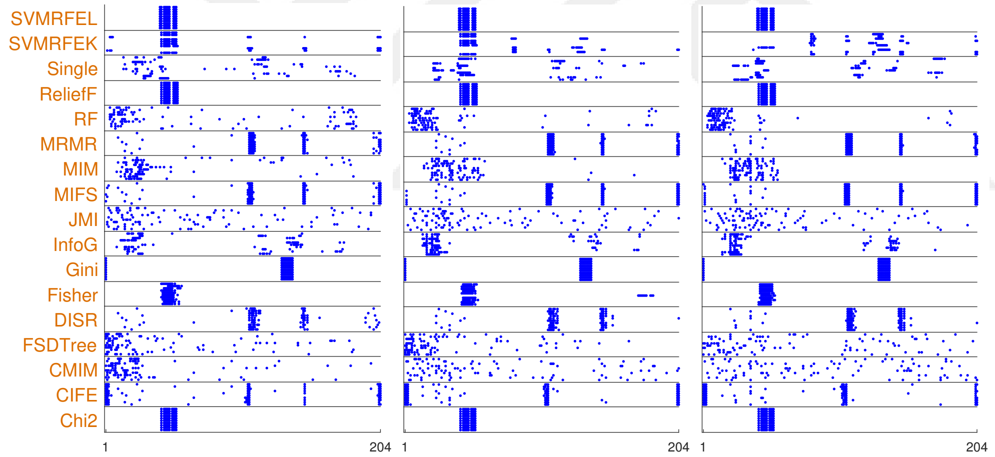
Figure 4.6 : Distribution of the top 10 features provided by FS methods on ten different realizations for University of Pavia dataset.

stable methods. This means that, the MI based feature selection methods are dependent on the training data and with changing the training samples, the selected features are also changed.



(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

Figure 4.7 : Distribution of the top 10 features provided by FS methods on ten different realizations for Salinas dataset.



(a) 10 Samples per class (b) 25 Samples per class (c) 50 Samples per class

Figure 4.8 : Distribution of the top 10 features provided by FS methods on ten different realizations for SalinasA dataset.

As mentioned before, the stability is a measure to rate the robustness of feature selection methods without taking into account any classifier. In other words, there is not any significant correlation between stability of feature selection methods and classification accuracy and vice versa. As an illustration, the Chi2 is an stable method which does not specify a good classification accuracy. In contrast, JMI and MIM which perform a higher classification accuracy, are not the stable ones. Nevertheless, for analysts, a good FS method is a method which is both stable and providing a high classification accuracy. For example, the SVMRFEL and SVMRFEK methods, which both are embedded methods, are a good example that have acceptable accuracy results as long as they are enhanced an adequate level of stability.

4.3 Ability of the Selected Features to Separate Classes

Classification methods are utilized to evaluate the feature selection methods, however, despite the fact that FS methods are profitable, they can not show high performance when the classification methods are not suitable for considered data. Hence, to detect the worthy feature selection methods, the analyses which are independent from any classification method are required. Investigating stability of the FS method is a classifier-independent way that is discussed in Section 4.2. Obviously, while the FS methods rank the features, the important features come first. In order to evaluate these ranked feature sets, the ability of selected features to separate classes is another way to examine the feature selection methods.

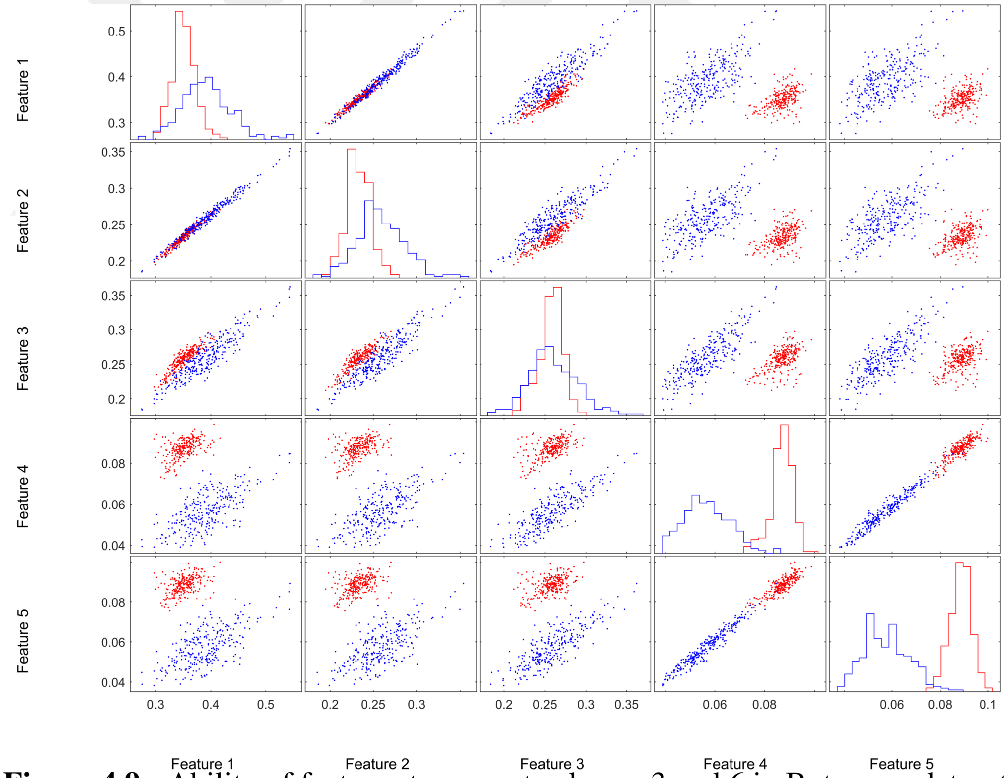


Figure 4.9 : Ability of features to separate classes 3 and 6 in Botswana dataset. Features 1 to feature 5 are the first five important features ranked by RF method.

Figure 4.9 shows the ability of the first five important features ranked by RF method to separate the classes 3 and 6 in Botswana dataset. As mentioned in section 3.1.1, these classes are the most difficult ones to separate. As can be seen, 1st and 2nd features are not able to distinguish two classes, however, they are first two important features from the view of RF method. Diagonal histograms are clearly show that the classes 3 and 6 are very hard to separate by using these two first features. Instead, feature pairs such as (4,1), (4,2), (4,3) can strongly isolate two mentioned classes as clear as possible. Similarly, the combination of 5th ranked feature with the first three features are another strong candidates that separate the classes.

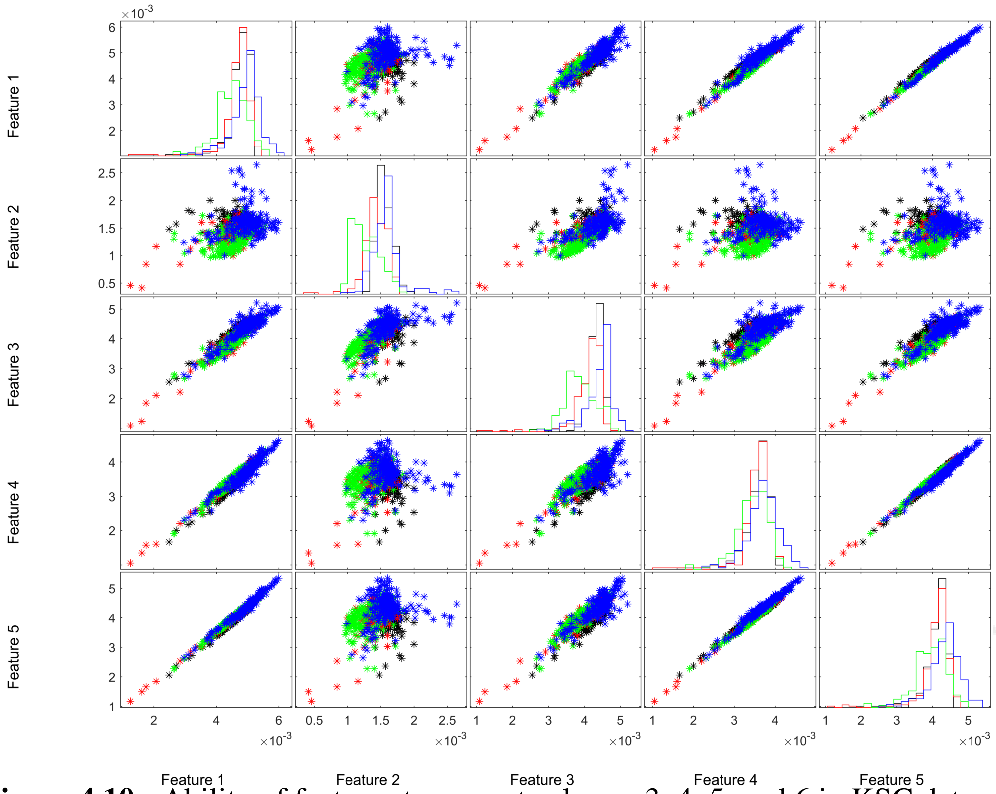


Figure 4.10 : Ability of features to separate classes 3, 4, 5, and 6 in KSC dataset. Features 1 to Feature 5 are first five important feature ranked by JMI method.

The discrimination of classes are not always easy like in the Botswana dataset. For example, in Figure 4.10 which shows the most complicated classes (Classes 3, 4, 5, and 6) for KSC dataset, the first five important features that ranked by JMI method can not separate the classes simply. In this dataset, the selected features are not able to distinguish the classes effortless, meaning that the classes of KSC dataset are suffering from a high degree of data correlation.

Figure 4.11 illustrates the median of first 50 features of four most complicated classes for KSC dataset that ranked by the JMI method. As can be resulted from the figure, while the medians are too close to each other, the features can not be able to separate

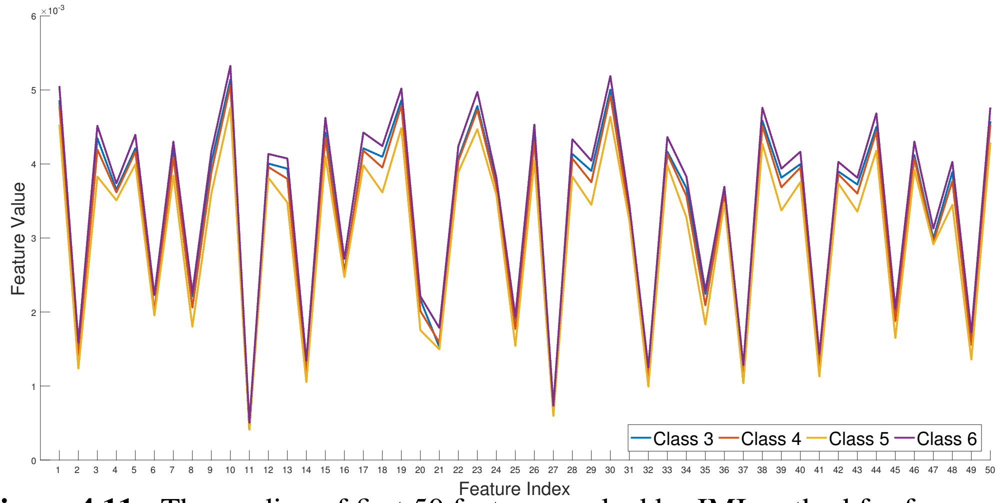


Figure 4.11 : The median of first 50 features ranked by JMI method for four most complicated classes of KSC dataset.

the classes as easy as possible. This inability means that if the features ranked by the JMI is used for classification, the classification accuracy are not significant.

4.4 Computational Time

Another key point to evaluate the performance of an FS method is its computational time. The computational time, is measured by taking account the time that a feature selection method needs to rank the features.

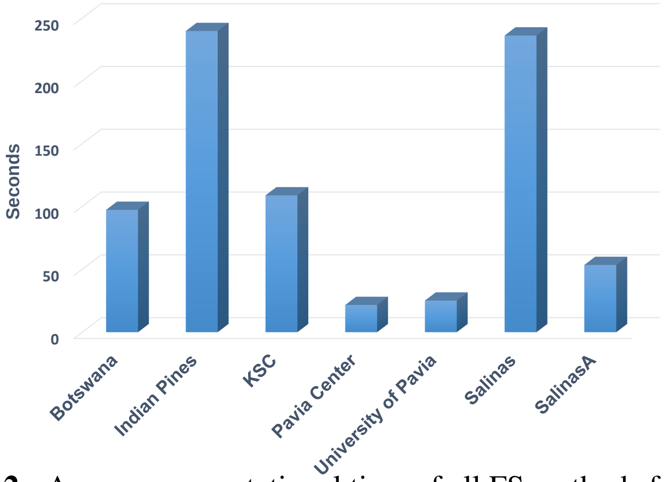


Figure 4.12 : Average computational time of all FS methods for different hyperspectral datasets.

The computational time of feature selection methods can be affected by the structure of each dataset. The number of classes, correlation between the features, and the level of redundancies can influence the time required to select the important features. Figure 4.12, illustrates the average CPU-time of all FS methods for each hyperspectral dataset.

As demonstrated in the figure, while Indian Pines and Salinas have more abundant features and classes than the other datasets, FS methods consume more computational time in these two HSI datasets than the others. Respectively, in case of Pavia datasets (Pavia Center and University of Pavia), FS methods are faster than the other datasets.

