1-1-2014

Spectral Band Selection for Ensemble Classification of Hyperspectral Images with Applications to Agriculture and Food Safety

Sathishkumar Samiappan

Spectral band selection for ensemble classification of hyperspectral images with applications to agriculture and food safety

By

Sathishkumar Samiappan

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Electrical and Computer Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

August 2014

Spectral band selection for ensemble classification of hyperspectral images with applications to agriculture and food safety

By

Sathishkumar Samiappan

Approved:

Lori M. Bruce (Co-Major Professor)	Nicolas H.Younan (Co-Major Professor)

Eric Hansen
(Minor Professor)

Robert J. Moorhead
(Committee Member)

John E. Ball
(Committee Member)

James E. Fowler
(Graduate Coordinator)

Jason M. Keith
Interim Dean
Bagley College of Engineering

Name: Sathishkumar Samiappan

Date of Degree: August 15, 2014

Institution: Mississippi State University

Major Field: Electrical and Computer Engineering

Major Professor: Lori M. Bruce, Nicolas H. Younan

Title of Study: Spectral band selection for ensemble classification of hyperspectral images with applications to agriculture and food safety

Pages in Study: 113

Candidate for Degree of Doctor of Philosophy

In this dissertation, an ensemble non-uniform spectral feature selection and a kernel density decision fusion framework are proposed for the classification of hyperspectral data using a support vector machine classifier. Hyperspectral data has more number of bands and they are always highly correlated. To utilize the complete potential, a feature selection step is necessary. In an ensemble situation, there are mainly two challenges: (1) Creating diverse set of classifiers in order to achieve a higher classification accuracy when compared to a single classifier. This can either be achieved by having different classifiers or by having different subsets of features for each classifier in the ensemble. (2) Designing a robust decision fusion stage to fully utilize the decision produced by individual classifiers.

This dissertation tests the efficacy of the proposed approach to classify hyperspectral data from different applications. Since these datasets have a small number of training samples with larger number of highly correlated features, conventional feature selection approaches such as random feature selection cannot utilize the variability in the correlation level between bands to achieve diverse subsets for classification. In contrast, the approach proposed in this dissertation utilizes the variability in the correlation between bands by dividing the spectrum into groups and selecting bands from each group according to its size. The intelligent decision fusion proposed in this approach uses the probability density of training classes to produce a final class label. The experimental results demonstrate the validity of the proposed framework that results in improvements in the overall, user, and producer accuracies compared to other state-of-the-art techniques. The experiments demonstrate the ability of the proposed approach to produce more diverse feature selection over conventional approaches.

DEDICATION

I dedicate this dissertation to my grandmother, Dhanammal, to my parents, Samiappan and Padma Samiappan, my wife, Bhanupriya and my sister, Kalaivani.

ACKNOWLEDGEMENTS

It was a great pleasure for me to work as a graduate research assistant at Geosystems Research Institute and as a graduate teaching assistant at the department of Electrical and Computer Engineering. I am grateful to all those people who have made this research work possible and because of whom my graduate experience has been joyous and productive. I thank my major professors, Dr. Lori M. Bruce and Dr. Nicolas H. Younan for their generous time and commitment, and for providing valuable guidance and encouragement over the course of my graduate program. I thank my minor professor, Dr. Eric A. Hansen for his valuable and time and support. I thank my dissertation committee members, Dr. Robert J. Moorhead II, Dr. John E. Ball for providing me valuable insight and guidance. I thank Dr. Saurabh Prasad and Mrs. Jane Moorhead for their constant support and guidance.

I also respectfully thank the following individuals: Dr. Haibo Yao for providing me with the experimental aflatoxin hyperspectral data; Dr. Mathew Lee for helping me with experimental data collection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

Duda *et al* [1] define pattern recognition as “the act of taking in raw data and making an action based on the ‘category’ of the pattern has been crucial for the survival of human beings, and over the past tens of millions of years humans have evolved with highly sophisticated neural and cognitive systems for such tasks”. In the past few decades, researchers continue to develop algorithms that mimic neural and cognitive evolutions that humans undergone. Such computer algorithms are known as machine learning algorithms. Arthur Samuel [2] defined machine learning as the “Field of study that gives computers the ability to learn without being explicitly programmed”. Modern day machines are intelligent, adaptive, and they can improve the performance through experiences. One of the decisive component in any intelligent system is the underlying learning algorithm. The applications of such algorithms are primarily in the fields of science and engineering. From automated speech recognition [3], fingerprint recognition [4], face recognition [5], and many other applications, it is very important to have a reliable and robust learning algorithms.

Learning algorithms can be broadly categorized as supervised and unsupervised. The former requires class labels or training samples to learn appropriate class conditional models and the latter assigns a label automatically based on the natural closeness among samples with respect to some underlying metric [6]. In situations where class labels are not available, unsupervised approaches can be used. In supervised learning, a known set of training samples are used to learn a model that better represents its global structure, then, this model is used for predicting new samples/classes for new data or forecasting.

Supervised learning problems can be further categorized as classification and regression. Regression has strong roots in statistical analysis but is integral part of supervised machine learning paradigms. Regression is mainly used for estimating the relationship among the variables. Classification techniques are very much part of machine learning where a label is assigned to an incoming unknown sample. Both classification and regression are used for predicting the category of an unknown sample from the experience of learning from known samples. There are two major differences between them. First, in the case of classification, the category or label is discrete whereas in regression it is continuous. Second, the regression can be used for forecasting analysis along with predicting the category of an unknown variable. In most of the remote sensing applications and especially with hyperspectral data, supervised techniques are of tremendous interest [7].

Supervised learning algorithms are very useful for performing the analysis on remotely sensed data. Remote sensing can be described as the process of obtaining information about an object or phenomenon without making any physical contact [8]. The applications of remotely sensed data are huge in Earth observation [9], robotic systems [10], medicine [11], and food security and safety [12]. In these applications, generally, a spectral imaging is employed to obtain the information about the object of interest. If the captured spectral image contains three bands that are visible to human eyes such as red, green, and blue, we call it visible image or a photograph. It is possible to extend the bands beyond the visible spectrum. It is also possible to capture any number of bands in the wide electromagnetic (EM) spectrum. Based on the number of bands, the image can be called multi-spectral or hyperspectral. In case of multi-spectral images, the number of bands is less than ten whereas hyperspectral images can have hundreds or even thousands of bands.

Hyperspectral imagery (HSI) provides a detailed description of materials ranging from visible to infrared regions of the electromagnetic spectrum. Such a wide spectral range of information has the potential to yield higher classification accuracies compared to that of multi-spectral counterpart. In HSI, the correlations between successive bands are often very high. So, the information contained in successive bands can be redundant. The key to the design of a powerful classification system lies in extracting pertinent features from the high-dimensional data and employing classifiers to exploit those features. In HSI, every pixel has multiple reflectance values corresponding to a wide range of bands. So, the classification problem becomes very high dimensional, and it requires a very large number of training samples for reasonable estimation of class conditional distributions. An increase in the dimensionality results in a decrease in the generalization capability and this can cause poor classification performance. In the literature, this is often referred to as Hughes phenomenon [13].

The Hughes phenomenon has been well studied for classifiers built for HSI. Statistical classification methods, in particular Gaussian Maximum Likelihood (ML), a traditional supervised pattern classification approach, often fails to classify HSI data accurately because of the high dimensionality of features and limited ground truth availability [14]. When the ratio of the number of samples to the number of bands in HSI is small, a parametric approach, such as the ML classifier, suffers from Hughes phenomenon and as a result, the classification accuracy is driven down. This happens mainly due to the inaccurate estimate of statistics such as the mean and covariance that play a major role in the design of statistical classifiers. For a fixed number of samples in a training set, this ratio decreases as the number of bands in HSI increases. An increase in the number of samples results in an increase in the accuracy up to a point after which the accuracy starts decreasing. So, the Hughes effect may not occur when the sample size is increased appropriately with an increase of additional bands. Similarly, for a training set with a large number of samples such that even all the bands in a HSI is used, The Hughes phenomenon may not be observed at all. In a practical scenario, acquiring a large number of training samples is not possible due to the difficulties and expense incurred in time and money in acquiring them. More importantly, it is expensive to collect large number of labeled samples (ground truth) and the process is highly time-consuming.

Various approaches can be employed to alleviate this effect. The commonly used approach is to apply some form of feature reduction prior to classification. Feature reduction techniques for HSI can be broadly classified into two categories: feature extraction and feature selection. With feature extraction, the HSI data is transformed in a way that represents the same information in a smaller number of features [15]. With feature selection, a subset of the original features is selected in such a way that the subset preserves the discrimination capability of the original data [16]. In statistical classification approaches on HSI, classifiers are usually preceded by feature reduction techniques such as Principal Component Analysis (PCA) and Fisher's Linear

Discriminant Analysis (LDA). The above techniques are well studied and they are widely used in the pattern recognition community. However, for HSI classification problems, PCA will discard useful discrimination information if the vectors are oriented in the directions of a small global variance [17]. LDA, on the other hand, is very good at preserving the discrimination information but fails under multi-model class distributions [18]. The common issue with PCA and LDA is that they suffer from the problem of ill conditioned covariance and scatter matrices when the number of training samples is insufficient. This affects the efficiency of learning true projections. The feature selection techniques are sub-optimal. Hence, it does not fully exploit the rich spectral information available in HSI.

Another approach to alleviate the Hughes effect is to use classifiers that are insensitive to this phenomenon. Kernel based techniques such as Support Vector Machines (SVM) are widely recognized as being efficient approaches to classify HSI without being concerned about the Hughes effect [19], [20]. Unlike the statistical algorithms such as ML, kernel based classification algorithms can learn the data without assuming any underlying statistical distribution. Kernel based algorithms were also reported to perform better than statistical techniques such as ML in terms of the overall accuracy and class accuracies. These algorithms have the ability to handle high dimensional data, even with only a small number of training samples. Although this is true to some extent, there is an uncertainty about the role of feature reduction for SVMs [21], [22]. This area of research is one of the main focuses of this dissertation.

SVM is based on structural risk minimization and it exploits a margin-based criterion. Other approaches are based on empirical risk, where the aim is to minimize the misclassification error on the training set. SVM guarantees the smallest possibility of misclassifying an unknown data sample randomly drawn from a fixed but unknown probability distribution. Furthermore, SVM seeks to find an optimal hyperplane that maximizes the margin between classes using only a small number of training points. These training points are called support vectors. It uses only a small number of training samples and the results of some studies may suggest that SVM may not be affected by an increase in the number of features. This alleviates the problem of the Hughes effect to some extent. Other studies have shown that the classification accuracy of SVMs can be increased by reducing the number of features [22]. Feature reduction has impacts on the speed of the classifier as well. It will also provide advantages in terms of low memory usage. Feature reduction for SVM, therefore, is a useful analysis tool both in terms of the accuracy and overall performance of the classifier. A detailed discussion of the SVM is provided in Section 2.1.2.

There is a growing interest among the research community in using multiple learners to classify the HSI data [23]. This generally yields an improved accuracy when compared to single learners provided that the diversity among the learners is established. The decisions of individual learners are then used to compute the final decision. This multi classification technique is often referred to as ensemble classification. In this work, a non-uniform random feature selection technique is proposed that employs a kernel density based decision fusion to exploit the rich spectral information contained in HSI.

The three key problems addressed in this research are:

1. Design an algorithm to create a diverse feature subset capable of providing good classification performance for a variety of applications.

2. Design a decision fusion technique that uses information about the separation of class distributions in the training data. This enables the decision fusion technique to assign weights to decisions of individual learners in an efficient way.
3. Finding a suitable diversity measure for HSI data in a multi-classifier setup with SVM. This further increases the utilization of rich spectral information available in HSI.

1.2 Motivation behind the proposed work – Spectral band selection techniques for classifying hyperspectral data using ensemble kernel based classifiers

Classifying hyperspectral images is a challenging problem because of the high dimensionality of the feature space and typically a very high degree of correlation between successive features. Therefore, it is important to make use of such information effectively in a manner that does not result in reduced performance. In order for a classifier to perform well, feature selection is an important step –particularly more so when the feature space is very high dimensional and the amount of training data available is limited (as is the case with hyperspectral images). In the last decade, Support Vector Machines (SVMs) have been shown to perform well for supervised classification of hyperspectral images. The traditional view of SVMs in the research community is that they can handle high feature space dimensionality in an efficient way and hence feature selection is not vital to successful deployment of such classifiers. Using multiple classifiers with HSI data yields a better classification performance and hence it is very popular. A multi-classifier setup with SVM as learner is the most obvious combination in many remote sensing applications especially with HSI. Recently, the interest among the remote sensing research community is to investigate the effects of some sort of feature selection prior to kernel classification such as SVM. Waske *et al* [21] demonstrated the sensitivity of SVMs to random feature selection (RFS) for hyperspectral data. It was shown that a feature selection algorithm for SVMs can improve the classification performance by eliminating features that cause confusion between classes. The system proposed in this work employs a multi-classifier setup with non-uniform random feature selection and kernel density based decision fusion to utilize the information in HSI.

1.3 Contributions of this work

This research work seeks to develop a robust feature selection algorithm for performing hyperspectral image classification by using a multi-classifier setup with SVM classifiers. The research seeks to validate the algorithm for urban, agricultural, and food safety applications. By using various band grouping techniques, the hyperspectral spectrum is partitioned into different regions based on the correlation among the bands. A novel non-uniform random feature selection is then employed on each band groups multiple times to arrive at a subset of bands. Then, the probability density of these subsets is computed to estimate a class score matrix that assigns a rank to every classifier with respect to its ability to distinguish every class from the other. The subsets are classified using a bank of kernel based classifiers. Each classifier in the bank produces a local class label. A kernel density fusion technique is used to fuse these local class labels to form the final class label.

The primary objectives of this dissertation are listed below

1. Design a scheme to perform a feature selection which creates a good diversity among the classifiers. This helps in alleviating the problem of the

curse of dimensionality by intelligently eliminating the redundant bands in the selected subset.

2. Determine the ability of each local classifier to distinguish a particular class from all other classes by developing a score matrix. This is very critical as the scores computed help in the decision fusion process.
3. Design a decision fusion system based on the class matrix and determine its ability in terms of the overall accuracy and class accuracies.
4. Determine a most suitable diversity measure for ensemble hyperspectral data classification. In particular, perform a comparative analysis on different diversity measures available for multi-classifiers and their effects on hyperspectral data.
5. Perform a case study with the proposed ensemble classification approach for a food safety application.
6. Create an aquatic plants dataset by using handheld hyperspectral sensor.

1.4 Outline of this dissertation

The outline of this dissertation is as follows: Chapter II discusses the background of supervised classification techniques, overview of parametric and non-parametric (statistical and kernel based) classification techniques, ensemble classification techniques, feature selection, and band grouping algorithms. This chapter also provides an overview of hyperspectral image analysis and its challenges.

Chapter III describes the methodology developed in this dissertation and the proposed system based on non-uniform random feature selection and kernel density based decision fusion for hyperspectral image classification. Chapter IV demonstrates the experimental setup and detailed results of four hyperspectral datasets. This chapter also shows the application of the proposed system to a food security problem of non-invasively identifying aflatoxins in corn kernels. Chapter V focuses on creating and measuring diversity among classifiers in an ensemble setup. This chapter also discusses the challenges involved in measuring diversity. This is followed by conclusions and future work in chapter VI.

1.5 References

[1] Richard O. Duda, Peter E. Hart, and D. G. Stork, Pattern Classification, Second Edi. Wiley Interscience, 2006.

[2] A. Samuel, “Programming Computers to Play Games,” Adv. Comput., vol. 1, pp. 165–192, 1960.

[3] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77, no. 2. pp. 257–286, 1989.

[4] A. K. Jain, Y. Chen, and M. Demirkus, “Pores and Ridges: High-Resolution Fingerprint Matching Using Level 3 Features,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, no. 1. pp. 15–27, 2007.

[5] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” Computer Vision and Pattern Recognition, 1991. Proceedings CVPR ’91., IEEE Computer Society Conference on. pp. 586–591, 1991.

[6] A. K. Jain, R. P. W. Duin, and M. Jianchang, “Statistical pattern recognition: a review,” Pattern Anal. Mach. Intell. IEEE Trans., vol. 22, no. 1, pp. 4–37, 2000.

[7] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, “A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification,” Selected Topics in Signal Processing, IEEE Journal of, vol. 5, no. 3. pp. 606–617, 2011.

[8] J. R. Jensen, Remote Sensing of the Environment. Pearson Education Inc, 2000.

[9] S. Platnick, M. D. King, S. A. Ackerman, W. P. Menzel, B. A. Baum, J. C. Riedi, and R. A. Frey, “The MODIS cloud products: algorithms and examples from Terra,” Geoscience and Remote Sensing, IEEE Transactions on, vol. 41, no. 2. pp. 459–473, 2003.

[10] M. Trierscheid, J. Pellenz, D. Paulus, and D. Balthasar, “Hyperspectral Imaging or Victim Detection with Rescue Robots,” Safety, Security and Rescue Robotics, 2008. SSRR 2008. IEEE International Workshop on. pp. 7–12, 2008.

[11] T. Vo-Dinh, “A hyperspectral imaging system for in vivo optical diagnostics,” Engineering in Medicine and Biology Magazine, IEEE, vol. 23, no. 5. pp. 40–49, 2004.

[12] M. Atas, A. Temizel, and Y. Yardimci, “Classification of aflatoxin contaminated chili pepper using hyperspectral imaging and artificial neural networks,” Signal Processing and Communications Applications Conference (SIU), 2010 IEEE 18th. pp. 9–12, 2010.

[13] B. M. Shahshahani and D. A. Landgrebe, “The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon,” Geoscience and Remote Sensing, IEEE Transactions on, vol. 32, no. 5. pp. 1087–1095, 1994.

[14] M. A. Lee, S. Prasad, L. M. Bruce, T. R. West, D. Reynolds, T. Irby, and H. Kalluri, “Sensitivity of hyperspectral classification algorithms to training sample size,” in Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS ’09. First Workshop on, 2009, pp. 1–4.

[15] L. Jimenez and D. A. Landgrebe, “Projection pursuit in high dimensional data reduction: initial conditions, feature selection and the assumption of normality,” Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century. IEEE International Conference on, vol. 1. pp. 401–406 vol.1, 1995.

[16] R. Solberg and T. Egeland, “Automatic feature selection in hyperspectral satellite imagery,” Geoscience and Remote Sensing Symposium, 1993. IGARSS ’93. Better Understanding of Earth Environment. International. pp. 472–475 vol.2, 1993.

[17] A. Cheriyadat and L. M. Bruce, “Why principal component analysis is not an appropriate feature extraction method for hyperspectral data,” Geoscience and Remote Sensing Symposium, 2003. IGARSS ’03. Proceedings. 2003 IEEE International, vol. 6. pp. 3420–3422 vol.6, 2003.

[18] T. Fung, F. Y. Ma, and W. L. Siu, “Hyperspectral data analysis for subtropical tree species recognition,” Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS ’98. 1998 IEEE International, vol. 3. pp. 1298–1300 vol.3, 1998.

[19] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” Geosci. Remote Sensing, IEEE Trans., vol. 42, no. 8, pp. 1778–1790, 2004.

[20] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 43, pp. 1351–1362, 2005.

[21] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data,” Geosci. Remote Sensing, IEEE Trans., vol. 48, no. 7, pp. 2880–2889, 2010.

[22] M. Pal and G. M. Foody, “Feature Selection for Classification of Hyperspectral Data by SVM,” Geosci. Remote Sensing, IEEE Trans., vol. 48, no. 5, pp. 2297–2307, 2010.

[23] J. A. Benediktsson, X. Ceamanos Garcia, B. Waske, J. Chanussot, J. R. Sveinsson, and M. Fauvel, “Ensemble Methods for Classification of Hyperspectral Data,” in Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International, 2008, vol. 1, pp. I–62–I–65.

CHAPTER II

CURRENT STATE OF KNOWLEDGE

2.1 Supervised Classification Techniques

Supervised learning contributes to the majority of research conducted in machine learning. These methods are extensively used in applications involving speech recognition, finger print recognition, face recognition, and also in remote sensing for classifying multi-spectral and HSI data. The defining property of supervised learning is the presence of labeled training data. These labeled data acts as a ‘supervisor’ that guides the learning system to induce models from labels [1]. These models can be later used to classify new unlabeled data. The quality of learning depends on the amount of the available training data and its diversity. Some supervised classifiers are parametric in nature and some are non-parametric. Parametric classifiers parameterize the model of each class with a finite number of parameters. For example, the maximum likelihood (ML) classifier, a parametric classifier, models each class with the mean and covariance matrices. Parametric techniques have very simple assumptions about the data. In many practical applications, these assumptions do not hold well. For example, ML assumes the class densities follow a unimodal Gaussian distribution whereas data from many practical applications are multimodal and non-Gaussian. The non-parametric approach estimates densities from sample patterns which can be substituted as true densities and sometimes the non-parametric approach has ways to estimate the posteriori probabilities directly.

Techniques such as Support Vector Machines (SVM) can assume decision functions directly rather than estimating probability densities [2]. In this dissertation, two categories of supervised techniques are used: statistical classifiers and kernel classifiers. In particular, a statistical approach, the Gaussian Maximum Likelihood (ML), and a kernel-based approach, SVMs, are of interest. ML is a parametric approach and SVM is a non-parametric approach. In this chapter, a description of these two techniques will be provided. These two techniques can be employed in the form of a single classifier system and Multi-Classifier Systems (MCS), which will be discussed later in this chapter. A description of different dimensionality reduction and feature selection techniques will also be presented. A brief overview of decision fusion, band grouping, and feature selection techniques are presented in this chapter. This chapter concludes with a brief overview of hyperspectral image analysis.

2.1.1 Statistical Methods

The research on pattern classification initially started with the development of artificial neural networks [3]. Due to the high number of bands in HSI data, statistical approaches have gained popularity among the remote sensing community. If the parameters of class-conditional probability densities are known, the problem of computing the distribution can be avoided. This is a basic approach followed in ML and Baysian estimation [4]. Statistical pattern classification methods are characterized by having a well-defined statistical model, which gives a probability that each instance belongs in every class.

2.1.1.1 Maximum Likelihood Classifier

The maximum likelihood classifier is a parametric classifier that uses the mean and covariance of a Gaussian probability density model for each class [5]. ML assumes that each class is Gaussian with a known mean and covariance so the classifier is optimal when the data satisfies this assumption. The discrimination function for each class is given by Equation 2.1

$$g_i(X) = p(X|w_i)p(w_i) \tag{2.1}$$

$$g_i(X) = \frac{p(w_i)}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} e^{-(1/2)(X-U_i)^T\Sigma_i^{-1}(X-U_i)} \tag{2.2}$$

where n is the number of bands or features in the HSI data, X is the n-dimensional pixel vector, and U_i and $\sum_i$ are the mean vector and covariance matrix of class i respectively. The mean vector and covariance matrix are estimated by unbiased estimators from the data available for training. Equation 2.1 can be expanded by substituting the Gaussian function to obtain Equation 2.2. This can be further modified by applying a natural logarithm and omitting the constants to obtain

$$g_i(X) = log_e\, p(w_i) - \frac{1}{2}\, log_e\, |\Sigma_i| - \frac{1}{2}\,(X - U_i)^T\Sigma_i^{-1}(X - U_i) \tag{2.3}$$

The first term in Equation 2.3 represents priori probabilities. This term can become a constant if priori probabilities are assumed to be equal and can be ignored. For every class, the second term will be a constant as well. During a classification process, the only thing that needs to be computed is term three. $g_i(X_i)$ is computed for each class in the data and the class with the largest value is decided as a class label.

2.1.1.2 Dimensionality Reduction Techniques

The high dimensionality of the HSI is beneficial in many ways but when it comes to classifiers, like ML, they suffer from the problem of Hughes effect when there is insufficient number of training samples. The covariance matrix in Equation 2.3 becomes ill-conditioned when the number of features is large. The generalization capability of the ML can go down as a result of higher dimensionality of features with insufficient number of training samples. To alleviate this issue, the data in the high dimensional space can be projected onto a lower dimensional space and perform the ML classification with a small number of features.

PCA and LDA are two common choices for performing dimensionality reduction. Each has a different approach towards transforming the features to a smaller dimensional space. Given an *n* dimensional feature space consisting of HSI bands, PCA can be used to transform this into a subspace of *m* dimensions, whose basis vectors are aligned in the direction of the maximum variance in the original space, where the transformed subspace has less number of features than the original space ($m < n$). Let W represent this transformation, then, the new reduced feature vectors can be defined as $y_i = W^T x_i, i = 1,2, \dots N$. The columns of this transformation matrix W are the eigenvalues e_i. The eigenvalues can be computed directly from $\lambda_i e_i = \sum e_i$, where $\sum = XX^T$ is the covariance matrix and λ_i is the eigenvalue of the vector e_i.

The approach followed by PCA is more suitable for compression because it seeks to best describe the data in a lower dimensional space. This may not be a good approach for HSI classification, where the main focus should be to find a subspace that best discriminates classes. That is the approach followed in LDA transformation. Here, provided a number of independent features relative to which the data is defined, LDA seeks to find a transformation that yields the largest mean differences between the desired classes. Two measures are defined to achieve this. They are called the within-class scatter matrix and between-class scatter matrix. Equations 2.4 and 2.5 represent the computation of these two measures.

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} \left(x_i^j - \mu_j\right)(x_i^j - \mu_j)^T \tag{2.4}$$

$$S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T \tag{2.5}$$

When the between-class measure is maximized and the within-class measure is minimized, optimal LDA transformation can be achieved. This can be done by maximizing the ratio $\frac{det|S_b|}{det|S_w|}$. This ratio is known as the Fisher's ratio.

In [6][7], it is mathematically shown that PCA can be detrimental for HSI classification applications. LDA, on the other hand, is clearly not suitable for classes with multi-modal distributions and hetroscedastic data classes (subsets of data having different statistical properties). In [8], it is mathematically and experimentally shown that LDA and its variations are only sub-optimal at best and may not be suitable for HSI classification applications.

2.1.2 Kernel-Based Classification

Kernel-based techniques have gained popularity in the past decade within the HSI research community [9][10]. Kernel-based methods map the data from the original space to a higher dimensional kernel feature space. A linear problem is then solved in the higher dimensional space. The learning algorithms can be designed and interpreted geometrically in the kernel space. Generally, the relationship between the kernel space and the original space is non-linear. The performance of this theoretically elegant technique is superior to that of other statistical techniques discussed in the previous sections. The dimensionality reduction techniques such as PCA and LDA can be easily extended in terms of a kernel space. Kernel PCA (KPCA) [11] and Kernel Fisher Discriminant Analysis (KFDA) [12] are successfully employed for the analysis of HSI data. However, the computational complexity makes them unsuitable to be used as a part of a statistical classification system such as ML.

Kernel-based classification problems are designed for two class problems and can be extended to multi-class. Considering *N* samples of a labeled training data, $\{(\boldsymbol{x}_1 y_1), (\boldsymbol{x}_2 y_2) \ldots (\boldsymbol{x}_N y_N)\}$, with $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \{+1, -1\}$, is generated from a probability distribution $P(\boldsymbol{x}, y)$ and is assumed to be independent and identically distributed. The problem is then to find a function *f* that minimizes the risk. This is given in Equation 2.6.

$$R(f) = \int Q[f(\boldsymbol{x}), y] dP(\boldsymbol{x}, y) \tag{2.6}$$

where *Q* is the predefined risk function of errors attributed by *f*. The minimum of risk can be approximated by the error in the training dataset. This is called the empirical risk, $R_{empirical}(f)$.

$$R_{empirical}(f) = \frac{1}{N} \sum_{i=1}^{N} Q[f(\boldsymbol{x_i}), y_i] \tag{2.7}$$

The empirical risk presented in Equation 2.7 converges to the actual risk only when *n* goes to infinity. In practical applications, it is not possible to get infinite training samples. With limited training samples in HSI, this may cause over fitting. To avoid this problem, the solution can be regularized. Regularization can be achieved by minimizing an *l* norm of the model parameters, $\boldsymbol{w}$. This adds an extra term to the above equation and minimizing this term gives smooth solutions with small weights. This function is called a regularized minimizing function and is given in Equation 2.8. ,

$$R_{regularized} = R_{emperical} + \lambda \,\|\mathbf{w}\|_l^2 \tag{2.8}$$

where λ is a tuning parameter. This is used to tune the tradeoff between the complexity of the model and training error minimization. Equation 2.8 can be effectively solved by a structural risk minimization principle and it says generalization can be improved by minimizing an upper bound of the generalization error. It is learned from statistical learning theory that simple learning is achieved by low complexity classifiers in infinite dimensional space ꞕ instead of $\mathbb{R}^n$. This can be achieved by introducing a mapping function φ which maps the $\mathbb{R}^n$ to the kernel space ꞕ.

