ADVANCED SPECTRAL SPATIAL APPROACHES FOR DIMENSIONALITY REDUCTION OF HYPERSPECTRAL DATA

Thesis
Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
by

DEEPA C
177010AM003

Under the guidance of

Dr. AMBA SHETTY
Professor,
Dept. of Water Resources and Ocean Engineering,

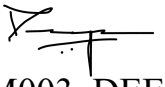
Dr. A.V. NARASIMHADHAN
Associate Professor,
Dept. of Electronics & Communication Engineering,
NITK, Surathkal

DEPARTMENT OF WATER RESOURCES AND OCEAN ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA SURATHKAL,
MANGALURU–575025
JANUARY, 2024

DECLARATION

By the Ph.D. Research Scholar

I hereby *declare* that the Research Thesis entitled **Advanced Spectral Spatial Approaches for Dimensionality Reduction of Hyperspectral Data** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfillment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Water Resources and Ocean Engineering** is a *bonafide report of the research work* carried out by me. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.



177010AM003, DEEPA C

(Register Number, Name & Signature of the Research Scholar)

Department of Water Resources and Ocean Engineering

Place: NITK, Surathkal

Date: 05/01/2024

CERTIFICATE

This is to certify that the Research Thesis entitled **Advanced Spectral Spatial Approaches for Dimensionality Reduction of Hyperspectral Data** submitted by **DEEPA C** (Register Number: 177010AM003) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

Prof. Amba Shetty

Dr. A.V. Narasimhadhan
08-01-2024.

(Research Guides)

(Name and Signature with Date and seal)

Vanija K
9.1.24
Chairman - DRPC

(Signature with Date and Seal)
Chairman (DRPC)
Dept. of Water Resources & Ocean Engineering

Department of Water Resources and Ocean Engineering
National Institute of Technology Karnataka, India

ACKNOWLEDGEMENTS

Developing the research thesis initially seemed extremely challenging due to the requirements for conducting research in the field of hyperspectral remote sensing. I would like to express my deepest gratitude to the exceptional individuals who supported and guided me throughout this transformative journey, enabling me to successfully complete Ph.D. despite several unforeseen obstacles that arose.

First and foremost, I would like to thank my research supervisors Prof. Amba Shetty and Dr. A V Narasimhadhan. Their exceptional academic expertise, coupled with commitment to support early career researchers truly make them a role model in the competitive realm of academia. The guidance and encouragement extended by them have been invaluable. Prof. Amba Shetty initiated the research in the world of remote sensing: first, planting the seed of curiosity in the area and then encouraging my research. Thank you for helping me to have carried out research and for your valuable advice and moral support in developing the thesis. Dr. A V Narasimhadhan played a key role in implementing the deep learning concepts. The encouragement, guidance and efforts put in by him have played a crucial role in shaping this work. Thank you for your availability and moral support.

I wish to thank Prof. Amai Mahesha, Prof. Amba Shetty, Prof. Dodamani BM and Dr. Varija K. the successive Heads of the Department of Water Resources and Ocean Engineering, NITK for their kind support, encouragement and providing the facilities required for research. I am thankful to the DRPC Secretary Dr. Subramanya K for his kind support and encouragement. I will always be grateful to all the faculty members of the Department of Water Resources and Ocean Engineering, for being a source of inspiration throughout my tenure as a research scholar in the department.

I would like to express my sincere gratitude to the RPAC members, Dr. D. Karmakar, Associate Professor, Dept. of WROE and Prof. Murigendrappa, Dept. of Mechanical Engineering. Thank you for your time, effort, valuable feedback and thought-provoking questions. Your insights have greatly enriched the quality of the research work. Furthermore, special thanks to Dr. Jeny Rajan, Associate Professor, Dept. of CSE who inspired me to carry out research in deep learning. I would also thank Dr. Shyam Lal, Associate Professor, Dept. of ECE for organizing workshops related to remote sensing which was a wonderful learning experience for me.

A special appreciation goes to my research colleagues at the Department of WROE and ECE. Engaging in stimulating academic exchanges with all of you has been truly enriching. Last but not the least, truly heartfelt and great thanks to my family and my loved ones for your blessings, amazing support and for always being on my side giving me the strength to continue and conclude this work.

My deepest gratitude and love belong to my parents, siblings, husband Dr. Ramesh H. and son Mst. Namith for their unconditional love and support throughout my research journey. To them, I owe all that I am and all that I have ever accomplished.

Sincere thanks to all,

DEEPA C

Dedication

To my parents, teachers and

my husband,

for their endless love and generous support.

ABSTRACT

Recent advances in sensor technology have enabled the collection of large data in hyperspectral remote sensing. Although rich spectral information is captured in hundreds of narrow contiguous bands, the hyperspectral data possess several limitations such as mixed pixels, high intraclass variability, interclass similarity, and the curse of dimensionality which restricts the potential of conventional machine learning classifiers. Dimensionality reduction (DR) and incorporation of spatial information can be taken into account to increase the interpretability of hyperspectral data. The thesis mainly focuses on the implementation of different approaches for DR of hyperspectral data to address the curse of dimensionality, limited samples and labelled data issues inherent in hyperspectral data.

First, a quality measure based on the co-ranking matrix has been proposed for the performance evaluation of 15 DR techniques for mineral exploration. The selection of appropriate techniques for a particular task is challenging due to the diversity and ever-increasing number of DR techniques. A few important aspects in this regard have been explored in detail. Clustering is performed using the K-means algorithm and the relationship between the quality index and clustering accuracy has been examined concurrently for the first time in hyperspectral remote sensing. Furthermore, the loss of quality in the process of DR has also been analyzed which provides sufficient input for the end-user to select an appropriate DR technique.

Second, the ability of the Convolutional Neural Network (CNN) for supervised learning of hyperspectral data is explored. A fast and compact hybrid CNN which combines the strengths of 3D and 2D convolutions to extract joint spectral-spatial information has been proposed to analyze the impact of different feature extraction techniques on classification performance. The effect of input patch size on final results has been well demonstrated. A detailed investigation of classification accuracy, execution time, and comparison with nine state-of-the-art approaches has been demonstrated.

Next, a novel deep feature selection strategy using autoencoders inspired by knowledge distillation has been implemented for the model compression and selection of informative bands. The potential of convolutional autoencoders has been well explored in selecting discriminative bands. Sensitivity analysis tests and different applications have been considered to verify the generalization capability of the proposed model. The potential of unsupervised learning schemes has been discussed in detail.

Finally, a generator model based on Generative Adversarial Networks (GAN) has been proposed for virtual sample generation and compact representation of hyperspectral data. The training instability issue in Vanilla GAN has been addressed by the effective implementation of deep convolutional GANs. By comparing the spectra of the generated hyperspectral images to the corresponding real ones, the quality of the images is assessed. The potential of augmented data for improvement in classification accuracy has also been investigated.

Keywords: Dimensionality Reduction, Hyperspectral Remote Sensing, Feature Extraction, Feature Selection, Knowledge Distillation, Convolutional Neural Networks, Generative Adversarial Networks.