Table 4.10 : The required computational times (in seconds) of FS methods to rank features for training data with size of 50 samples per class for each dataset.

Methods	Botswana	Indian Pines	KSC	Pavia Center	University of Pavia	Salinas	SalinasA
Chi2	<1	<1	<1	<1	<1	<1	<1
CIFE	49	88	54	17	18	94	36
CMIM	48	89	51	16	16	94	36
FSDTree	2411	6200	2537	443	548	5837	1098
DISR	131	247	140	43	45	253	96
Fisher	<1	<1	<1	<1	<1	<1	<1
Gini	15	17	4	6	6	17	5
InfoG	<1	<1	<1	<1	<1	<1	<1
JMI	48	87	52	15	16	96	35
MIFS	48	86	52	16	16	95	35
MIM	48	85	52	16	16	94	35
MRMR	48	84	52	16	16	93	34
RF	<1	<1	<1	<1	<1	<1	<1
ReliefF	4	4	2	1	1	5	<1
Single	107	161	139	41	44	174	95
SVMRFEK	15	19	13	5	5	21	1
SVMRFEL	1	2	2	<1	<1	1	<1

However, structure of the HSI datasets does not effect the computational time of FS methods as much as the configuration and procedure of FS methods. Table 4.10 shows the computational time of all feature selection methods for training data with the size of 50 samples per class. In order to achieve the reliable values, the computational time is measured for ten different realization of training data and an average of CPU-time is reported.

Evidently, the FSDTree method expends more CPU time than the other methods to rank the features. This means that, in all datasets, the FSDTree is a costly method. However, this method performs the significant classification accuracies and is found to be one of the accurate methods for dimensionality reduction of remotely sensed hyperspectral datasets. The DISR is another method that spends more computational time due to the fact that it uses two major characteristic of feature selection [68]. However, the DISR method achieves a good level of classification accuracy (like FSDTree) for datasets when the number of samples and classes of datasets are high.

It is easy to see that the methods CMIM, JMI, MIM, MIFS, and MRMR have almost the same computational time because they apply MI measurements as a common framework [44] [26].

As it has been pointed out in the Table 4.10, the methods Chi2, Fisher, InfoG, and RF are the methods that show the lowest computational time and therefore, are very fast methods. From all these methods, the RF is a successful method in terms of both the classification accuracy and the computational time. The SVMRFEL is another method that spends fewer time than the other methods to rank the features because this method uses a linear kernel with SVM classifier. However, SVMRFEL has one of the lowest performances in classification accuracies.

5. CONCLUSIONS

Feature selection is an effective and advantageous research domain in the area of hyperspectral images in remote sensing. By using feature selection on such a big data that contains lots of irrelevant and redundant features, the complicity of remotely sensed hyperspectral datasets as well as redundancy of data might be decreased without incurring much loss of information. However, selecting an appropriate FS method for considered HSI dataset is a dilemma.

There are three general categories for feature selection methods: filter, wrapper, and embedded methods. In this work, a review of seventeen feature selection methods is enforced on 7 well-known hyperspectral remotely sensed datasets. From these FS methods, twelve of them are filter-based, two are wrapper, and three of them are embedded. These methods are demonstrated with intention of studying their performance and outcomes. Besides, two famous classification methods are conducted to compare the influence of ranked features and to select a convenient FS method: the SVM classifier with a linear kernel and the KNN classifier as a non-linear classification method. In order to assess the FS methods that are used in this work, four evaluation criteria are considered: classification accuracy, stability of feature selection methods, ability of selected features to separate classes, and computational time of each method. These assessment measures are examined for three different size of training data over all seven hyperspectral remotely sensed datasets.

As classification accuracy evaluation criteria, the methods are investigated from different aspects such as the effects of datasets, size of training samples, classification method, and number of considered features on classification procedure. The experimental results show that the methods JMI, MIM, MIFS, CMIM, and MRMR as filter-based methods obtain higher classification accuracy than the other filter-based methods. These mentioned methods use mutual information measurements to rank the feature sets. Besides, FSDTree as a wrapper and RF as an embedded method, perform higher accuracy even greater than filter-based methods. These results are not unexpected because both the RF and FSDTree, use a classifier as an evaluator.

In case of stability of feature selection methods, the results demonstrate that Chi2, Fisher, Gini, ReliefF, SVMRFEK, and SVMRFEL are the most stable feature selection methods meaning that these methods select almost the same features when slightly changed training data is used, therefore they are more reliable than the others. However, the classification accuracy achieved by these methods is not satisfying. In contrast, the methods with higher classification accuracy than the other FS methods, that are FSDTree, RF, and MI based methods, are not found to be very stable methods, meaning that these methods should be used for each training data set.

As a third key point, the ability of features ranked by FS methods are examined to separate the most complicated classes of a dataset. The achieved results show that the selected features are adequately strong to separate the classes. However, in the datasets, which have high correlated data, separating the classes is challenging indeedly.

The forth evaluation criteria demonstrates the computational time that a feature selection method required to distinguish the effective features. The results clarify that the methods FSDTree, DISR, and Single occupied more CPU time than the other methods. The FSDTree also shows the highest computational time. However, in the case of classification accuracy, this method is the significant one. In contrast, the MI based methods like CIFE, CMIM, JMI, MIFS, MIM, and MRMR achieve acceptable computational time. It is worthy to say, the RF which is an embedded method, shows very low computational time, however, it is an embedded method.

In light of the results illustrated in this work, the MI based filter methods are suggested to settle the classification problems and analysis of hyperspectral images in the area of remote sensing.

- Firstly, these methods are independent from any classifiers, and they do not need to set any parameter that must be setted. Hence, the implementation of these methods are quite easy.

- Secondly, the computational time of these methods is lower than the wrapper and embedded methods.

- Thirdly, the MI based methods have a good generalization ability for different samples and data.

The RF method can be another alternative to tackle the dimensionality reduction problems of HSI datasets. This method demonstrates a high level of classification accuracy as well as low computational time. However, the implementation of the RF is more difficult that filter-based methods.

The FSDTree is a powerful method in case of classification accuracy, but spend more computational time and is known as a tardy method. In case where the computational time is not important, this method can achieve significant results.

REFERENCES

[1] **Li, Q. and Bernal, E.A.** (2016). Hybrid tenso-vectorial compressive sensing for hyperspectral imaging, *Journal of Electronic Imaging*, *25*(3), 033001.

[2] **Landgrebe, D.A.** (2003). *Signal theory methods in multispectral remote sensing*, Wiley series in remote sensing, Wiley, Hoboken, N.J., Chichester.

[3] **Grahn, H. and Geladi, P.** (2007). *Techniques and applications of hyperspectral image analysis*, Chichester, England ; Hoboken, J. Wiley.

[4] **Chang, C.I.** (2003). *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, volume 1, Springer Science and Business Media.

[5] **Wang, Y.**, **Chen, G. and Maggioni, M.** (2016). High-Dimensional Data Modeling Techniques for Detection of Chemical Plumes and Anomalies in Hyperspectral Images and Movies, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *PP*(99), 1–9.

[6] **Chen, T.**, **Yuen, P.**, **Richardson, M.**, **Liu, G. and She, Z.** (2014). Detection of Psychological Stress Using a Hyperspectral Imaging Technique, *IEEE Transactions on Affective Computing*, *5*(4), 391–405.

[7] **Fukunaga, K. and Hayes, R.R.** (1989). Effects of Sample Size in Classifier Design, *IEEE Trans. Pattern Anal. Mach. Intell.*, *11*(8), 873–885.

[8] **Hughes, G.** (1968). On the mean accuracy of statistical pattern recognizers, *IEEE Transactions on Information Theory*, *14*(1), 55–63.

[9] **Chi, M.**, **Feng, R. and Bruzzone, L.** (2008). Classification of hyperspectral remote-sensing data with primal {SVM} for small-sized training dataset problem, *Advances in Space Research*, *41*(11), 1793–1799.

[10] **Jain, A. and Zongker, D.** (1997). Feature selection: evaluation, application, and small sample performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158.

[11] **Liu, H.**, **Dougherty, E.**, **Dy, J.**, **Torkkola, K.**, **Tuv, E.**, **Peng, H.**, **Ding, C.**, **Long, F.**, **Berens, M.**, **Parsons, L.**, **Zhao, Z.**, **Yu, L. and Forman, G.** (2005). Evolving feature selection, *Intelligent Systems, IEEE*, *20*(6), 64–76.

[12] **Navot, A.**, **Gilad-Bachrach, R.**, **Navot, Y. and Tishby, N.** (2006). Is Feature Selection Still Necessary?, volume3940 of *Lecture Notes in Computer Science*, Springer.

[13] **Guyon, I. and Elisseeff, A.** (2003). An Introduction to Variable and Feature Selection, *The Journal of Machine Learning Research*, *3*, 1157–1182.

[14] **Kohavi, R. and John, G.H.** (1997). Wrappers for feature subset selection, *Artificial Intelligence*, *97*(1), 273–324.

[15] **Loughrey, J. and Cunningham, P.**, (2005). Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets, pp.33–43.

[16] **Pal, M. and Foody, G.** (2010). Feature Selection for Classification of Hyperspectral Data by SVM, *IEEE Transactions on Geoscience and Remote Sensing*, *48*(5), 2297–2307.

[17] **Ghamisi, P.**, **Couceiro, M.S. and Benediktsson, J.A.** (2015). A Novel Feature Selection Approach Based on FODPSO and SVM, *IEEE Transactions on Geoscience and Remote Sensing*, *53*(5), 2935–2947.

[18] **Patra, S.**, **Modi, P. and Bruzzone, L.** (2015). Hyperspectral Band Selection Based on Rough Set, *IEEE Transactions on Geoscience and Remote Sensing*, *53*(10), 5495–5503.

[19] **Hossain, M.A.**, **Jia, X. and Benediktsson, J.A.** (2016). One-Class Oriented Feature Selection and Classification of Heterogeneous Remote Sensing Images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(4), 1606–1612.

[20] **Persello, C. and Bruzzone, L.** (2016). Kernel-Based Domain-Invariant Feature Selection in Hyperspectral Images for Transfer Learning, *IEEE Transactions on Geoscience and Remote Sensing*, *54*(5), 2615–2626.

[21] **Zabalza, J.**, **Ren, J.**, **Yang, M.**, **Zhang, Y.**, **Wang, J.**, **Marshall, S. and Han, J.** (2014). Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, *93*, 112–122.

[22] **Bolón-Canedo, V.**, **Sánchez-Maroño, N. and Alonso-Betanzos, A.** (2013). A review of feature selection methods on synthetic data, *Knowledge and Information Systems*, *34*(3), 483–519.

[23] **Pohjalainen, J.**, **Räsänen, O. and Kadioglu, S.** (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits, *Computer Speech and Language*, *29*(1), 145–171.

[24] **Ang, J.C.**, **Mirzal, A.**, **Haron, H. and Hamed, H.N.A.** (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(5), 971–989.

[25] **Vergara, J.R. and Estévez, P.A.** (2014). A review of feature selection methods based on mutual information, *Neural Computing and Applications*, *24*(1), 175–186.

[26] **Brown, G.**, **Pocock, A.**, **Zhao, M.J. and Luján, M.** (2012). Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection, *The Journal of Machine Learning Research*, *13*, 27–66.

[27] **G.Vijouyeh, H. and Taşkın, G.** (2016). A comprehensive evaluation of feature selection algorithms in hyperspectral image classification, *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp.489–492.

[28] **Liu, H. and Setiono, R.** (1995). Chi2: Feature Selection and Discretization of Numeric Attributes, ***J. Vassilopoulos**, editor, Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, November 5-8, 1995*, IEEE Computer Society, pp.388–391.

[29] **Lin, D. and Tang, X.** (2006). Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion., *ECCV (1)*, volume3951 of *Lecture Notes in Computer Science*, Springer, pp.68–82.

[30] **Fleuret, F.** (2004). Fast Binary Feature Selection with Conditional Mutual Information, *The Journal of Machine Learning Research*, *5*, 1531–1555.

[31] **Jakulin, A. and Bratko, I.** (2003). Quantifying and Visualizing Attribute Interactions, **Technical Report**.

[32] **Meyer, P.E.**, **Schretter, C. and Bontempi, G.** (2008). Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity, *IEEE Journal of Selected Topics in Signal Processing*, *2*(3), 261–274.

[33] **Duda, R.**, **Hart, P. and Stork, D.** (2001). *Pattern Classification*, John Wiley & Sons, New York, 2 edition.

[34] **Aggarwal, C.C.** (2014). *Data Classification: Algorithms and Applications*, Chapman & Hall/CRC, 1st edition.

[35] **Forman, G.** (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification, *The Journal of Machine Learning Research*, *3*, 1289–1305.

[36] **Mitchell, T.** (1997). *Machine Learning*, McGraw-Hill.

[37] **Cover, T.M. and Thomas, J.A.** (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience.

[38] **Roobaert, D.**, **Karakoulas, G. and Chawla, N.V.** (2006). *Feature Extraction: Foundations and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg.

[39] **Yang, H.H. and Moody, J.** (1999). Data Visualization and Feature Selection: New Algorithms for Nongaussian Data, *in Advances in Neural Information Processing Systems*, MIT Press, pp.687–693.