2.1.2.1 Support Vector Machines (SVM) Classifier

The effectiveness of SVMs for HSI data has been shown in [13] and has gained popularity over the last decade. They often provide high classification accuracies compared to other non-parametric and statistical approaches. SVM classifiers are particularly useful to classify heterogeneous classes with a limited number of training samples. A detailed tutorial of SVMs can be found in [2] . SVMs are intrinsically designed as binary classifiers; however, multi-class SVM classifiers can be constructed by using the original SVMs as basic blocks. One–vs-all and hierarchical tree based approaches are popular techniques for constructing multi-class SVM classifiers. A more detailed explanation for constructing multi-class SVMs can be found in [14][15].

The SVMs strategy is to separate the training samples belonging to each class by tracing the maximum margin hyper planes in a higher dimensional kernel space. The training samples are implicitly mapped to a kernel space which is usually higher dimensional. This is shown in Figure 2.1

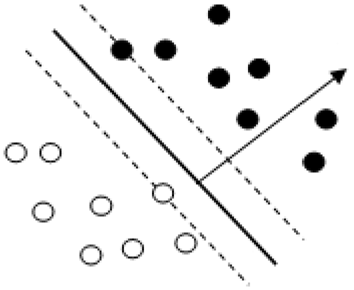

Figure 2.1 Optimal hyper plane for linearly separable classification problem

Maximizing the distance from the decision hyper plane to the samples can be achieved by minimizing the norm of **w**. So minimizing the l_2-norm of weights $||w||_2^2$ becomes the first term in the minimizing function. Therefore, the SVM method solves

$$\min(w, \xi_i, b) \left\{\frac{1}{2} ||w||_2^2 + C \sum_i \xi_i\right\} \tag{2.9}$$

Constrained to

$$y_i(\varphi^T(x_i).w + b) \geq 1 - \xi_i, \quad for\ i = 1,2,..N \tag{2.10}$$

$$\xi_i \geq 0,\ for\ i = 1,2,..N \tag{2.11}$$

where **w** is the normal to the optimal decision hyper plane and it represents the nearest distance to the origin of the underlying coordinate system. This defines a linear classifier

($\hat{y_i} = f(x_i) = \varphi^T(x_i).w + b$) in the kernel space ɥ. The non-linear mapping function φ guarantees the linear separability in ɥ. This property comes from Cover's theorem [16]. C is the parameter that controls the generalization capabilities of the classifier and ξ_i are called positive slack variables and this allows the classifier to deal with permitted errors. The optimal hyper plane and slack variables are shown in Figure 2.2.

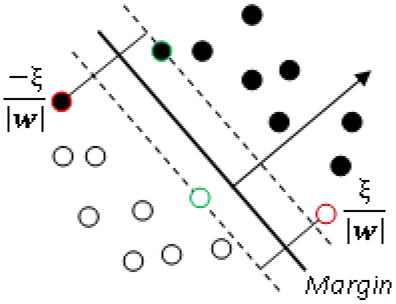

Figure 2.2 Optimal hyper plane for non-linearly separable classification problem with slack variables

Since the vector variable **w** lies in a kernel feature space ɥ, the best way to solve this equation is to solve a primal function given in Equation 2.9 through its Lagrangian dual problem, which consists of maximizing

$$Q_d \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \left(\varphi(x_i).\varphi(x_j)\right) \tag{2.12}$$

constrained to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0, for\ i = 1,2,..,N,$ where the auxiliary variables are the Lagrange multipliers corresponding to restrictions in Equation 2.10. In this way, the explicit usage of **w** can be avoided and Equation 2.12 can be optimized with respect to the variables $\alpha_i$instead. All φ mappings in the SVM learning occur in the form of inner products. This allows one to define a kernel function

$$K(x_i, x_j) = \varphi(x_i).\varphi(x_j) \quad (2.13)$$

without explicitly computing the mapping φ, a nonlinear SVM can be defined.

The pair { Ӈ , φ} will only exist if the kernel function K satisfies Mercer’s conditions. Linear kernels, polynomial kernels, and radial basis function kernels are by far the most popular functions that satisfy the Mercer’s conditions.

Linear kernel:

$$K(x_i, x_j) = x_i.x_j \quad (2.14)$$

Polynomial:

$$K(x_i, x_j) = (x_i.x_j + 1)^d, d \in \mathbb{Z}^+ \quad (2.15)$$

Radial Basis Function kernel (RBF):

$$K(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \sigma \in \mathbb{R} \quad (2.16)$$

After the dual problem, 2.9 is solved, $w = \sum_{i=1}^{n} y_i\, x_i \varphi(x_i)$, and the decision function for any test vector x is given by 2.17

$$\hat{y} = f(x) = sgn(\textstyle\sum_{i=1}^{n} y_i\, \alpha_i K(x, x_i) + b) \quad (2.17)$$

where *b* in Equation 2.17 can be computed by using the primal-dual relationship, with only samples with nonzero Lagrange multipliers α_i account in the solution. This leads to a concept of sparsity, i.e., the solution is expressed as a function only of the most critical training samples in the distribution, namely support vectors (SV). In this study, RBF kernel functions are used with SVMs.

2.2 Conventional Single classifier system

Since the introduction of artificial neural networks (ANN) for pattern classification [3], there are many approaches proposed over the years. *Chen* proposed a classification system based on back propagation neural networks (BPNN) [17]. This single classifier system that was developed to classify water, grasslands, and buildings proved to be encouraging with overall accuracies in the range of 80s with 20 meter spatial resolution AVIRIS data. A radial basis function (RBF) neural network-based classifier proposed in [18] and is shown to perform better than BPNN with similar data. Later, statistical techniques, such as Gaussian ML, are shown to be superior compared to ANN counterparts [5]. Few studies even compared the merits and demerits of the above mentioned approaches. The advancement of kernel-based classifiers for the classification of remote sensing data is relatively new [9]. The inceptions of all these classifiers are based on using them as lone methods for classifying the input patterns. With a single classification setup, there is a decade of research involving the improvement of these techniques in terms of training speed and accuracy. In the early stage of this development, most of the applications were to solve fairly small problems that can be solved easily with single classifiers. The aim of the research at that time was to focus on exploring new single classifier systems.

2.2.1 Limitations and Challenges

Later, the advancement of sensor technology is vastly improved as a result of that the classification problems become more and more challenging for single classifiers to handle. Most of the aforementioned single classifiers are far from being optimal. Most of these techniques cannot handle complex datasets containing random errors or insufficient training samples. So, the generalization capability of these approaches is not very good. The inherent nature of these approaches later became a limiting factor for improving the performance of single classifiers. The instability of these approaches hampered the development of better algorithms for the analysis of HSI data. As a result, the research in the last decade shifted its focus towards Multi-Classifier System (MCS) for HSI data. This will be discussed in the next section.

2.3 Ensemble Classification Systems

The concept of combining the predictions of multiple classifiers to produce a single classifier has been proposed by various researchers in the past [19][20]. In the literature, this concept is referred to as ensemble classifiers or MCS. The resulting MCS is generally more accurate when compared to the individual classifiers that form MCS. An effective MCS is one where the individual classifiers in the MCS are accurate and make their classification errors on different parts of the input space. Combining the predictions of identical classifiers will not have any improvement. So, it is useful only when there is a disagreement among the individual classifiers. In [21], Krogh et al proved that the overall classification error can be divided into a quantity which is the average generalization error of each classifier and a quantity proportional to the disagreement among the classifiers. From [22], [23], it can be concluded that an ideal MCS should consist of classifiers that have the highest disagreement possible. Bagging [24] and boosting [25] are very popular methods used to create diversity among classifiers. In [26], an improved approach called attribute bagging was introduced by Bryll et al. This followed the development of many wrapper based MCS approaches. Each classifier is trained with independently randomly selected feature subsets. The outputs are expected to be diverse and can be combined to form a final decision. Breiman [27] introduced a decision tree (DT)-based classification approach with Random Forests (RF). Min [28] proposed a dynamic subspace approach and Waske proposed the construction of SVM ensemble using Random Feature Selection (RFS) [29] . These are some of the approaches that are inspired from the basic idea of bagging and boosting classifiers successfully used in hyperspectral applications. In [30] Jacobs proposed an approach with mixture of experts followed by [31]. In [32] S Kumar et al. demonstrated the effectiveness of this technique with binary classifiers for a multiclass problem for hyperspectral data. In their work, partitioning of groups of classes is achieved by binary classifiers at different levels. Figure 2.3 shows a typical MCS setup.

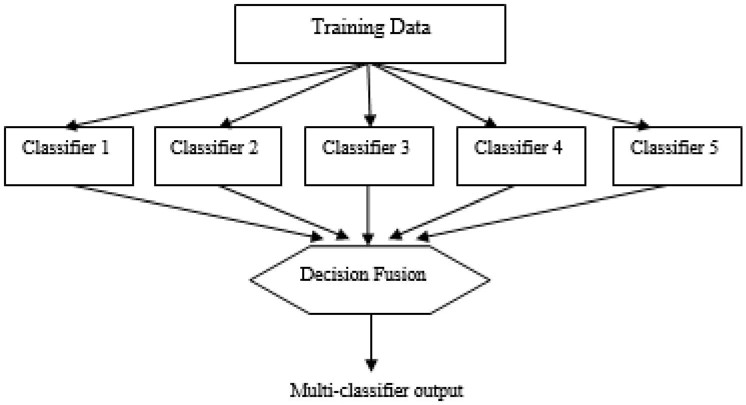

Figure 2.3 Typical MCS setup

Until recently, HSI classification has been performed with single classifiers. In recent work, to improve the performance of conventional single classifiers, MCS have been developed [33]–[35]. MCS are often referred to as ensemble classifier systems, and they potentially perform better than single classifiers when diversity is established among the classifiers. The diversity among the classifiers can be established in different ways [36], [37]. In chapter 4, the topic of diversity is discussed in more detail. Prasad et al. [34], [35] , demonstrated that, with an MCS setup with ML as classifier, the performance can be improved when compared to single classifiers, and there was a potential to further improve such a system by incorporating non-linear SVM classifiers.

Recently, a MCS based on Random Feature Selection (RFS) proposed by Waske et al [29] and a dynamic subspace approach [28] proposed by Min et al were shown to perform well with HSI data. Techniques such as random forests [38] and RFS perform well because they create diversity among the classifiers by resampling the spectral bands at the inputs of the classifier. As proposed in [39], diversity can also be created in other ways. In [24] , Breiman has demonstrated the diversity creation by re-sampling. Strategies related to this approach such as bagging [26] and boosting [25] are also shown to be effective.

2.3.1 Decision Fusion

After the training phase in MCS, each classifier in the system generates its own class decision. Assuming the individual classifiers or the features used by these classifiers are diverse, the class decisions will be expected to be different. It is important to combine these class decisions to decide the final class label. This process is often referred to as decision fusion.

The most straightforward and obvious approach for combining these decisions uses simple averaging. After training, the individual outputs are summed and divided by the number of classifiers. This simple approach is shown to be very effective in many cases [40]. This approach treats the output of each classifier equally as it weighs each outputs decision with the weight of the inverse of the number of classifiers. This approach is particularly useful when we have classifiers with different local minima. The simplicity of equal weights is the drawback of this approach. This approach fails to utilize the classifiers that make more contribution to output generalization. So, they are not preferred for HSI classification as often we have diverse classifiers with different strengths. To alleviate this issue, the weights can be set to be unequal. This approach is referred to as weighted averaging. The total weight sums to one and each classifier decision is multiplied by a fraction of the total weight according to its performance or the diversity of features.

For HSI data, voting is explored in [40]. It takes more than half the number of classifiers to agree on a class decision for it to be accepted as a final decision. This is called majority voting. Linear opinion polls and logarithmic polls are some variations of this approach. The problem with this approach is that majority voting ignores the fact that, in some cases, bad classifiers do produce results in such a way that can influence the overall decision. This defeats the purpose of having diverse classifiers and MCS itself. To avoid this problem, a ranking approach could be used. Here, along with the class decisions, the classifiers produce a list of choices ranked according to the likelihood in-terms of probability. The decision can be taken by using additional information about the data (usually the training data). In this dissertation, a new ranking approach called kernel decision fusion is proposed. This technique uses the distance between class probability density functions as likelihood. This approach is shown to perform better than majority voting.

2.4 Classical Feature Selection Techniques

The Maximum Likelihood (ML), a traditional supervised pattern classification approach, often fails to classify HSI data accurately because of (a) the high dimensionality of features, (b) multimodal distribution of data, and (c) limited ground truth availability. In order to solve the problem of high dimensionality, there are several existing approaches based on the concepts of dimensionality reduction and feature selection [1]. The Principal Component Analysis (PCA) and Fisher's Linear Discriminant Analysis (LDA) are popular dimensionality reduction techniques. Feature selection can also be performed using metrics such as the Bhattacharya Distance (BD), Jeffries-Matusita (JM), entropy, etc. The Gaussian ML classifier assumes that the classes are Gaussianly distributed. This is a limitation for the majority of the practical HSI datasets. Algorithms based on Gaussian Mixture models [41] have been proposed in the past to accommodate multi-modal distributions. An alternative approach that has become more popular recently with HSI data is the use of Support Vector Machines (SVM). Finally, there are techniques to solve the limited ground-truth availability such as sample interpolation and adaptive classifiers [42][43].

2.4.1 Feature Selection Techniques for SVM

SVMs are generally well suited for datasets that have a high dimensional feature space. Hence, they are naturally well-suited to be employed within a MCS framework. An SVM based MCS framework with Random Feature Selection (RFS) has been shown to outperform conventional single-classifier approaches for HSI applications. Random feature selection and random forests are well known techniques in the machine learning community. They are often used in situations where it is important to avoid noise or outlier features during the training phase of a classifier. These techniques can also be used in MCS algorithms to provide diversity between classifiers. Diversity within MCS is achieved by RFS with a uniform sampling distribution for the feature selection process. SVMs are generally thought to be insensitive to the dimensionality of the feature set, hence, at first glance, feature selection may seem unnecessary. Simple experiments with SVMs on HSI data have demonstrated that SVMs can perform better when a feature selection algorithm is employed before classification [44][45]. Recently, a few methods have been proposed for selecting features for SVM within a MCS setup.

2.4.2 Random Feature selection

The selection of features in [29] is a uniform random feature selection (RFS). In [44], the possibilities of using a non-uniform RFS (NURFS) based MCS with SVMs is explored. It is found that a diverse classifier ensemble for a classification problem need not always come from a RFS, as proposed in [29][28]. In [44], it is demonstrated that NURFS can provide better performance than uniform RFS. As extension, a fully automated MCS with NURFS using SVM, is presented in [46]. It is assumed that a diverse set of features leads to higher classification accuracies. Although the diversity can be defined in many ways [39] for the purposes of this study, a diverse set of spectral bands is defined as follows:

1. Bands are selected from multiple spectral regions across the entire spectrum of signature.
2. Cross-correlation between selected bands is minimized.

The approach proposed here combines the following methods to create diversity within a pool of classifiers and to ensure that strengths and weaknesses of individual classifiers are incorporated into the final decision making: a) re-sampling features in the data through RFS; b) manipulation of input features through NURFS; and c) manipulation of output classes through scores computed from Kernel Density estimation. The approach uses a spectral band grouping [47] to perform NURFS and it uses kernel density scores to perform decision fusion. To verify the effectiveness of this approach, experiments are performed to compare the overall accuracies, improvement in user and producer accuracies of SVM, RFS, NURFS, SVM with kernel density fusion, and NURFS with kernel density fusion. The sensitivity of the above mentioned approaches to the number of samples required to train them is also studied.

2.5 Band Grouping

Hyperspectral data has hundreds to thousands of reflectance values associated with every pixel. Because of this, HSI demands huge storage space and computing power. In addition, with supervised classification, HSI suffers from Hughes phenomenon. So, a feature reduction or band selection step is inevitable. Traditional ML-based statistical approaches include projection based dimensionality reduction techniques such as Principal Component Analysis (PCA) or Fisher's Linear Discriminant Analysis (LDA) followed by a supervised classifier. Spectral band grouping is another way to alleviate some of the problems caused by high dimensionality of hyperspectral data. HSI signature with narrow spectral spacing exhibits a high degree of correlation between each other. This is true for most of the HSI data and can be easily exploited to reduce the number of bands. A contiguous band grouping is shown to produce the desired results. In [48], Lee *et al*. have proposed a band grouping strategy that utilizes the spatial and spectral information to find band partitions. In [47] , Prasad proposed a supervised method that partitioned the HSI spectrum into multiple subspaces, each with a group of contiguous spectral bands. This technique is designed to exploit the MCS configuration. In [49] and [50], the authors propose a segmented PCA algorithm that employs a spectral band grouping prior to implementing PCA on each identified group separately. Band grouping based on Dirichlet Process Variable Clustering (DPVC) is gaining popularity in the area of spectral clustering and classification is explored in [51][52].

2.6 Hyperspectral Image Analysis

Hyperspectral imagery consists of three dimensions, two spatial and one spectral. Each spatial pixel can have hundreds to thousands of spectral reflectance values. HSI data is spectrally over determined so each pixel has ample amount of information. This rich information does not always guarantee the ability to identify and distinguish different materials. So, a robust feature elimination or feature selection step is very important to utilize the rich information to achieve higher classification accuracies. Reflectance can be defined as the percentage of the light reflected from the material of interest. Reflectance does not take into account the amount of light that is absorbed or transmitted by that material.

Figure 2.4 shows the reflectance spectrum of vegetation under various levels of herbicide stress. The hyperspectral sensor used in this case is an Analytical Spectral Devices (ASD) sensor. This figure shows the reflectance of the vegetation measured across a range of wavelengths (400nm to 2400nm). Different materials have varying levels of reflectance and absorption properties, so they can be used to uniquely determine the material. Commercial sensors offer various spectral and spatial resolutions, more number of bands can have more information in the data to distinguish between different materials. As an example, if a corn crop is sprayed with herbicide with different concentration, then it can be observed that the signatures share a common shape with slight difference in reflectance values across the spectrum. Although it is easy to distinguish different levels of herbicides, algorithmically it is a challenging problem as there are only subtle differences between each signature. The visible spectrum is from 400nm to 700nm. In this figure, a sharp edge can be observed at 700nm where the visible red region ends and near-infrared (NIR) begins. This characteristic can be observed in most vegetation types. The differences between these classes are little in the visible spectrum and gets larger at the higher wavelengths. This plays a vital role in classifying the HSI data compared to that of multi-spectral counterpart.

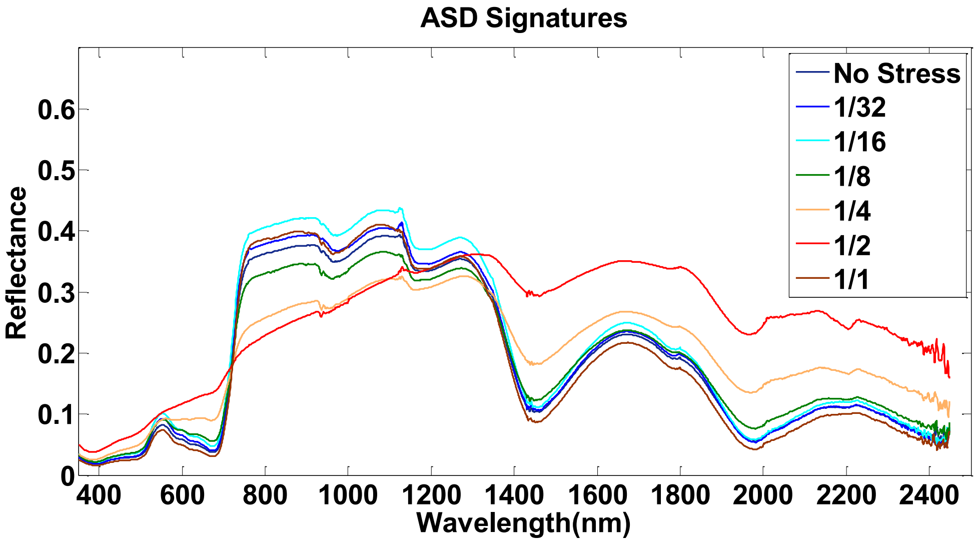

Figure 2.4 Hyperspectral signatures of corn under various levels of herbicide stress

Figure 2.5 shows a typical HSI remote sensing system for ground cover classification where the ground scene is captured by charge coupled devices (CCD) sensors.

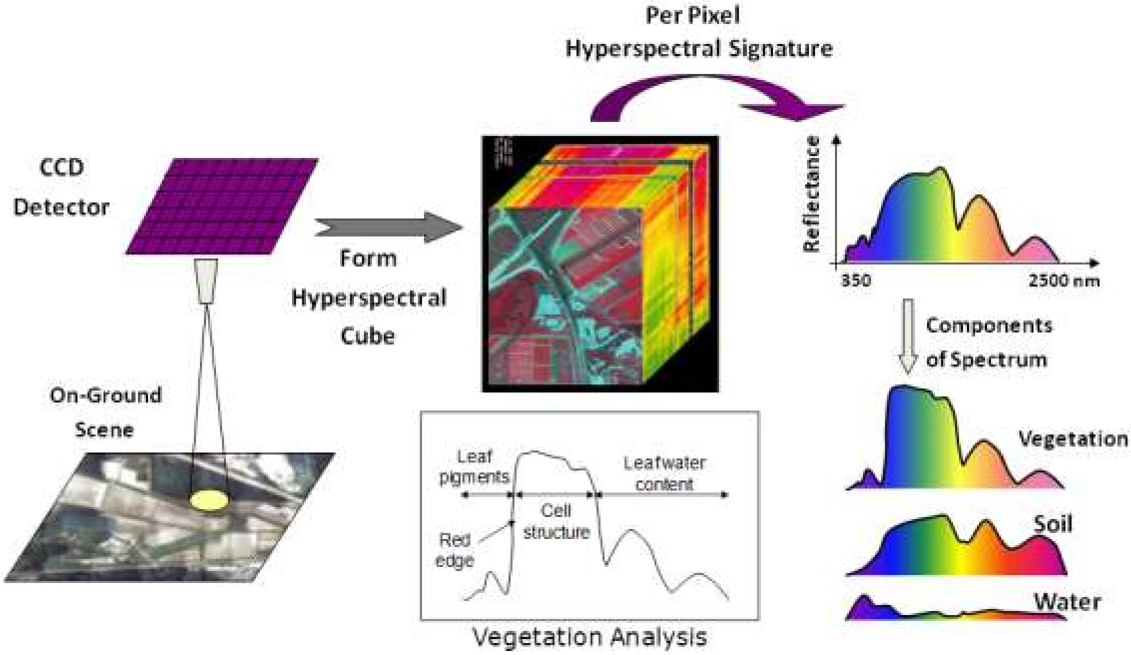

Figure 2.5 Hyperspectral Remote Sensing System

Image Courtesy – Lori M. Bruce, Mississippi State University

The reflectance values are then pre-processed and converted into a HSI cube. Red edge, leaf pigment, cell structure, and water content of the material are shown in the vegetation analysis. It can be observed that the vegetation has a unique spectral signature compared to water and soil. The difference is obvious in the case of classifying vegetation from soil and water but the problem could be a little harder when it comes to classifying different vegetation groups. In [53], Landgrebe discusses the inability of conventional multi-spectral methods to handle large number of bands and shows a way to perform the analysis with HSI data. An orthogonal subspace projection approach is proposed in [54] for the classification and dimensionality reduction of HSI data. Haertel investigated the problem of classifying subtly varying classes with HSI data [55]. Jimenez proposed a projection pursuit approach in [56]. This technique explored the possibility of bypassing the problem of a small number of training samples by making all computations in the lower dimensional space.

2.7 References

[1] A. K. Jain, R. P. W. Duin, and M. Jianchang, “Statistical pattern recognition: a review,” *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 22, no. 1, pp. 4–37, 2000.

[2] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition ,” *Data Min. Knowl. Discov.*, vol. 2, pp. 167–212, 1998.

[3] R. P. Lippmann, “An introduction to computing with neural nets,” *ASSP Magazine, IEEE*, vol. 4, no. 2. pp. 4–22, 1987.

[4] Richard O. Duda, Peter E. Hart, and D. G. Stork, *Pattern Classification*, Second Edi. Wiley Interscience, 2006.

[5] J. D. Paola and R. A. Schowengerdt, “A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 33, no. 4. pp. 981–996, 1995.

[6] A. Cheriyadat and L. M. Bruce, “Why principal component analysis is not an appropriate feature extraction method for hyperspectral data,” *Geoscience and Remote Sensing Symposium, 2003. IGARSS ’03. Proceedings. 2003 IEEE International*, vol. 6. pp. 3420–3422 vol.6, 2003.

[7] S. Prasad and L. M. Bruce, “Limitations of Principal Components Analysis for Hyperspectral Target Recognition,” *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 4. pp. 625–629, 2008.

[8] T. Fung, F. Y. Ma, and W. L. Siu, “Hyperspectral data analysis for subtropical tree species recognition,” *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS ’98. 1998 IEEE International*, vol. 3. pp. 1298–1300 vol.3, 1998.

[9] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 43, no. 6, pp. 1351–1362, 2005.

[10] B. Kuo, C. Li, and J. Yang, “Kernel Nonparametric Weighted Feature Extraction for Hyperspectral Image Classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1139–1155, Apr. 2009.

[11] M. Fauvel, J. Chanussot, and J. A. Benediktsson, “Kernel Principal Component Analysis for Feature Reduction in Hyperspectrale Images Analysis,” *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*. pp. 238–241, 2006.

[12] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach.,” *Neural Comput.*, vol. 12, no. 10, pp. 2385–404, Oct. 2000.

[13] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[14] D. J. Sebald and J. A. Bucklew, “Support vector machines and the multiple hypothesis test problem,” *Signal Process. IEEE Trans.*, vol. 49, no. 11, pp. 2865–2872, 2001.

[15] H. Chih-Wei and L. Chih-Jen, “A comparison of methods for multiclass support vector machines,” *Neural Networks, IEEE Trans.*, vol. 13, no. 2, pp. 415–425, 2002.

[16] T. M. Cover, “Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition,” *Electronic Computers, IEEE Transactions on*, vol. EC-14, no. 3. pp. 326–334, 1965.

[17] P.-F. Chen and T. C. Tran, “Hyperspectral imagery classification using a backpropagation neural network,” *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 5. pp. 2942–2947 vol.5, 1994.

[18] Q. Du and C.-I. Chang, “An interference rejection-based radial basis function neural network for hyperspectral image classification,” *Neural Networks, 1999. IJCNN ’99. International Joint Conference on*, vol. 4. pp. 2698–2703 vol.4, 1999.

[19] E. Alpaydin, “Multiple networks for function learning,” in *Neural Networks, 1993., IEEE International Conference on*, 1993, pp. 9–14 vol.1.

[20] R. T. Clemen, “Combining forecasts: A review and annotated bibliography,” *Int. J. Forecast.*, vol. 5, no. 4, pp. 559–583, Jan. 1989.

[21] J. V. Anders Krogh, “Neural Network Ensembles, Cross Validation, and Active Learning,” *Advances in Neural Information Processing Systems*, vol. pp 231–238. MIT Press, 1995.

[22] J. W. S. David W. Opitz, “Generating Accurate and Diverse Members of a Neural-Network Ensemble,” *Advances in Neural Information Processing Systems*. MIT Press, 1996.

[23] J. W. S. and O. S. David W. Opitz, “Actively Searching for an Effective Neural-Network Ensemble,” *Conn. Sci.*, vol. 8, pp. 337–353, 1996.

[24] L. Breiman, “Bagging Predictors ,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[25] Y. F. and R. E. Schapire, “Experiments with a new boosting algorithm,” in *13th International conference on Machine Learning*, 1996.

[26] R. Bryll, R. Gutierrez-Osuna, and F. Quek, “Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets,” *Pattern Recognit.*, vol. 36, no. 6, pp. 1291–1302, 2003.

[27] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[28] J. Yang, B. Kuo, P. Yu, C. Chuang, Y. Jinn-Min, K. Bor-Chen, Y. Pao-Ta, and C. Chun-Hsiang, “A Dynamic Subspace Method for Hyperspectral Image Classification,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2840–2853, Jul. 2010.