Table of Contents

List of Figures

List of Tables

Acronyms

AVIRIS	Airborne Visible Infrared Imaging Spectrometer
BS	Band Selection
CCA	Curvilinear Components Analysis
CNN	Convolution Neural Network
DA	Data Augmentation
DL	Deep Learning
DR	Dimensionality Reduction
EO	Earth Observation
FE	Feature Extraction
FS	Feature Selection
GAN	Generative Adversarial Networks
GPU	Graphical Processing Units
GSD	Ground Sample Distance
HSI	Hyperspectral Imaging
HRS	Hyperspectral Remote Sensing
HVS	Human Visual System
HyMap	Hyperspectral Mapper
ICA	Independent Component Analysis
ISOMAP	Isometric Feature Mapping
JPL	Jet Propulsion Laboratory
KD	Knowledge Distillation
KNN	k-nearest Neighbour
KPCA	Kernel Principal Components Analysis
LDA	Linear Discriminant Analysis

LEM	Laplacian Eigenmaps
LLE	Local Linear Embedding
LTSA	Local Tangent Space Analysis
MDS	Multidimensional Data Scaling
MNF	Non-negative Matrix Factorization
MSE	Mean Squared Error
MVU	Maximum Variance Unfolding
NMI	Normalized Mutual Information
NN	Neural Networks
PCA	Principal Component Analysis
ProbPCA	Probabilistic PCA
PSNR	Peak Signal to Noise Ratio
ROSIS	Reflective Optical System Imaging Spectrometer
SE	Schrodinger Eigenmaps
SNE	Stochastic Neighbour Embedding
SPE	Stochastic Proximity Embedding
SRP	Sparse Random Projection
SSIM	Structural Similarity Index Metric
SVM	Support Vector Machines
TSFS	Deep Teacher Student Feature Selection
t-SNE	t-distributed Stochastic Neighbour Embedding
PU	Pavia University

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Numerous applications of machine learning, data mining, and image processing result in the collection of high dimensional data. Even though the data is high dimensional, it is intrinsically low dimensional since it is located in a subspace or a manifold. Dimensionality Reduction (DR) transforms the data in high dimensional space to a low dimensional space. The transformation generates a compressed version of the data while retaining a few of its original characteristics. DR is regarded as a pre-processing step in the visualization, analysis and modelling of data. It has become the focus of study in areas such as document analysis, gene expression array analysis, combinatorial chemistry, and medical image processing. Removing extra features from data not only allows the search algorithm to run faster and more accurately, but it also leads in high accuracy and better computing efficiency. Recent advancements in remote sensing technology enable the simultaneous record of hundreds of spectral wavelengths for each image pixel. The extensive spectral information provided by hyperspectral sensors facilitates to discriminate between different physical substances, potentially leading to more accurate categorization and opening the door to a plethora of new applications. Since hyperspectral data exhibits a substantial amount of spectral redundancy, DR is appropriate in majority of the applications.

1.1.1 Overview of Hyperspectral imaging (HSI)

Hyperspectral remote sensing (HRS) combines image and spectroscopy as sensing modalities. Based on reflected or transmitted electromagnetic radiation, an imaging system captures a remote scene of an area under observation. Spectroscopy evaluates the variation in power as a function of wavelength of light, revealing information about the chemical composition of the materials being studied. An imaging spectrometer or hyperspectral sensor is the instrument used to capture the corresponding spectral information. The sensor captures the information from the area of interest in many narrow contiguous spectral bands with fine spectral and spatial resolution as compared to multispectral data (Green et al. 1998). The rich spectral information facilitates to distinguish spectrally similar features. Each material absorbs light at a specific

wavelength range, which aids in identifying them based on their spectral signatures. The distinctiveness is the consequence of electromagnetic radiation interacting with the atoms and molecules of the material, resulting in absorption features in the reflectance spectrum. (Kruse et al. 2003).

Over the past few years, hyperspectral imaging (HSI) has gained interest in wide area of applications such as medical imaging, food processing, environment monitoring, vegetation, mineralogy, astronomy, land cover mapping, surveillance, military and so on (Vane et al. 1988). On the other hand, it has also created unique challenges for researchers working on HSI data for representation, exploitation and analysis of such voluminous data in an efficient way.

In a hyperspectral image, each pixel represents a high dimensional vector containing values corresponding to the reflectance spectrum, with the size of the vector equal to the number of spectral bands, as illustrated in Figure 1.1. In other terms, a hyperspectral image can be thought of as a three-dimensional hyperspectral data cube that stacks numerous grey-scale images corresponding to different spectral channels from the same scene together. For hyperspectral images, hundreds of spectral bands are typically available. Since different materials have unique spectral signatures, the abundance of spectral information accessible for each pixel of an image increases the likelihood of accurately differentiating different physical materials.

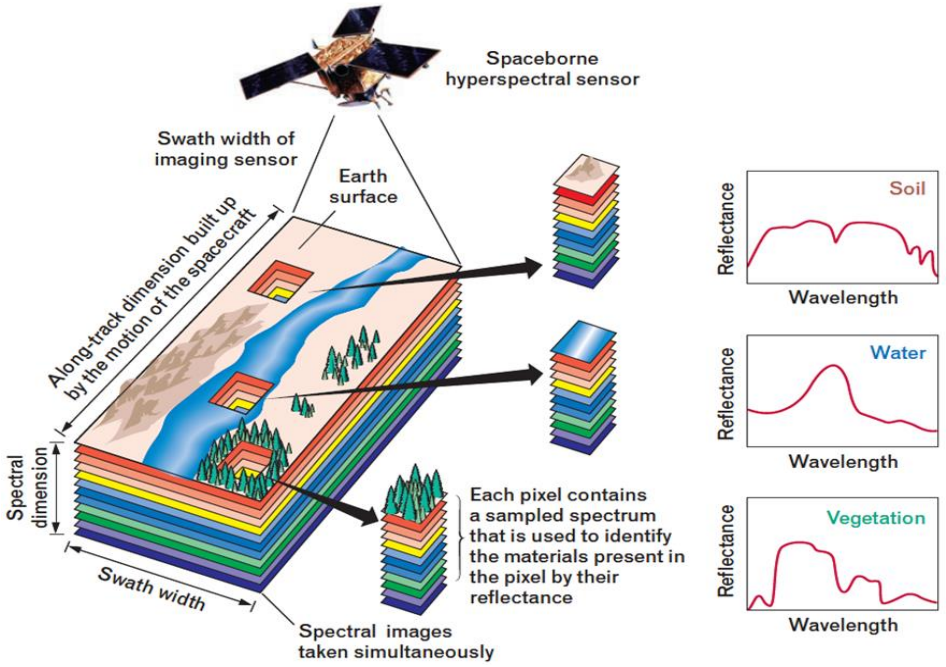


Figure 1.1 Hyperspectral data cube (Green et al. 1988)

1.1.2 General workflow of hyperspectral image processing

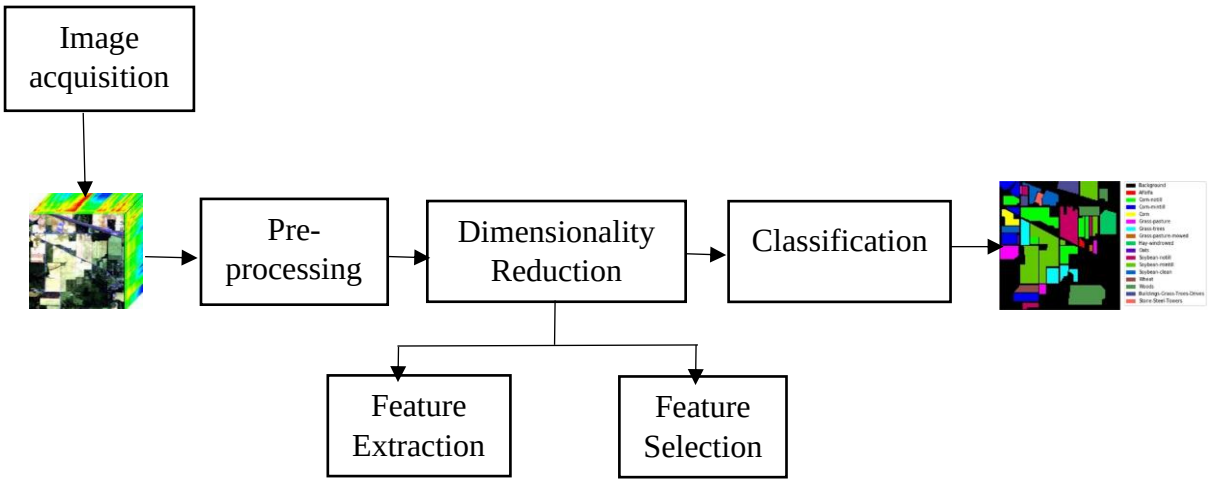


Figure 1.2 General workflow of hyperspectral image processing illustrating various stages in the processing chain