[40] **Battiti, R.** (1994). Using mutual information for selecting features in supervised neural net learning, *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

[41] **Lewis, D.D.** (1992). Feature Selection and Feature Extraction for Text Categorization, *Proceedings of the Workshop on Speech and Natural Language*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp.212–217.

[42] **Cassarà, P. and Rozza, A.** (2016). A Novel Mutual Information-based Feature Selection Algorithm, *CoRR*, *abs/1607.07186*.

[43] **Wells, W.M.**, **Viola, P.**, **Atsumi, H.**, **Nakajima, S. and Kikinis, R.** (1996). Multi-modal volume registration by maximization of mutual information, *Medical Image Analysis*, *1*(1), 35–51.

[44] **Peng, H. Long, F.D.C.** (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.

[45] **Kononenko, I.**, **Šimec, E. and Robnik-Šikonja, M.** (1997). Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF, *Applied Intelligence*, *7*(1), 39–55.

[46] **Freeman, C.**, **Kuli, D. and Basir, O.** (2015). An evaluation of classifier-specific filter measure performance for feature selection, *Pattern Recognition*, *48*, 1812–1826.

[47] **Kira, K. and Rendell, L.** (1992). The feature selection problem: Traditional methods and a new algorithm, *aaai*, 129–134.

[48] (2008). *Computational Methods of Feature Selection*, Chapman and Hall.

[49] **Breiman, L.**, **Friedman, J.H.**, **Olshen, R.A. and Stone, C.J.** (1984). Classification and regression trees. Wadsworth & Brooks, *Monterey, CA*.

[50] **Baraniuk, R.G.** (2007). Compressive Sensing [Lecture Notes], *IEEE Signal Processing Magazine*, *24*(4), 118–121.

[51] **Liaw, A. and Wiener, M.** (2002). Classification and Regression by randomForest, *R News*, *2*(3), 18–22.

[52] **Breiman, L.** (2001). Random Forests, *Machine Learning*, *45*(1), 5–32.

[53] **Breiman, L.** (1996). Bagging Predictors, *Machine Learning*, *24*(2), 123–140.

[54] **Guyon, I.**, **Weston, J.**, **Barnhill, S. and Vapnik, V.** (2002). Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, *46*(1-3), 389–422.

[55] **Shieh, M.D. and Yang, C.C.** (2008). Multiclass SVM-RFE for product form feature selection, *Expert Systems with Applications*, *35*(1—2), 531–541.

[56] **Nie, L.**, **CHU, H. and Korostyshevskiy, V.R.** (2008). Bias reduction for nonparametric correlation coefficients under the bivariate normal copula assumption with known detection limits, *The Canadian Journal of Statistics*, *36*(3), 427–442.

[57] **Melgani, F. and Bruzzone, L.** (2004). Classification of hyperspectral remote sensing images with support vector machines, *IEEE Transactions on Geoscience and Remote Sensing*, *42*(8), 1778–1790.

[58] **Vapnik, V. and Kotz, S.** (2006). *Estimation of Dependences Based on Empirical Data*, Information Science and Statistics, Springer.

[59] **Vapnik, V.N.** (1998). *Statistical Learning Theory*, Wiley-Interscience.

[60] **Cortes, C. and Vapnik, V.** (1995). Support-Vector Networks, *Machine Learning*, *20*(3), 273—-297.

[61] **Bhatia, N. and Vandana** (2010). Survey of Nearest Neighbor Techniques, *CoRR*, *abs/1007.0085*.

[62] **Rajan, S.**, **Ghosh, J. and Crawford, M.M.** (2008). An Active Learning Approach to Hyperspectral Data Classification, *IEEE Transactions on Geoscience and Remote Sensing*, *46*(4), 1231–1242.

[63] **Landgrebe, D.A. and Biehl, L.L.**, (1994), MultiSpec - A Tool for Multispectral-Hyperspectral Image Data Analysis, `https://engineering.purdue.edu/~biehl/MultiSpec/index.html`.

[64] **Oguslu, E.**, **Zhou, G. and Li, J.** (2013). Hyperspectral image classification using a spectral-spatial sparse coding model, *Proc. SPIE*, *8892*, 88920R–88920R–6.

[65] **Dalla Mura, M.**, **Villa, A.**, **Benediktsson, J.A.**, **Chanussot, J. and Bruzzone, L.** (2011). Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis, *IEEE Geoscience and Remote Sensing Letters*, *8*(3), 542–546.

[66] **NASA**, (2011), AVIRIS Data Portal, `http://aviris.jpl.nasa.gov/`.

[67] **Burges, C.J.** (1998). A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

[68] **Bennasar, M.**, **Hicks, Y. and Setchi, R.** (2015). Feature selection using Joint Mutual Information Maximisation, *Expert Systems with Applications*, *42*(22), 8520–8532.

APPENDICES

1. **APPENDIX A:** Classification Accuracy
   - **Appendix A.1 :** Tables
   - **Appendix A.2 :** Graphs
2. **APPENDIX B:** Standard Deviation

APPENDIX A

Appendix A.1

Table A.1 : Classification accuracy (%) for Botswana dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	64.4	73.7	77.6	69.9	79.9	80.8	75.9	83.5	85.3	75.1	83.4	85.1	80.2	87.9	90	83.5	89.6	92.3
CIFE	56.7	64.3	75.5	63.3	70.5	80	70.7	76.4	84.2	67.8	74.4	82.9	73.7	80.3	87.5	79	84.2	89.9
CMIM	71.4	76.6	79.3	77.8	84	84	79.5	86.9	86.9	79.1	84.5	86.4	84.1	89.7	90.7	86.3	92.3	92.9
FSDTree	77.6	78.9	80.6	83.7	83.6	83.8	85.5	87	87.5	83.8	85.6	86.3	89	90.4	90.6	90.4	92.1	92.4
DISR	71.6	77.5	80.9	78.3	82.9	83.8	79.7	84.2	86.6	78.9	85.2	86.9	85.1	89.5	90.8	85.5	90.1	92.5
Fisher	52	70.1	75	51.5	76.3	78.2	51.2	78.9	80.7	62.2	77.8	81.1	62.3	83.6	85.9	62.9	85.1	87.9
Gini	54	69.7	69.9	58.3	74.2	74.8	59.5	76	76.6	61.1	77.5	78.6	65.1	83	84.4	67	85.9	87.3
InfoG	63.6	72.8	77.3	69.3	75.9	79.9	70.8	77.4	80.1	70.6	80	84.5	76.5	83.6	87	78.5	85.6	88.3
JMI	72.6	77.9	80	78.8	83.2	84.3	84.3	86.9	87.6	81	85.8	86.3	86.1	90.3	90.9	89.7	92.2	92.9
MIFS	61.2	68.4	76.2	64.8	71.8	80.4	68.2	75.7	83.3	72.4	76.8	82.9	74.9	80.9	87.3	77.8	83.3	89.2
MIM	75.8	78.6	79.1	78	81.4	81.5	78.7	84.1	85.5	82.4	85.2	86.4	85.3	89.3	90.8	84.5	89.8	92.5
MRMR	57.1	68.5	76.8	61	68.4	78.2	63.3	70.3	80.4	68.7	76.9	83.3	70.5	78.4	86.1	71.3	80.2	88.3
RF	75.6	79.4	79	78.9	82.6	84.5	79.3	84.1	86.6	81.1	85.2	86.3	84.6	88.2	90.4	84.7	89.6	92.3
ReliefF	55.5	73.8	75.8	68.8	79.5	80.7	72.1	83	84.7	69.7	83.2	84.6	77.9	87.6	89.6	79.7	89.3	91.1
Single	64.3	75.4	79.3	67.4	75.5	82.1	63.6	75.8	84.2	71.3	81.6	84.8	74.2	83.3	88.9	71.7	84.2	90.1
SVMRFEK	46.6	55.5	73.3	49.8	58.9	77.2	49.1	61.9	79.7	50.3	63	79.2	53	67	83.9	52.1	69.4	86
SVMRFEL	67.9	76.2	77	73.8	80.8	82.2	76.3	83.9	85.4	75.4	83.5	85	80.8	87.5	89.5	82.2	89.3	90.8
All feature		80.8			85.2			88.1			87.5			91.9			94.1	

Table A.2 : Classification accuracy (%) for IndianPines dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	32.5	41	46.1	35.1	46	53.5	37	50.1	56.4	36.1	48.7	54.6	36.7	53.5	60.2	39.4	55.5	62.7
CIFE	30.4	35.7	37.7	31.7	40.9	40.3	38.3	48.1	45.7	32.4	45.4	50.1	35.9	51.3	57	39.2	54.9	60.8
CMIM	42.2	47.2	45.4	39.5	52.5	52.2	48.1	57.8	56.9	46.9	55.1	52.2	46	59.2	61.1	55.7	66.4	65.8
FSDTree	46.9	48.2	48.1	53.9	55.6	54.7	58.5	59.8	59.7	54.5	55	54.8	62	64.2	63	64.7	67.8	67.5
DISR	43.3	45.7	46.4	49	54.1	52.9	51.9	56.6	56.4	52.1	50.9	52.8	57.9	61.6	61.4	59.4	63.9	64.5
Fisher	35.7	39.2	42	37.5	43.7	47.1	43.5	47.8	53.2	39.2	49.4	54.2	42	54.6	60.4	43.8	53.1	62
Gini	27.8	30.9	32.6	32.4	36.5	39.8	37.7	43	46.3	36	42.3	51	41.2	49.5	56.1	44.3	52.9	58.6
InfoG	36.5	40.3	41.7	38.7	44.6	46.9	41.6	47.3	50.4	44.2	49.1	52.1	47	56	58.1	47.4	59.9	62.9
JMI	42.6	47.6	48.3	48	53.2	53.8	56.2	58.3	57.6	49.5	51.8	53.4	55.9	61.1	61.1	60.3	63.7	63.9
MIFS	33.2	37.2	38.7	31.4	37	39.6	35.9	42.1	44	34.9	45.8	50.7	34	49.2	57.6	37.2	50.5	60.5
MIM	41.7	45.5	47.9	40.5	49.9	51.6	41.1	52.7	55.1	46.5	52.9	55.4	47.1	54.7	59	45.4	57	59
MRMR	29.2	36.2	35.5	29.9	38.2	36.9	32.9	42.8	42.3	34.6	47.4	51	33.3	49.1	55.1	36.4	50.9	58.2
RF	44.9	49.8	47.9	40.6	53.9	54.4	45.1	58	59.9	51.3	56.8	52.2	48.1	63.3	63.8	48.7	65.8	67.6
ReliefF	33.4	42.8	45.8	36.6	48.5	51.4	38.9	52.7	54.2	36.7	48.7	51.2	38.7	52.4	55.7	41.4	56	56.9
Single	35.6	40	42.1	38.1	43.1	46.6	43.1	48.4	51.5	41.2	47.9	51.7	40.1	50.8	56.2	48.2	59.4	62.7
SVMRFEK	24.6	36.1	43.9	28.6	39.8	47.9	32.4	45.3	53.1	28.2	41.4	51.9	30.1	44	53.8	35.3	45.5	56.3
SVMRFEL	29.5	43.7	46.8	32.5	49.6	53.4	35.1	53.7	56.9	36.5	47.7	52.6	38.1	50.8	55	39.5	52.5	57.1
All feature		49.9			54.9			58.4			53.8			63.1			68.4	

Table A.3 : Classification accuracy (%) for KSC dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

	KNN									SVM								
Method	**10 SpC**			**25 SpC**			**50 SpC**			**10 SpC**			**25 SpC**			**50 SpC**		
	10F	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**
Chi2	56.6	63.4	64.4	65.7	70.4	71.1	67.1	76.3	77.1	41.9	41.2	43.8	52.1	56.7	59.7	51.9	64	66.7
CIFE	70	75.3	74.7	78.6	81.4	81.6	84.3	85.5	85.3	51.7	59.6	59.2	51.5	55.2	54.1	59.1	50.6	54.4
CMIM	74.3	75	73.7	75.2	77.1	77.8	79.6	81.7	82.6	55.4	57.9	57.5	57.5	60.1	58.4	66.8	71.6	74.4
FSDTree	72.2	72.4	72	78.7	77.7	77.7	82.7	83.2	82.7	55.9	56.7	55.9	54.9	54.9	58.9	65.2	71.9	77.1
DISR	69.5	71.3	71.5	77.5	79	78.8	82.6	82.4	82	49.7	52.2	52.5	56.5	58.8	63.4	65.1	74.4	77
Fisher	50.4	61.5	65.7	57.4	71.7	74	61.9	76.7	78	38.6	41.8	45.9	36.3	44.8	60.3	42.9	60.1	67.1
Gini	49.4	63.2	65.7	54.8	68.2	71	57.5	72.3	75.4	42.5	50.3	52.8	42.6	55.5	58.6	44.3	63.3	66.6
InfoG	61.8	66.9	68.9	68.4	74.2	75.5	71	77.5	79.7	46.5	49.7	52.8	42.3	54	60.9	47	63.4	72
JMI	69.6	70	70.2	77.6	78.6	78.4	81.2	81.8	81.9	50.5	52.1	52.6	56.6	61.8	63.3	65.1	74.3	76.9
MIFS	72.2	75.4	75.4	77.4	80.9	80.4	78.2	83.5	83.7	50.2	58.1	59	49.2	54.2	53	58.1	52.6	49.7
MIM	58.7	62.4	65.1	62.1	68.6	71.1	67.6	73.3	75.7	41.5	43.6	46.3	45.5	52.1	61.5	50.3	60.7	66.1
MRMR	71.9	74.2	74.5	75.2	78.8	79.8	79.2	82.1	84	53.6	57.9	58	49.8	52.6	53.2	61	54.8	56.7
RF	73.9	75	75.1	79.6	83	82.6	82	85.8	85.4	54.1	55.7	55.6	53.8	57.8	57.9	56.8	68.1	73.1
ReliefF	53.9	62.6	64	61	69.3	70.6	68.9	76.6	77	41.1	43.6	45.5	49	54.2	59.9	56.2	63.7	66.6
Single	62	67.3	68.4	72.1	74.9	75.8	71.2	78.4	79.9	47.2	51.6	52.9	49.3	53.9	55	52.8	65.9	72.3
SVMRFEK	57.3	60.7	66.9	63.3	67.4	74.5	67.4	72.7	79.9	36.9	38.5	41.8	46	51.7	59.2	52.8	60.5	68.7
SVMRFEL	55	63	63.1	61.8	70.3	70.9	66.2	75.5	75.2	41.4	44.2	45.3	51.2	57.1	59.1	55.2	63.6	65.7
All feature		72.4			78.4			82.8			50.6			66.3			81	