[29] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2880–2889, 2010.

[30] M. I. J. R.A Jacobs S.J Nowlan and G.E Hinton, “Adaptive mixtures of local experts,” *Neural Comput.*, vol. 3, pp. 79–87, 1991.

[31] M.I. J. R.A Jacobs, “Adaptive mixtures of local experts and the EM algorithm,” *Neural Comput.*, vol. 6, pp. 79–87, 1991.

[32] J. G. and M. M. C. S Kumar, “Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis,” *Springer verlag Pattern Anal. Appl.*, pp. 210–220, 2002.

[33] J. A. Benediktsson, X. Ceamanos Garcia, B. Waske, J. Chanussot, J. R. Sveinsson, and M. Fauvel, “Ensemble Methods for Classification of Hyperspectral Data,” in *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 2008, vol. 1, pp. I–62–I–65.

[34] S. Prasad, L. M. Bruce, and H. Kalluri, “A Robust Multi-Classifier Decision Fusion Framework for Hyperspectral, Multi-Temporal Classification,” in *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 2008, vol. 2, pp. II–273–II–276.

[35] S. Prasad, L. M. Bruce, R. I. Hammoud, and L. B. Wolff, “A Divide-and-Conquer Paradigm for Hyperspectral Classification and Target Recognition Optical Remote Sensing,” vol. 3, Springer Berlin Heidelberg, 2011, pp. 99–122.

[36] M. S. Haghighi, A. Vahedian, and H. S. Yazdi, “Creating and measuring diversity in multiple classifier systems using support vector data description,” *Appl. Soft Comput.*, no. 0, 2011.

[37] Gavin Brown, Jeremy Waytt, Rachel Harris, and X. Yao, “Diversity Creation Methods: A Survey and Categorisation,” *J. Inf. Fusion*, vol. 6, 2005.

[38] J. Ham, C. Yangchi, M. M. Crawford, and J. Ghosh, “Investigation of the random forest framework for classification of hyperspectral data,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 43, no. 3, pp. 492–501, 2005.

[39] R. Ranawana and V. Palade, “Intelligent Multi-Classifier Design Methods for the Classification of Imbalanced Data Sets - Application to DNA Sequence analysis,” University of Oxford, 2007.

[40] L. O. Jimenez, A. Morales-Morell, and A. Creus, “Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 3. pp. 1360–1366, 1999.

[41] S. G. Beaven, D. Stein, and L. E. Hoff, “Comparison of Gaussian mixture and linear mixture models for classification of hyperspectral data,” in *Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. IEEE 2000 International*, 2000, vol. 4, pp. 1597–1599 vol.4.

[42] B. Demir and S. Erturk, “Increasing hyperspectral image classification accuracy for data sets with limited training samples by sample interpolation,” in *Recent Advances in Space Technologies, 2009. RAST ’09. 4th International Conference on*, 2009, pp. 367–369.

[43] Q. Jackson and D. A. Landgrebe, “An adaptive classifier design for high-dimensional data analysis with a limited training data set,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 39, no. 12, pp. 2664–2679, 2001.

[44] S. Samiappan, S. Prasad, and L. M. Bruce, “Automated hyperspectral imagery analysis via support vector machines based multi-classifier system with non-uniform random feature selection,” in *2011 IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 3915–3918.

[45] S. Samiappan, S. Prasad, L. M. Bruce, and E. A. Hansen, “Branch and bound based feature elimination for support vector machine based classification of hyperspectral images,” in *International Geoscience and Remote Sensing Symposium IGARSS*, 2011, pp. 2523–2526.

[46] S. Samiappan, S. Prasad, and L. M. Bruce, “Non-Uniform Random Feature Selection and Kernel Density Scoring With SVM Based Ensemble Classification for Hyperspectral Image Analysis,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, pp. 1–9, 2013.

[47] S. Prasad and L. M. Bruce, “Decision Fusion With Confidence-Based Weight Assignment for Hyperspectral Target Recognition,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 46, no. 5, pp. 1448–1456, 2008.

[48] M. A. Lee, L. M. Bruce, and S. Prasad, “Concurrent spatial-spectral band grouping: Providing a spatial context for spectral dimensionality reduction,” *2011 3rd Work. Hyperspectral Image Signal Process. Evol. Remote Sens. WHISPERS*, pp. 1–4, 2011.

[49] X. J. X. Jia and J. A. Richards, “Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 37, pp. 538–542, 1999.

[50] Q. Du, W. Zhu, Y. H. Yang, and J. E. Fowler, “Segmented Principal Component Analysis for Parallel Compression of Hyperspectral Imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, pp. 713–717, 2009.

[51] N. Bouguila and D. Ziou, “A Dirichlet process mixture of dirichlet distributions for classification and prediction,” *2008 IEEE Work. Mach. Learn. Signal Process.*, 2008.

[52] S. Kim, M. G. Tadesse, and M. Vannucci, “Variable selection in clustering via Dirichlet process mixture models,” *Biometrika*, vol. 93, pp. 877–893, 2006.

[53] D. A. Landgrebe, “A perspective on the analysis of hyperspectral data,” *Geoscience and Remote Sensing Symposium, 1993. IGARSS ’93. Better Understanding of Earth Environment., International*. pp. 1362–1364 vol.3, 1993.

[54] J. C. Harsanyi and C.-I. Chang, “Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 32, no. 4. pp. 779–785, 1994.

[55] V. Haertel and D. A. Langrebe, “On the classification of classes with nearly equal spectral response in remote sensing hyperspectral image data,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 5. pp. 2374–2386, 1999.

[56] L. O. Jimenez and D. A. Landgrebe, “Hyperspectral data analysis and supervised feature reduction via projection pursuit,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 6. pp. 2653–2667, 1999.

CHAPTER III

NON-UNIFORM RANDOM FEATURE SELECTION, DECISION FUSION BASED ON KERNEL DENSITY ESTIMATION FOR HYPERSPECTRAL IMAGE CLASSIFICATION

3.1 Introduction

Hyperspectral imagery (HSI) provides the potential for high classification accuracies of subtly different ground-cover classes, but the key to its success is effectively extracting pertinent features from the high-dimensional datasets and effectively designing classifiers to exploit those features. Over the last decade, Support Vector Machines (SVMs) and Multi-Classifier Systems (MCS) have gained significant popularity compared with conventional statistical classification methods. Traditional statistical approaches, such as the Maximum Likelihood (ML), fail to classify these HSI data accurately because of the high data dimensionality or the multimodal distributions of the data. A major drawback of statistical approaches is that they often perform poorly under limited availability of training data, as is often the case with HSI data. Dimensionality reduction algorithms such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Stepwise-LDA (S-LDA) can provide partial solutions to this problem, but the resulting performance is typically still not at par with techniques such as SVMs.

MCS based classifiers can perform better than single classifiers if the diversity among the classifiers is established effectively. Even in a MCS setup, if the individual learners are ML classifiers, then the overall classifier will again suffer from similar drawbacks. SVMs are generally well suited for datasets that have a high dimensional feature space. Hence, they are naturally well-suited to be employed within a MCS framework [1]. Random feature selection and random forests [2] are well known techniques in the machine learning community. They are often used in situations where it is important to avoid noise or outlier features during the training phase of a classifier. These techniques can also be used in MCS algorithms to provide diversity between classifiers. In [1], diversity within MCS is achieved by RFS with a uniform sampling distribution for the feature selection process. SVMs are generally thought of to be insensitive to the dimensionality of the feature set, hence at first glance, feature selection may seem unnecessary. Simple experiments with SVMs on HSI data have demonstrated that SVMs can perform better when a feature selection algorithm is employed before classification [3]. Recently, a few methods have been proposed for selecting features for SVM within a MCS setup [4][5].

3.2 Non-Uniform Random Feature Selection

In [1], the performance of uniform RFS is shown to be better than that using single SVMs and other traditional approaches. Based on some experiments [3], it is found that a diverse classifier ensemble for a classification problem need not come from a uniform random selection and partitioning of the feature space. Considering the nature of the hyperspectral data, it is reasonable to believe that a non-uniform or spectrally-constrained RFS will provide a more diverse classifier setup, resulting in further improvements.

In the conventional RFS based multi-classifier system, the randomly selected subset of features provided to each classifier in the ensemble is generated by uniformly selecting the subsets from all features. In the proposed approach, features are selected for any classifier to come from one spectral region resulting in a non-uniform random feature selection

3.2.1 Preliminaries

The hyperspectral dataset is assumed to have *n* classes, each represented as μ_i. N_i is the number of samples in μ_i. Samples in μ_i are denoted as $C_{\mu_i} = \{c_i^1, c_i^2, \dots c_i^k\}$, where c_i^k is the k^th^ sample of the class μ_i. Samples from different classes can be grouped together to form a super class and is represented by $\Omega = \{C_{\mu_1}, C_{\mu_2}, \dots, C_{\mu_n}\}$. A feature vector $\bar{X}$ is *d* dimensional and each feature is represented by x_i, i.e., $\bar{X} = (x_1, x_2, \dots, x_d)$. We define the normalized distance between any two set of samples with respect to its feature vector $\bar{X}$ as *P* and P^n, where *P* is the distance between two classes μ_i and μ_j and P^n is the distance between two sets μ_i and Ω.

3.2.2 Spectrum Partitioning

The proposed approach is arrived after the following experiments. Initially, the feature space is divided into *m* distinct (but contiguous) regions with *NRi* being the numbers of features selected from each region by a uniform random feature selection. The first region always starts from the first band and the last region ends at the last band. The size of each region (*R*) is represented by R_i. Figure 3.1 shows how the feature vector is divided into *m* different regions. Since a uniform RFS is performed in each region separately, this approach can be thought of as a piece-wise uniform RFS.

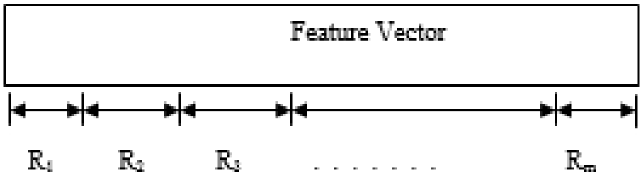

Figure 3.1 Dividing the feature vector into *m* regions

Since the HSI data often has high correlation between successive bands, there is a high chance of consecutive or nearby bands getting grouped into different learners when using conventional uniform RFS for MCS. This would clearly reduce the diversity of the ensemble and would result in reduced robustness of the MCS approach. However, in the proposed feature selection approach, features for individual classifiers in the resulting MCS are drawn in a non-uniform fashion. This creates greater diversity among the classifiers compared to selecting features from a uniform random selection. In particular, this approach can result in better ensembles that are less correlated, owing to the fact that the probability of bands that are spectrally close to each other being sent to different learners is very low. After these initial experiments with encouraging results, non-uniform RFS (NURFS) is used for the ensemble classification system. The above-mentioned approach follows a manual way of finding region boundaries in order to automate a band grouping strategy. The decision fusion used for these initial experiments is averaging and majority voting. In the ensemble classification system proposed in this work, a novel kernel density-based decision fusion strategy is employed. Band grouping and kernel density fusion are discussed later in this chapter.

In a RFS based multi-classifier system [1], a subset of features are selected by random sampling from a complete set of features whose indices tend to follow a uniform distribution. Figure 3.2(a) illustrates two examples of equally likely uniformly distributed spectral band feature selection where $[d_1]$ has highly correlated bands as compared to $[d_2]$. An obvious way to avoid this situation is, as shown in Figure 3.2(b), to divide the spectrum uniformly into smaller regions and perform feature selection within each region and concatenate the selected features. The outcome of this approach depends on the choice of the number of partitions and partition boundaries. Features can still be correlated with this approach. As one progresses their way along the spectrum, the bands in a hyperspectral signature are typically more highly correlated if they are adjacent. In addition, from one set of contiguous bands to another, the rate of change in the correlation of neighboring bands varies. An intelligent way of partitioning the spectrum would be to place the partition at a point in the feature set where the correlation of neighboring bands changes drastically. This will result in a non-uniform partitioning of spectral bands and bands selected from these non-uniform regions, which are expected to be less correlated. This is shown in Figure 3.2(c).

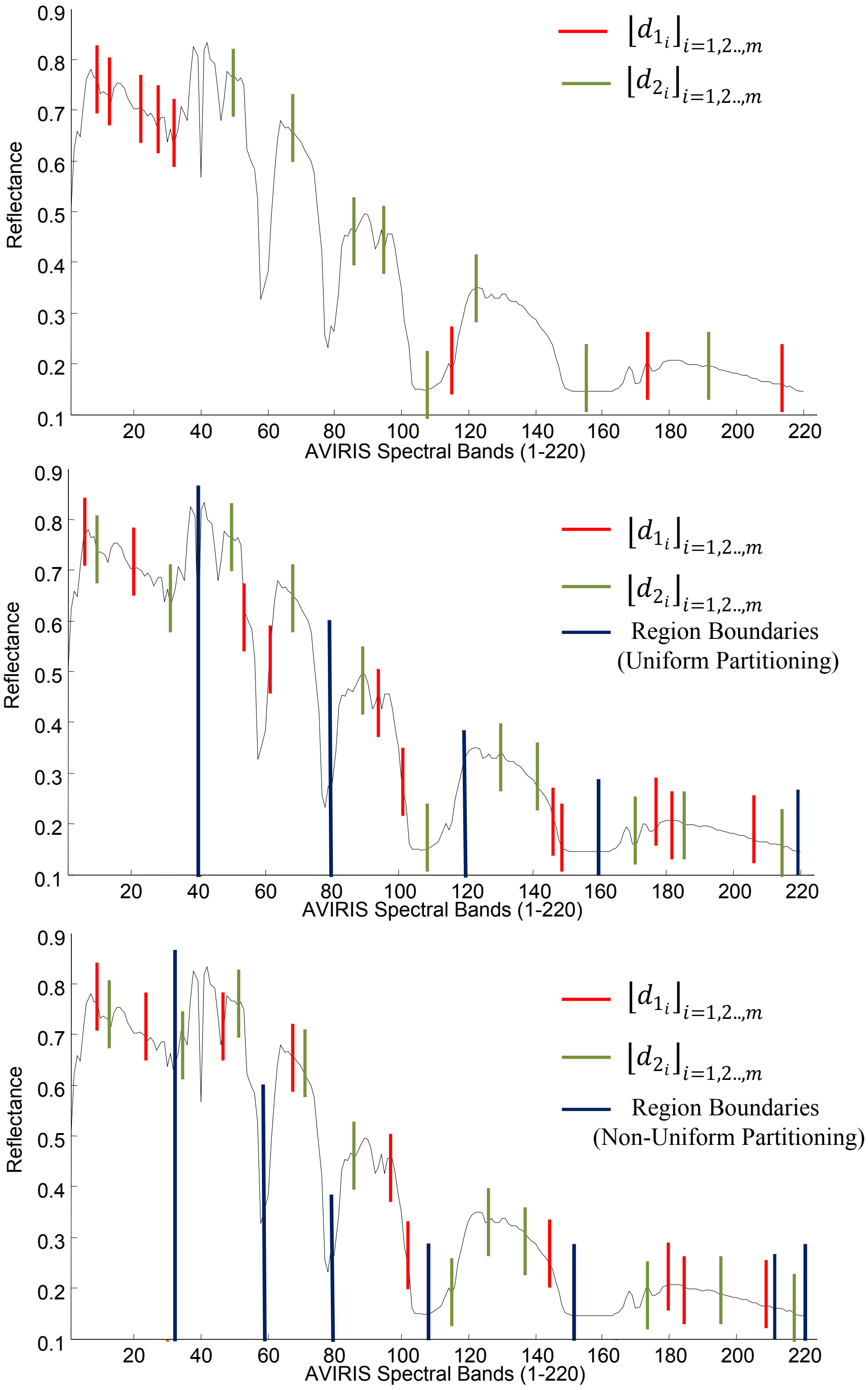

Figure 3.2 Partitioning of Spectral Bands

(a) Two examples of equally likely uniformly distributed spectral band feature selection where d1 has highly correlated band compared to d2. (b) Example of uniform partitioning of spectral band (shown in blue). (c) Non-uniform partitioning of spectral bands with uniformly distributed feature selection per partition

3.2.3 Band Grouping

To obtain an optimal set of partition boundaries in terms of correlation, an automatic band grouping strategy is used. In [6], an intelligent spectral partitioning technique that groups highly correlated bands into distinct contiguous subspaces and then use those partitions with a MCS is proposed. This is often called spectral band grouping or subspace partitioning. An intelligent (non-random) band grouping was performed to partition the spectrum into subsets. In this approach, the diversity among the classifiers is gained by breaking up the spectrum into smaller groups. The region boundaries are automatically selected based on a bottom-up band grouping strategy. In this approach, starting with the first band, each successive band is added to the group. If this addition does not change the performance metric employed, then, the growth of that group is stopped and a new group is started, resulting in a contiguous partitioning of the spectra. The metric employed for band grouping in this work is the product of the Bhattacharya distance and correlation.

3.2.4 Proposed Non-Uniform Random Feature Selection Approach

In the proposed approach, the feature space is divided into *m* distinct but contiguous regions in such a way that, in each region, it attempts to maximize the class separation and minimize the statistical dependence by band grouping. Let R_i be the size of each region, V be the total number of features in the data, and v_i be the number of features that are selected from each region, which is directly proportional to R_i. Then, the total number of features selected for each classifier is

$$V_{NU-RFS} = \sum_{i=1}^{m} v_i \tag{3.1}$$

Figure 3.3 illustrates this setup. Since uniform RFS is performed in each region separately, this approach can be thought of as a piece-wise uniform RFS. Since the HSI data has high correlation between consecutive bands, there is a good chance of consecutive bands getting grouped into different classifiers when using uniform RFS in a MCS. These highly correlated bands would clearly affect the diversity of the ensemble and, then, result in reduced robustness of the MCS approach. However, in the proposed NURFS, features for individual classifiers in the resulting MCS are drawn in a non-uniform fashion thereby creating greater diversity among the classifiers compared to selecting features from a uniform random selection.

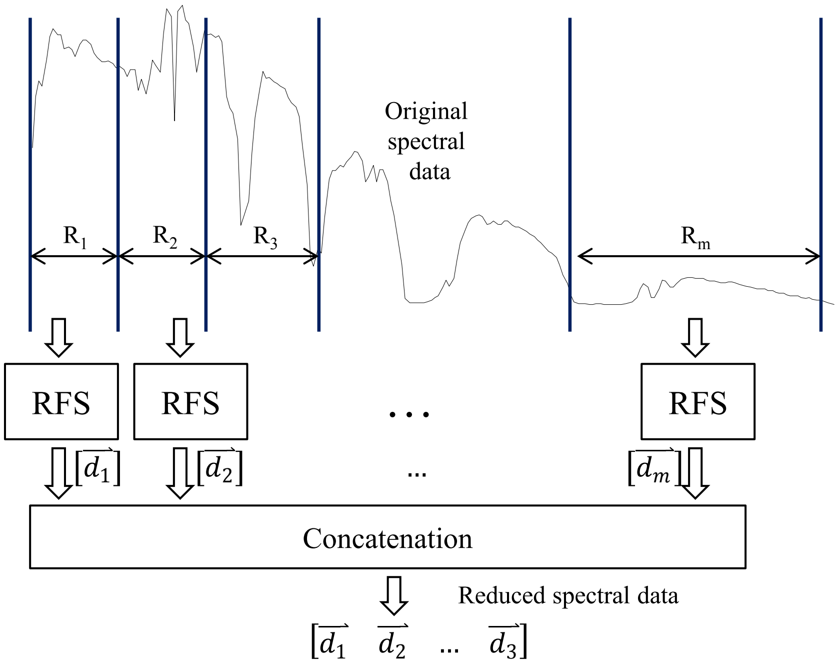

Figure 3.3 NURFS Feature Selection from original data

Experimentally, it is observed that this approach can result in better ensembles where features are less correlated, because the probability of features that are spectrally close to each other being sent to different learners is very low. The above said is applied to each classifier in the MCS separately unlike [1]. With initial experiments presented in [3], it is found that the size of the region R_i plays an important role in the performance of the overall classification. As recommended in [7], $\mathrm{V}_{\mathrm{NU-RFS}} \approx \frac{\mathrm{V}}{2}$. The proposed MCS system is discussed in Section 3.4 The random subspace selection demonstrated in [1][5][7] for MCS aims to create diversity among classifiers. The aforementioned techniques propose to construct ensembles by bagging and boosting variants.

In the proposed NURFS, it is believed that the optimal subset to create maximum diversity need not come from a uniform RFS because there is a very good chance that the nearby spectral bands get grouped into the same classifier. This is clearly the case of classifiers having correlated features. This situation of similar grouping of features can sometimes affect the diversity by forcing the classifiers to commit similar errors. By using the proposed approach described in Section 3.4, this approach attempts to alleviate this issue. The proposed approach is compared against regular SVM, RFS, and NURFS using band grouping only.

3.3 Kernel Density Based Decision Fusion

NURFS produces a group of features to be trained by each classifier. Each of these sets of features has a unique class separation capability since they form a different combination of the original feature set. In order to make use of this uniqueness in our system, we estimate a set of scores for each classifier that is proportional to the ability of the classifier to classify each class from all the other classes. For example, if there are z classifiers and the data has n classes, then, we generate a score matrix of size $z \times n$. These scores are computed by kernel density estimation across all the features. Oh *et al.* proposed an approach to estimate the class separation [8] to perform hand writing recognition. This proposed decision fusion approach is inspired from the algorithm in [8].

3.3.1 Estimation of Density Using Kernel Function and Distance between Classes

NURFS results in group of selected bands for each classifier. A probability density for a class for a feature vector $\bar{X}$ is estimated. A probability distribution for the class μ_i can be computed by,

$$f\left(\mu_i \middle| \overline{X}\right) = \frac{1}{hN_i} \sum_{i=1}^{N_i} K\left(\frac{\bar{X}-c_i}{h}\right) \tag{3.2}$$

where, $K(\bar{X})$ is the kernel function and $h > 0$ is the smoothing parameter. In the proposed NURFS, the rectangular, normal, triangular, and Epanechnikov kernel functions [9] are tested and compared. Let P be the distance between any two classes and P^n be the distance between any given class and all other classes (Ω) can be computed by Equations 3.3 and 3.4 respectively.

$$P\left(\mu_i, \mu_j, \bar{X}\right) = \int_{R_d}^{\cdot} \left| f_{\mu_i}(\bar{X}) - \; f_{\mu_j}(\bar{X}) \right| \, d\bar{X} \tag{3.3}$$

$$P^n\left(\mu_{i,}\Omega, \bar{X}\right) = \sum_{i \not\subset \Omega} P(\mu_i, \Omega, \bar{X}) \tag{3.4}$$

where f_{μ_i} and f_{μ_j} are class distributions of μ_i and μ_j respectively. When there is a complete overlap between distributions, Equation 3.3 gives a minimum (≈ zero) and no overlap gives maximum. i.e., when there is a complete overlap, $\bar{X}$ cannot distinguish two classes whereas it can distinguish the best when there is no overlap. Thus, it defines the ability of $\bar{X}$ to differentiate between any two classes of μ_i and μ_j. Equation (3.4) is computed for every class μ_i. Then the values are averaged over all the selected features resulting in an array of scores of seperability of each class with respect to the selected features.

3.3.2 Computation of Rank Matrix

These scores are sorted in descending order where the higher the score, the higher the ability to classify a class. This step is shown as a compute class score in the proposed system. This process is repeated for every classifier in MCS resulting in a $z \times n$ score matrix representing the ability of each classifier to distinguish a particular class μ_i from Ω. These scores are denoted as ρ. Now, each row of this matrix corresponds to the ability of each classifier to distinguish n classes where the higher the value of ρ, the higher the chance of distinguishing that class from all other classes (Ω). Although estimating the class probability density function is a harder problem than classification, the aim of estimating these scores is to get a coarse estimate of separation which can be used during decision fusion.

	μ_1	μ_2	μ_3	μ_4	μ_5
1	3	5	5	1	1
2	5	1	1	3	3
3	1	3	4	4	4
4	2	4	3	5	2
z=5	4	2	2	2	5

Figure 3.4 Kernel Density Score Matrix

3.3.3 Decision Fusion

After estimating the score matrix, the actual classification is performed with all the SVM classifiers in MCS resulting in class labels for every test sample from each classifier. Let Q be the number of test samples with z being the number of classifiers in MCS, then, the resulting class labels can be represented as a $z \times Q$ matrix. Each column of this matrix (l_i) holds the prediction of each test sample from z different classifiers in MCS. In the hard decision fusion scenario, the final classification decision can be obtained by a majority vote over each classifier. The final decision of the i^{th} test sample L_i can be obtained from l_i via Equation 3.5

$$L_i = mode(l_i) \tag{3.5}$$

$$\eta = \begin{cases} 0 & for\ 0 \geq \rho \geq 0.3 \\ z/3 & for\ 0.3 > \ \rho \geq 0.5 \\ z/2 & for\ 0.5 > \ \rho \geq 0.7 \\ z & for\ 0.7 > \ \rho \geq 1 \end{cases} \tag{3.6}$$

Mathematically, mode gives the most frequently occurring event. By means of a majority vote, a hard decision fusion is obtained. This only uses the predictions of z classifiers in MCS. The voting scheme described in Equation 3.5 is uniform voting. i.e., each classifier in MCS has equal strength in deciding the final class label. The proposed system has a voting mechanism based on scores ρ where the strength of each classifier is modified according to its ability to classify a particular class μ_i from all other classes (Ω). This can be achieved by creating a modified class label column matrix $\widehat{l_\iota}$ for each test sample based on the corresponding ρ. This is done by appending l_i with an array of length η. The elements of the appended array will have the class label corresponding to the highest ρ, where the length of $\widehat{l_\iota}$ can vary depending on ρ as given in Equation 3.6. From experiments with various HSI datasets, Equation 3.6 is arrived. This $\widehat{l_\iota}$ is then used to perform a majority vote. The decisions of MCS are not modified when $\rho \leq 0.3$. These scores will bias the majority voting decision based on the strengths and weaknesses of each classifier.

3.4 Proposed System for Hyperspectral Image Classification

A non-uniform random feature selection based ensemble classification system with kernel density decision fusion is depicted in Figure 3.5. NURFS, as demonstrated in Figure 3.3, is performed on the HSI data *z* times.

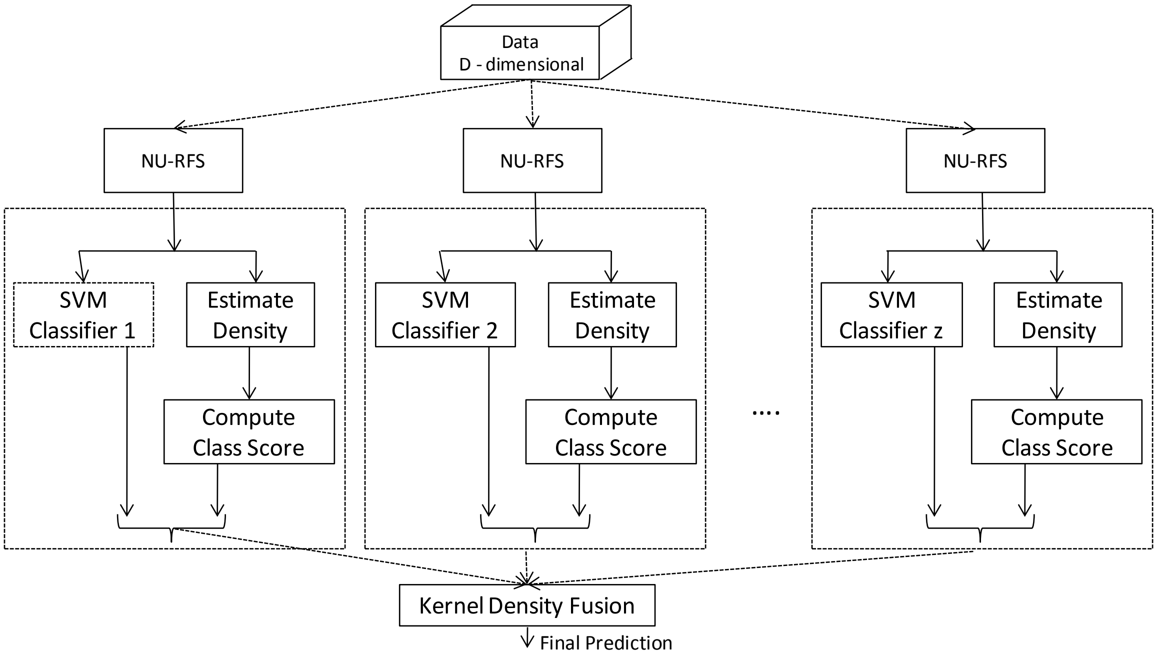

Figure 3.5 Proposed NURFS based Multi-Classifier System

The value of *z* is usually arrived after experimentation with the underlying training data. In this system, SVM classifiers are used as learners and each training dataset is used to estimate the separation between classes. The separation estimates are used to form a class score matrix, as shown in Figure 3.4. This matrix gives the ability of each of the *z* classifiers to distinguish a particular class from others. This estimate is later used to guide the decision fusion process. The decision fusion process combines the class scores and the class labels produced by SVM classifiers. The detailed explanations of these steps are presented in Sections 3.3.1 to 3.3.3.