Figure 1.2 displays the overall architecture of hyperspectral image processing. It mainly consists of four stages: image acquisition, pre-processing, DR, and classification. Initially, hyperspectral images are captured using hyperspectral sensors (e.g., Earth observation (EO)-1 Hyperion, Airborne Visible Infrared Imaging Spectrometer (AVIRIS), Reflective Optical System Imaging Spectrometer (ROSIS), and Hyperspectral Mapper (HyMap)) over a broad wavelength range spanning the visible spectrum to the near-infrared region, providing precise spectral information about ground objects in multiple continuous spectral bands (from tens to hundreds).

Hyperspectral data captured by imaging spectrometer is prone to errors caused by variations in the viewing geometry, the atmosphere, platform motions, and other factors. The errors can be categorized as atmospheric, radiometric, and geometric errors. Errors can be reduced during the pre-processing stage by employing several correction strategies. By eliminating atmospheric interference, atmospheric correction recovers surface reflectance from remotely sensed imagery. In radiometric corrections, pixel grey values are transformed into radiance values that represent radiation reflected or emitted from the surface. Geometric correction includes the adjustment of captured image to enable the acquisition of the scale and projection characteristics of a specific map projection. The benchmark hyperspectral datasets are pre-processed and publicly available. The collection of spectra is further affected by sensor noise, fluctuating illumination, and climatic conditions. Consequently, the hyperspectral images

typically feature a few noisy and water absorption bands. Before further processing, the corrupted bands have to be eliminated.

Due to the existence of several bands (features) and a small number of labelled samples, classifying hyperspectral images is a tedious task. The performance of the classifier is influenced by the coupling between the number of training samples and features. Huge volumes of data produced by the large spectral bands contribute to the "curse of dimensionality" or “Hughes phenomenon” (Hughes, 1968) which can cause a considerable decline in classification performance if there are not enough labelled training samples. A hyperspectral image has enormous features which makes it impossible to accurately classify without sufficient training samples. Additionally, there is typically a substantial correlation between the adjacent bands resulting in spectral redundancy. When there are only a few training samples available, the curse of the dimensionality problem can be mitigated by a minimal number of features. Hence DR is a crucial step in the processing of hyperspectral images.

The classification of hyperspectral data aims to provide a distinct label to each test pixel in the image given a set of training samples. Each pixel of a hyperspectral image is represented by a vector. The reflectance of the object is represented by each pixel, and the length of the vector is equal to the number of discrete spectral bands. A wide range of classifiers are employed in the literature including k-nearest neighbour (KNN), Support Vector Machines (SVM), Multinomial Logistic Regression, Neural Networks (NN), Relevance Vector Machines, and Extreme Learning Machines.

1.2 DIMENSIONALITY REDUCTION FOR HSI DATA

The spectral curves of the different classes have similar shapes with different reflectance values. The spectra of classes strongly overlap in certain wavelength ranges, leading to harder discrimination. Furthermore, the spectral curves of all classes are highly correlated, resulting in a high degree of similarity between classes. Due to the lower spectral distance between the two neighbouring bands in the hyperspectral image and their extremely strong correlation, significant redundancy is prevalent. As a result, even though the spectral resolution is enhanced, no useful information is provided. Hence the complete spectral bands are not essential for classification. An optimal number of bands from a given entire set of bands can be accurately selected depending on the application. As the curse of dimensionality is predominant, a minimal number of features can ameliorate the problem. DR is a method of

reducing the number of bands and transforming the image from the original high dimensional data space to the lower dimensional data space, where the necessary information from the original data can be well retained. In general, there are two approaches for reducing the dimensionality of the hyperspectral images namely, Feature Extraction (FE) and Feature Selection (FS).

The FE approach transforms the original high dimensional feature space to low dimensional feature space which reduces the physical significance of bands but preserves more discriminative information needed for further analysis (Huilin Xu et al. 2019; Zhao et al. 2015; Romero et al. 2015). In the FS/band selection approach, a set of informative bands is selected according to the criteria where the physical significant characteristics of the original spectral bands can be preserved (Zhang et al. 2017; Han K et al. 2018; Sheikhpour et al. 2017). The extraction and selection of informative features in the classification of voluminous hyperspectral image are highly crucial tasks. In recent years, numerous approaches have been proposed for the extraction of suitable features as well as selection of the most informative bands. Although the existing FE approaches demonstrate significant performance, the emphasis of the conventional strategies is on raw spectral features rather than exploiting more complementary information from the bands of the hyperspectral data. In the current study, novel FE and FS based DR algorithms are proposed and developed that significantly improves the performance of hyperspectral image classification.

1.3 MOTIVATION

Although the large spectrum dimensionality of hyperspectral images improves pattern recognition precision, it challenges the computing capacity and performance of traditional signal processing techniques. A hyperspectral image contains millions of data points. The sheer volume of information creates complications in data processing and interpretation. Furthermore, in hyperspectral image analysis, numerous spectral bands are associated, implying the processing of irrelevant information. As a result, one of the primary steps in the hyperspectral data processing chain is DR, which allows the elimination of redundant information that could significantly limit classifier performance. In this way, the goal of DR in hyperspectral image processing is to reduce computing costs and minimize resource utilization while maintaining information quality. The proper selection of a subset of spectral information derived from the original data set has a direct impact on system speed and efficiency.

1.4 THESIS OUTLINE

The organization of remainder of the thesis is as follows.

Chapter 2 includes a complete analysis of the literature surveyed to gain and provide a better understanding of the study concepts, followed by problem description and research objectives.

Chapter 3 contains detailed information about the study area, data, and methods related to the research. A step-by-step description of the technique and analysis done is provided, which is organized in relation to the overall workflow.

Chapter 4 describes various techniques for FE and band selection, their properties and crucial factors to be considered in the DR process. The performance evaluation of 15 DR techniques for mineral exploration has been carried out and evaluated based on co-ranking criteria.

Chapter 5 presents a study on the impact of different DR techniques on hybrid CNN architecture in detail. The influence of different patch sizes and dimensions on classification performance has also been clearly explained.

Chapter 6 introduces deep FS using teacher-student networks inspired by knowledge distillation. The proposed scheme has been tested for both supervised and unsupervised scenarios to explore the generalization capability.

Chapter 7 describes the effectiveness of GAN in generating virtual samples and data augmentation in mapping vegetation data with compact representation.

Chapter 8 presents a summary and draws few important conclusions from the implemented objectives of the research work simultaneously providing recommendations for future research.

References to literature referred to in the thesis follow.