Table A.4 : Classification accuracy (%) for PaviaCenter dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

	KNN									SVM								
Method	**10 SpC**			**25 SpC**			**50 SpC**			**10 SpC**			**25 SpC**			**50 SpC**		
	10F	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**
Chi2	62.3	77.4	81.6	61.4	82.7	87	61.8	85.7	89.9	67.4	88	93.1	66.6	90.1	94.9	66.8	92.2	95.6
CIFE	82.6	91.1	91.6	84.8	93.1	93.5	83.9	93.2	94.4	88.1	94	94.7	89.1	95.7	96	89.3	95.7	96.3
CMIM	71.5	87.9	91.1	87.8	93.5	94.3	92.9	94.9	95.1	79.4	92.9	94.3	93	96	96.1	95.5	96.6	96.6
FSDTree	90.5	91	91.2	93.2	93.7	94	95.1	95	95.2	94.5	94.8	94.8	95.9	96.1	96.2	96.6	96.8	97
DISR	81.7	87.3	89.7	88.9	92.1	92.8	91.6	94.7	94.6	88.4	91.5	93.1	93.3	95.3	95.7	94.5	96.3	96.3
Fisher	72.4	81.1	85.4	65.4	81.8	87.4	66.9	83.3	89.2	77.7	89.2	93.9	72.9	89.7	94.8	73.7	89.7	95.6
Gini	68.2	79.5	80.2	73.2	86	86.3	75.7	87.9	88.5	81.1	91.4	92.4	85.8	93.8	94.1	87.4	94.8	95.2
InfoG	81.3	87.6	90	78.3	91.5	92.9	86.3	94.2	94.5	88.2	93.5	94.5	84.7	95.4	96.2	90.7	96.8	96.8
JMI	84.7	91.7	91.3	90.8	93.9	94.1	93.9	95.3	95.3	91.1	94.5	94.7	94.1	96	96.2	96.1	96.6	97
MIFS	87.8	91	91.5	90.8	92.6	93.6	94	94.2	93.9	91.7	94	94.5	94.3	95.6	96	95.8	95.8	96.1
MIM	86.5	91.3	91.2	86.4	92.7	93.4	79	92.5	94.4	90.7	95.1	95	92.2	96	96.3	85	96.5	96.8
MRMR	86.4	90.8	91.2	92.7	94	94	94.3	95.1	94.9	92.1	94.2	94.6	95.3	96.1	96.1	96.3	96.3	96.5
RF	90.8	91.7	91.6	90.6	93.6	93.9	92.5	94.7	95.2	94.4	94.9	95	94.8	95.9	96.2	95.8	96.5	96.8
ReliefF	62	70.4	79.6	61.5	74.6	86.5	61.8	78.4	88.5	67.2	81	92.7	66.4	83.3	94.7	66.8	85.3	95.4
Single	82.8	87.8	90.9	79.9	90.7	92.4	78.1	93.1	94.6	88.7	93.2	94.4	87.6	95.2	96	84.7	95.7	96.5
SVMRFEK	48.3	71.5	88.9	51.8	71.9	91.2	51.3	74.9	91.9	51.7	80.3	92	53.1	83.4	95.2	55.3	85.8	95.2
SVMRFEL	62.2	69.6	80.1	61.5	74.6	86.5	63.7	78.4	88.5	68.4	80.9	92.9	66.4	83.3	94.7	69.1	85.3	95.4
All feature		91			94.1			95.4			95.1			96.4			97	

Table A.5 : Classification accuracy (%) for PaviaUniversity dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	48.4	54.7	60.5	49.2	60.2	66.6	51	64.3	69.9	48.8	55.6	62.1	50.9	60.8	69.7	53.5	62.9	71.1
CIFE	59.9	62.8	61.7	62.1	66.2	67.8	65	69.6	71	61.9	67.6	68.2	65.4	74.1	77	67.9	76.3	78.6
CMIM	49.1	56	60.8	63.5	66.5	68.2	68.2	70.9	70.8	50.6	59.8	66.4	65.9	75.8	76.3	70.8	77.8	77.1
FSDTree	62.1	61.4	62.3	66.4	67	67.5	69.4	70.1	70	66.7	68	68.6	72.3	76.4	78.4	73.8	78.1	79.1
DISR	51.4	58.4	60.2	64	67.6	67.6	68.8	72	72	52.7	60	64.3	70.8	77.8	80	74.8	79.5	80.4
Fisher	47.9	55.4	60.4	48.7	59.6	65.8	47.8	62.6	69.6	47.6	56.1	62.2	48.7	61	71.1	48.5	61.5	72.2
Gini	55.2	57.6	58.3	57.8	63.8	64.2	61.1	66.7	67.5	58.6	64.9	66.9	65.7	75.6	77.6	69.9	77.1	79.3
InfoG	55	59.8	60.8	58.2	64.7	66.9	63.4	68.5	71	57.4	65.1	69.2	64.3	72.7	77.3	66	73.2	75.9
JMI	54.5	59.3	61	62.2	65.5	66.3	68.1	69.9	71	54	66	67.1	67.9	75.7	79.1	72.9	79.8	80.2
MIFS	58.6	61.7	61.1	64.3	66.9	67.4	67.7	70.3	70.6	61.9	66.5	66.8	71.1	76.2	77.9	71.9	77.1	77.9
MIM	53.1	60.1	60.9	65.6	66.7	68.1	66.6	70.3	70.7	59	66.7	67.5	73.6	79.2	80.3	73.8	77.9	79.3
MRMR	56.6	61.8	61.5	65.1	68.1	67.5	69.7	71.7	72.2	58.9	67.3	69.5	73	78.5	79.5	73.5	79.7	80.4
RF	57.5	61.5	62	62.7	64.9	66.7	65.4	69.4	70.8	64.1	66.1	67.5	71.4	76.8	79.9	70.5	79	79.8
ReliefF	51.3	58.7	60.3	51.9	63	63.3	56.6	66.8	70	53.9	61.5	65.1	56.8	72	75.3	59.6	73.4	76.1
Single	55.8	59.9	59.9	56	66.3	66.5	58.4	68.7	71.5	58.9	65.8	68.6	61	75.5	77.7	59.7	72	76.4
SVMRFEK	53.6	57.8	58.9	56.3	61.8	62.9	61	65.5	68.5	57.2	63.6	63.7	63.1	70.7	72.4	67.1	74.3	75.7
SVMRFEL	48.3	57	59.3	53	63.5	66.9	57.7	69.7	70.1	56	65.3	67.6	63.2	74.9	78.5	65.6	76.1	80
All feature		62.4			68.3			71.8			68.2			80.3			80.7	

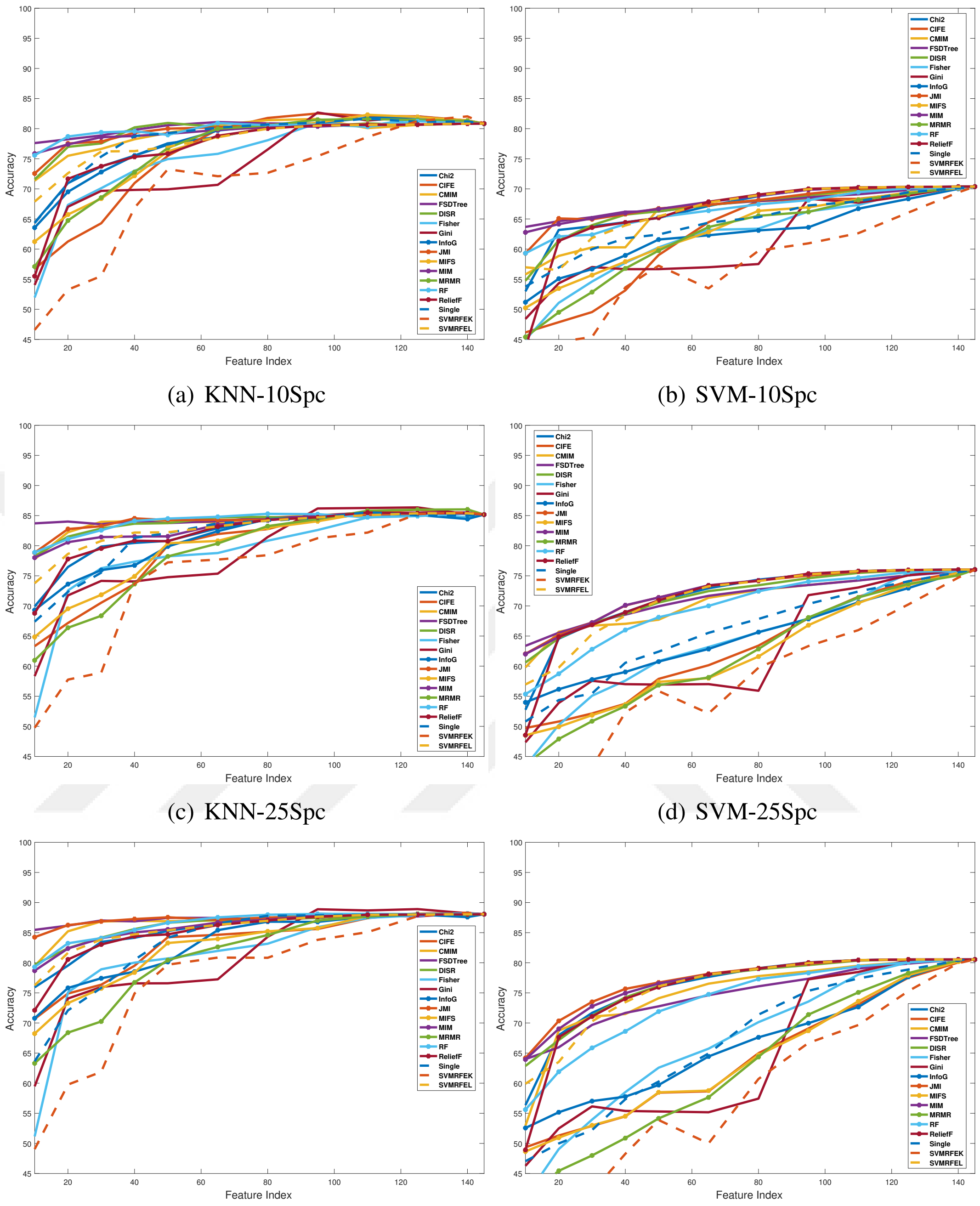
Table A.6 : Classification accuracy (%) for Salinas dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	64.3	76.6	78.9	67.2	78.9	80.7	69.8	81	82.9	72.7	80.1	81.9	75.9	84.2	85.9	76.9	85.1	86.9
CIFE	64.3	75.3	78.6	73.5	80.3	82.2	74.9	82.5	83.8	68.1	78.6	80.9	75.8	83.7	85	79	86.3	87.4
CMIM	78	80.3	80.4	80.9	81.6	82.8	83.6	84.3	84.6	79.9	82.5	83	84.7	87.1	87.3	88.1	89.2	89.3
FSDTree	79.7	80	80.6	80.8	82.3	82.3	83.6	84.2	84.1	80.9	82.7	83	85.3	87.5	87.5	87.5	88.8	89.4
DISR	76.2	80.1	81.2	80.3	82.3	83	82.6	84.3	84.6	78.9	82	83.3	85.3	87.5	87.7	86.8	88.7	89.2
Fisher	66.3	73	77.2	69.1	78	80.2	71.2	80	82.1	72.2	79.1	81.1	76.3	83.9	86	77.4	84.7	87.2
Gini	51.4	72.5	75.9	59.3	75.7	78.2	63.2	77.7	80.2	61	77	79.3	68.9	81.8	84	71.7	83.7	85.8
InfoG	74.2	78.8	79.5	77.4	80.9	81.5	80.6	83.2	83.3	77.9	81.4	81.8	83.7	86.4	86.9	84.7	87.7	88
JMI	79.3	79.5	79.6	80.8	81.4	81.4	82.9	83.2	83.3	81.2	82.1	82.8	84.9	87.3	87.2	88.2	88.8	88.8
MIFS	64.2	72.4	78.1	64.7	75.6	81.2	66.4	78.7	83.2	70.5	76.1	80.5	74	82.1	84.5	75.9	84.8	87.3
MIM	74	77.7	78.9	71.7	79.3	80.6	73	81.4	82.7	79.4	81.2	82	79.4	84.9	85.9	79.9	86.2	86.9
MRMR	64.3	75.1	79	68.5	77	81.4	73.2	79.6	83.1	70.5	78.5	81.5	75.7	84.2	85.5	78.8	85.7	87.3
RF	77.5	79.2	79.5	79.9	81.3	81.5	82.6	83.2	83.6	80.4	82.9	82.9	85	87.5	87.5	85.8	88.1	88.6
ReliefF	64.4	76.7	78.9	67.6	79.3	80.6	70.1	81.2	82.9	72.3	80.5	81.7	76.4	84.5	86	77.1	85.8	87.1
Single	76.6	78.1	79	78.1	80.2	80.6	79.5	81.8	82.6	80.4	81.2	81.8	82.7	85.7	87	83	86.6	88.2
SVMRFEK	63.7	64.2	74.4	65.4	68.5	77	67.4	72.1	79.8	57.6	73.3	79	61.4	77.8	83	66.5	80.9	85.7
SVMRFEL	65	76.2	79.4	69.3	80	81.2	72.2	82.2	83.1	74.3	81.2	82.2	78.1	85.7	85.9	79.3	86.5	87.1
All feature		80.6			82.9			84.6			84			87.1			88.4	