3.5 Practical Application of Non-Uniform RFS MCS

Aflatoxin contamination is a concern for all classes of livestock. Aflatoxins are produced by certain mold fungi: Aspergillus flavus and Aspergillus parasiticus. Aflatoxin in food is hazardous for humans and animals. In this work, we propose a non-invasive system for detecting aflatoxin and classifying corn kernels based on the aflatoxin contamination levels. Fluorescence hyperspectral images of single corn kernels were used. Single and multi-classifier configurations of support vector machines are used to classify single corn kernels on a per-pixel basis. The performance of SVM classification with and without feature selection is assessed. Confusion matrices of different configurations are used for comparison, demonstrating that the multi-classifier system with non-uniform feature selection performs well, achieving an overall accuracy of 84%.

3.5.1 Problem Definition

Aflatoxin is an extremely toxic chemical produced primarily by two molds, namely, Aspergillus flavus (A. flavus) and A. parasiticus. The toxins are produced when the molds invade grain crops that are stressed by heat and drought [10]. Aflatoxin contamination of corn in particular is a serious problem because consumption of the toxin in food or feed can lead to deleterious health effects for human beings as well as animals [11]. Additionally, the financial loss to farmers due to rejection and disposal of infected grain can be devastating. In the United States, the Food and Drug Administration (FDA) regulates aflatoxin levels, allowing 20 parts per billion (ppb) in food and 100 ppb in feed as a general guideline [12].

The presence of aflatoxin can be tested by several methods [13]. Traditional detection methods include the black light presumptive test followed by thin layer chromatography (TLC) for quantification. More recently developed methods include mini column tests, rapid test field and laboratory kits, and enzyme-linked immuno assay (ELISA) kits. Although convenient, these techniques have various drawbacks such as lack of quantitative ability, being time consuming, costly, invasive in nature, and most require destruction of samples. The most reliable detection and quantification methods such as high performance liquid chromatography (HPLC) or mass spectrometry coupled with HPLC (MS-HPLC) are not only very expensive but also instrument and interpretation intensive [14].

Hyperspectral imaging is a useful technology in determining contamination in food especially corn [15]. By exciting the corn kernels with ultra violet (UV) radiation, the emitted fluorescence is captured by a hyperspectral image sensor with bands covering visible to near-infrared regions. In the past, the applicability of fluorescence hyperspectral imaging for estimating aflatoxin content in individual corn kernels was studied in [15]. The performance of the spectral angle mapper classification technique was evaluated in [16].

3.5.2 Proposed System for Classifying Aflatoxins in Corn Kernels

Due to the high dimensionality nature of the data, statistical classifiers, such as maximum likelihood, require a feature reduction step. The performances of such classifiers are shown to perform poorly when compared to classifiers based on Support Vector Machines (SVM). In [17], the suitability of SVMs for hyperspectral image classification was presented. Multi-classifier systems (MCS) are shown to perform better than their single classifier counter parts [18][19]. MCS are often referred to as ensemble classifiers, where more than one classifier is employed and each perform classification of the same data with different classification algorithms or different features of the data. The decisions of individual classifiers are combined to produce the final decision. In [5] and [20], the effectiveness of MCS is demonstrated with hyperspectral data. In [1] and [3], it is shown that the SVM classification performance can be improved by performing feature selection on hyperspectral data. Uniform and non-uniform random feature selection (RFS) techniques are shown to perform better than single classifiers.

The proposed non-invasive detection system for aflatoxin in corn, using SVM classifiers under different configurations is illustrated in Figure. 3.5.

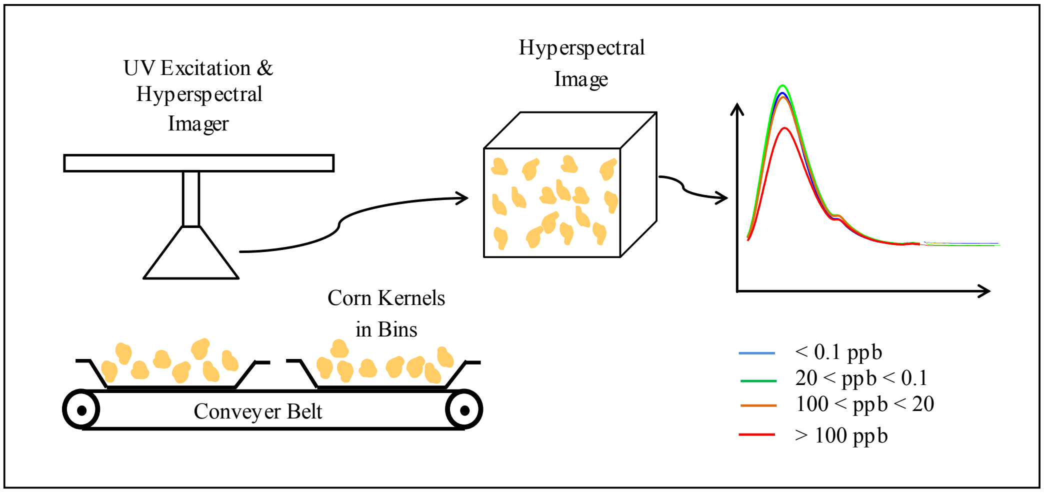

Figure 3.6 Non-Invasive detection of Aflatoxins in corn kernels

The proposed non-invasive system for classifying four aflatoxin levels in corn kernels is shown in Figure 3.5, with aflatoxin levels being divided into <0.1ppb (parts per billion), 20>ppb>0.1, 100>ppb>20 and >100ppb as Class1, Class2, Class3 and Class4 respectively . It is assumed that the corn kernels sampled are passing on a conveyer belt through the UV-excitation and hyperspectral imager in small bins. Then, a hyperspectral image is captured for each bin. This is a pixel based classification system, and it produces a map of acquired images with different chromic representation for each aflatoxin concentration.

In this study, several feature selection and classification schemes for the proposed system have been trained and tested. The references of all 504 kernel pixels are obtained from VICAM chemical testing process [12]. The hyperspectral signatures of the training pixels are used to train candidate feature selection and classification methods, including 1) Single SVM classifier, 2) SVM-MCS with uniform RFS, and 3) SVM-MCS with non-uniform RFS. Then, each method is tested, with each pixel in the image classified, resulting in classification maps, or images. The resulting classification maps are then compared to the reference maps and confusion matrices are created. Instead of dividing the spectrum into regions with uniform length, an intelligent approach could be employed, computing the correlation between consecutive bands and splitting the regions where there is a steep change in correlation. This can be achieved by a band grouping strategy as described in [21].

For this study, the product of the Bhattacharya distance and correlation has been used in band grouping. Epanechnikov kernel is used in the estimation of the density functions. Detailed results are presented in Chapter 4. A non-invasive system for classifying corn kernels based on the contamination levels of aflatoxin has been proposed. From the study, it is evident that SVM classifies are well suited to automatically analyze the UV excited hyperspectral pixels of corn kernels. Further, the study shows that the SVM multi-classifier with non-uniform random feature selection provides better classification accuracies than single SVM and multi-classifier SVM with RFS. In future work, a classification accuracy map of corn palates can be used to show the potential of the proposed system. In addition, post processing such as morphological processing, can be performed on classification maps to eliminate outliers.

3.6 References

[1] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2880–2889, 2010.

[2] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[3] S. Samiappan, S. Prasad, and L. M. Bruce, “Automated hyperspectral imagery analysis via support vector machines based multi-classifier system with non-uniform random feature selection,” in *2011 IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 3915–3918.

[4] M. Pal and G. M. Foody, “Feature Selection for Classification of Hyperspectral Data by SVM,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 5, pp. 2297–2307, 2010.

[5] J. Yang, B. Kuo, P. Yu, C. Chuang, Y. Jinn-Min, K. Bor-Chen, Y. Pao-Ta, and C. Chun-Hsiang, “A Dynamic Subspace Method for Hyperspectral Image Classification,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2840–2853, Jul. 2010.

[6] S. Prasad and L. M. Bruce, “Decision Fusion With Confidence-Based Weight Assignment for Hyperspectral Target Recognition,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 46, no. 5, pp. 1448–1456, 2008.

[7] T. K. Ho, “The Random Subspace Method for Constructing Decision Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[8] I.-S. Oh, J.-S. Lee, and C. Y. Suen, “Analysis of Class Separation and Combination of Class-Dependent Features for Handwriting Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 1089–1094, 1999.

[9] V. A. Epanechnikov, “Non-Parametric Estimation of a Multivariate Probability Density,” *Theory Probab. its Appl.*, vol. 14, no. 1, pp. 153–158, 1967.

[10] G. A. Payne and N. W. Widstrom, “Aflatoxin in maize,” *CRC. Crit. Rev. Plant Sci.*, vol. 10, no. 5, pp. 423–440, Jan. 1992.

[11] G. A. Payne and J. Richard, “Mycotoxins: Risks in Plant, Animal, and Human Systems,” 2003.

[12] U. S. D. of A.-F. Grain, *Aflatoxin Handbook*. 2009, pp. 8.1–8.27.

[13] L. Sweets and J. Wrather, “Aflatoxin in Corn,” 2009.

[14] I. K. Cigic and H. Prosen, "An Overview of Conventional and Emerging Analytical Methods for the Determination of Mycotoxins," *Int. J. Mol. Sci.*, vol. 10, no. 1, pp. 62–115, 2009.

[15] Z. H. H Yao R Kincaid, R.L Brown, T.E. Cleveland, D. Bhatnagar, "Correlation and Classification of Single Kernel Fluorescence Hyperspectral Data with Aflatoxin Concentration in Corn Kernels Inoculated with Aspergillus flavus Spores," *J. Food Addit. Contam.*, vol. 27(5), pp. 701–709, 2010.

[16] H.Yao, Z. Hruska, R. Kincaid, A. Ononye, R.L. Brown, and T. E. Cleveland, "Spectral Angle Mapper classification of fluorescence hyperspectral image for aflatoxin contaminated corn," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, 2010, pp. 1–4.

[17] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Geosci. Remote Sensing, IEEE Trans.*, vol. 42, no. 8, pp. 1778–1790, 2004.

[18] J. V. Anders Krogh, "Neural Network Ensembles, Cross Validation, and Active Learning," *Advances in Neural Information Processing Systems*, vol. pp 231–238. MIT Press, 1995.

[19] J. A. Benediktsson, X. Ceamanos Garcia, B. Waske, J. Chanussot, J. R. Sveinsson, and M. Fauvel, "Ensemble Methods for Classification of Hyperspectral Data," in *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 2008, vol. 1, pp. I–62–I–65.

[20] C. Mingmin, K. Qian, J. A. Benediktsson, and F. Rui, "Ensemble Classification Algorithm for Hyperspectral Remote Sensing Data," *Geosci. Remote Sens. Lett. IEEE*, vol. 6, no. 4, pp. 762–766, 2009.

[21] M. A. Lee, L. M. Bruce, and S. Prasad, "Concurrent spatial-spectral band grouping: Providing a spatial context for spectral dimensionality reduction," *2011 3rd Work. Hyperspectral Image Signal Process. Evol. Remote Sens. WHISPERS*, pp. 1–4, 2011.

CHAPTER IV

EXPERIMENTAL SETUP AND RESULTS

4.1 Experimental Hyperspectral Datasets

In this dissertation, five hyperspectral datasets are used to analyze the efficacy of the proposed NURFS feature selection and kernel-based density decision fusion techniques. Among the five datasets, two of them represent agricultural problems. The first is a corn herbicide stress dataset where each class belongs to HSI signatures with varying levels of herbicide stress, The second is a standard Indian pines agricultural dataset acquired over northern Indiana using AVIRIS [1]. The third dataset represents a problem of classifying aquatic plant species. It is a simulated HyspIRI dataset [2] and is simulated from ASD [3] signatures. The fourth is another standard dataset representing an urban classification problem. It was acquired over Pavia, Italy using a ROSIS sensor [4][5]. The fifth and final dataset is a laboratory fluorescence HSI data representing a problem of classifying aflatoxins in corn kernels.

4.1.1 Corn Herbicide Stress Dataset - Agricultural Data

The corn herbicide stress experimental HSI dataset was acquired over North Mississippi's Blackbelt Experiment Station agricultural test site in June 2008. The dataset has seven classes, each representing chemical stress on a corn crop [6]. The corn crop, grown under controlled conditions, was induced with varying degrees of chemical stress.

The crop was sprayed with seven different concentrations of Glufosinate herbicide diluted with water, where the seven classes or concentrations were (control) 1/32, 1/16, 1/8, 1/4, 1/2, and 1 times the labeled rate of the herbicide concentrations.

This dataset is acquired by using handheld Analytical Spectral Devices (ASD) sensor resulting in HSI datasets with 2151 bands. Note that all seven classes in this dataset represent the same species under varying degrees of stress, thus resulting in a very challenging classification problem. Table 4.1 shows the class names and number of samples considered for this experiment from each class. The site where the data acquired is shown in Figure 4.1(b) and (c) along with the color map in 4.1(a). The hyperspectral signatures of stressed and healthy crop are shown in Figure 4.1(d)

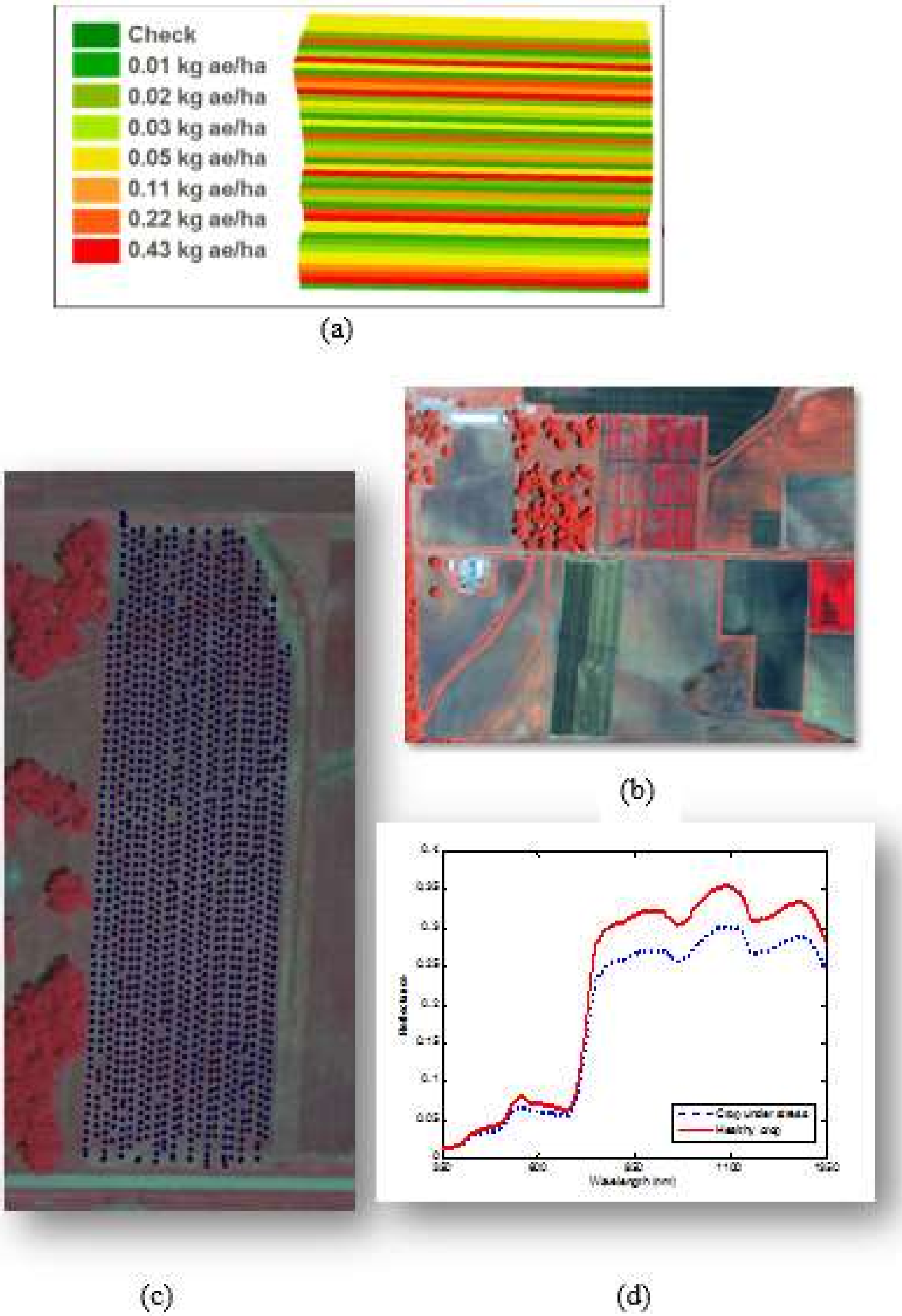

Figure 4.1 Corn herbicide dataset

(a) color map showing concentrations of herbicide (b) corn field (broad view) (c) corn field (narrow view) with ground truth points marked as blue dots (d) mean signature of crop under stress and healthy crop (Image Courtesy: Lori M. Bruce, Mississippi State University)

Table 4.1 Class names of corn stress data with number of samples from each class

Class Name	Control	1/32x	1/16x	1/8x	1/4x	1/2x	1x
Number of Total Samples	212	148	166	178	194	158	164

4.1.2 Indian Pines Dataset - Agricultural Data

The second experimental HSI dataset employed was acquired using NASA's AVIRIS sensor and was collected over the Northwest Indiana's Indian Pine test site in June 1992. The image represents a vegetation-classification scenario with 145x145 pixels and 220 bands in the 400 to 2450nm region of the visible and infrared spectrum. This dataset has 16 classes.

Figure 4.2 (a) shows the pseudo colored version of the site, Figure 4.2(b) shows the ground truth, and 4.2(c) shows the class names for each regions in the ground truth. This is a standard dataset used to compare the performance of the proposed approach with the state-of-the-art techniques.

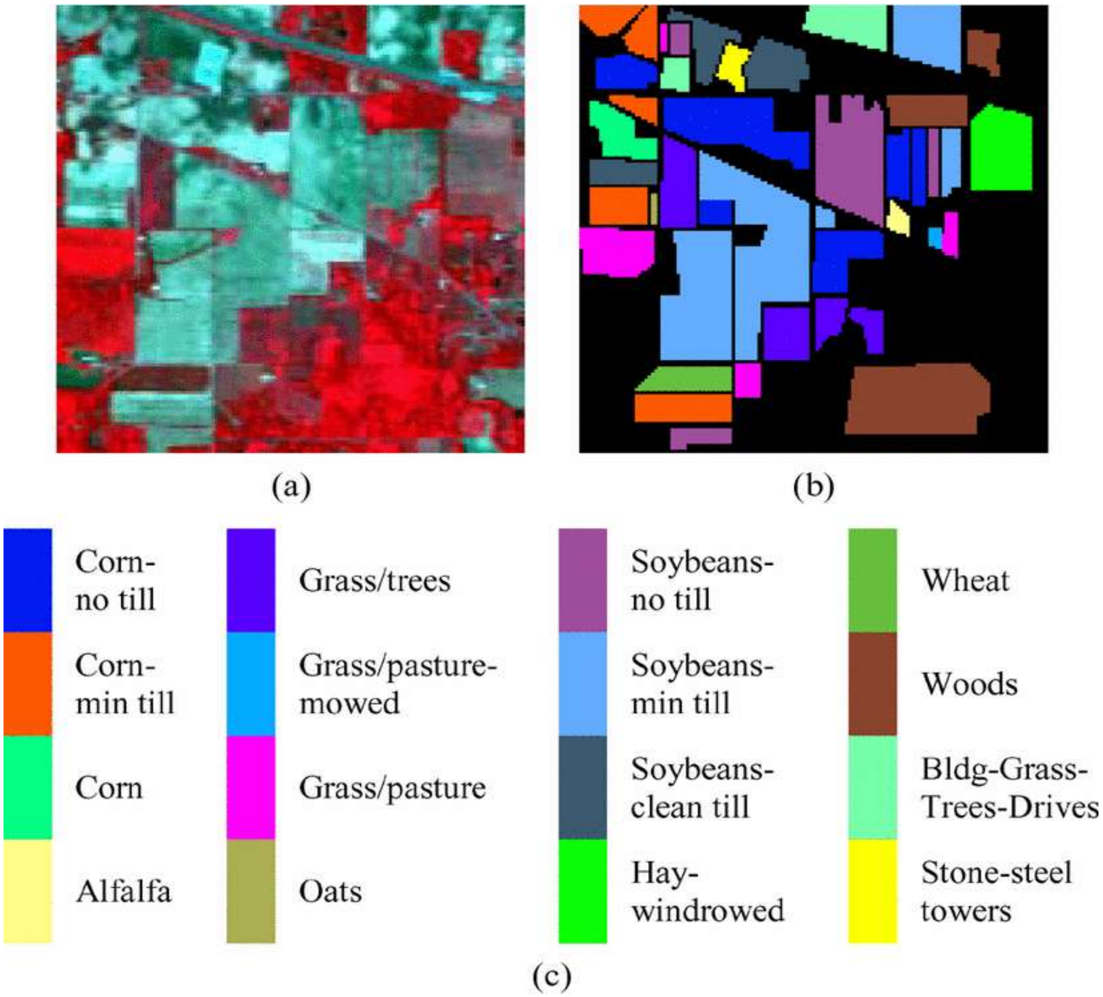

Figure 4.2 Indian pines dataset, ground truth and labels

(a) Indian pines pseudo colored RGB image (b) Indian pines ground truth (c) Indian pines class labels

4.1.3 Aquatic Plants Dataset - Simulated HyspIRI data

The third dataset represents an aquatic invasive and native vegetation species dataset. The aquatic dataset consists of signatures from five different classes – Nelumbo (*Nelumbonales*), Waterhyacinth (*Eichhornia crassipes*), Duckweed (*Araceae*), Salvinia (*Salviniaceae*) and water. These aquatic species were grown under controlled conditions at Mississippi State University and the water samples are collected from the Oktibbeha

County Lake near Mississippi State University. Collecting these water samples is one of the contribution in this research work. The signatures of water are collected by using an ASD handheld hyperspectral sensor carried in a motor boat. The white reference values are recorded at multiple times and the water signatures are captured at various depths of the lake. This dataset represents a typical aquatic species (native/invasive) mapping/detection task. Figure 4.3 illustrates the average sample proxy HyspIRI signatures from this dataset. The aquatic dataset has approximately 28 samples in each class and leave–one-out cross validation is employed for this dataset.

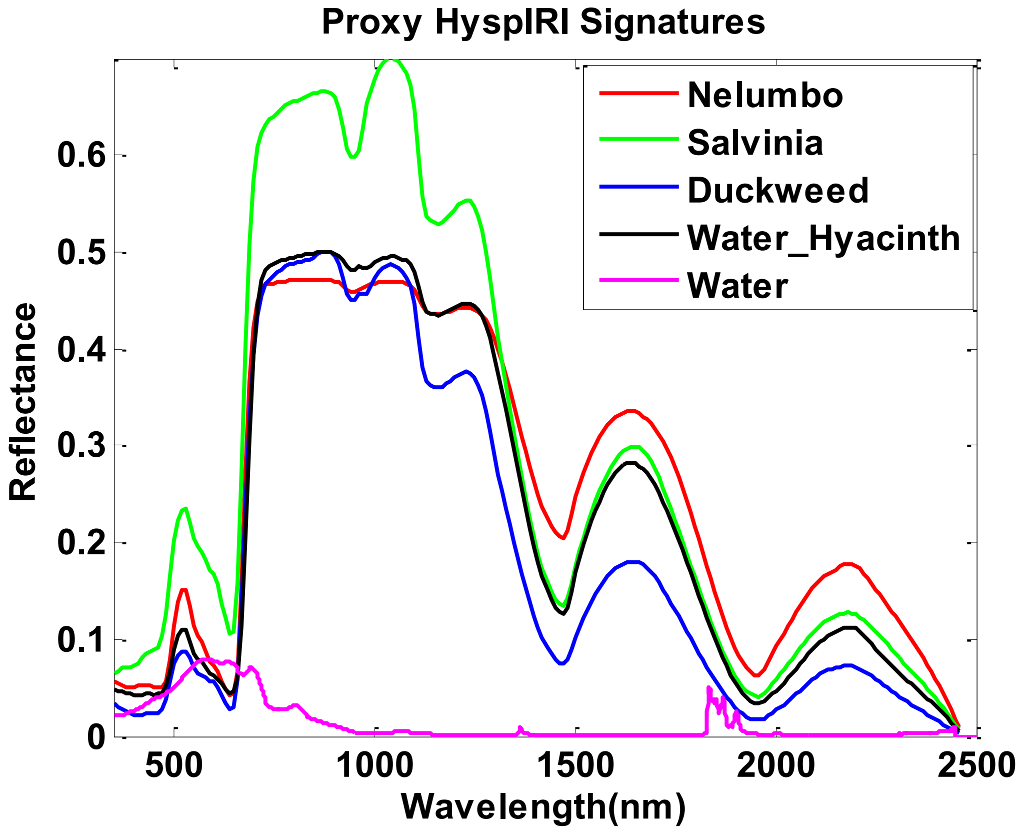

Figure 4.3 Mean signatures of various aquatic species and water

To simulate the HyspIRI data, existing ASD spectral signatures are mapped to the spectral specifications of HyspIRI. The spectral range of ASD is 350nm to 2500nm with a sampling interval of 1.4nm to 2nm. These specifications of ASD are suitable for simulating the HyspIRI signatures. The spectral range of HyspIRI is 380nm to 2500nm with a sampling interval of 10nm. A Gaussian-weighted averaging of every 21 successive ASD bands is performed to produce one proxy HyspIRI band. Detailed specifications of ASD and HyspIRI can be found in [2], [3]. Table 4.2 shows the class names and number of samples from each class used for the experiments. It can be observed that the training sample size is very small compared to other datasets.

Table 4.2 Class names of aquatic species and number of available samples from each class

Class Name	Nelumbo	Salvinia	Duckweed	Water Hyacinth	Water
Number of Total Samples	30	27	30	30	100

4.1.4 Pavia, Italy Dataset - Urban Data

The fourth dataset used has 102 spectral bands acquired by the ROSIS sensor over Pavia, Northern Italy [5][4]. This is a standard dataset used by the HSI research community and is used here to check the efficacy of the proposed approach with state-of-the-art techniques from other research work. This data has 9 classes. The classes are Water, Trees, Asphalt, Self-Blocking Bricks, Bitumen, Tiles, Shadows, Meadows, and Bare soil. For this dataset, considering the very high number of samples from each class, the model selection is conducted on a subset of training samples rather samples from each class. These model parameters are used to train the SVM classifiers.

4.1.5 Corn Aflatoxin Dataset - Laboratory Data

This dataset was created by Dr. Haibo Yao of Mississippi State University as part of the project funded by Gates foundation. The corn cobs were grown in Tifton, GA, USA. The toxigenic inoculant AF13 strain of A.flavus was obtained from the United States Department of Agriculture Southern Regional Research Center (USDA-SRRC) facility in New Orleans, LA, USA. The toxins are inoculated (field inoculation) into corn ears with a 12 gauge stainless needle. The inoculum was injected into each side of corn ears through the husk at an early stage. The corn kernels were harvested, dried, shucked and shelled into individual kernels. The kernels located near the inoculated point were selected as possibly contaminated and kernels located on the opposite side of the same corn ear were selected as control or reference. These kernels are then used for imaging. All these kernels were subjected to Aflatest analytical method from VICAM, Milford, MA, USA to measure the true aflatoxin levels and this enables them to use these HSI signatures as training points. A total of 504 corn kernels were collected for this study.