The following chapter reviews the literature adopted in the current study to formulate the proposed objectives.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter provides a detailed review of the relevant literature reviewed for the research. The following sections justify the use of hyperspectral data, DR techniques applicable to hyperspectral data.

2.2 HYPERSPECTRAL REMOTE SENSING

In the 1980s, the Jet Propulsion Laboratory (JPL) and NASA launched a program to build devices that could produce images of the earth's surface with unprecedented levels of spectral information due to the spectrum resolution limitations of traditional multispectral remote sensing. The new instruments were able to capture data in 200 or more extremely accurately defined spectral regions, as opposed to the few widely defined spectrum regions that were collected by earlier multispectral sensors (Kruse et al. 2003).

The word "Hyper" in hyperspectral data stands for "too many," which refers to the numerous measured wavelength bands. Since hyperspectral images are spectrally overdetermined, they can extract information more precisely than any other type of conventional remotely sensed data. The spatial coordinates provide the first two dimensions and spectral bands define the third dimension of a three-dimensional cube representation of hyperspectral data. As it operates on the spectroscopy concept, HRS is frequently referred to as "imaging spectroscopy." The process of creating spectra, separating their individual wavelengths, and employing them for chemical or physical research, as well as the identification of energy levels and molecular structure is known as spectroscopy. When the energy levels of atoms and molecules fluctuate, electromagnetic radiation either emits or absorbs creating a spectrum. The sort of energy levels involved and, consequently the surface and substance being viewed, determine the frequency of reflection or radiation. Numerous materials with distinctive reflectance spectra have been identified and mapped using hyperspectral imaging. Geologists, for instance utilized hyperspectral images to map the distribution of minerals and to identify salinity, moisture content, and organic matter in soil. Hyperspectral imaging has been effectively employed by

vegetation scientists to classify specific plant species (Goodenough et al. 2004; Bachmann et al. 2004).

2.3 DIMENSIONALITY REDUCTION OF HSI DATA

Numerous feature extraction (FE/DR) algorithms have been proposed recently for data visualization. In order to represent high dimensional data in a scatter plot or point cloud that enables field professionals to analyze the structure and distribution of massive amounts of data in a user-friendly and accessible manner, high dimensional data are first converted to two or three dimensional (2D/3D) vectors using a DR approach. The visualization has to accurately reflect the high dimensional data structure and distribution of the original data.

In general, HSI data has tens or hundreds of spectral bands. If the ground sample distance (GSD) is close to the objects or regions of interest, the bands are probably correlated, and it is likely that the pixels are spatially associated as well. As a result, redundant information is undoubtedly present in HSI (Li et al. 2011, Luo et al. 2013, Zhao et al. 2016). The high-dimensional spectrum space for HSI data is largely empty and the data occurs mainly in a subspace (Koren et al. 2004). In order to uncover relevant structures in high dimensional multivariate data, DR attempts to create a low-dimensional representation or embedding (Koren et al. 2004). Additionally, by removing redundant features from analysis, low dimensional embeddings can reduce the space and time complexity of analysis process (Lee et al. 1993; Prasad et al. 2008).

Additionally, DR can effectively extract valuable features from HSI (DeMers et al. 1992). As a result, it might be able to increase the precision of pixel-level categorization and helps to visualize the distribution of different classes (Koren et al. 2004; Vlachos et al. 2002).

Most of the HSI processing tasks intends to achieve two primary goals.

i) To recognize and categorize each pixel in the scene.

ii) To minimize the data volume/dimensionality while preserving important information (Harsanyi et al. 1994).

The literature review for DR is organized into two themes namely:

I. Feature extraction.

II. Feature selection.

2.3.1 Theme I: Feature Extraction

Without DR which is frequently accomplished by using linear transformations like PCA (Rodarmel et al. 2002; Du et al. 2007), MNF (Green et al. 1988), etc., conventional classification approaches may not be useful. For statistical pattern recognition, Hughes originally outlined the curse of dimensionality problem. The results demonstrated that recognition accuracy initially rises with more measurements on a pattern. However, as the dimensions crosses an optimal value, the accuracy declines.

Principal Component Analysis (PCA) is a widely utilized DR technique. The proposed technique for DR of hyperspectral data was thoroughly explained by Rodarmel et al. in 2002. The authors examine PCA as a pre-processing method for classification. The study not only demonstrated the usefulness of PCA but also revealed the data would only include noise after the first ten components. Along with evaluating accuracy, a comparison of computation times was also made which provides sufficient evidence that performing classification on data after pre-processing with PCA requires less computations. PCA generates data in a new uncorrelated coordinate system by computing orthogonal projections that maximize variance (Plaza et al. 2005). However, hyperspectral data does not always agree with these projections (Kaewpijit et al. 2003). As a result, numerous DR algorithms have been put forth in recent years.

According to linear DR, the data exists in or relatively close to a linear subspace of the high dimensional space. The most widely used linear DR approach is PCA since it is straightforward, effective and generates a set of uncorrelated axes that are arranged in decreasing variance by orthogonal projection. Principal component significance is assessed using the eigenvalues of the covariance matrix of original data and DR is achieved by preserving only a few components that correspond to the largest eigenvalues.

Another widely used linear DR technique is Fisher's Linear Discriminant Analysis (LDA) (Zhang et al. 2007). Unlike PCA, supervised LDA requires class labels for the data. When LDA is used, a loss function is minimized, resulting in less distance between samples from the same class. The class conditional distributions are presumed to be multi-variate Gaussian, which poses a significant constraint for both PCA and LDA (Martinez et al. 2001). However, real-world HSI data is frequently not Gaussian. Therefore, in such cases PCA and LDA will not produce meaningful embeddings. Independent Component Analysis (ICA) is a different linear DR technique that does not rely on the Gaussian assumption (Lennon et al. 2001). Data is

transformed into a different space where each element is statistically independent. In contrast to PCA, ICA not only removes all higher order dependencies but also decorrelates second-order statistics. However, similar to PCA and LDA; ICA has the drawback of not taking spatial dependencies or correlations between different pixels into account.

Nonlinear DR approaches differ from linear DR methods as they do not assume that data in high dimensional space occupies a linear subspace. Rather, several nonlinear DR techniques assume that the data are located on or close to a manifold that may have nontrivial curvature. Linear approaches will produce poor low dimensional representations if the data is on a nonlinear manifold and will typically overestimate the manifold inherent dimensionality. Nonlinear DR techniques are capable of integrating both the spatial and spectral information found in HSI and concentrate on computing low dimensional representations that preserve the manifold structure.

A variety of nonlinear approaches to DR have been investigated with respect to applications in HSI including Local Linear Embedding (LLE), Isometric Feature Mapping (ISOMAP), Kernel Principal Components Analysis (KPCA), Laplacian Eigenmaps (LEM), Diffusion Maps, Stochastic Proximity Embedding (SPE) (Agrafiotis 2003), Local Tangent Space Analysis (LTSA), Linear Local Tangent Space Alignment (LLTSA) (Kumar et al. 2016), t-distributed Stochastic Neighbour Embedding (t-SNE), Curvilinear Components Analysis (CCA) (Demartines et al.1997), Maximum Variance Unfolding (MVU) (Weinberger et al. 2006) and Schrodinger Eigenmaps (SE) (Benedetto et al. 2012).