Table A.7 : Classification accuracy (%) for SalinasA dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**
Chi2	84.8	93.3	95.4	90.5	95.7	96.2	93.1	97.1	97.5	94.6	96.2	96.7	96.2	97.5	97.6	97.6	98.3	98.4
CIFE	84.9	95.3	96.6	91.1	95.9	97	91	96.6	97.6	86.4	95.5	96.8	94.1	96.6	97.5	94.4	97.3	97.7
CMIM	94.7	96.3	96.7	95.9	96.7	96.9	97.5	97.8	97.9	94.8	96.9	97.1	97.6	97.9	97.8	98.3	98.6	98.6
FSDTree	95.4	95.9	96	95.8	96.7	97.1	97.1	97.6	97.9	95.4	97.1	97.2	96.2	97.7	97.8	98.2	98.5	98.4
DISR	89.7	95.1	95.6	87.8	96.3	96.7	91.9	97.1	97.7	81.5	97.4	97.8	87.8	97.9	98.2	91.7	98.4	98.8
Fisher	83.7	92	93.1	89.5	94.4	95	91.9	96.3	96.8	94.2	95.5	96.3	95.7	97.6	97.6	97.1	98.2	98.3
Gini	80.2	86.8	90.2	84.6	90.6	93.5	89.9	93.8	96	73.6	91	95	94.6	97.9	98.1	96.5	98.5	98.9
InfoG	90.1	94.5	94	94	95.1	95.4	95.6	96.4	96.7	90.9	96.7	97.1	94.4	96.6	97.5	95.3	97.9	98.4
JMI	95.8	96.3	96.5	96.7	96.7	96.6	97.5	97.6	97.7	96.8	97.2	97	97.6	97.8	97.8	98.3	98.4	98.6
MIFS	82.7	91.9	96.3	84.1	91.4	96.9	86.5	93.2	97.5	78.7	94	96.4	87	96.1	97.7	91.6	97.2	98
MIM	94.6	95.6	96.1	95.6	95.9	96.3	97.1	97.3	97.5	96.4	96.8	97	97.1	97.8	97.8	98.1	98.4	98.5
MRMR	86.5	94.8	95.8	84.4	94.1	97	87.1	95.9	97.6	73.7	94.4	97.3	80.6	97.5	98	82.6	98	98.4
RF	95.5	96.6	96.9	94.5	96.6	97	95.7	97.5	97.8	96.1	96.9	96.9	95.7	97.5	97.7	96.6	98.3	98.5
ReliefF	85.9	92.5	95.6	90.7	95.6	96.6	93.7	97.2	97.5	94.9	96	96.7	96.6	97.8	97.6	97.7	98.4	98.4
Single	91.6	93.5	93.9	89.8	94.3	95.7	91.5	96.4	97.3	92.8	96.3	96.4	92.6	96.6	97.6	89.6	97.8	98.4
SVMRFEK	83.6	91.7	94.5	89.8	95.3	96.1	93	96.9	97.5	88.1	93.3	96.2	92.2	97.5	97.6	92.5	98.3	98.7
SVMRFEL	84.5	95.4	96.2	91	96	96.5	93.1	97.4	97.4	94.7	96.6	96.9	96.2	97.5	97.7	97.6	98.5	98.5
All feature		96.9			97			98			97.3			97.9			98.7	

Appendix A.2



(a) KNN-10Spc

(b) SVM-10Spc

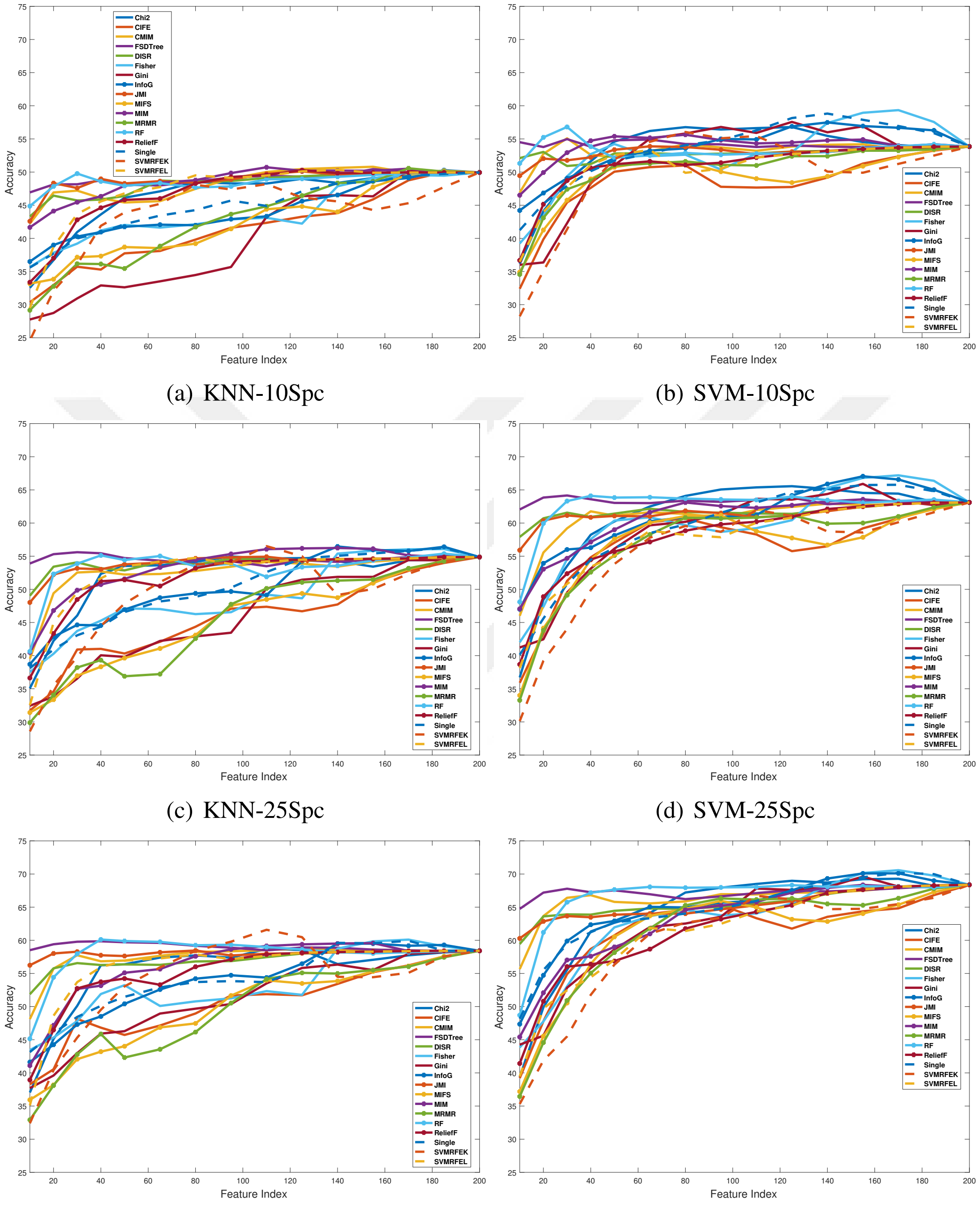
(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.1 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over Botswana dataset. The results are averaged over ten realizations.



(a) KNN-10Spc

(b) SVM-10Spc

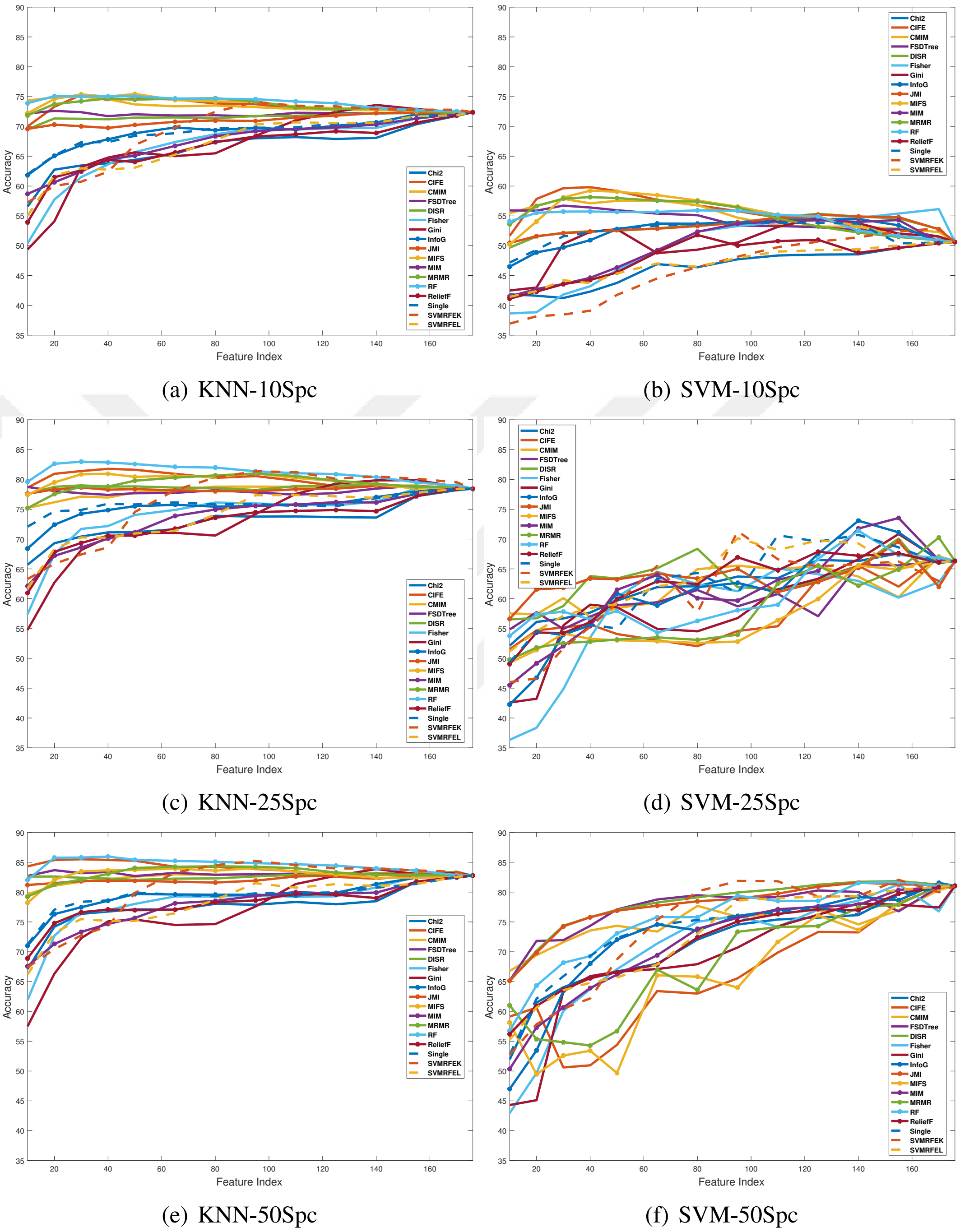
(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.2 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over IndianPines dataset. The results are averaged over ten realizations.