As shown in Figure 4.4 (a), the individual kernels were placed on a flat plate for hyperspectral imaging. Each plate contained a maximum of 30 kernels and this plate is illuminated with a UV light source with a wavelength centered at 365nm. The UV lamp (Model XX-15A) used in this experiment was manufactured by Thermo Fisher Scientific Inc, Waltham, MA, USA. Figure 4.4 (b) show the class label truth map obtained through VICAM test.

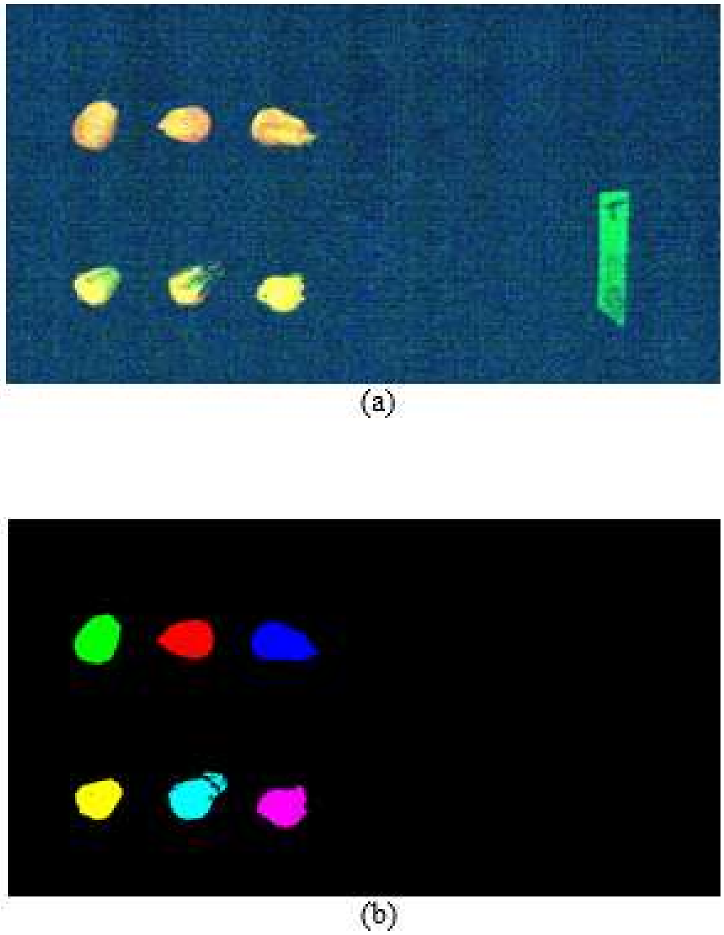

Figure 4.4 Corn aflatoxin tray and ROIs

(a) Example image of corn kernels under UV excitation on ceramic plate (b) different levels of aflatoxins represented as color map.

A 14-bit PCO 1600 charge-coupled device (CCD) camera manufactured by Cooke Corporation, Romulus, MI, USA and Imspector v10E spectrograph manufactured by Spectral Imaging Ltd, Oulu, Finland are used for imaging. The spectrograph had 30 micron entrance slit and a 35mm lens. Push broom line scanning was used and the HSI pixels were captured with a patented focal plane scanning method. The size of captured HSI cubes is 800x425x183 in spectral range of 400-600nm. Image preprocessing steps included dark current subtraction, wavelength assignment, and spectral low pass filtering for noise removal. A region of interest (ROI) was created for each kernel representing the spatial area of a particular kernel. The spatial size of each image contained 800x425 pixels [7]. Table 4.3 shows the range of aflatoxin classes and number of samples from each class considered for experiments.

Table 4.3 Class names of corn aflatoxin with number of available samples from each class

Class Name	0 – 0.1	>0.1 - 20	>20 - 100	>100
Number of Total Samples	11,193	23,169	11,884	24,799

4.2 Experimental Setup

The classification is performed using SVM with Gaussian Radial Basis Function (RBF) kernel for all experiments. The model selection for the SVM is performed using cross validation and grid search. An example of a grid search plot for the model selection of C and γ is shown in Figure 4.5. Optimization showed for Indian pines data with 10% training for the proposed NURFS approach. For all the datasets, the RBF parameters C and γ are estimated by selecting 10% of the training samples from each class and performing a grid search using cross validation, except for the Pavia dataset where we used 5% of the training data for the model selection. A confusion matrix for every classification problem is computed and the user, producer, and overall accuracies of different algorithms are compared. The statistical classifiers are standard Gaussian maximum likelihood with PCA and Fisher's LDA.

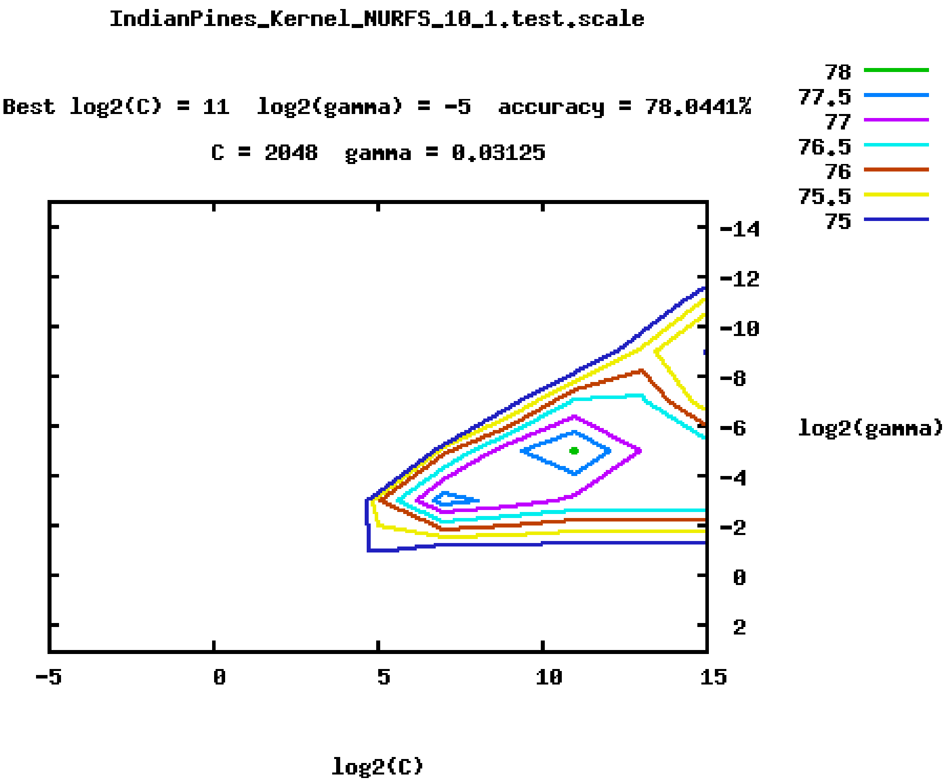

Figure 4.5 Example of grid search plot for the model selection of C and γ.

Optimization showed for Indian pines data with 10% training for the proposed NURFS approach

4.3 Experimental Results

The efficacy of the proposed approach is studied for the datasets described in Section 4.2. In Section 4.3.1, the state-of-the-art statistical classification techniques are studied. This establishes the bench marking of the statistical techniques with these datasets. A Gaussian maximum likelihood classifier (ML) is used with popular dimensionality reduction techniques such as principle component analysis (PCA) and linear discriminant analysis (LDA). In Section 4.3.2, the performance of the proposed NURFS and kernel density-based decision fusion is compared with other state-of-the-art single and multi-classifier configurations. In this study, the performance of a single SVM classifier is profiled to demonstrate the compelling reasons to choose ensemble classifier configurations. On the ensemble classifier side, random feature selection (RFS) and configurations involving NURFS and kernel density decision fusion are studied independently. This demonstrates the effect of these two individual techniques on the overall, user, and producer accuracies.

For all of the aforementioned algorithms, the effect of the user and producer accuracies are studied and compared with the proposed approach. In Section 4.3.3, the effect of various kernel density estimation windows are studied for the proposed system. This is to demonstrate the effect of different kernels on the overall classification. All of these experiments are conducted for a different number of training data. This demonstrates the requirement of the amount of training samples for every classifier configuration under study. Training data from 10% to 50 % is of particular interest in all the experiments except for the aquatic species data. For aquatic data, the number of samples is not adequate to perform jack knifing, thus leave *n* out cross validation is used. In particular, the value of *n* varying from one to five is studied. The experiments that use random feature selection are repeated 10 times and the average of the overall, user, and producer accuracies are reported. The deviations of these individual runs are reported as tables. Manhattan bars are used to report the change in the user and producer accuracies and line plots are employed for other studies. The results are reported with 95% confidence intervals marked as error bars.

4.3.1 Comparison of state-of-the-art statistical single classifier techniques.

The experimental results demonstrate mostly a similar trend with LDA-ML showing superior results compared to that of PCA-ML. For the corn stress data, the overall accuracy with 10% of training is around 65% to 70%. With non-overlapping confidence intervals and almost equal number of training samples from each class, it is very clear that this is a problem of small sample size. Kernel-based classifiers such as SVM could handle this dataset better than the statistical ones. Figure 4.6 demonstrates the sensitivity of the number of training samples for PCA-ML and LDA-ML with corn herbicide stress dataset.

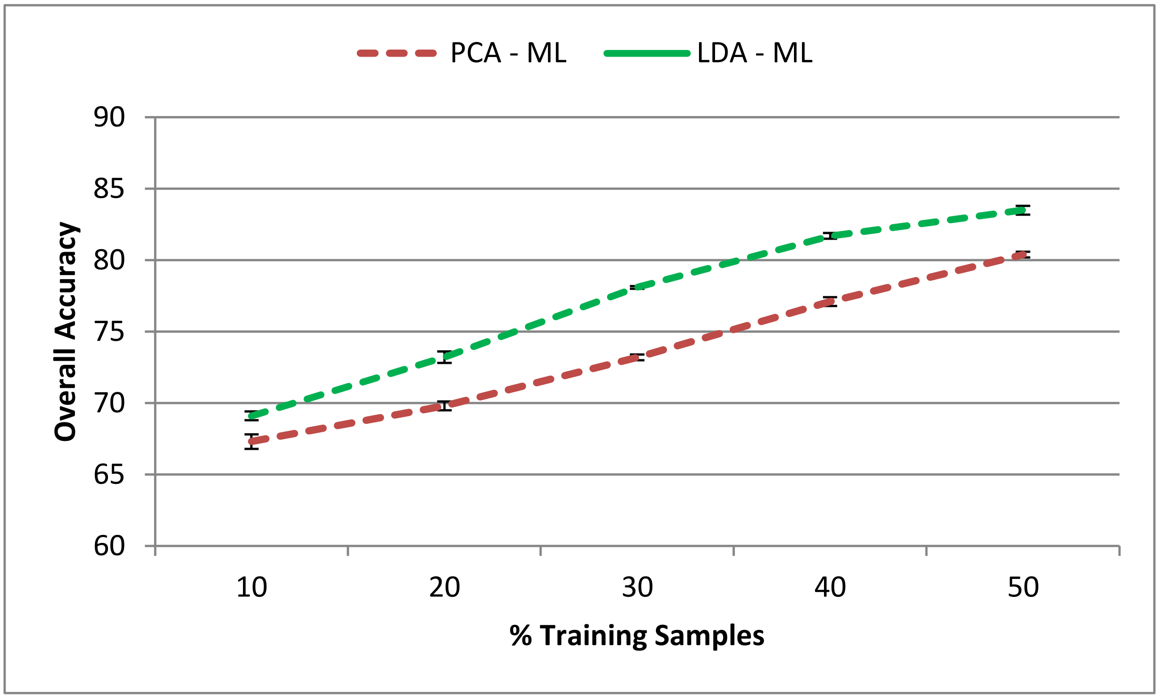

Figure 4.6 Comparison of the state-of-the-art statistical techniques with corn stress data

A similar trend in terms of the overall accuracy is observed with the Indian pines dataset. With 220 features, this AVIRIS data has more number of training samples than others. However, there is an unequal number of training samples from each class. The classes such as alfalfa and oats have a total of 46 and 20 samples respectively. With 10% of training from each class, the class probability densities have to be estimated with trifling number of training samples (4 and 2 respectively). This is clearly not adequate to estimate the densities of the class. As a result, the class accuracies suffer and thus reducing the overall accuracy to the range of 66% to 69%. The accuracies of the classes with a higher number of training samples are better so that the overall accuracy could reach 69%. Figure 4.7 shows the sensitivity of the ML with PCA and LDA feature reduction techniques to the number of training samples from each class.

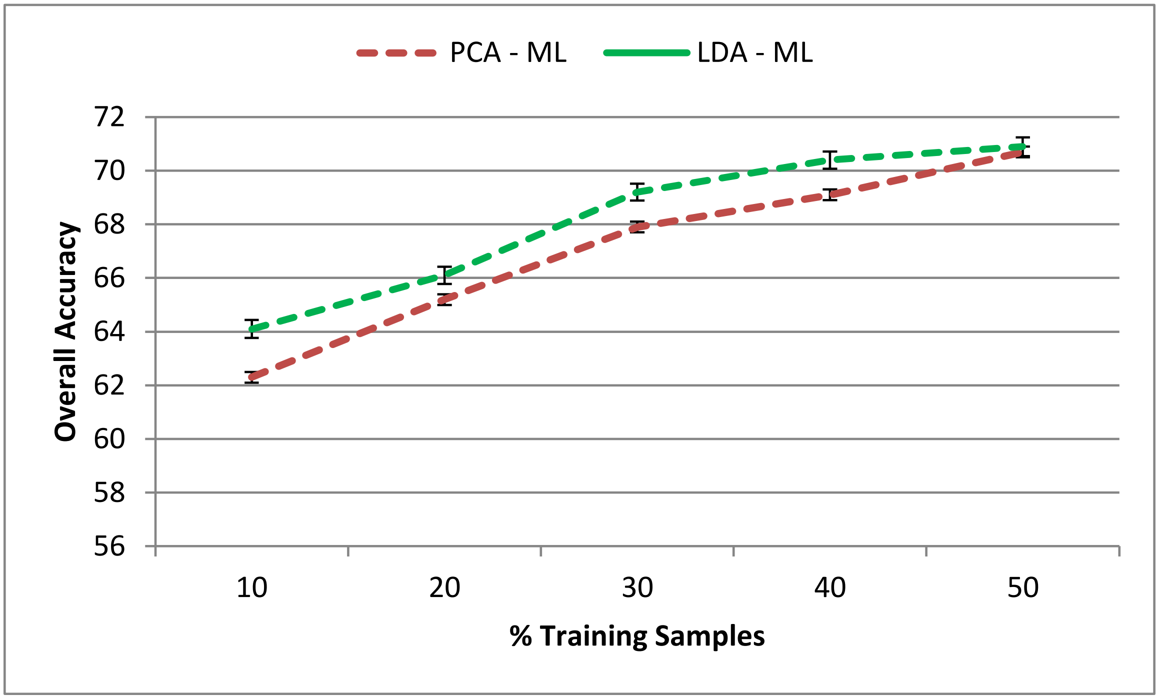

Figure 4.7 Comparison of the state-of-the-art statistical techniques with Indian pines data

With the aquatic species dataset, there are not enough samples from each class to perform jack knifing. Hence, leave *n* out cross validation is used to study the performance. The accuracies are in the range of 72% to 73% with leave-one-out. The leave-one-out process is repeated for many times and the average results are reported in Figure 4.8. The confidence intervals are understandable large with a small number of samples. It is very difficult to conclude a clear winner among PCA-ML and LDA-ML because of the overlapping confidence intervals. However, this forms a clear understanding of the nature of the data and problem.

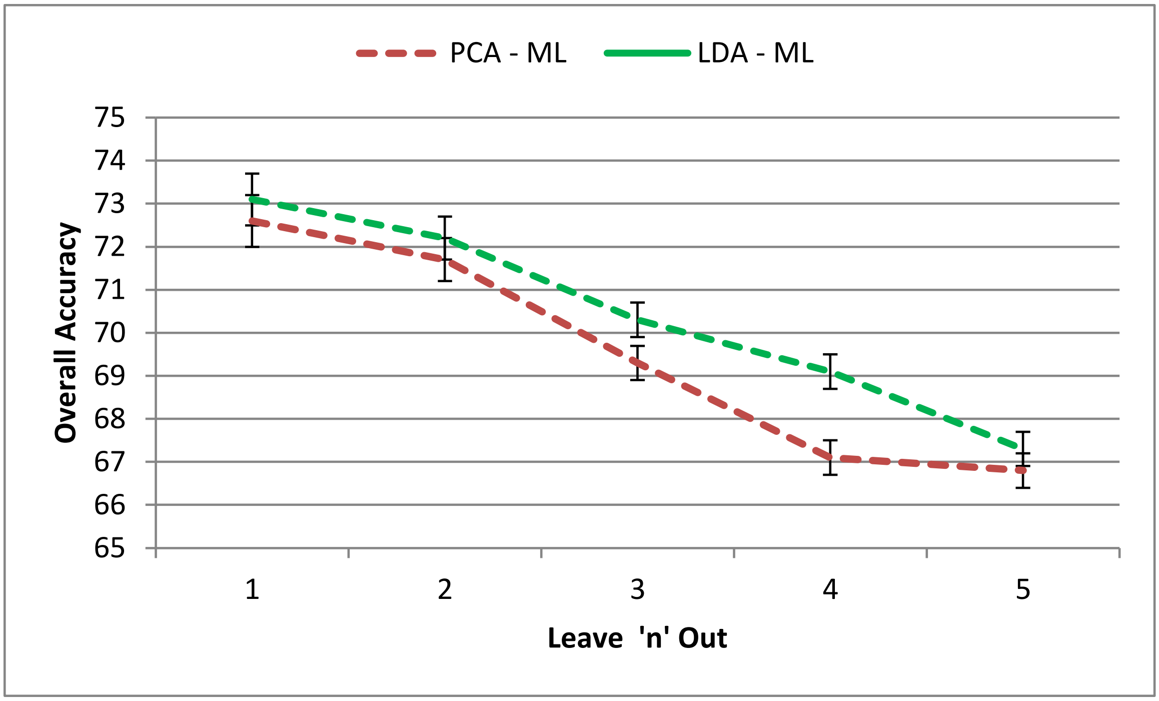

Figure 4.8 Comparison of state-of-the-art statistical techniques with aquatic species data

Unlike the Indian pines data, the ROSIS Pavia data has almost an equal number of samples in every class. This is adequate enough to compute the class densities to a considerable amount of accuracy. This data exhibits an overall accuracy of 72% to 74% with 10 % of training data. This problem is relatively an easier one compared to the

Indian pines data. With Indian pines, the classes are representatives of similar agricultural crops but here it is an urban setup where the class densities are relatively easier to distinguish. This sensitivity study is shown in Figure 4.9.

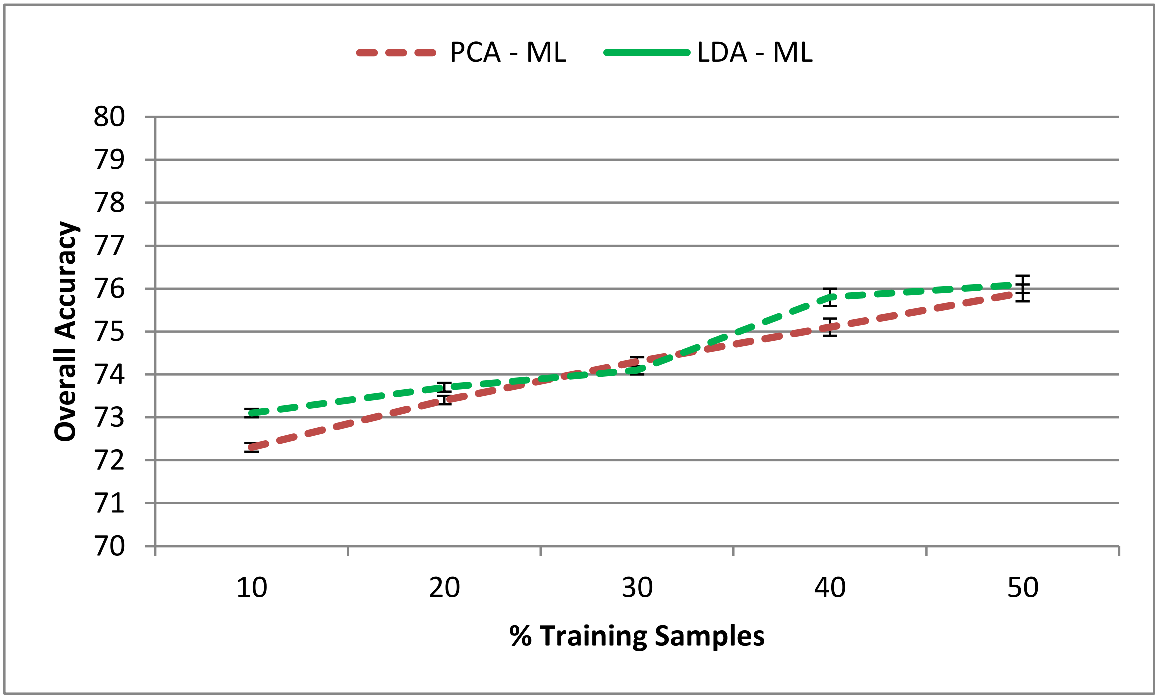

Figure 4.9 Comparison of state-of-the-art statistical techniques with Pavia data

Sensitivity experiments of statistical techniques with corn aflatoxin dataset yield poor results in terms of the overall accuracy. The reason for this is mainly because of the unequal number of training samples and difficulty of the problem.

With the USDA regulations guidance, the levels of toxins are separated as different classes in a non-linear way and this makes some classes with a wide range of aflatoxins. Figure 4.10 shows the study with PCA-ML yielding an overall accuracy of 52% with 10% of training data.

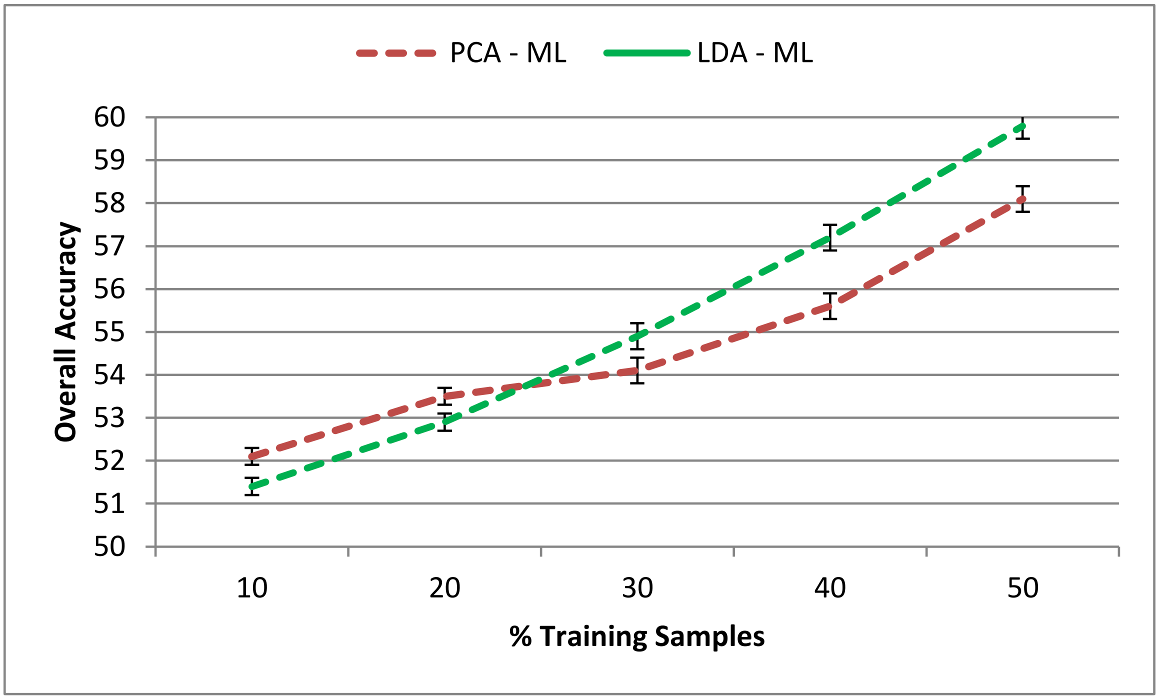

Figure 4.10 Comparison of state-of-the-art statistical techniques with corn aflatoxin data

4.3.2 Comparison of the proposed approach with state-of-the-art single and multi-classifier kernel techniques

The study of the sensitivity of various kernel based classifiers against different training sample sizes reveals an interesting pattern. The proposed approach handles the small sample size situation better than other approaches. Figure 4.11 shows a comparison study of the overall accuracy versus the number of training samples. Systems based on NURFS exhibit a 1.5 to 3% increase in the overall performance. It is worth pointing out that the performance of the kernel scoring NURFS algorithm is above 99% with a sample size of 10%, where the single SVM and original RFS algorithms produce an accuracy of approximately 93% and 95% respectively. Figures 4.12 and 4.13 illustrate the user and producer accuracies for each class in the Corn stress dataset. A similar increase in the user and producer accuracies is observed as with many classes. The standard deviation is shown as error bars for the user and producer accuracies. The deviation is approximately 0.1% for both the user and producer accuracies. Table 4.4 show the details of different classifiers, feature selection and decision fusion techniques that are used in this comparison study. SVM and RFS are included as state of the art techniques, 'Band grouping NURFS' and 'SVM & kernel density scoring' are included in the study to test the efficacy of proposed feature selection and decision fusion techniques respectively. These four techniques are compared against the proposed 'Kernel density NURFS'.