One of the more prominent nonlinear DR techniques, Kernel PCA (KPCA) uses the same linear algebra as traditional PCA. KPCA initially performs a nonlinear mapping of the original process data into a high dimensional feature space where a linear data structure is more likely to exist. Further, linear PCA is performed on this feature space, and the resulting principal components are capable of capturing nonlinearities in the original data space. The advantage of KPCA is that it might be able to manage a variety of nonlinearities by utilizing various kernel functions. Despite this freedom in the choice of kernel function, research frequently focuses on common kernels like the polynomial, Gaussian, or hyperbolic tangent.
Developments in manifold learning provide an alternate approach to kernel selection. Manifold learning algorithms seek to unravel a high dimensional data manifold into a meaningful low dimensional space. To identify a function that maps the manifold to the low dimensional space,

this collection of algorithms makes the assumption that the data is situated on or near to an embedded low dimensional manifold in the original high dimensional space. A topological space that is locally identical to an Euclidean space is referred to as a manifold. Common manifold learning algorithms include Isomap, Diffusion maps (DM), LEM, LTSA and LLE. The majority of the techniques seek to achieve nonlinear DR while preserving certain local neighborhood structures in the data. These algorithms can be described as KPCA with the choice of a specially constructed kernel matrix. Unlike the Gaussian kernel, the kernels corresponding to these algorithms are data-driven and hence can effectively discover the intrinsic nonlinear structure hidden in the high dimensional data.

LLE maps close observations in the input data to nearby points in the low dimensional representation. Rather than assuming that all observations are linearly related as in PCA, LLE assumes that each observation is linearly related to a set of its nearest neighbors. On the other hand, global manifold approaches like Isomap attempt to preserve geometry at all scales and thus also attempt to map distant points to distant points. However, it is computationally intense, which is a concern for applications in real-time monitoring. Local approaches such as LLE involve only sparse matrix computations, resulting in substantial computational savings when the number of observations in the input data is large.

In order to reduce the difference between high-dimensional and low-dimensional probabilities produced from distances, probability-based DR approaches like SNE, symmetric SNE (SSNE), and t-SNE are used. The difference is calculated as the Kullback-Leibler divergence between the two distributions P and Q, which denotes the distributions for the high-dimensional and low-dimensional data respectively. When high dimensional data is visualized as clusters, the t-SNE probability-based technique produces best results.

According to recent studies, a spectral-based FE system may benefit from integrating spatial information (Chen et al., 2014). The advancement of sensor technology has enabled hyperspectral sensors to provide high spatial resolution. As a result, precise spatial data is easily accessible. According to research, spectral-spatial FE approaches significantly improve classification performance. Incorporating spatial information into classifier can be done using two approaches. The first category aims to account separately for spectral and spatial features. Advanced techniques such as morphological operations, attribute profiles and entropy (Sun et al. 2018; Bruni et al. 2022) are used to provide spatial information, which is then combined

with spectral data for pixel-wise categorization. The other group combines spectral and spatial data to obtain joint features. For example, wavelets (Yang et al. 2019) and Gabor filters (Chen et al. 2017) are built at different scales to simultaneously extract spectral-spatial features for classification.

In contrast, handcrafted FE techniques extract only shallow features and rely on a deep level of subject expertise when designing features. Deep learning (DL) has rapidly grown to be a research hotspot over the past few years as parallel computing techniques have been constantly improving. End-to-end models (i.e., feature extraction/learning and classification) based on deep learning have been frequently utilized to automatically learn the low and high-level representation of HSI in a hierarchical fashion in order to get around these restrictions. Numerous studies report the efficacy of deep learning for HSIC (Banerjee et al. 2022; Jia et al. 2023). Standard techniques include CNN, stacked autoencoders (Chen et al., 2014), deep belief network (Li et al., 2014) and recurrent neural network (Mou et al., 2017). With the exception of the CNN, majority of the techniques in the aforementioned models employ vector inputs, which ignores the spatial contextual interactions between pixels. CNN is pre-dominant in HSIC to effectively extract spatial information. HSIC built on CNN enhances generalization and prediction performance. Due to the significant performance gains, recently CNN based HSIC designs have received a lot of attention in hyperspectral remote sensing (Yang et al. 2018). A semi-supervised 2D CNN model with the encoder, corrupted encoder and decoder components was proposed (Liu et al. (2017). A semi-supervised nonlocal graph CNN is proposed (Mou et al. 2020) for classification. The network offers a new perspective for HSIC by accepting the entire hyperspectral image as input rather than just its local components such as pixels and patches. However, high computational and GPU memory overheads result in restrictions for large-scale classification tasks.

Since 2D convolution focuses solely on spatial features while ignoring important spectral information; it is incompetent to acquire discriminative features and provides poor performance in majority of the applications. However, 3D convolution extracts more discriminative spatial-spectral information from hyperspectral images. A 3D CNN network was developed that stacked several 3D convolutional layers without the pooling layer (Hamida et al. 2018). The proposed model effectively captures the local signal changes in spectral-spatial data. The 3D CNN-based approaches undoubtedly have more parameters to be trained than the 2D CNN-based approaches. As a result, compared to 2D CNN approaches the 3D CNN approaches have

substantially higher model complexity and memory requirements. The major limitation of 3D convolution is, as the generated feature maps grow in size, the convolution operation becomes significantly complex and demands more computational power.

Although CNN based techniques have demonstrated significant improvement for HSI FE, for supervised classification CNN tends to overfit and the hyperparameters fine-tuning process still requires sufficient samples with labels to aid in the training process. However, ground truth, training samples are limited and manual labelling is time-consuming, cost-effective process. To address these issues Generative Adversarial Networks (GAN) are proposed to extract and generate features in HRS (Makhzani et al. 2015; Zhang M et al. 2018)

The summary of literature corresponding to various FE techniques is presented in Table 2.1.

Table 2.1 Literature summary of various FE techniques

Method	Algorithm	Author and year	Significant findings	Inference
Linear methods	Principal Component Analysis	Rodarmel et al. 2002	Pre-processing for categorization using PCA	Computational effort is reduced
		Michael et al. 2005	Impact of PCA on target detection performance	Robust for detecting difficult targets
		Wu et al. 2016	Implementation of PCA in a distributed and parallel way by cloud computing technologies.	Significant speed compared to serial version
	Linear Discriminant Analysis	Sumithra et al. 2015	LFDA as a dimensionality reduction tool for complex nonlinear classifiers	Significantly outperforms conventional techniques
	Independent Component Analysis	Lenon et al. 2001; Vaddi et al. 2017	ICA for unsupervised analysis of hyperspectral images	Suitable for non Gaussian hyperspectral datasets
	Isomap	Bachmann et al. 2005	Global nonlinear technique that operates on geodesic distances between data sets.	Compared to PCA, Isomap extracts more structural information about the data.