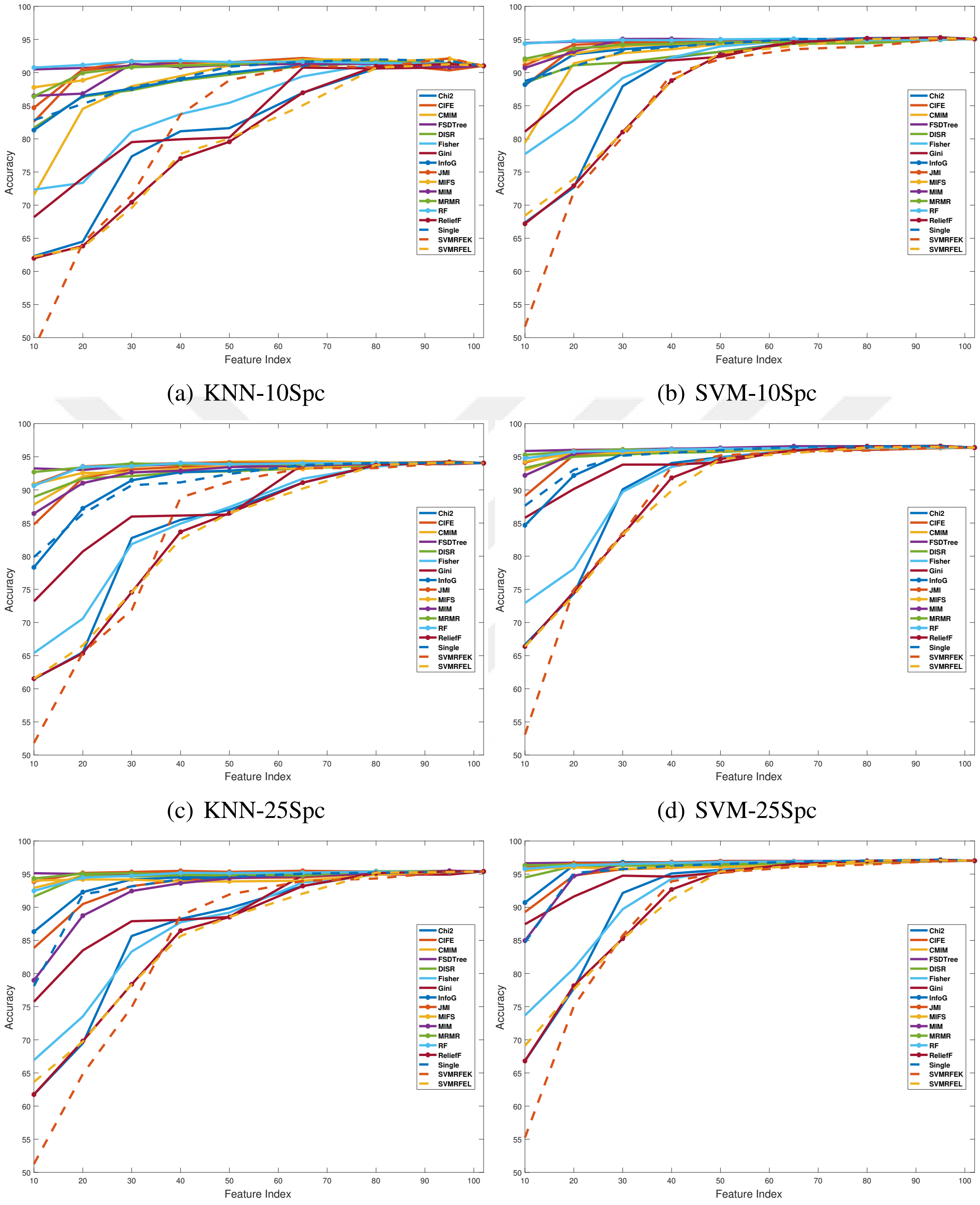


(a) KNN-10Spc (b) SVM-10Spc

(c) KNN-25Spc (d) SVM-25Spc

(e) KNN-50Spc (f) SVM-50Spc

Figure A.3 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over KSC dataset. The results are averaged over ten realizations.



(a) KNN-10Spc

(b) SVM-10Spc

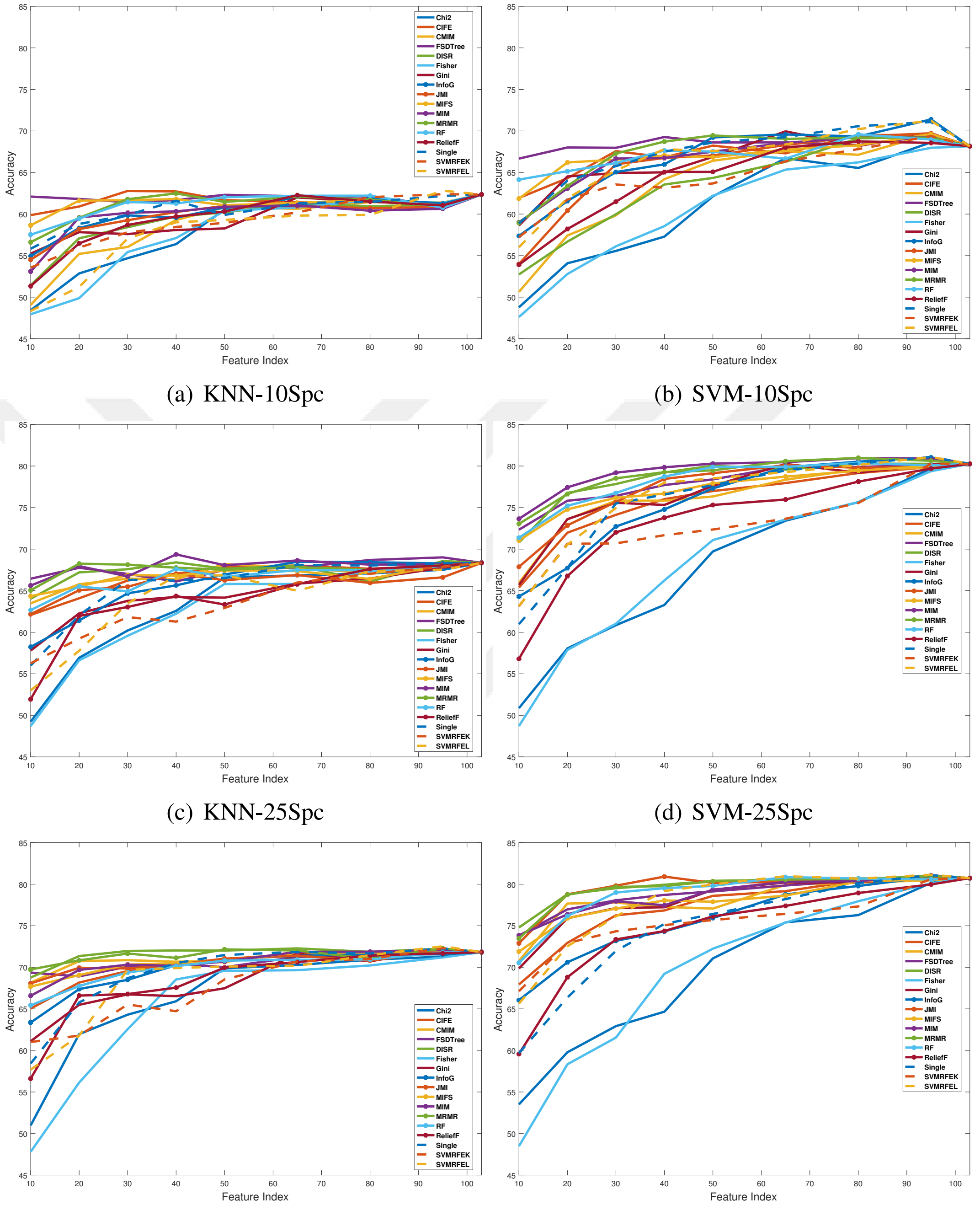
(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.4 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over PaviaCenter dataset. The results are averaged over ten realizations.



(a) KNN-10Spc

(b) SVM-10Spc

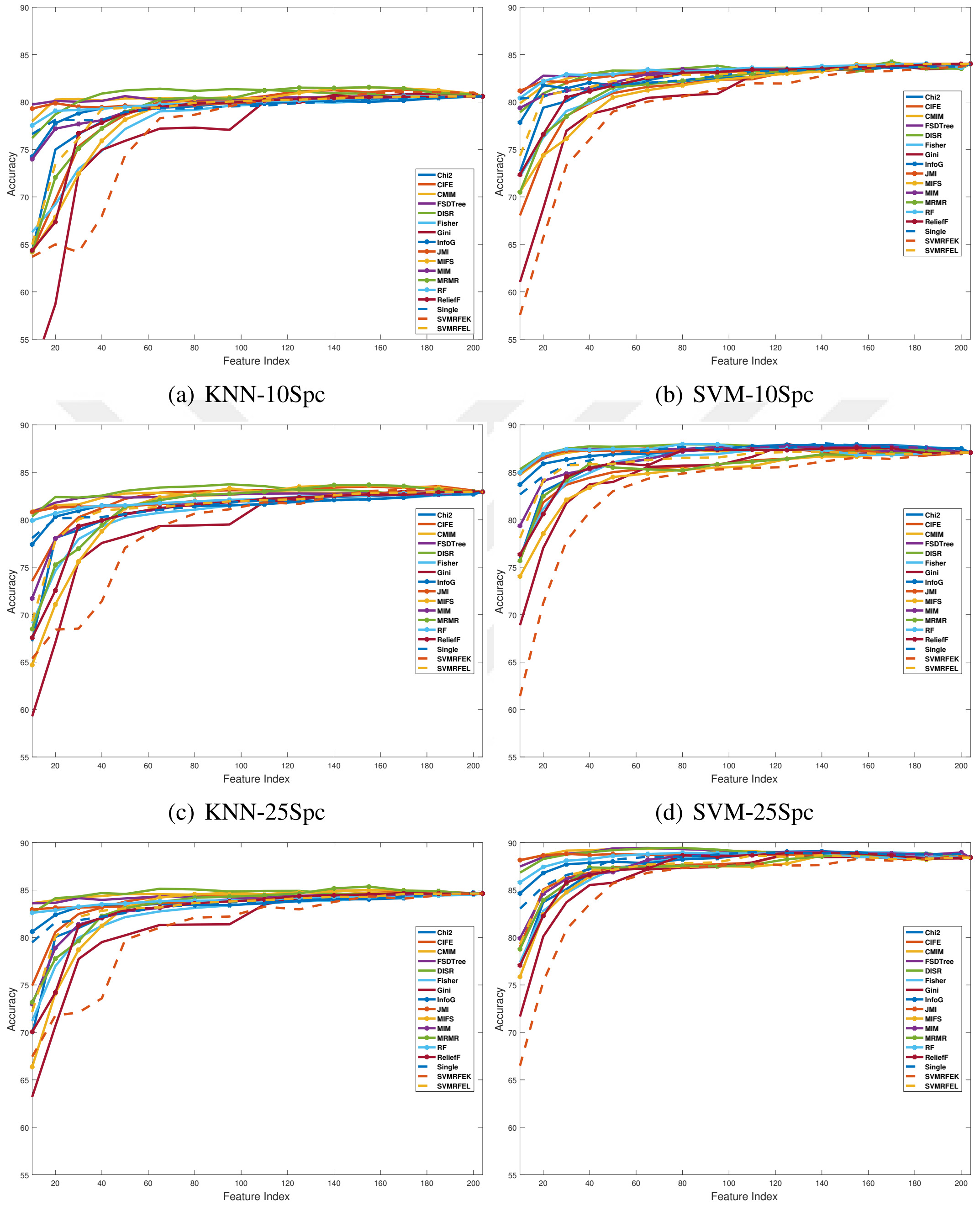
(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.5 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over PaviaUniversity dataset. The results are averaged over ten realizations.



(a) KNN-10Spc

(b) SVM-10Spc

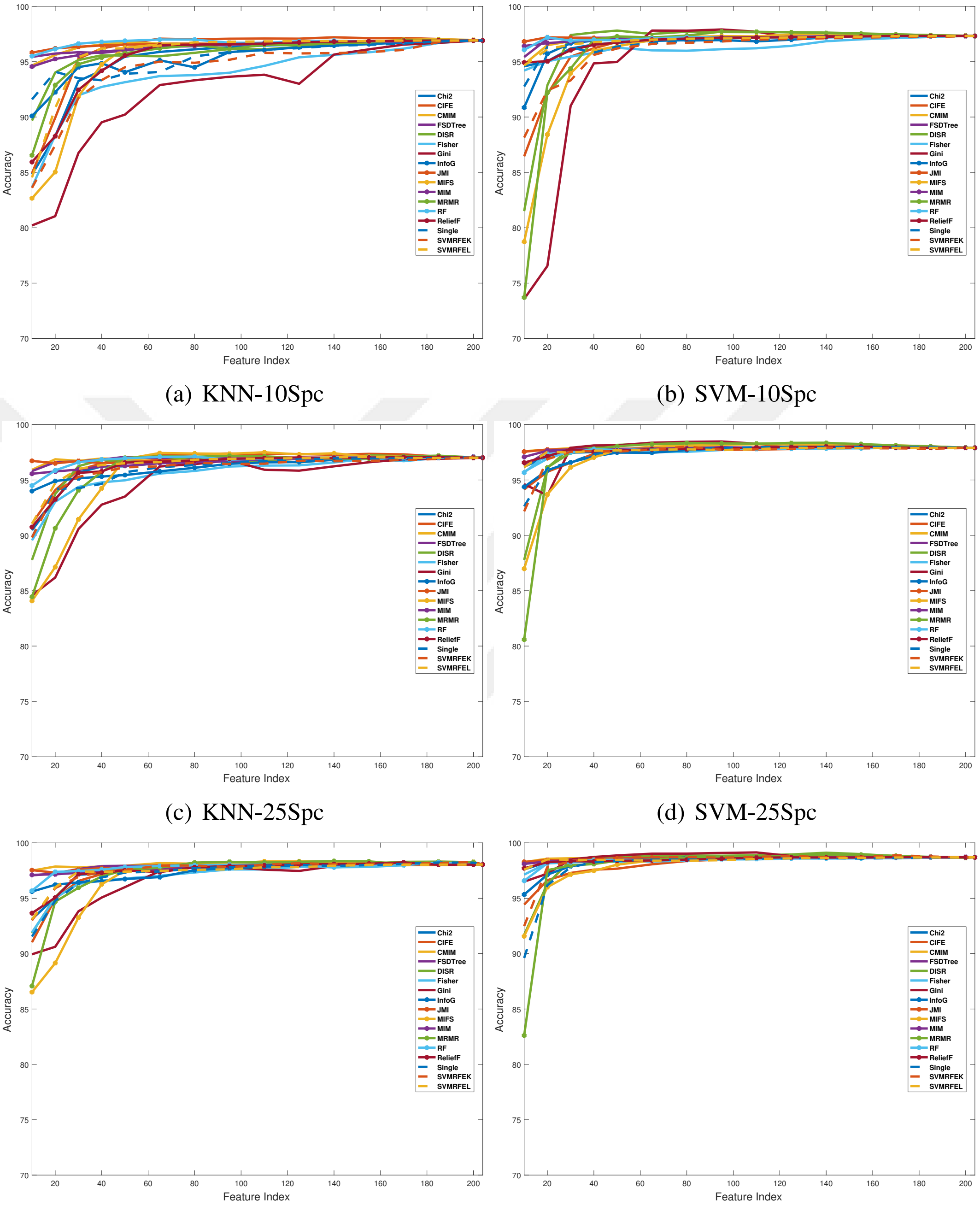
(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.6 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over Salinas dataset. The results are averaged over ten realizations.