Table 4.4 Classifiers, feature selection and decision fusion techniques used in the comparison study

Classification Algorithm	Feature Selection	Decision Fusion	Classifier Type
SVM	No	N/A	SVM - Single classifier
RFS	RFS	Majority vote	SVM - Ensemble classifier
Band grouping NURFS	NURFS	Majority vote	SVM - Ensemble classifier
SVM & kernel scoring	RFS	Kernel scoring	SVM - Ensemble classifier
Kernel density NURFS	NURFS	Kernel scoring	SVM - Ensemble classifier

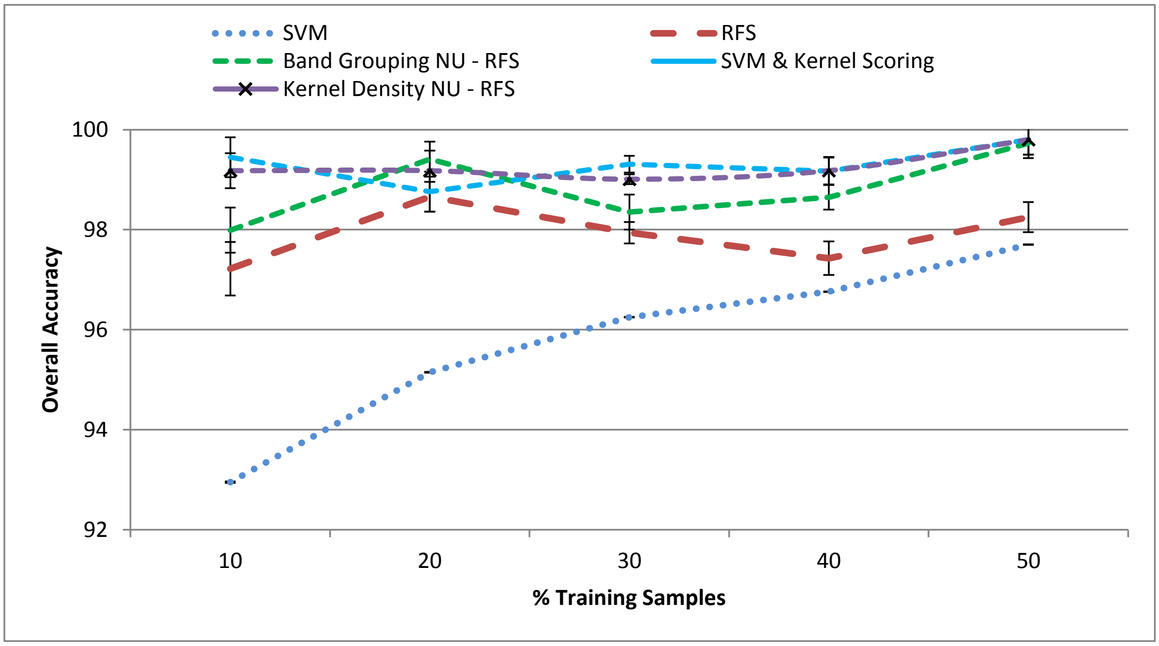

Figure 4.11 Comparison of the proposed approach with its variations and other kernel-based techniques with corn stress data

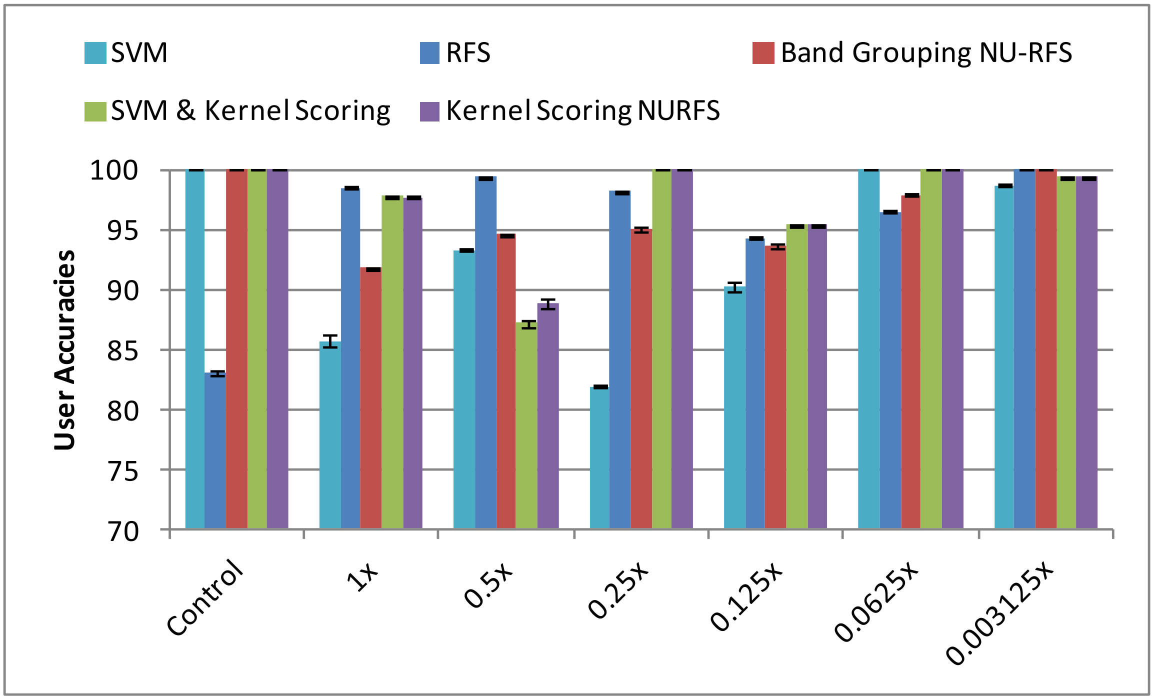

Figure 4.12 User Accuracies of different classes of corn herbicide stress data with various algorithms

For each class, first bar corresponds to SVM, second bar corresponds to RFS, third bar corresponds to band grouping NURFS, fourth bar corresponds to SVM & kernel density scoring and the fifth bar corresponds to kernel scoring NURFS.

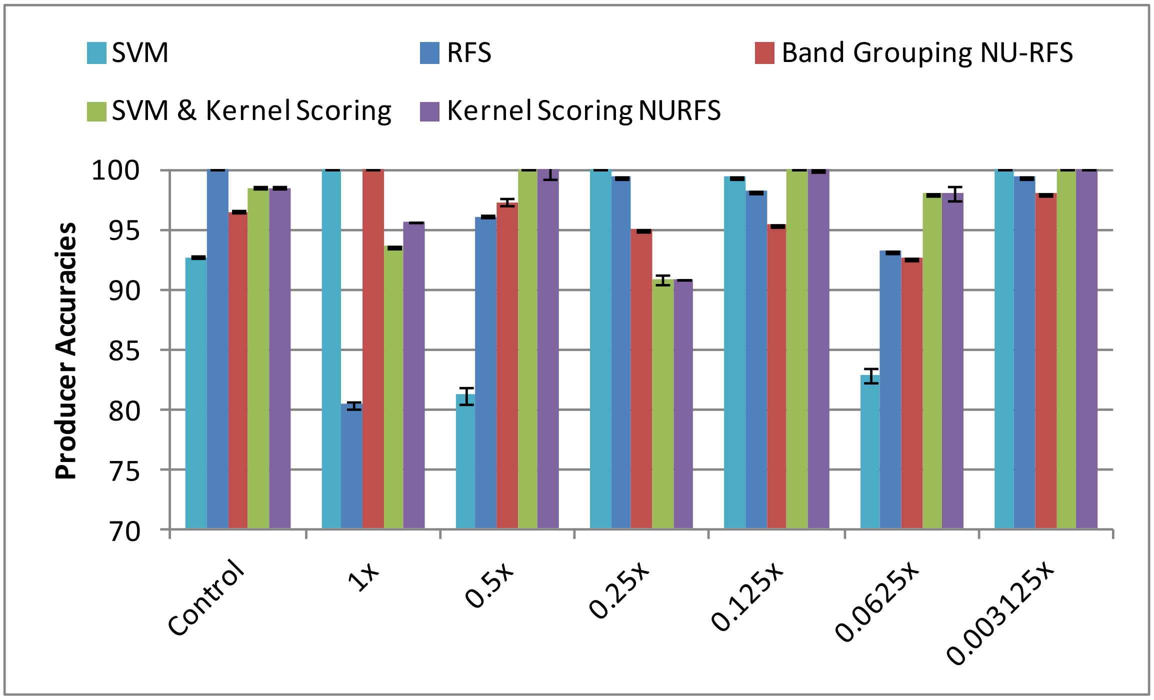

Figure 4.13 Producer Accuracies of different classes of corn herbicide stress data with various algorithms

Both the user accuracies (UA) and producer accuracies (PA) are improved with the proposed feature selection. The overall accuracies of other feature selection approaches and other kernels are also shown.

The SVM based classifier gives the advantage in terms of handling the small sample size. Further improvement is observed with the proposed NURFS approach and kernel density-based decision fusion independently.

When combining both approaches, the overall accuracy is further improved and in very few cases unchanged. In Figure 4.14, a rectangular kernel is used to compute the density.

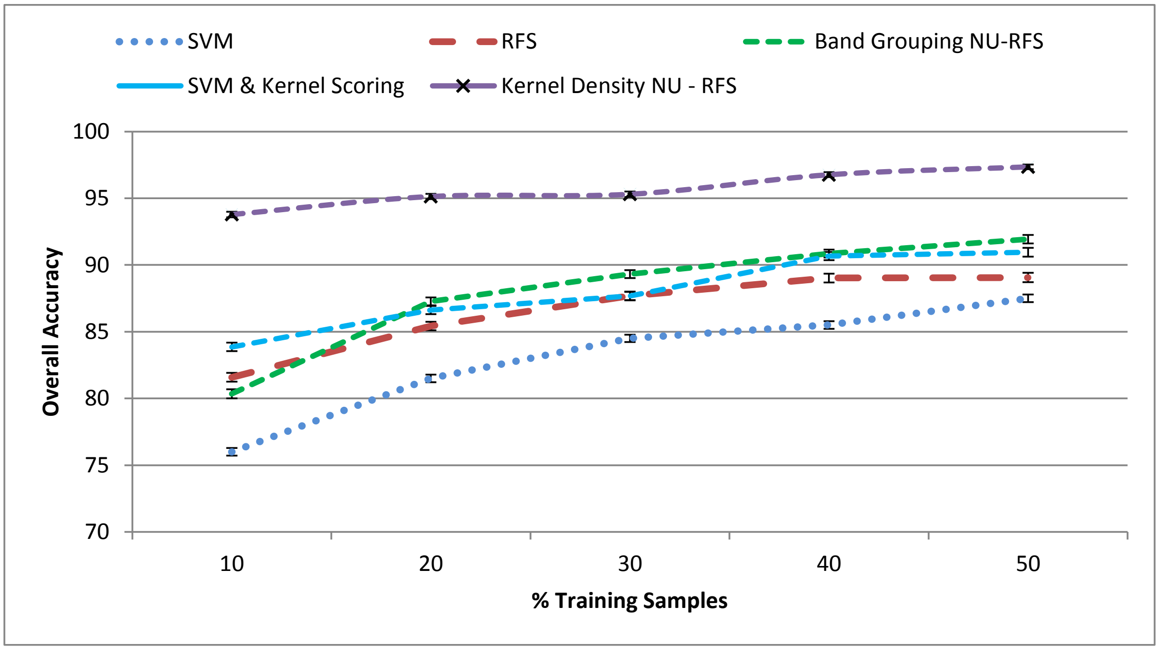

Figure 4.14 Comparison of the proposed approach with its variations and other kernel-based techniques with Indian pines data

With the Indian pines data, experimental results demonstrate the superiority of the proposed approach compared to SVM and RFS. The study of the overall accuracy for various numbers of training samples is shown in Figure 4.14. At 10% training, NURFS with a kernel density-based fusion achieved an overall accuracy of 93.7% with a rectangular kernel, and RFS and NURFS achieve 81.5% and 80.3% respectively. Interestingly, SVM with kernel scoring performs better than RFS and Band grouping-based NURFS. The proposed kernel scoring based NURFS outperforms other approaches by 10%. The maximum overall accuracy achieved is 97.3% with 50% training. For all the experiments, an ensemble size of $z=10$ is used. It is observed that increasing the ensemble size does not provide any significant improvement beyond 8. This is similar to an observation made by Waske *et al.* [8] when using a simple RFS.

Figures 4.15 and 4.16 illustrate the user accuracies (UA) and producer accuracies (PA) for each class of the Indian Pines dataset. It is observed that consistent improvement (~ 2-40%) in both the user and producer accuracies throughout all classes occur when employing the proposed kernel density-based scoring approach. This is expected as the confusion between the classes is reduced via the proposed scoring approach. The standard deviation is shown as error bars for the user and producer accuracies. The deviation is approximately 0.8% for both the user and producer accuracies.

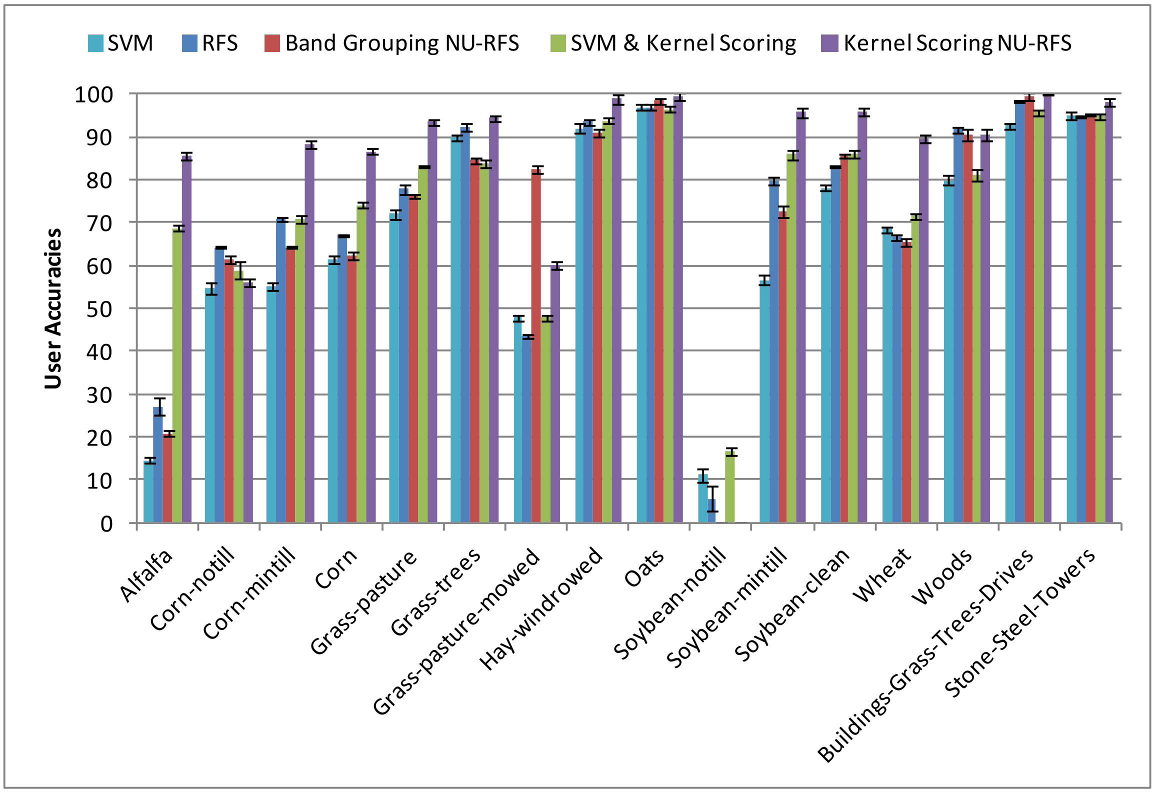

Figure 4.15 User Accuracies of different classes of Indian pines data with various algorithms

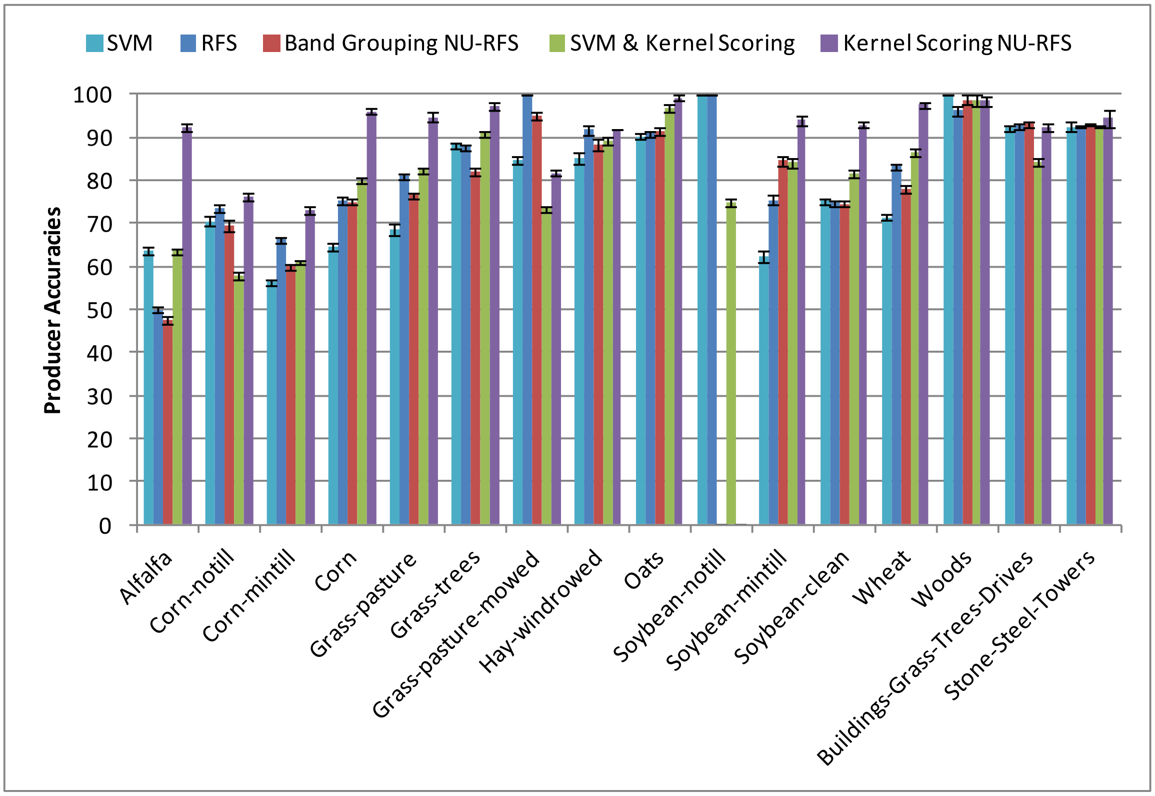

Figure 4.16 Producer Accuracies of different classes of Indian pines data with various algorithms

As already mentioned, the aquatic species dataset has a smaller number of features per class compared to other datasets. The small sample size is obviously better handled by SVM. The proposed approach has an increase in the accuracy by 4% for leave one out. The gain of NURFS and kernel density decision fusion is not significant when considered independently. Figure 4.17 shows the comparison of various SVM based methods with the proposed approach. The accuracies reported in Figure 4.17 are arrived after averaging the accuracy of every leave *n* out run.

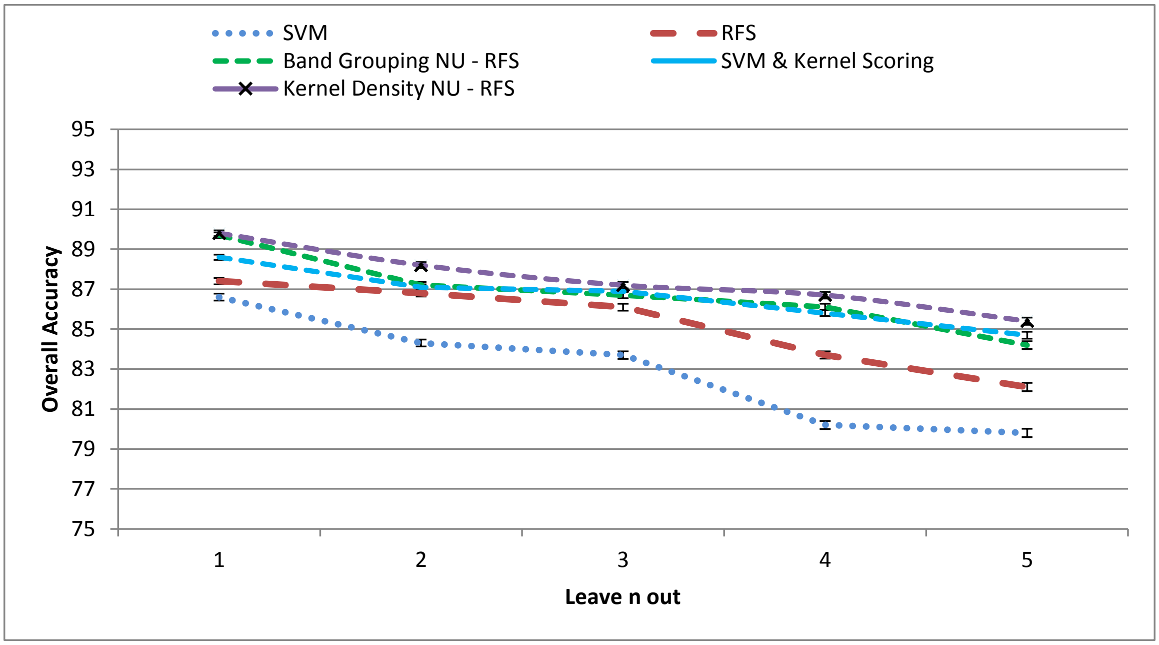

Figure 4.17 Comparison of the proposed approach with its variations and other kernel-based techniques with aquatic species data

The experimental results with the Pavia, Italy dataset show an improvement in the overall classification accuracy compared to other algorithms. Figure 4.18 shows the performance of the proposed approach under various percentage of the number of training samples. Kernel density-based NURFS achieves a gain of 7% and also performs well under limited training samples. Both Kernel density-based approaches combined with SVM and NURFS show superior performance over all the other approaches, and this shows the effectiveness of the proposed decision fusion approach.

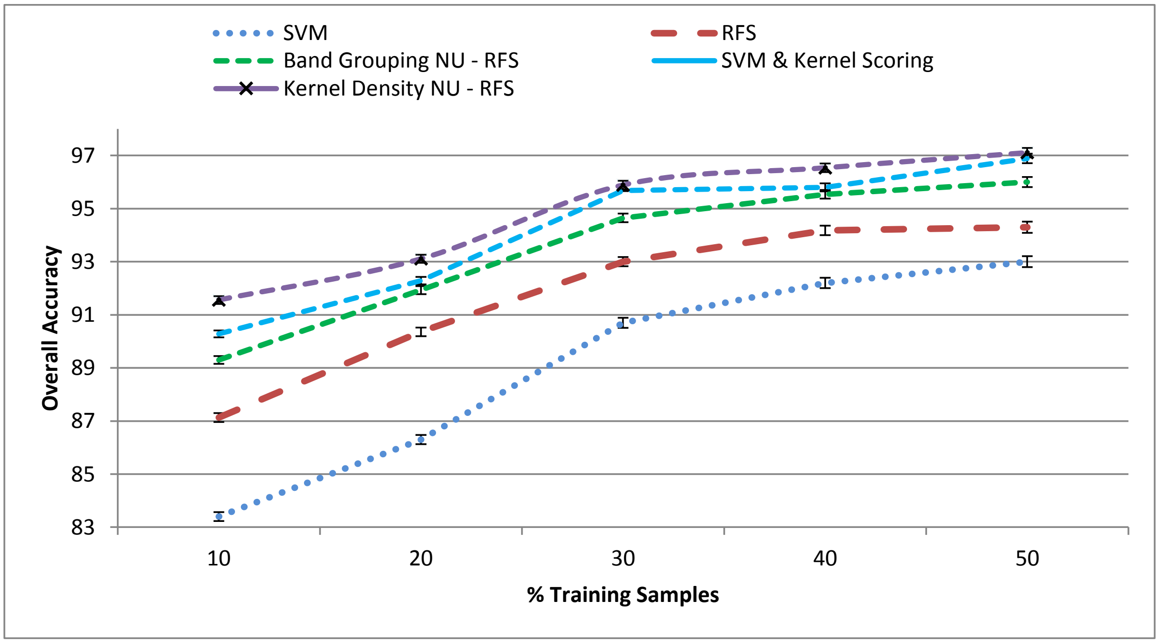

Figure 4.18 Comparison of the proposed approach with its variations and other kernel-based techniques with pavia data

Figures 4.19 and 4.20 show the user and producer accuracies for each class of the Pavia dataset. Water, trees, bitumen, tiles, and bare soil classes gained an improvement of 1 to 5%. This improvement can be seen from other kernel scoring techniques without feature selection. The standard deviation is shown as error bars for the user and producer accuracies. The deviation is approximately 0.2% for both the user and producer accuracies.

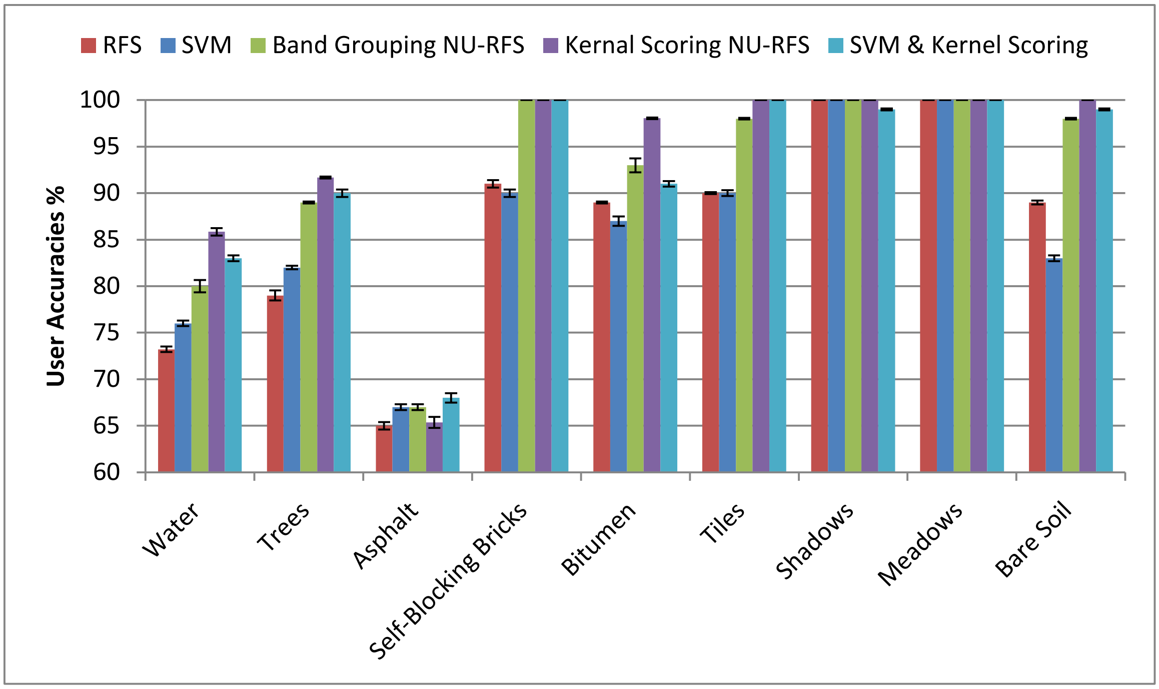

Figure 4.19 User accuracies of different classes of pavia data with various algorithms

For each class, first bar corresponds to RFS, second bar corresponds to SVM, third bar corresponds to band grouping NURFS, fourth bar corresponds to kernel scoring NURFS and the fifth bar corresponds to SVM & kernel density scoring.

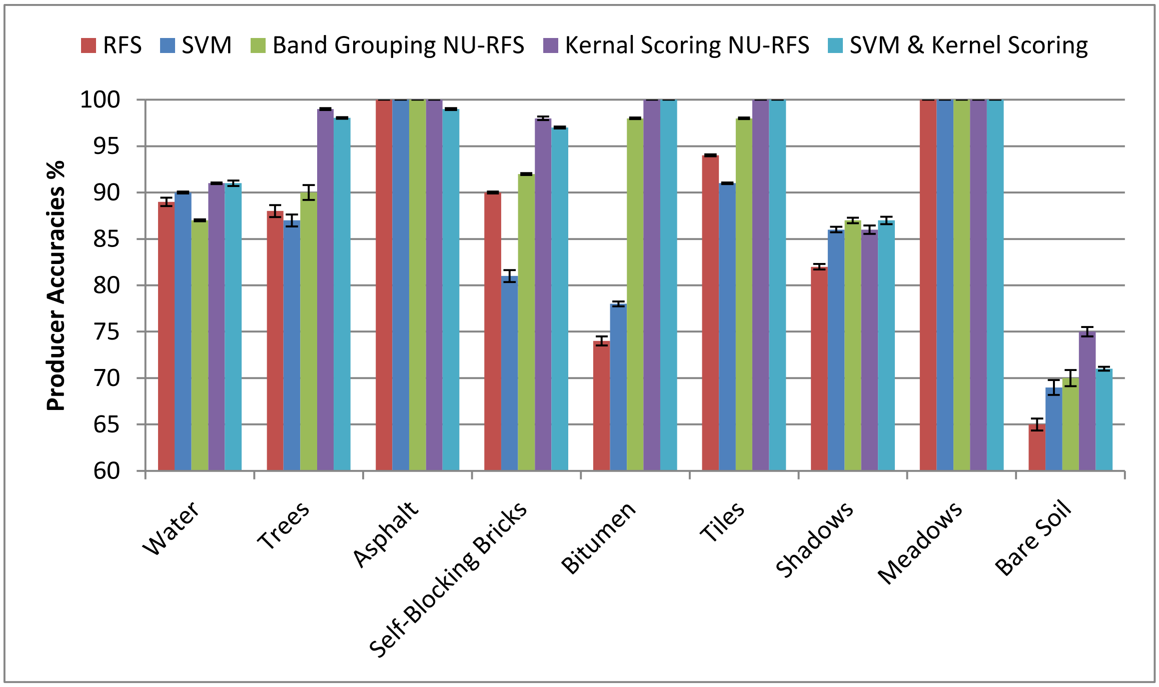

Figure 4.20 Producer accuracies of different classes of pavia data with various algorithms

The most challenging among all the datasets is the corn aflatoxin. Because of the reasons already discussed, statistical algorithms perform very poorly with corn aflatoxin data. The overall accuracy and user and producer accuracies are greatly improved with SVM classification. With the proposed approach, the overall accuracy is close to 75% with 10% of training data, which corresponds to approximately an increase of 6% from the single SVM. NURFS and kernel density decision fusion have good impact on this particular dataset with improved accuracies for all the classes. The sensitivity to different training size is shown in Figure 4.21. The improvements in the user and producer accuracies are presented in Figures 4.22 and 4.33 respectively.