Global nonlinear methods	Diffusion maps	Coifman et al. 2006	Creates a graph of patterns using Markov random walks.	Robust to noise, low computa-tional cost
	MDS	Borg et al. 1997	It only uses pairwise distance matrix between patterns	Demands less computing and storage requirements.
	Kernel PCA	Vaddi et al. 2017	Efficiently captures nonlinear relationships	Ideal for describing higher order complex and nonlinear distributions
	GDA	Park et al. 2004	Reformulation of traditional LDA using kernel trick.	Performs better than LDA for high dimen-sional datasets.
	SNE	Hinton et al. 2003	Probability-based stochastic selection of similar neighbors.	Problem of loc-al minima is avoided with the help of gradient desce-nt optimisation
	Sym.SNE	Lavander Maaten et al. 2008	Uses a pairwise similarity matrix to preserve neighbor identity.	Symmetric cost function involved speeds up optimization
	t-SNE	Lavander Maaten et al. 2008	Uses student distribution with heavier tail to avoid crowding problem.	Better for visualising high dimensional datasets
Local non linear methods	LLE, LTSA, LEM	Vaddi et al. 2017	Local non-linear method that produces a number of local mappings.	Performs better than Isomap
DL methods	Autoencoders	Chen. 2014	Efficient unsupervised feature extraction scheme	Performs better than PCA
	CNN	Hamida et al. 2018; Yang et al. 2018; Liu et al. 2017	Incorporates spatial information in analysis	CNNs are able to extract more prominent features for classification
	GAN	Mukherjee et al. 2019; Audebert et al. 2018	The generator model can be effectively utilized for virtual sample generation	Robust than CNN to reduce overfitting

2.3.2 Theme II: Feature Selection

Irrelevant and redundant bands are ignored in the FS/Band Selection (BS) process of hyperspectral data since they do not include relevant and suitable information for classification. The term band subset generation refers to the creation of a candidate subset for evaluation in the search space. For determining the nature of the band subset creation process, two simple criteria are taken into account. First, the band subset generation determines the starting point of the search process, which influences the search direction. The search starting points might be chosen using scoring, forward, backward, or random procedures. Second, the band is chosen using a specified approach such as a sequential or exhaustive search. A new band subset is examined according to predefined criteria. In the literature, numerous evaluation criteria for determining the adequacy of the candidate subset of features have been demonstrated. Stop criteria must be determined to end the selection process. BS procedures are classified based on the subset evaluation criteria, the availability of prior information, and the selection strategy used to create a band subset.

The BS approaches are classified as filter approaches, wrapper approaches, and hybrid approaches based on the subset evaluation criteria. The filter technique selects bands based on criteria that are independent of the classifiers used to classify the data (Geng et al., 2014; Yuan et al., 2015; Yang et al., 2017). The wrapper strategy selects bands depending on the classification performance of a specific classifier, such as maximum likelihood, support vector machines, k-nearest neighbour, and logistic regression (Medjahed et al. 2016). The hybrid BS strategy combines the filter and wrapper approaches. Since a lower computational cost is incurred, filter techniques are frequently faster than wrapper approaches. In contrast, wrapper approaches typically outperform filter approaches as they choose more representative bands from the initial band set. The effectiveness of a band subset is assessed using particular evaluation criteria. The criteria are either dependent on the learning process or independent of it. Generally, the filter approach uses independent evaluation criteria such as, information measures (entropy or mutual information) (Yang et al., 2017; Xie et al., 2017), distance measures (Bhattacharya distance, Kullback-Leibler divergence, Jeffries-Matusita distance, Hausdorff distance, and Spectral Angle Mapper (SAM)) (Medjahed et al., 2016), and dependency measures (correlation measures, similarity measures) (Zhang et al. 2018). The wrapper technique searches for a predetermined learning process.

The BS procedures are divided into supervised BS (Cao et al., 2016), semi-supervised BS (Feng et al., 2015), and unsupervised BS (Shukla et al. 2018; Xie et al., 2019) based on the availability of prior knowledge. A collection of labelled data is required for supervised BS methods, which is a very expensive and time-consuming process. The evaluation criteria used by the supervised BS methods maximize the class separability of training data samples with known class labels. Since different ground objects have varied spectral characteristics, numerous training samples exhibit divergent characteristics. As a result, the chosen subset of bands is unstable. Unsupervised BS procedures are more practical since collecting class information a priori is a costly and time-consuming process.

Previous research on unsupervised BS has focused on four methods: ranking, clustering, searching and embedding learning. Ranking-based methods evaluate the relevance of bands based on specific indicators to choose the top-ranked bands; however the selected subset typically suffers from information redundancy since it ignores band correlation. By grouping original data, the clustering-based BS algorithms aim to extract representative bands from each cluster (Tang et al. 2021), and the selected bands comprise the subset. To avoid redundancy, the strategy can both minimize and maximize interclass variance. The searching-based approaches select a subset by exploring band combinations based on a given criterion function, converting BS into an optimization issue. Bands are chosen using embedding-based methods that optimize certain application models such as classification, target identification and spectral separation (Beirami et al. 2020).

There are two basic techniques to determine the best band subset based on selection strategies. The first technique comprises individual band evaluation, whereas the second strategy incorporates band subset evaluation. Individual evaluation approaches include clustering-based approaches (Zhao et al., 2011; Cao et al., 2016; Yuan et al., 2015; Zhai et al., 2019) and ranking-based approaches (Wang et al., 2016; Jia et al., 2016; Zhu et al., 2017). The score of an individual band is measured in the individual band evaluation based on its degree of relevance. Certain search strategies such as exhaustive search, greedy search and combinatorial or metaheuristic optimization procedures are used in the band subset assessment to generate candidate band subsets (Su et al., 2014; Medjahed et al. 2016; Su et al., 2017).

The summary of literature corresponding to various FS techniques is presented in Table 2.2. The advantages and disadvantages of few FE and FS techniques have been reported in Table 2.3.

Table 2.2 Literature summary of few BS techniques

Band selection category			Author and year	Techniques used	Inference
Band subset Evalua-tion criteria	Prior informa-tion	Selection strategy			
Filter	Unsupe-rvised	Ranking	Wang et al. 2016, Gao et al.2019, Tang et al. 2021	Rank of each band is calculated and high rank bands are sorted to form subset.	Correlation between bands is not evaluated
		Clustering	Yuan et al. 2015, Yang et al.2017, Beirami et al. 2020	Band clusters are generated by increasing the inter-cluster variance and decreasing the intra-cluster variance	Efficient to provide less correlated bands but sensitive to initial conditions
		Exhaustive search	Zhan et al. 2017	Verifies all possible band combinations	High computation complexity
	Supervised	Greedy search	Yang et al. 2017,	Sequential search strategy with labelled samples is used	Fails to identify discriminating bands with limited samples
		Ranking	Feng et al. 2017	Unlabelled samples are also used	Highly correlated and stable band subsets are identified
	Semi-supervised	Exhaustive search	Bai et al. 2015	All possible combinations are tested	High computational cost
		Clustering	Jiao et al. 2014	Unlabelled samples, similarity measures	Low correlated bands
Wrapper	Supervised	Exhaustive search	Li et al. 2016	All possible combinations are tested	High computational cost
		Greedy search	Serpico et al. 2007	Sequential search	Classifier dependent and more complex
	Unsupervised	Greedy search	Sui et al. 2015	BS by integrating overall accuracy and redundancy	Highly correlated and stable band subsets
	Semi-supervised	Greedy search	Cao et al. 2017	Sequential search with both labelled and unlabelled samples	High computational cost

Hybrid		Clustering/ Meta-heuristic search	Medjahed et al. 2016, Feng et al. 2016	Combination of filter and wrapper approa-ches	Sensitive to initial cluster centers and more complex

Table 2.3 Advantages and disadvantages of few FE and FS techniques

Method	Technique	Merits	Demerits
FE	PCA	Prevents overfitting, removes correlated features, improves visualization	Information loss, hard to interpret, not applicable for nonlinear data
	LDA	Improved discrimination ability	Computationally complex, less efficient
	t-SNE, Isomap, LEM	Works well for strongly nonlinear data, better visualization	Can be inefficient for large data
	NMF	Easy interpretation of results	Computationally expensive
FS	Filter	Lower risk of overfitting, computationally less complex	Ignores feature dependencies, no interaction with classification model
	Wrapper	Better generalization capability, considers feature dependencies	Computationally infeasible, high risk of overfitting
	Embedded	Faster running time, interacts with classification model for FS	Identification of smaller subset of features is tedious

2.4 SIGNIFICANCE OF DR IN HRS

The interpretation of hyperspectral images is challenging even though the data is highly informative. Providentially, processing every band of a hyperspectral image is not always essential. Majority of the objects under investigation have few selected bands where they exhibit particular properties, leaving the other bands unnecessary. However, to obtain high classification accuracy, the number of training samples for each class needs to increase with the intensifying dimensionality of the hyperspectral image. Additionally, when data grows sparser in size, both supervised and unsupervised learning may face significant challenges. Consequently, DR is a crucial stage in the HSI pre-processing stage.