(a) KNN-10Spc

(b) SVM-10Spc

(c) KNN-25Spc

(d) SVM-25Spc

(e) KNN-50Spc

(f) SVM-50Spc

Figure A.7 : Classification accuracies obtained by the classifiers for the features ranked by different feature selection methods for all training data size (SpC: samples per class) over SalinasA dataset. The results are averaged over ten realizations.

APPENDIX B

Table B.1 : Standard deviation of ten different realizations of classification accuracy for Botswana dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	7.7	3	1.6	5.5	2	1.9	2.8	0.9	0.7	3.9	2	1.9	3.6	1	0.9	2.1	0.6	0.4
CIFE	4.9	4.5	4.7	6.7	4.7	1	6.1	4.9	1.9	3.6	3.1	3	4.1	2.4	0.9	5.1	2.6	1
CMIM	6.8	3.4	1.8	2.6	1.2	1.7	4.1	0.8	0.5	5.3	2.5	2.1	2.5	0.9	1	3.7	0.7	0.8
FSDTree	2.6	2.3	2	2	1.7	1.8	1	1.4	0.7	2.6	2.5	2.1	0.8	0.7	0.6	0.7	0.6	1
DISR	5.4	2.3	1.5	3.8	1.3	1.3	2.3	1.1	0.6	3.8	1.8	1.9	3.2	0.8	0.7	1.8	1.1	0.7
Fisher	9.9	5.1	1.7	7.5	1.5	1.1	1	1.4	0.8	8.1	2.5	1.3	5.4	0.6	0.4	0.6	1.2	0.7
Gini	2.2	1.6	1.9	0.8	1	1.7	1	1	1	1.9	1.4	1.3	1.2	0.8	0.9	0.7	0.7	0.6
InfoG	7.8	4	3.9	7	1.5	2.8	7.5	0.9	1.9	5.3	2.5	2.6	4.8	1.1	1.8	5.4	0.9	1
JMI	5.7	2.9	2.6	2.8	1.9	1.9	1.7	0.7	0.4	3.1	1.7	1.8	1.1	0.7	0.7	1.2	0.6	0.6
MIFS	5.8	4.5	2.6	7.1	4.6	2.1	6.7	5.4	1.4	5.2	3.2	2.5	5.4	3	0.9	5.4	2.9	0.9
MIM	3.2	2.4	3	2.5	2	2.2	2.4	1.2	0.4	2.6	2	1.9	2.1	0.7	1	2.2	1.3	1
MRMR	4.8	3.3	3.2	6.7	5	1.7	7.6	6.2	2.6	5.2	2.9	2.4	5.6	3	1.4	6.1	3.7	1.3
RF	2	2.1	2.5	4.1	3	2.7	2.1	1.9	1.8	2	2.5	2	2.9	2.2	1.4	2.3	1.9	1.3
ReliefF	6.3	3.1	2.7	7.6	1.7	2.1	7.3	0.6	0.8	3.1	1.6	1.9	4.5	1.2	0.9	4.5	0.6	0.7
Single	10	5.2	2.6	7.9	6.3	2.8	9.3	5.4	3	9.7	3.5	1.9	6	5	1.6	7.5	3.8	1.5
SVMRFEK	3.2	2.7	1.8	3.5	0.9	1.5	1	1.3	1.2	4.5	1.5	1.8	4.8	1.2	0.7	0.9	1	0.7
SVMRFEL	2.3	2.1	2.1	0.8	1.4	1.4	0.7	0.8	0.5	1.7	2	2	0.8	1	0.5	0.9	0.8	0.7
All feature		2.1			1.8			0.8			2			0.5			0.3	

Table B.2 : Standard deviation of ten different realizations of classification accuracy for IndianPines dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	2.1	2.9	4.5	1.3	2.2	1.2	0.8	1.5	2.3	1.8	3	3.2	1.8	2.1	2.6	2.5	1.9	1.5
CIFE	3.3	5.2	3.8	1.7	5.1	2.1	6.9	6.1	1.8	6	3.9	2	3.7	2.2	1.8	7.1	3.1	2.5
CMIM	7.8	4.5	3.2	5	2.2	2.2	5.7	3	2.4	5.9	3.4	3.8	3.4	3.1	3.6	6.5	2.5	2.1
FSDTree	4	3	3.9	2.6	1.5	1.7	1.7	1.4	1.6	2.2	1.1	3	2.1	1.9	2.6	2.8	3	2.9
DISR	5.5	4.2	2.8	5	2.4	2.5	6.8	3.1	2.8	3.9	4.1	3.1	3.4	2.1	2.3	5.5	1.9	2.4
Fisher	4.2	4.3	4.3	2.4	2.5	3.6	1.8	1.9	3.3	3	5	6.5	2.6	3.5	3.5	1.6	2.4	2.2
Gini	2.8	1.8	3	1.8	1.3	1.8	1.3	0.8	1.3	4.8	4.3	3.1	3.1	1.2	1.3	2.7	2.1	1.8
InfoG	2	1.5	2.4	1.9	2.7	2.4	2.1	2.4	2.9	5	3.5	2.8	6.4	3.9	3	4.7	3.4	2.8
JMI	3.6	4.3	3.8	4	2.9	2.8	1.8	1.3	2	4.3	3.1	2.6	4.1	3.2	1.6	4.2	2.1	1.9
MIFS	3.3	3.2	3.3	2.7	2	2.1	1.9	1.6	2.5	4	4	2.9	3.7	1.8	2.2	3	1.7	1.4
MIM	1.7	4.2	3.2	2.7	1.8	2.3	1.5	1.5	2.4	4.5	2.1	2.7	3.4	2.3	2.6	1.8	1.9	1.7
MRMR	4	3.1	4.3	1.4	2.2	4.2	1.7	1.6	4.9	4.8	3.1	3.9	2.6	2	2.5	2.8	2.3	1.8
RF	4.7	3.2	2.1	3.2	2.7	1.5	3.8	3.3	2.7	4.7	3.5	3.2	4.2	2.2	2.1	4.3	2.4	2
ReliefF	2.4	3.6	3.7	1.5	2.2	1.7	2.3	1.6	2.1	2.1	2.3	3.6	3.2	1.5	2	4.2	1.4	1.7
Single	4.2	3.3	3.7	3.1	3.1	3.1	2.2	2.8	2	5.5	3.6	3	4.7	4.1	2.9	3.1	2.8	2.7
SVMRFEK	3.5	3.1	1.9	2.7	3	1.2	3.7	3	1.5	2.7	2.3	1.5	1.8	1.9	1.4	1.8	1.8	1.5
SVMRFEL	2.3	2.8	2.5	1.8	1.6	1.2	1.5	1.5	1.1	2.3	2.2	1.4	1.9	1.4	2	1.3	1.7	1.7
All feature		3.2			1.8			1.9			2			1.5			1.6	

Table B.3 : Standard deviation of ten different realizations of classification accuracy for KSC dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	5.6	2.1	2.1	3.9	1.9	2.2	7.4	0.9	1.1	4.6	5.1	5.9	4.3	3.3	3.4	12.1	3.5	3.5
CIFE	2.7	1.4	1.4	2	2.2	1.9	1.3	0.9	0.8	6.6	7	6.2	6.7	8.7	9.4	2	8.7	11.2
CMIM	2.9	2.8	2.6	3.4	2.4	2.5	2.7	1.7	1.5	7.5	5.1	4.2	3.6	7.4	10.3	2.5	6.2	6.5
FSDTree	3	2.3	2.6	2.3	2.3	2.4	1.7	1.7	1.4	3.2	3.4	3.5	4.5	6	8.8	4.9	6.7	2.1
DISR	4.3	3	2.4	2.3	1.6	1.6	1.1	0.9	0.9	4.2	4.2	3.6	6.7	7.5	9.1	4.5	1.2	1.4
Fisher	2.8	3.9	3.8	1.9	2	1.5	0.8	1.3	1.1	3.3	4.9	3.9	1.5	7.1	6.6	2.5	3.4	3.5
Gini	2.6	1.7	1.9	1.1	1.4	1.2	1.1	0.9	1	3.6	3.2	2.6	1.2	6.2	5.9	1.9	2.5	1.7
InfoG	5.9	3.9	2.3	5.2	2.7	1.7	2.6	1.4	1	6.7	5.8	4	5.3	6.7	8.8	7.2	3.5	3.2
JMI	2.4	3.4	2.9	1.2	1.6	1.6	1.2	0.8	1	4.3	3.4	3.4	6.4	6.3	9	4	1.1	1.4
MIFS	3.5	2.5	1.9	3.2	2.6	1.8	1	0.5	0.7	6.3	5.8	6.2	4.6	8.1	9	4.8	8.6	11.2
MIM	3.4	3.3	3.3	3.7	1.7	2.3	2	1.4	1	2.8	3.5	3.5	4.6	6.8	1.5	2.6	2.4	3.2
MRMR	2.5	1.9	2.4	1.5	1.7	1.6	1.5	0.9	0.9	7.2	5.9	6.4	6.4	8.8	8.9	4.8	12.6	15.8
RF	2.6	1.7	1.8	3.7	1.7	1.7	3.6	1	1.1	4.4	4.1	4.6	6.7	6.3	7.8	9.4	11.7	11.8
ReliefF	4.7	3.2	2.7	4	2.1	1.5	6	0.7	1.2	4.4	4.4	5.2	6.5	6.2	3.1	6	3.5	3.6
Single	4.9	2.2	1.8	3.9	2	1.8	4.6	1.5	0.9	5	4.2	3.9	8	5.7	7.8	6.4	5.3	2.7
SVMRFEK	3.2	2.5	2.5	1.8	2.1	1.8	1.8	1.2	1.3	3.1	3.6	5.2	5.3	7.4	8.8	3.8	2.8	4.5
SVMRFEL	4.5	2.5	2.6	3.6	1.3	1.6	3.2	2.1	1.6	5	6.3	6.9	2.5	2.4	2.7	3.4	3	3.1
All feature		2			1.7			1.2			6.8			14.6			1.4	

Table B.4 : Standard deviation of ten different realizations of classification accuracy for PaviaCenter dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	3.4	4.8	5.5	3	3.3	2	3.1	2.5	1.6	4.6	2.1	1.3	1.8	1.4	0.8	2	0.6	0.4
CIFE	6.5	2.1	2.6	4.9	1.2	1.2	3.8	2	1	5.6	1.4	0.7	4.6	0.5	0.5	3.5	0.5	0.3
CMIM	10.6	3.5	1.5	4.9	1.2	0.9	2.4	0.8	0.7	8.4	1.7	1	2.3	0.4	0.4	1.4	0.4	0.3
FSDTree	3	2.8	3.2	1	1.2	0.9	0.5	0.7	0.6	1.1	0.8	0.7	0.7	0.5	0.4	0.2	0.2	0.3
DISR	4.7	3.2	3.2	5	3	1.9	8.8	1.2	1.4	3.6	2.6	1.8	2.9	1.7	1.3	4.9	0.5	0.6
Fisher	8.3	6	5.2	6.3	3.8	2.4	5.9	3.9	1.5	8.4	4.7	1.4	6	3.5	1.2	6.4	4.7	0.6
Gini	4.2	4.2	3.8	1.9	1.9	1.6	1.6	1.4	1.3	4.4	1.8	1.4	2.4	1.3	1.6	1.4	0.9	0.7
InfoG	9.6	5.5	2.8	11.9	2.5	1.5	9.9	1.2	0.9	7.4	2.4	1.1	11.4	1.5	0.7	10.3	0.3	0.3
JMI	3.6	2.3	2.7	4.2	1.1	0.8	1.6	0.6	0.7	2.6	0.8	0.9	3.2	0.5	0.4	0.5	0.4	0.2
MIFS	2.3	2.1	2.6	3.7	2	1.2	1	0.6	1.9	1.7	1.2	0.9	2.7	0.9	0.5	0.3	0.6	0.4
MIM	4.7	2.6	2.6	7.7	1.4	1.1	11.1	1.8	0.9	3.7	0.5	0.7	5.6	0.9	0.5	9.6	0.6	0.3
MRMR	5.8	2.6	2.8	2.1	1.1	1.2	1.3	0.7	0.8	1.9	1.5	1	1.3	0.4	0.5	0.5	0.5	0.3
RF	2.7	2.5	2.3	3.2	1	0.8	2.2	0.7	0.7	1.4	0.8	0.8	1	0.6	0.4	1	0.6	0.4
ReliefF	3.2	6	4.9	2.9	4	1.9	3.1	2.2	2	4.6	2.7	1.6	2	1.8	1.1	2	2.2	0.5
Single	8	5	2.8	11.1	3.4	1.6	11.6	2.1	1.2	8	2.2	1.3	8.1	1.4	1	9.3	1.7	0.6
SVMRFEK	6.6	9.7	2.9	2.7	2.4	1.3	4.2	4.5	1.6	5.5	5.7	1.8	2.3	2.8	0.9	5.5	3.2	0.6
SVMRFEL	3.6	5.5	4.8	2.9	4	1.9	2.4	2.2	2	3.1	2.5	1.6	2	1.8	1.1	3.3	2.2	0.5
All feature		2.4			0.9			0.6			0.6			0.4			0.2	