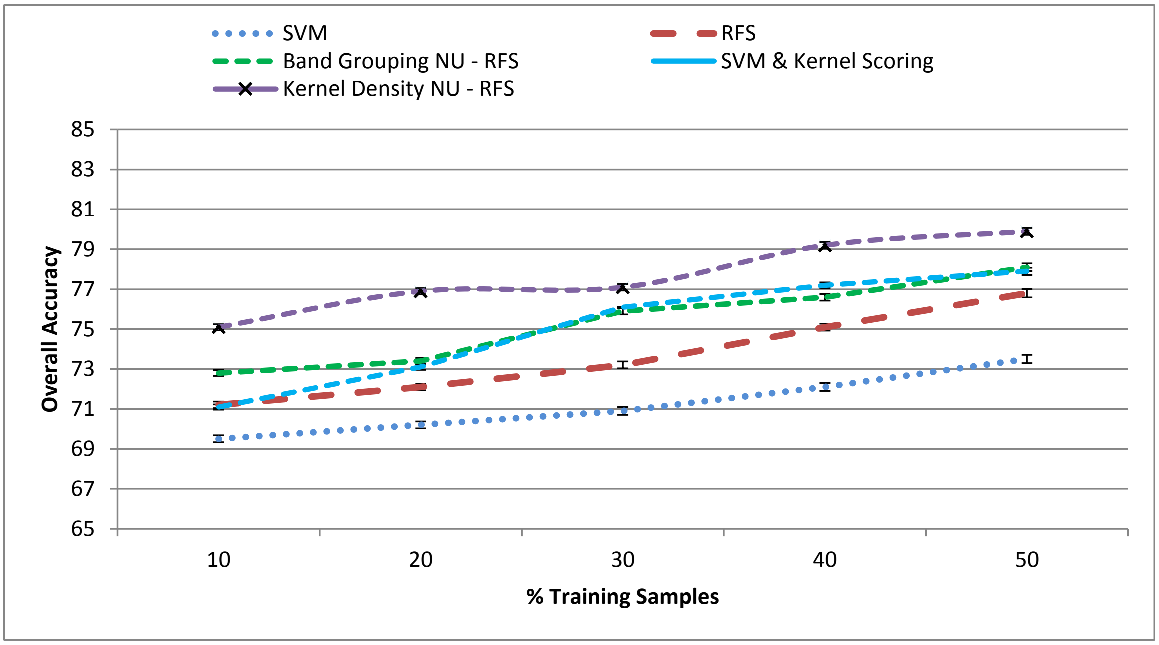

Figure 4.21 Comparison of the proposed approach with its variations and other kernel-based techniques with corn aflatoxin data

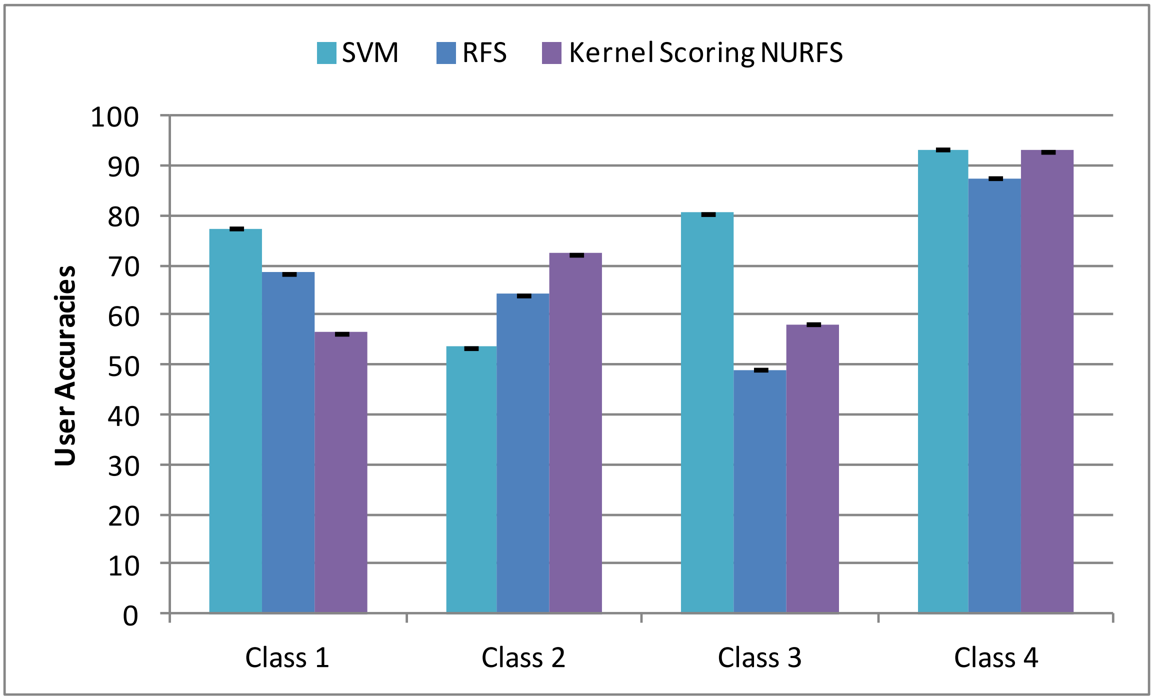

Figure 4.22 User accuracies of different classes of corn aflatoxin data with various algorithms

For each class, first bar corresponds to SVM, second bar corresponds to RFS and the third bar corresponds to SVM & kernel density scoring.

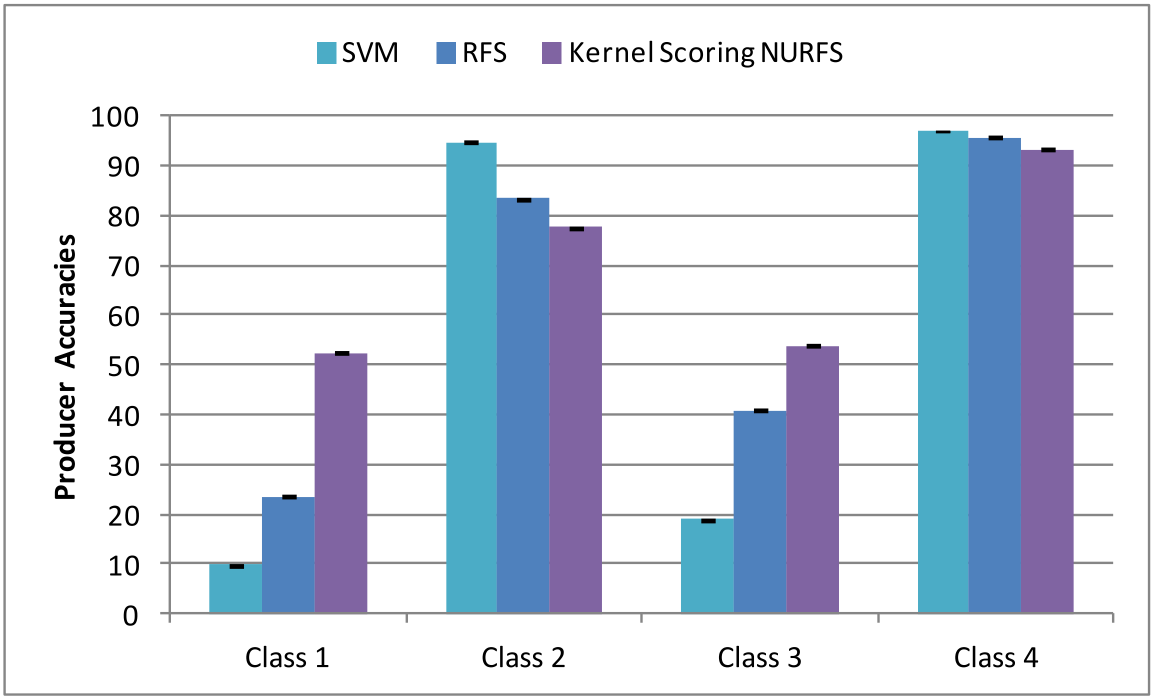

Figure 4.23 Producer accuracies of different classes of corn aflatoxin data with various algorithms

An increase in the class producer accuracies can be observed in Figure 4.23 in class 1 and class 3. There is an increase of 30% for class 1 and approximately 14% for class 3. A better performance is clearly achieved with NURFS and kernel density decision fusion combination.

4.3.2.1 Impacts of NURFS and Kernel Density based decision fusion on the MCS

The MCS proposed in Figure 3.5 has two different techniques to improve the efficacy of RFS. First, a NURFS system is proposed where the features are selected in a non-uniform fashion. This is depicted in Figure 3.3 and described in Section 3.2. Second, a decision fusion scheme based on kernel density estimation is described in Section 3.3. With various datasets described in Section 4.1, the experiments are conducted to study the performance of the proposed approach compared to state-of-the-art techniques. It is interesting to study the impact of the proposed NURFS and decision fusion schemes when they are considered individually. The experimental results presented in Section 4.3 demonstrate the above discussed impacts. In those plots, band grouping NURFS represents the effect of NURFS alone and SVM and kernel scoring represent the effect of the proposed decision fusion scheme alone in terms of the overall, user, and class accuracies.

With corn stress data, the performance of NURFS has an improvement of about 4% and the proposed decision fusion scheme gives an improvement of 5% with 10% training data. The combination of both techniques offers a little more than 5% increase in the overall accuracy. It is also interesting to note that the user and producer accuracies of individual classes are improved for most of the cases consistently with the proposed approach. These amounts of improvement with these techniques depend on two factors, 1) the amount of diversity in NURFS features and 2) the quality of kernel density estimation. Particularly, with decision fusion, the knowledge of the training data separability drives the decision of the final class label. So, the amount of available training data and its generalization capability play a vital role. This is evident from the results with the Indian pines and Pavia datasets. With almost an equal number of samples, the aquatic species dataset exhibits similar user and producer accuracies for all the classes (the user and producer accuracies are not reported since the accuracies are almost the same for each class).

Since the proposed feature selection approach is random in nature, each experiment is repeated for 10 times and the mean overall accuracy is reported. Tables 4.5 and 4.6 show the mean and standard deviation of these runs for different datasets.

Table 4.5 Mean and standard deviation of the NURFS classifier overall accuracy

Training Sample Size in %	Corn herbicide stress		Indian pines		Pavia		Corn Aflatoxin	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
10	96.95	0.2038	92.88	0.7206	91.71	0.5282	75.23	0.7242
20	97.36	0.0994	94.84	1.5072	93.39	0.3673	76.50	0.4389
30	99.53	0.1299	95.17	0.5217	95.52	0.3903	76.93	0.5696
40	99.14	0.2408	96.07	0.5186	96.31	0.2979	79.07	0.2540
50	99.11	0.1152	97.24	0.3213	96.94	0.3809	78.92	0.7346

Table 4.6 Mean and standard deviation of the NURFS classifier overall accuracy (aquatic dataset)

Number of samples for testing	1	2	3	4	5
Mean	88.69	88.47	87.12	85.72	84.88
Standard Deviation	1.0928	0.3335	0.4732	0.5181	0.6321

4.3.3 Sensitivity to density estimation kernels

Four kernel window functions are used in this study to determine their effect on the proposed approach. Although there is not much of a difference in class accuracies (the class accuracies are not reported here as there is no significant improvement), the overall accuracy in some cases increased and/or decreased by 3% for different kernel functions. This is particularly significant for Indian pines and aquatic species datasets. Figures 4.24 to 4.28 demonstrate the effect of different kernels on the overall accuracy with different training sample sizes. The kernel corresponding to the best overall accuracy at 10% training is reported in Figures 4.11, 4.14, 4.17, 4.18, and 4.21.

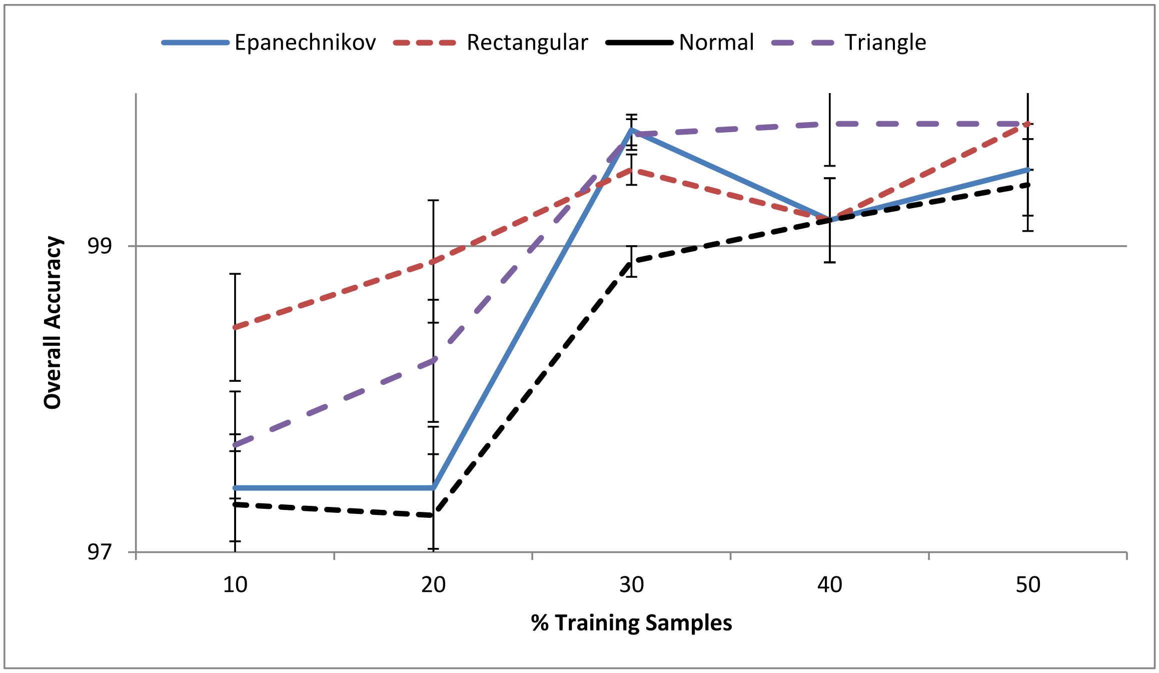

Figure 4.24 Sensitivity of the proposed approach with different kernel functions to the number of training samples with Corn Stress data

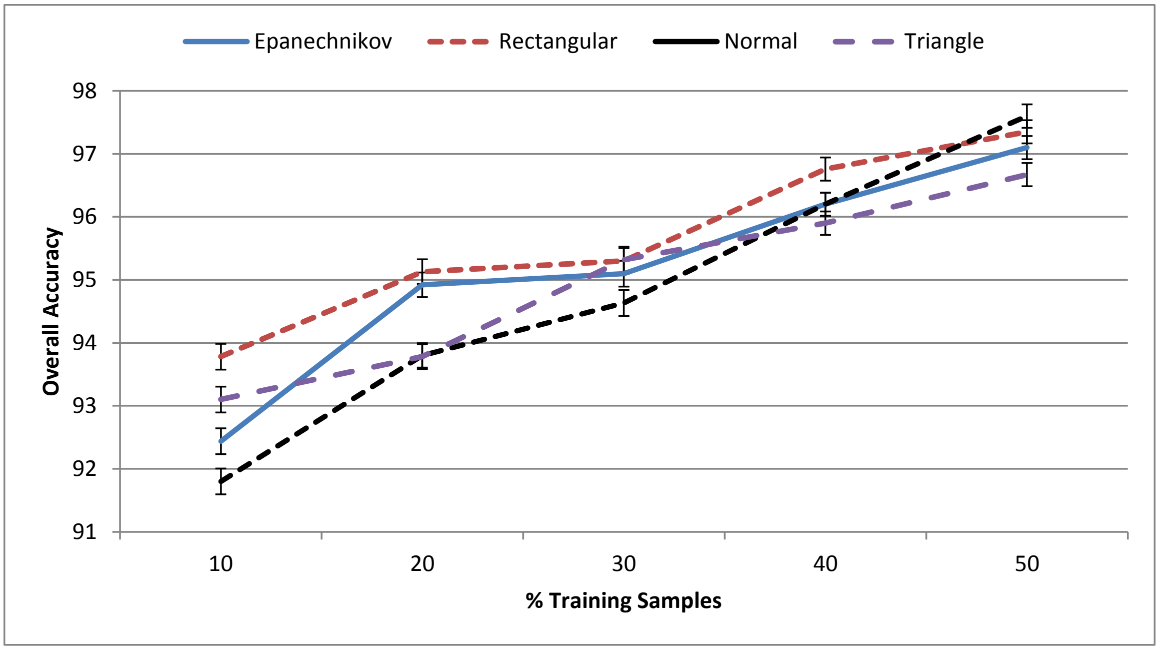

Figure 4.25 Sensitivity of the proposed approach with different kernel functions to the number of training samples with Indian pines data

The study of the performance of various algorithms with respect to the sample size follows a very interesting trend. The proposed approach clearly outperforms other techniques. Figure 4.25 shows the comparison of various kernel functions with respect to the number of training samples used for training. In this case, the rectangular and triangular kernels perform almost equally well compared to normal and Epanechikov. The error bars shown correspond to 95% confidence intervals. From experimentation, it is found that the performance of the classifier increases with an increase in the number of band groups, *m* initially, and it decreases after reaching a particular value. In order to maintain uniformity among various experiments, the value of *m=4* is used, this yields the best performance for all the datasets. From experimentation, it is found that the performance of the classifier increases with increasing *m* initially and it decreases after reaching a particular value.

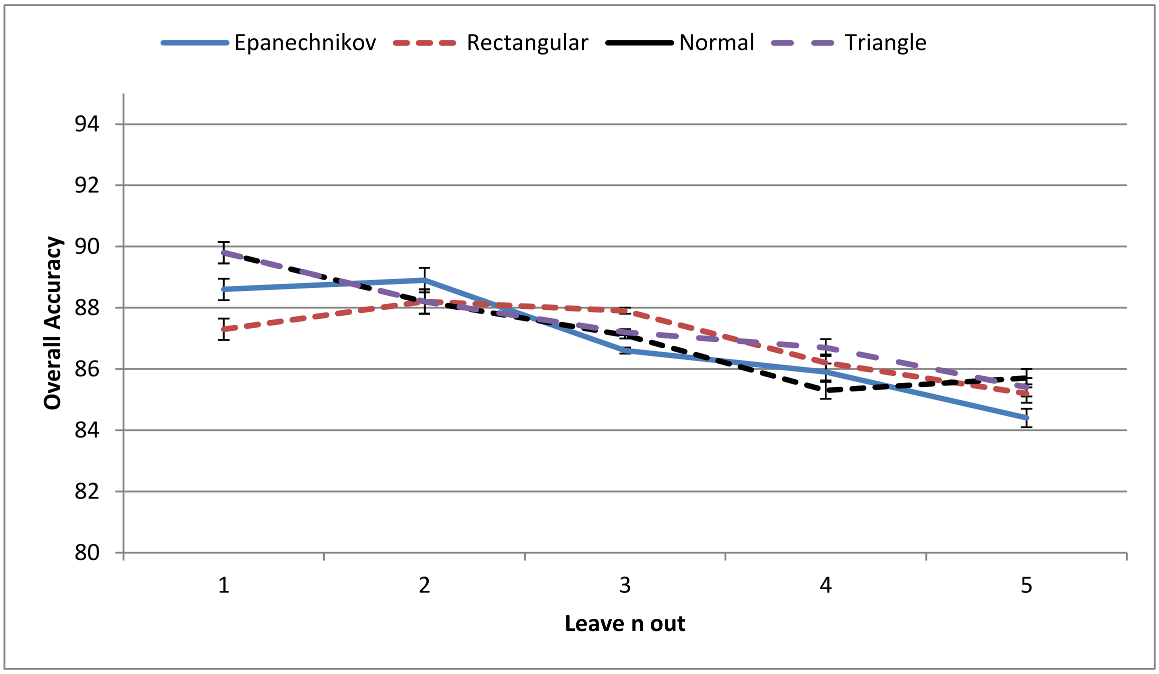

Figure 4.26 Sensitivity of the proposed approach with different kernel functions to the number of training samples with aquatic species data

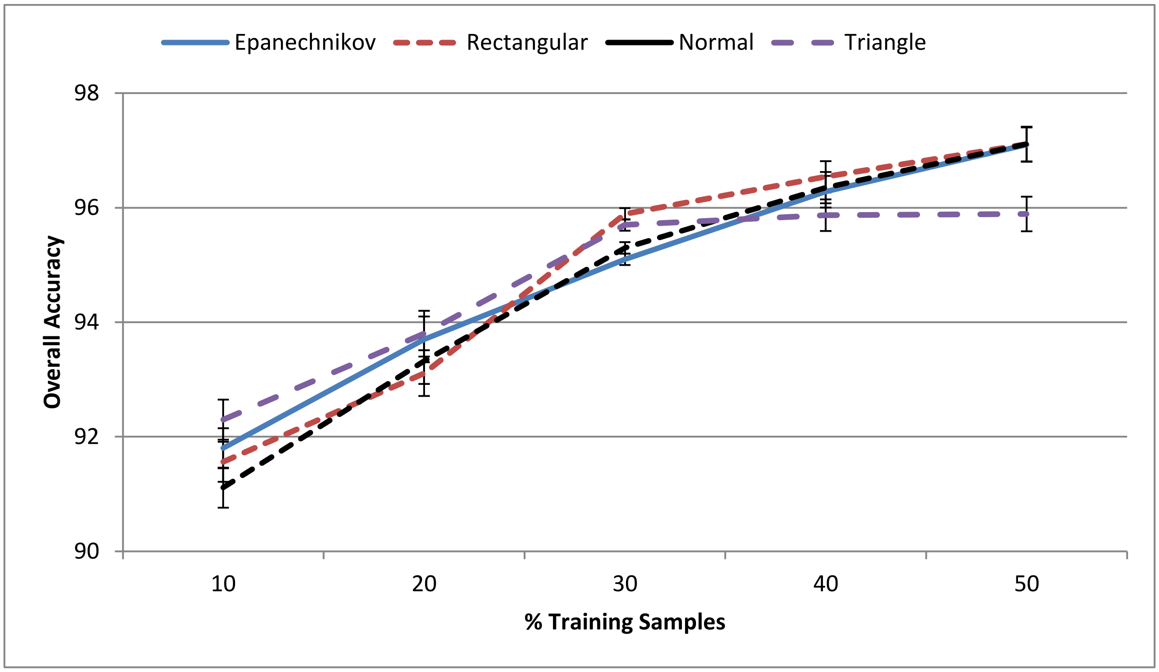

Figure 4.27 Sensitivity of the proposed approach with different kernel functions to the number of training samples with pavia data

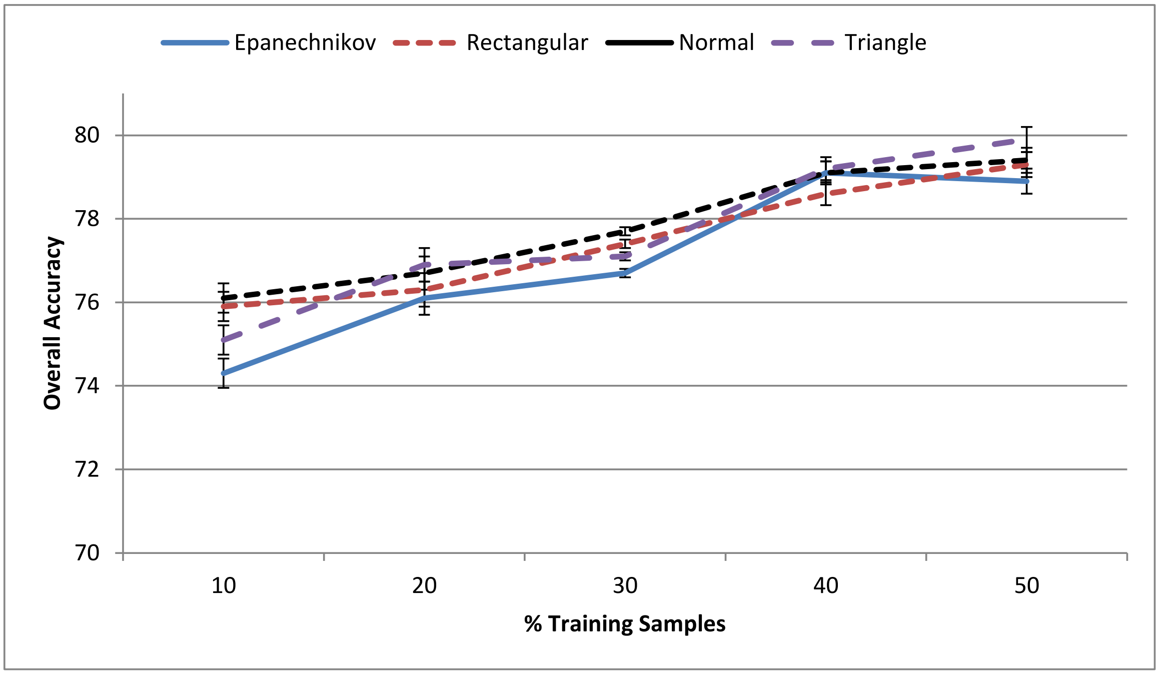

Figure 4.28 Sensitivity of the proposed approach with different kernel functions to the number of training samples with corn aflatoxin data

4.4 References

[1] Purdue University, “Indian Pines AVIRIS Data Set,” 1992. [Online]. Available: https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html. [Accessed: 01-May-2014].

[2] NASA, “HyspIRI Mission Study.” [Online]. Available: http://hyspiri.jpl.nasa.gov/. [Accessed: 01-May-2014].

[3] ASD, “Spectral Remote Sensing for Hyperspectral and Multispectral Imagery Analysis.” [Online]. Available: http://www.asdi.com/applications/remote-sensing/spectral-remote-sensing. [Accessed: 01-May-2014].

[4] DLR, “ROSIS Hyperspectral Data Product.” [Online]. Available: http://messtec.dlr.de/en/technology/dlr-remote-sensing-technology-institute/hyperspectral-systems-airborne-rosis-hyspex/index.php. [Accessed: 01-May-2014].

[5] “Hyperspectral Remote Sensing Scenes.” [Online]. Available: http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. [Accessed: 01-May-2014].

[6] M. A. Lee, S. Prasad, L. M. Bruce, T. R. West, D. Reynolds, T. Irby, and H. Kalluri, “Sensitivity of hyperspectral classification algorithms to training sample size,” in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS ’09. First Workshop on*, 2009, pp. 1–4.

[7] Z. H. H Yao R Kincaid, R.L Brown, T.E. Cleveland, D. Bhatnagar, “Correlation and Classification of Single Kernel Fluorescence Hyperspectral Data with Aflatoxin Concentration in Corn Kernels Inoculated with Aspergillus flavus Spores,” *J. Food Addit. Contam.*, vol. 27(5), pp. 701–709, 2010.

[8] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2880–2889, 2010.

CHAPTER V

ROLE OF DIVERSITY IN ENSEMBLE CLASSIFICATION OF HYPERSPECTRAL DATA

5.1 Introduction

For pattern classification problems, employing more than one classifier is well established and a widely used technique for improving the performance. This technique is often referred to multi classifier systems (MCS), ensemble classification, and mixture of experts or committees of learners. The development of ensemble classifiers for HSI data is discussed in Section 2.3 of this dissertation. The most important reason for the success of ensemble classification is the diversity of classifiers or learners. For example, an ensemble consists of two similar classifiers producing the same classification label for every sample has no serious benefit over a single classifier. Methods for achieving this diversity can be broadly grouped into two categories: 1) an ensemble with different classifiers and same training data, and 2) an ensemble with the same classifier but different subsets of training data. The focus of this dissertation is on the second category. Subset creation methods based on bagging, boosting, and its variants are shown to perform well in terms of the overall classification accuracy [1]–[4]. However with HSI data, the scope of these techniques are shown to be rather limited [5], [6]. Since the HSI data often has more features or spectral bands than the number of training samples, having an ensemble with the same classifier and different spectral subsets of features is of obvious interest.

The MCS proposed in [7] is shown to have the diverse subset of features through random feature selection (RFS). In [8], a multi classifier decision fusion system is proposed with different maximum likelihood classifiers learning different part of HSI bands and shown to perform very well when compared to other statistical techniques. In this approach, the diversity is achieved by having totally different bands fed to each classifier. Every ensemble classification technique described in [9] achieves this diversity by one of the above mentioned techniques. Krogh *et al.* [10] recognize diversity as an inevitable characteristic of an ensemble neural network classifier. Cunningham *et al.* [11] demonstrate the importance of monitoring diversity in ensemble classifiers by using entropy measures. Theoretical issues, implementation of classifier combinations, and diversity are discussed by Lam [12]. So, it is very clear that diversity plays a vital role in any ensemble classification system.