2.5 CHALLENGES IN DR OF HSI DATA

- The volume of data that needs to be processed might be overwhelming since hyperspectral images typically have hundreds of bands and data cubes are frequently hundreds of megabytes. Since spectral bands are compact and adjacent, there is typically a strong correlation between adjacent bands. As a result, the number of bands

is substantially higher than the subspace dimensionality where the hyperspectral data is located. Most of the DR techniques are specifically applied to classification problem. However, it is extremely hard to generalize effective DR technique for numerical prediction or visualization purposes.

- Plethora of algorithms exists for DR. The proliferation of DR techniques, as well as their broad use in many applications, makes it difficult for end users to comprehend how to select a better methodology for a given use scenario. While the scientific community has been focusing on the development of novel nonlinear DR approaches, the subject of quality evaluation has largely gone unresolved.
- For a given data, different techniques result in qualitatively very different visualizations. As a result, it is unclear which DR technique is best suited for the task at hand. Furthermore, almost all current approaches include settings that regulate the preservation strategy for the embedding. As a result, depending on the settings used, even a single DR approach can produce diverse results. Furthermore, due to the random characteristics of the method, many nonlinear DR algorithms do not produce a unique solution. They can instead produce distinct outputs in each run, corresponding to multiple local optima of the objective. As a result, a single procedure with a single set of model parameters may provide qualitatively diverse solutions.

2.6 RESEARCH GAPS

- Labelling of hyperspectral data is tedious and time-consuming, hence acquiring enough labelled samples for the learning algorithm to adequately capture the scene appearance is challenging. Hence, there is a need for the development of techniques based on unsupervised learning that do not require large amounts of labelled data.
- Though numerous techniques are available for DR, the lack of methodology to assess and compare the performances of different DR methods on hyperspectral data is a challenging issue, and it is not yet well explored in the literature. In addition, it is highly difficult to separate the performance of feature extractor and classifier in hyperspectral processing chain.
- From a theoretical perspective, DR is an ill-posed problem; not all the structure and relations that exist in high dimensional data can be appropriately represented in the low dimensional space, and it is unclear which relations should be retained. The application

task decides the parameters to be chosen. A different goodness measure is required to complete the task. This measure is anticipated to be simple, applicable to the majority of algorithms and datasets, resistant to the presence of outliers, and resilient against incorrect tuning parameter selection.

- 2D CNN focuses mostly on spatial features while ignoring important spectral information, it is unable to acquire discriminative features. However, 3D convolution extracts more discriminative spatial-spectral information from hyperspectral images with increased complexity. In addition, most of the studies utilize PCA as a standard pre-processing technique for deep learning models without exploring the capabilities of other well-known DR techniques in HSIC.
- In BS, conventional techniques with hand crafted features do not explore the correlation among bands well and also results in more complex models for exhaustive search leading to sub-optimal band subsets. Hence for efficient BS, there is a need for models capable of selecting optimal bands with less complexity.
- Although DL-based algorithms have made significant progress in HSI classification, large training samples are required for model training. Deep models trained with limited samples lead to overfitting. The studies related to generative models are limited.

2.7 RESEARCH OBJECTIVES

The high dimensionality of the data, which leads to high intrinsic information redundancy and thus to the Hughes phenomenon, is still an open issue. The proposed research work aims at investigating and defining advanced spectral-spatial approaches for DR of hyperspectral data. In particular, the focus is on the implementation of strategies, based on the use of linear and nonlinear DR techniques and deep learning models. Aiming at overcoming the aforementioned issues and limitations that affect the analysis of hyperspectral data analysis, the following objectives are defined:

1. To explore conventional feature extraction techniques, application on hyperspectral mineral data and its evaluation based on the co-ranking framework.
2. To analyze the impact of feature extraction strategies on hybrid CNN and design an efficient model for compression of hyperspectral imagery.
3. To compress deep neural networks using knowledge distillation and develop an integrated model for deep feature selection of hyperspectral data.

4. To design a DL model based on GANs for virtual sample generation and compact representation of hyperspectral data.

The description of the study area and datasets under investigation, software tools and overall methodology is outlined in the succeeding chapter.

CHAPTER 3

DATA AND METHODOLOGY

3.1 INTRODUCTION

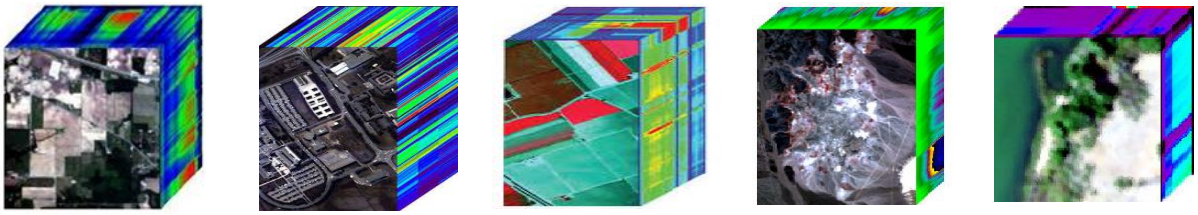
The current chapter provides an overview of the study area, dataset and methodology adopted in the research.

3.2 STUDY AREA AND DATASET

The following real-world hyperspectral datasets are used to test the efficacy of the proposed techniques. Table 3.1 provides a detailed description of the datasets used which are publicly available for research and can be downloaded from https://rslab.ut.ac.ir/data <2018> website.

Table 3.1 Description of datasets

Dataset	Bands	Size	Range(nm)	Width(nm)/GSD(m)	Classes
Indian Pines	200	145x145	400-2500	10/20	16
Pavia University	103	610x340	430-860	10/3.7	9
Salinas	204	512x217	360-2500	4/1.3	16
Cuprite	188	250x190	370-2480	10/20	12
Samson	156	95x95	401-889	3.13	3



(a) (b) (c) (d) (e)

Figure 3.1 Pre-processed hyperspectral data cubes of different scenes captured by various sensors: (a) Indian Pines (b) Pavia University (c) Salinas (d) Cuprite (e) Samson

Indian Pines (IP): The scene comprises a spatial dimension of 145x145 pixels and 224 spectral reflectance bands in the wavelength range of 0.4 to 2.5 µm, was captured by the AVIRIS sensor over the Indian Pines test site in northwest Indiana. Two-thirds of the Indian Pines scene is made up of agricultural, and one-third is made up of forest or other types of natural perennial flora. There are sixteen classes of ground truth. The bands [104-108], [150-163], and 220 that

cover the water absorption region have been eliminated, reducing the total number of bands to 200.

Pavia University (PU): The scene was captured by ROSIS sensor while flying over Pavia, Italy in a campaign. It comprises 103 spectral bands with spatial dimensions of 610x610 pixels and a geometric resolution of 1.3m. The ground truth consists of 9 land cover classes. A few empty samples have been discarded before analysis.