Table B.5 : Standard deviation of ten different realizations of classification accuracy for PaviaUniversity dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	6	4.3	2.9	3.4	2.6	3.7	3.6	2.3	1.8	4.5	2.5	5.3	2.1	2.8	3	1.1	3.7	3
CIFE	6	5	6	4.5	3.8	3.1	3.2	3.3	2.5	3	5.9	5.5	2.9	2.8	2.8	5.5	2.7	2.1
CMIM	4.6	3.6	4	5.5	3.8	2.8	2.7	2.3	2.9	4.9	4.9	4	4.9	2.8	2.7	4.8	2.5	2.3
FSDTree	4.5	4.3	4.2	3.1	3.2	2.9	2	1.9	2.6	5.9	5.4	4.2	4.8	3.5	2.7	3.5	2.8	1.7
DISR	5.6	5	4.5	6	3.7	3.9	3	1.7	1.8	8.5	6.6	4.5	6.2	2.3	2	2.7	2.1	1.5
Fisher	6.5	5.1	4.2	5.1	3	2.9	4.2	2.8	1.3	7.9	4.5	5.3	6	3	2.6	2.4	3.3	2
Gini	3.7	4.9	5.7	3.2	4.3	4	2.7	2	3.3	5.4	5	4.8	4.2	2.2	2.3	1.7	1.3	1.4
InfoG	8.5	4.6	5	6.2	3.6	2.2	6.3	1.8	1.7	9.1	6.6	5	7.3	4.3	2.5	6.1	3	3.1
JMI	5.9	5.3	5.1	4.7	4.5	4	2.5	2.7	1.9	4.5	2.5	4.4	3.8	1.6	1.2	3.1	1.5	1.5
MIFS	3.4	5.5	5.5	3.6	4	4.1	3	2.8	2.7	5.6	5.1	6.5	3.9	3.5	2.9	3.9	1.8	1
MIM	5.4	3.9	3.1	3.3	2.6	3	3.4	2.1	2	5	5.7	4.8	5.3	2.2	1.7	3.8	2.5	2
MRMR	6.6	4.9	4.3	3.2	3.7	3	3.5	2.7	2.1	5.7	5.2	4.7	2.7	2	2.3	3.8	1.5	2
RF	7.3	5.8	3.9	4.8	3.7	2.6	2.9	3.3	2	5.5	4.5	3.3	3.6	3.5	2.1	4.9	2.5	2.3
ReliefF	5.3	3.2	6.2	3.4	3.2	6.1	4.6	3.1	1.5	5.6	5	5.1	4.3	3.3	2.1	4.9	3.1	2.3
Single	6.2	3.1	3.9	6.2	3.2	4.2	8.2	1.8	1.4	5.3	6.6	4.7	9.6	3.5	3.3	8.5	3.2	2.3
SVMRFEK	5.5	3.4	3.4	6.3	2.8	3.3	6.2	2.6	2.5	6.2	5.1	4.7	8.3	2.2	1.8	6.7	1.6	1.6
SVMRFEL	4.5	3.4	5.1	3.6	4	2.3	2.6	1.5	2.5	5.6	6.2	6.8	3.6	2.5	1.6	3.3	2.1	1.7
All feature		4.7			3.1			2			6.5			1.6			1.9	

Table B.6 : Standard deviation of ten different realizations of classification accuracy for Salinas dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

Method	KNN									SVM								
	10 SpC			25 SpC			50 SpC			10 SpC			25 SpC			50 SpC		
	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F	10F	30F	50F
Chi2	2.6	1.2	1.3	2.2	1.3	0.8	0.8	0.8	0.8	3.3	3	3.5	1.8	1.5	1.4	0.9	0.6	1.3
CIFE	5.3	2.5	1.4	4.8	1.7	1.4	2.7	1.5	0.7	4.8	2.5	2.5	2.9	1.8	2	1.2	0.8	0.7
CMIM	2.8	1.6	1.5	1.4	1	0.8	0.7	0.5	0.6	3.8	2.3	2.4	1.4	1.6	1.6	0.7	0.9	0.8
FSDTree	1.5	1.6	1.3	1.2	1.1	0.7	1.3	0.7	0.5	3.3	1.6	2.2	0.9	1.5	1.6	1.3	0.9	0.7
DISR	2.3	2.4	1.8	1.4	1.2	1.2	1.1	0.6	0.5	3.4	2.5	2.2	1.5	1.5	1.8	1.6	1.4	1.1
Fisher	5.8	4	3.5	3.8	1.7	1.2	2.5	1.8	0.8	6.4	3.3	3.6	2.7	1.7	1.9	2.9	1.3	1
Gini	4.2	1.6	1.9	1.7	0.8	1	0.9	1	0.8	2	1.7	2.1	1.5	1.7	1.1	1.6	1.1	1
InfoG	3.8	1.2	1.2	1.2	1.5	1	1.2	0.5	0.5	4.8	3.1	2.3	2.1	0.9	1.2	1.3	0.7	0.6
JMI	1.6	1.4	1.5	1.3	1.1	1.1	0.8	0.6	0.6	2.7	3.2	2.3	1.8	1.4	1.7	1	1	0.9
MIFS	4.5	1.5	1.6	1.6	1.7	1.4	1.6	1.3	0.8	4.5	2.6	2.5	1.5	1.6	1.4	1.1	0.9	0.8
MIM	5.2	1.1	1.3	4.2	1.4	0.9	3.9	1.1	0.8	2.8	2.9	2.7	2.1	1.6	1.4	2.7	1.1	1.3
MRMR	3.6	2.2	2	1.9	1.5	1.3	1.8	0.9	0.8	3.4	3.4	3.2	0.8	1.3	1.8	2.4	0.9	0.8
RF	1.9	1.6	1.6	1.2	1.1	1.4	0.5	0.3	0.4	3.7	2.1	2.5	1.3	1.6	1.5	1.1	0.9	0.9
ReliefF	2	1.7	1.3	1.6	1.3	0.9	1	0.6	0.8	3.8	2.9	3.1	1.4	1.9	1.4	0.9	1.1	1.2
Single	2.6	2.1	1.5	3.1	1.2	1	3.5	0.7	0.7	2.2	3.3	2.9	2	1.4	1.3	2.8	1.3	1.3
SVMRFEK	2.5	1.2	2.4	1.7	2.4	1.2	0.8	2.2	0.7	4.3	1.6	2.2	4.1	1.7	1.7	2.7	1.2	1.1
SVMRFEL	2.2	1.6	1.4	1.8	1.3	0.9	1	1.1	0.8	2.5	3.5	2.5	1.6	2.1	2	0.7	1.3	1.3
All feature		1.6			0.9			0.9			1.7			1.8			0.9	

Table B.7 : Standard deviation of ten different realizations of classification accuracy for SalinasA dataset. "10F", "30F", and "50F" show the number of ranked features are selected. For instance, 10F means first 10 ranked features are chosen. "10SpC", "25SpC", and "50SpC" represent the size of training data. For example, 10SpC means 10 samples were considered for each class as a training data.

	KNN									**SVM**								
Method	**10 SpC**			**25 SpC**			**50 SpC**			**10 SpC**			**25 SpC**			**50 SpC**		
	10F	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**	**10F**	**30F**	**50F**
Chi2	3.9	1.5	0.9	3.3	1.7	1.2	1.1	0.6	0.6	1.6	1.7	1	1.3	1	0.8	0.3	0.6	0.6
CIFE	5.2	1.1	0.6	3.6	1	0.6	2.4	0.7	0.3	4.2	1.6	0.8	2.8	1.1	0.9	2.2	0.8	0.7
CMIM	1.3	0.8	0.6	1.7	1	0.9	0.7	0.5	0.5	3.2	0.9	0.7	0.6	0.4	0.8	0.6	0.5	0.6
FSDTree	1.4	1.1	1	1.1	0.9	0.5	0.5	0.6	0.3	1.2	0.7	0.8	1.5	0.8	0.8	0.6	0.4	0.5
DISR	5.6	1.3	1.2	1.3	1.4	1.4	1.3	0.8	0.5	8.9	0.7	0.5	5.6	0.5	0.4	4.3	0.6	0.4
Fisher	3.5	1.8	1.5	2.9	2.2	1.9	1.3	0.5	0.5	2.2	2	1.6	1.2	0.5	0.7	0.5	0.6	0.5
Gini	3.5	1.4	1.7	2	1.4	1.3	1.5	1.3	0.7	4.3	9.2	7.8	1.3	0.5	0.5	0.3	0.4	0.4
InfoG	5.9	1.9	3.4	1.2	1.2	1	0.8	0.3	0.5	7.3	1.2	0.8	1.5	1	1	1.6	1.1	0.7
JMI	1	0.9	0.7	1.1	1.1	1.1	0.5	0.3	0.5	0.6	0.6	0.9	0.7	0.9	0.9	0.6	0.6	0.6
MIFS	5.9	2.5	0.7	3.5	1.7	0.8	3.7	1.2	0.5	11.3	3	1.1	4.6	0.9	0.7	3.4	0.9	0.5
MIM	1.5	1	0.8	1.7	1.5	1.3	0.5	0.5	0.6	1.3	1.2	1.1	1.5	0.9	0.9	0.5	0.6	0.7
MRMR	5	1.7	1.1	2.4	2.3	0.8	3.3	1	0.6	5	7.9	0.6	5.7	0.4	0.3	6.6	0.6	0.4
RF	1.5	0.8	0.8	1.4	0.8	0.8	1.4	0.5	0.5	2.3	1.2	1.3	2.3	1	0.9	2.1	0.8	0.6
ReliefF	3.5	1.4	0.9	3.4	1.7	1	1.1	0.4	0.6	1.8	1.6	0.9	1.2	0.5	0.8	0.4	0.6	0.6
Single	5.3	4.2	4.2	4.6	1.5	1.3	2.5	0.9	0.4	6.6	1.7	1.6	7.4	2	0.8	9.5	1	0.5
SVMRFEK	6.1	3.5	1.1	1.9	1.9	1.3	2.2	0.9	0.4	11.2	7.1	1	5.9	0.7	1	5.8	0.5	0.3
SVMRFEL	3.6	1.4	0.9	2.2	1.2	0.9	1.3	0.6	0.7	1.5	1.5	1.1	1.2	1.2	0.8	0.4	0.6	0.5
All feature		0.6			0.7			0.5			0.6			0.7			0.5	

CURRICULUM VITAE

Name Surname: Hamed Gholami Vijouyeh

Place and Date of Birth: Iran - 1983

E-Mail: gholamivijouyeh@itu.edu.tr
hamedghv@yahoo.com

EDUCATION:

- **Associate Deg. in Computer Software:** 2004, Islamic Azad University of Shabestar, Engineering Faculty, Department of Computer Engineering

- **B.Sc. in Computer Engineering – Software:** 2007, Islamic Azad University of Tabriz, Engineering Faculty, Department of Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2015-2016 Scholar in TUBITAK (The Scientific and Technological Research Council of Turkey) project: Random Sampling High Dimensional Model Representation and Feature Selection. A Case Study on Hyperspectral Satellite Images.

- 2016 Satellite Earth Observation Applications (Multi-Dimensional Imaging Radar & Hyperspectral) Summer School, IEEE GRSS Turkey.

- 1997 Selected in provincial step of the "Research and Study" competition in East Azerbaijan.

- 1997 1st Rank in regional step of the "Research and Study" competition in Tabriz.

- 1995 Selected in provincial step of the "Research and Study" competition in East Azerbaijan.

PUBLICATIONS:

- H. Gholami Vijouyeh, 2014. A Comparison between Software Engineering Five Implementation Models. *1st Conference of Electrical and Computer Engineering*, June 12, 2014 Anzali, Iran.

PUBLICATIONS ON THE THESIS:

- H. G. Vijouyeh and G. Taşkın, 2016. A comprehensive evaluation of feature selection algorithms in hyperspectral image classification, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 10-15, 2016, Beijing, China.