In spite of recognizing the importance of diversity, there is no clear definition for this term. There are some simple techniques for measuring diversity such as a Hamming distance between two classifiers. The Hamming distance is directly proportional to the disagreement between any two vectors. This could be useful in some simple cases and with only two classifiers. However, there is no clear way to measure the diversity for MCS with more than two classifiers. There are some pairwise dissimilarity measures described in the literature and these are shown to be useful in coarsely estimating the diversity [13]. Among many statistics described in [13], the simplest ones were chosen to study the diversity between the proposed NURFS and RFS. Yule’s Q-statistics [14], correlation coefficient [14][15], disagreement measure [16], [17], and double fault measure [18] are popularly used for measuring the diversity of a pair of classifiers. The overall diversity can be computed by averaging. In this chapter, various pair wise diversity measures are employed to study and compare the ability of the RFS and proposed Non-Uniform RFS methods to produce diverse HSI subsets.

5.2 Measures of Diversity

As mentioned earlier, there is no clear definition for diversity in the literature. Various researchers use statistical dissimilarity or similarity measures to estimate this diversity. To define these statistics, let us consider the super class Ω described in Section 3.2 that consists of labeled training for training the classifier. The output of any classifier in ensemble Z_i can be represented as a vector of length N, where the size of Ω is represented as N, which is given by Equation 5.1

$$N = \sum_{i=1}^{n} N_i \qquad (5.1)$$

For each sample that is recognized correctly by a classifier Z_i , a vote will be added to the matrix shown in Table 5.1 and no vote will be added for incorrect classification. This is represented in Table 5.1 for any pair of classifiers Z_i and Z_k.

Table 5.1 Vote table for pair wise ensemble diversity computation

	Z_k correct (vote = +1)	Z_k incorrect (no vote)
Z_i correct (vote = +1)	N^{11}	N^{10}
Z_i incorrect (no vote)	N^{01}	N^{00}

The total number of training samples *N* is given by,

$$N = N^{11} + N^{10} + N^{01} + N^{00} \tag{5.2}$$

Yule's Q statistic (*Q*) can be defined for two classifiers Z_i and Z_k as

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{5.3}$$

The lower the value of *Q*, the greater the diversity between classifiers. The range for *Q* is from -1 to 1. The correlation coefficient (*ρ*) for any two classifiers is defined as,

$$\rho_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}} \tag{5.4}$$

The lower the value of *ρ,* the greater the diversity between classifiers. The disagreement measure (*DIS*) used in [16], [17] for any two classifiers is defined as

$$DIS_{i,k} = \frac{N^{01}+N^{10}}{N} \tag{5.5}$$

This measure is an estimate of the ration of the number of disagreements to the total number of samples. With DIS, the higher the disagreements between the two classifiers is, the greater the diversity.

The double fault measure (*DF*) used in [18] for any two classifiers is defined as

$$DF_{i,k} = \frac{N^{00}}{N} \tag{5.6}$$

Equations 5.3 to 5.6 provide pair wise measures and for *Z* classifiers in the ensemble ($Z \geq 3$), the average measure can be calculated by Equation 5.7

$$Measure_{average} = \frac{2}{Z(Z-1)} \sum_{i=1}^{Z-1} \sum_{k=i+1}^{Z} Measure_{i,k} \tag{5.7}$$

5.3 Experimental Results

To investigate the effect of RFS and NURFS feature selection techniques on the diversity of individual classifiers, $Q_{average}$, $\rho_{average}$, *and* $DIS_{average}$ are computed for ensembles of size 10. The corn herbicide stress dataset and Indian pines dataset are considered for this study with different training sample sizes. It is found that the *Q* statistic is very consistent for almost all the cases. With the corn and Indian pines data having 7 and 16 classes respectively, it is very difficult to achieve the best diversity of -1 as the higher the classes are, it will be more difficult to achieve best diversity [13]. Except for two cases, the diversity exhibited by the proposed NURFS is better than the regular RFS algorithm. Among the 5 experiments conducted on each data, the average of *Q* across these experiments is approximately -0.6 for the corn herbicide data and -0.5 for the Indian pines data. These values are slightly higher than that of the regular RFS. This is demonstrated in Figures 5.1 and 5.2.

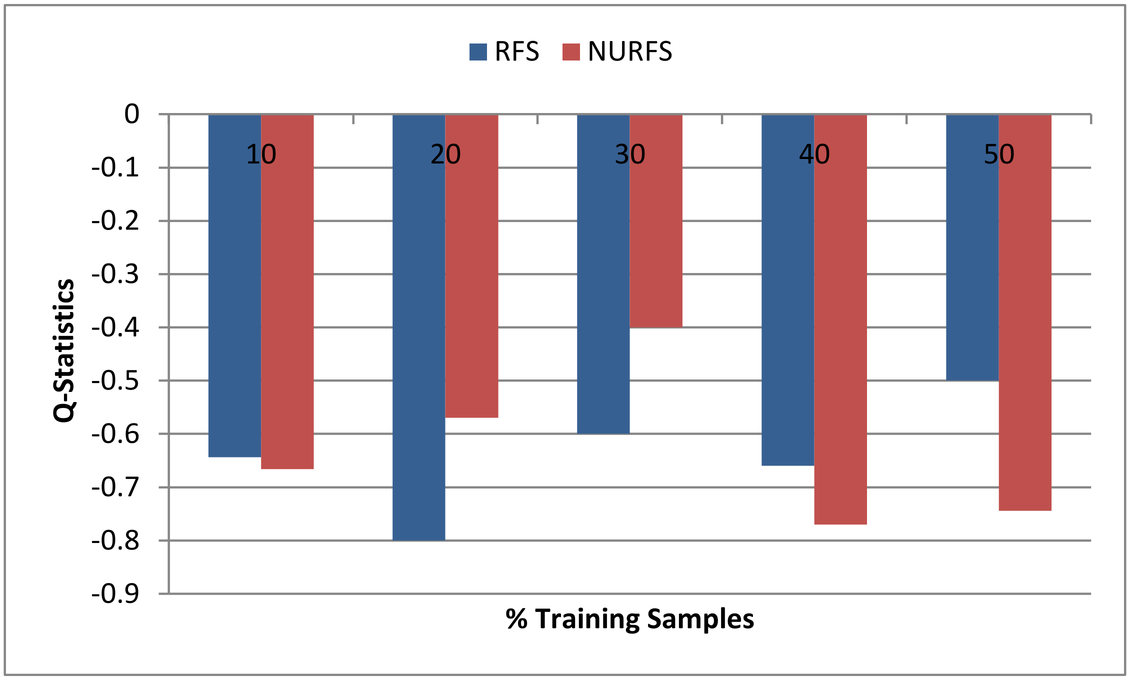

Figure 5.1 Q-Statistics measure for the corn herbicide data

Lower score indicates greater diversity

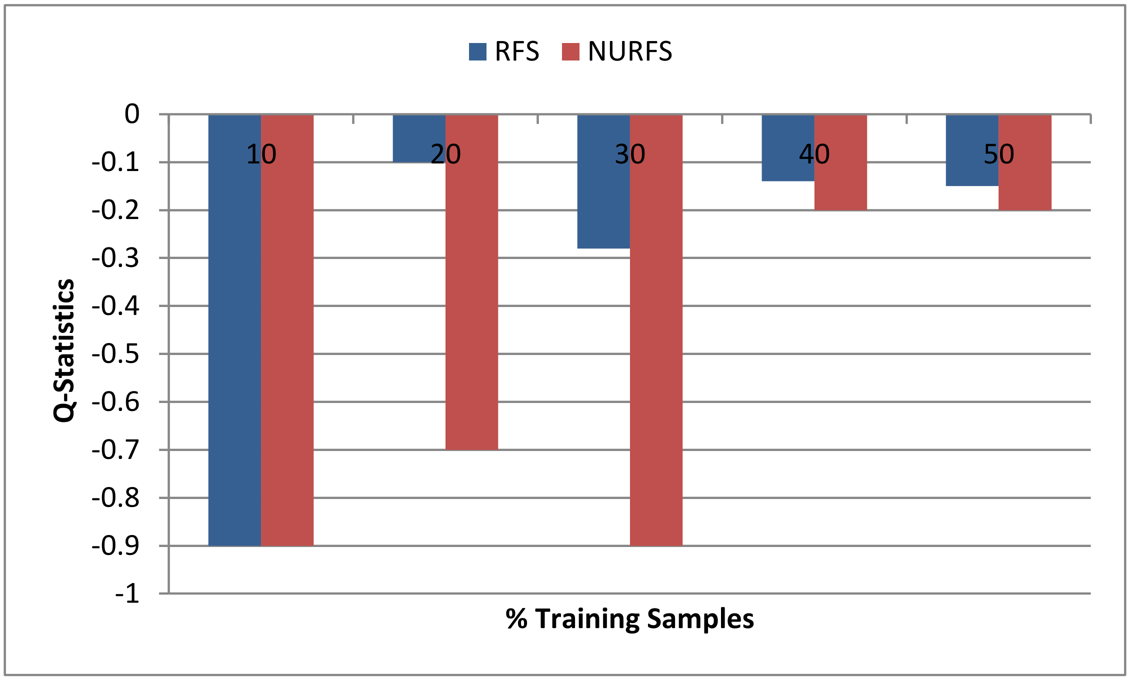

Figure 5.2 Q-Statistics measure for the Indian pines data

Lower score indicates greater diversity

Figures 5.3 and 5.4 show the plots of the correlation coefficient measure for the corn herbicide and Indian pines data. The results show that the NURFS algorithm provides better diversity than that of RFS. The results are mostly consistent for almost all the experiments.

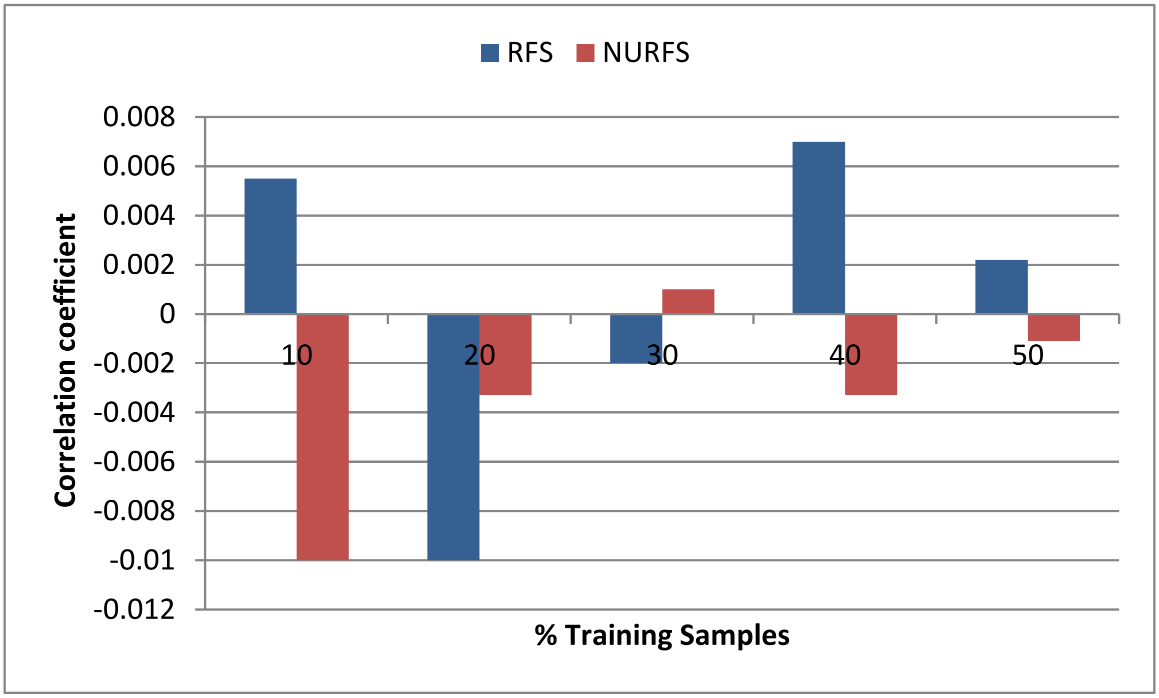

Figure 5.3 Correlation coefficient measures for the corn herbicide data

Lower score indicates greater diversity

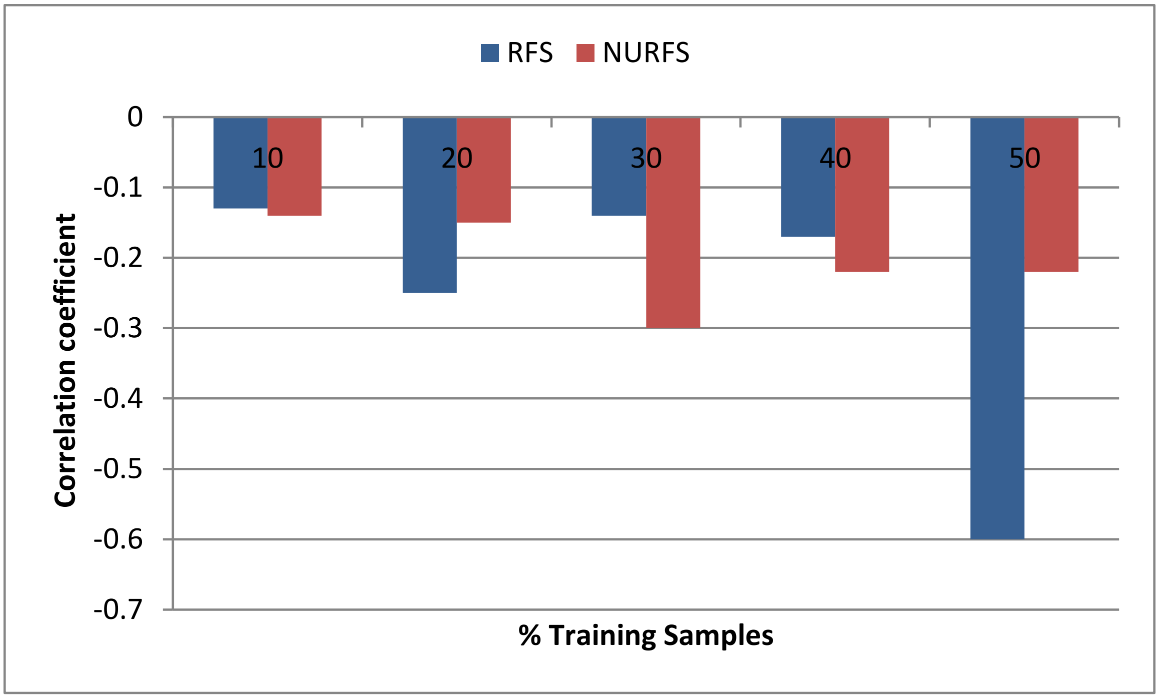

Figure 5.4 Correlation coefficient measures for the Indian pines data

Lower score indicates greater diversity

Figures 5.5 and 5.6 show the plots of disagreement measure for the corn herbicide and Indian pines data, respectively. This measure gives the ratio of classifiers disagreeing to the total number of classifiers, so the higher the value of this measure, the greater the diversity between classifiers.

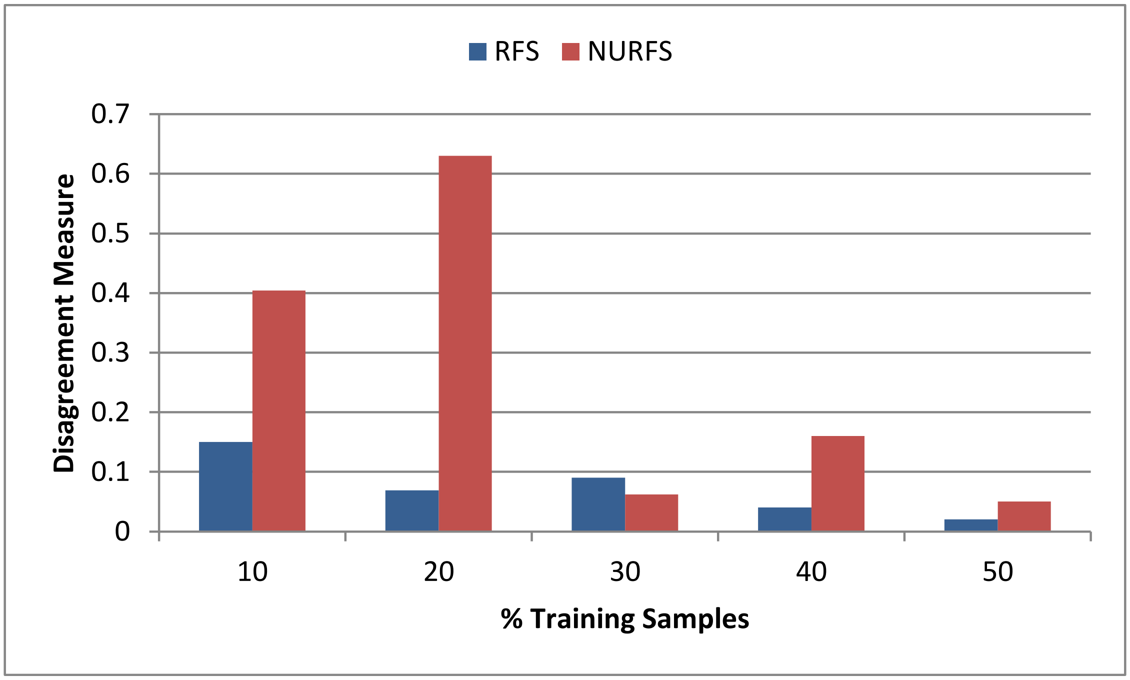

Figure 5.5 Disagreement measures for the corn herbicide data

Higher score indicates greater diversity

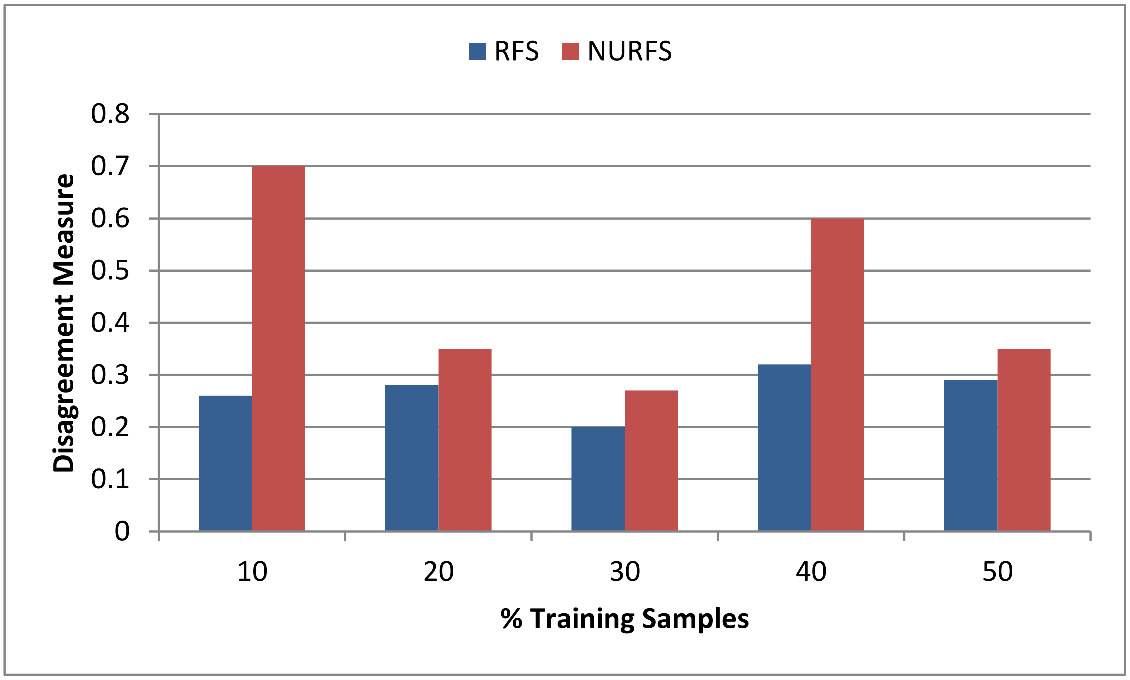

Figure 5.6 Disagreement measures for the Indian pines data

Higher score indicates greater diversity

The results clearly show that the proposed NURFS gives better ensemble diversity than the regular RFS algorithm. All the three measures are mostly consistent with the results. Unlike other measures, there is no practical upper limit for this measure as it depends on the number of samples and number of classes.

5.4 Difficulties and challenges

The diversity measures discussed in Section 5.2 are statistical approaches that rely on a simple accuracy of each classifier with respect to a given test sample. This may not represent the diversity in a true sense. Optimizing the individual classifiers in a MCS based on diversification is still not possible because that requires some techniques that can connect the accuracy of classifiers with ensemble diversity. The techniques discussed in this section provide a crude estimate of the classifier diversity and this is independent of the overall accuracy of the individual classifiers and overall accuracy. The main drawback with the current setup is that the diversification does not necessarily guarantee an optimized ensemble. This is largely due to disconnect between diversity and accuracy. This is observed through experiments and is reported in Section 5.3. In certain cases, improvement of class accuracies in the confusion matrix is of great interest. For such cases, a class wise diversity measure could be really useful. In the literature, there is no such a method available. This is an important area in ensemble classification that requires lot of research. An approach similar to the kernel density score matrix discussed in Section 3.3 is proposed by Ranawana [19] for estimating class wise specialties of the ensemble. This approach assigns a rank to a collection of classifiers based on its ability to distinguish different output classes. The problem again with this approach is its ability to connect the diversification of classifiers with class accuracies.

5.5 References

[1] Y. F. and R. E. Schapire, “Experiments with a new boosting algorithm,” in *13th International conference on Machine Learning*, 1996.

[2] R. Schapire, “Theoretical Views of Boosting and Applications,” in *Algorithmic Learning Theory SE - 2*, vol. 1720, O. Watanabe and T. Yokomori, Eds. Springer Berlin Heidelberg, 1999, pp. 13–25.

[3] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, “Boosting and Other Ensemble Methods,” *Neural Comput.*, vol. 6, no. 6, pp. 1289–1301, Nov. 1994.

[4] E. Bauer and R. Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Mach. Learn.*, vol. 36, no. 1–2, pp. 105–139, 1999.

[5] X. Ceamanos, B. Waske, J. Benediktsson, J. Chanussot, and J. Sveinsson, “Ensemble Strategies for Classifying Hyperspectral Remote Sensing Data,” in *Multiple Classifier Systems SE - 7*, vol. 5519, J. Benediktsson, J. Kittler, and F. Roli, Eds. Springer Berlin Heidelberg, 2009, pp. 62–71.

[6] P. Du, W. Zhang, and H. Sun, “Multiple Classifier Combination for Hyperspectral Remote Sensing Image Classification,” in *Multiple Classifier Systems SE - 6*, vol. 5519, J. Benediktsson, J. Kittler, and F. Roli, Eds. Springer Berlin Heidelberg, 2009, pp. 52–61.

[7] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, “Sensitivity of Support Vector Machines to Random Feature Selection in Classification of Hyperspectral Data,” *Geosci. Remote Sensing, IEEE Trans.*, vol. 48, no. 7, pp. 2880–2889, 2010.

[8] S. Prasad, L. M. Bruce, and H. Kalluri, “A Robust Multi-Classifier Decision Fusion Framework for Hyperspectral, Multi-Temporal Classification,” in *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, 2008, vol. 2, pp. II–273–II–276.

[9] J. Benediktsson, J. Kittler, and R. Fabio, Eds., *Multiple Classifier Systems*. Reykjavik: Springer, Lecture notes in Computer Science, 2009.

[10] A. Krogh and J. Vedelsby, “Neural Network Ensembles, Cross Validation, and Active Learning,” *Adv. Neural Inf. Process. Syst. MIT Press*, 1995.

[11] P. Cunningham and J. Carney, “Diversity versus Quality in Classification Ensembles Based on Feature Selection,” in *Machine Learning: ECML 2000 SE - 12*, vol. 1810, R. López de Mántaras and E. Plaza, Eds. Springer Berlin Heidelberg, 2000, pp. 109–116.

[12] L. Lam, “Classifier Combinations: Implementations and Theoretical Issues,” in *Multiple Classifier Systems SE - 7*, vol. 1857, Springer Berlin Heidelberg, 2000, pp. 77–86.

[13] L. Kuncheva and C. Whitaker, “Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy,” *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, 2003.

[14] A. Afifi and S. Azen, *Statistical Analysis : A Computer Oriented Approach*, Second Edi. Academic Press, 1979.

[15] L. Kuncheva, *Fuzzy Classifier Design, Studies in Fuzziness and Soft Computing*. Springer Verlag Heidelberg, 2000, p. 315P.

[16] D. B. Skalak, “The Sources of Increased Accuracy for Two Proposed Boosting Algorithms,” in *American Association for Arti Intelligence, Integrating Multiple Learned Models Workshop*, 1996, pp. 120–125.

[17] T. K. Ho, “The Random Subspace Method for Constructing Decision Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[18] G. Giacinto and F. Roli, “Design of effective neural network ensembles for image classification purposes,” *Image Vis. Comput.*, vol. 19, no. 9–10, pp. 699–707, Aug. 2001.

[19] R. Ranawana, “Intelligent Multi-Classifier Design Methods for the Classification of Imbalanced Data Sets: Application to DNA Sequence analysis,” University of Oxford, 2007.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

A new multi classifier system is proposed with a non-uniform feature selection and kernel density-based decision fusion. The proposed feature selection approach divides the HSI spectrum into different regions based on a band grouping algorithm. The features are then randomly selected from each group and the features selected from every group are concatenated to form the final subset. The number of features selected from any band group is proportional to the size of the group. The probability density of individual training classes are estimated using kernel density estimation and this information is used to determine the distance between classes. This information is used to form a score matrix where each score represents the ability of each classifier to distinguish a particular class from all other classes. The scores are then used for decision fusion. The proposed framework was tested on four different datasets each representing a different kind of a classification problem. With all the datasets, it was shown that the non-uniform random feature selection combined with kernel density decision fusion offers very high classification accuracies when compared to other state-of-the-art techniques. The impacts of the proposed feature selection and decision fusion techniques are studied individually. These studies revealed the impact of the two proposed techniques. A non-invasive system for classifying the corn kernels based on different aflatoxin levels is also proposed in this work. Diversity is an essential part of every ensemble classification system. The diversity of classifier outputs are studied using different measures. Although these measures cannot be used for optimizing the system design, they provide a coarse understanding of diversity created by the proposed feature selection approach when compared to a simple random feature selection.

To compare the performance of state-of-the-art statistical techniques with different HSI data, a maximum likelihood classifier is employed with linear discriminant and principal component analysis. This demonstrates the level of difficulty of the four classification problems under study. Then, the proposed approach is compared with state-of-the-art kernel classifiers. A single SVM, RFS, and variations of the proposed NURFS are compared. To demonstrate the suitability of the NURFS framework, three different types of experiments are conducted: 1) Study the sensitivity of the proposed approach to the amount of training samples, 2) Study the improvement of individual class accuracies, and 3) Study the effect of different kernel functions used to estimate the probability density function of classes in decision fusion. The proposed MCS' performance is shown to be superior compared to other methods. Especially with a small number of training samples for all the four HSI data under study, the kernel density NURFS outperforms the other techniques. It was evident that this system is better at handling smaller training sample sizes. The class accuracies from the confusion matrices showed a considerable improvement across all the classes of different HSI data under study. More importantly, in very few cases the class accuracies showed an inferior performance compared to other standard techniques. The experiments with different kernel functions for estimating the density revealed an interesting pattern. There is no common kernel function that works well with every data. Although this is data dependent, the difference in the overall accuracy is minimal for different kernel functions.

Experimental results with corn aflatoxin detection showed promising results with target class achieving very high accuracies. One can conclude that this approach is very effective for different data sets with a small number of training samples.

6.2 Suggested Future Work

The proposed kernel density NURFS system is very powerful for problems with a small training sample size. It would be interesting to study the performance of this approach to different multi spectral data from remote sensing and medical applications. In this dissertation, the proposed approach is tested for data with pure pixels. It would be interesting to study the performance of the proposed approach under mixed pixel situations.

Incorporating the spatial/vicinal information in classification applications such as aflatoxin detection and identifying herbicide stress. This can be achieved by post processing on class labels or classification map.

The band grouping approach used in this dissertation is based on metrics such as correlation and Bhattacharya distance. The studies could be extended with more advanced band grouping techniques such as Dirichlet process variable clustering.