Salinas Full scene (SA): The dataset was captured by AVIRIS sensor with 224 bands over Salinas valley, California characterised by high spatial resolution of 3.7m pixels. The image covers a spatial dimension 512x217 pixels from spectral range of 360-2500 nm. Salinas ground truth includes 16 classes primarily covers vegetables, bare soils and vineyard fields. The bands [108-112], [154-167], 224 are discarded due to water absorption providing 204 bands for analysis.

Cuprite: The Cuprite dataset serves as the most important benchmark for studies on hyperspectral unmixing in Las Vegas, Nevada, United States comprising 224 channels with wavelengths between 370 and 2480 nm. After eliminating the noisy channels (1-2 and 221-224) and water absorption channels (104-113 and 148-167), 188 channels are left for analysis. There are 14 different mineral types in a spatial subset of 250 ×190 pixel area with 14 minerals considered for analysis. The number of endmembers (pure spectral signatures) is reduced to 12, which are outlined as follows: Alunite, Andradite, Buddingtonite, Dumortierite, Kaolinite1, Kaolinite2, Muscovite, Montmorillonite, Nontronite, Pyrope, Sphene, and Chalcedony.

Samson: There are 952x952 pixels in the Samson dataset. There are 156 channels used to record each pixel and spans the wavelength range of 401 nm to 889 nm. The spectral resolution is quite good and reaches 3.13 nm. A portion of 95 × 95 pixels is used instead of the original image due to large size and high computing costs. In the original image, it begins at (252,332) pixel location. The blank channel or channels with excessive noise have no negative effects on the data. The three ground truth classes are soil, tree, and water.

3.3 SOFTWARE TOOLS

- MATLAB© (MATrix LABoratory) developed by Mathworks is a proprietary multi-paradigm programming language. It provides an excellent numeric computing environment to analyze data, develop algorithms and create mathematical models. A

group of application-specific solutions known as toolboxes are available in MATLAB. The current study employs MATLAB R2017a with DR toolbox.

- Google Colaboratory (Colab) is used for writing Python scripts which is an integrated development environment that runs on the cloud. The main packages for statistical analysis and visualisation are pandas, matplotlib and seaborn. For scientific computation and implementation of machine learning algorithms, numpy and scikit-learn is used. Keras with TensorFlow backend is preferred as a library for building neural networks and applying DL.

3.4 OVERALL METHODOLOGY

Figure 3.2 depicts the overall methodology involved in DR of HSI data. The primary step includes the collection of pre-processed hyperspectral data with highly correlated and irrelevant features. A number of FE algorithms: linear, nonlinear and DL models are applied on the data in order to extract meaningful features. The corresponding objective functions are optimized

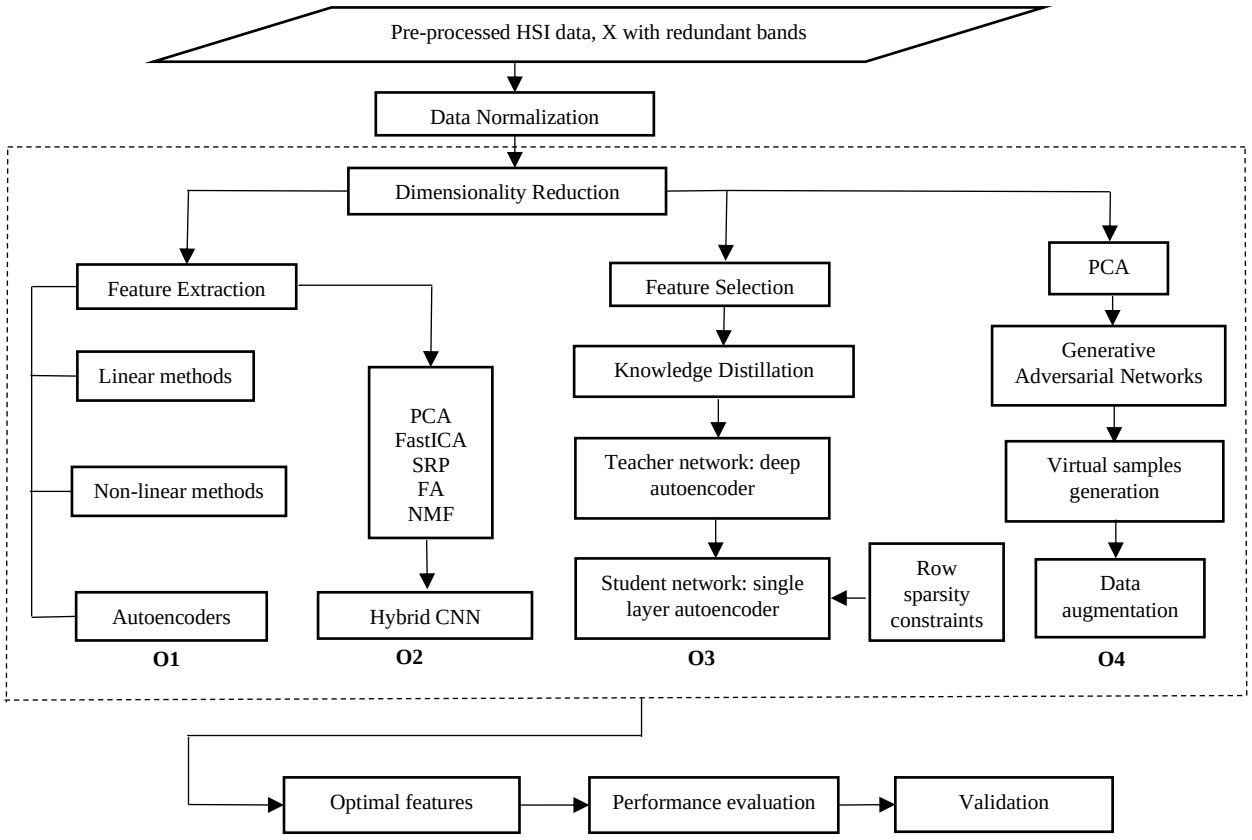


Figure 3.2 Overall methodology for DR of HSI data

(**O1**-objective 1, **O2**-objetive 2, **O3**-objective 3, **O4**-objective 4)

by the proposed techniques. According to the analysis, each technique yields a different set of outcomes. In the next step, the performance of all the techniques is evaluated based on certain criteria like topology preservation, classification, clustering or quality of reconstruction. The optimum technique for a certain application can be chosen based on experimental validation.

The next chapter provides an exhaustive survey and performance evaluation of 15 DR techniques for mineral exploration. A detailed investigation on quality of DR process and its relation with clustering has been brought out for the first time in HRS.

CHAPTER 4
QUALITY ASSESSMENT OF DR TECHNIQUES

4.1 INTRODUCTION

This chapter provides an essential background of DR techniques for a better intuitive understanding of concepts and further exploration of other techniques. DR is the statistical process of reducing the number of dimensions (or variables) required to describe a dataset. Hyperspectral datasets are of high dimensionality as the reflectance at each wavelength for a single pixel can be interpreted as a separate dimension. A new criterion for assessment of DR techniques has been proposed for the first time in HRS based on co-ranking matrix and mutual information for mineral exploration. A few popular methods for DR are organized in groups and explained. A majority of them have been traditionally applied to DR. Figure 4.1 depicts the taxonomy of DR techniques. Various available methods for DR are described as follows:

4.2 LINEAR DR TECHNIQUES

A) Principal component analysis (PCA): PCA (Rodarmel et al. 2002) is a very popular statistical algorithm for exploratory data analysis which is also used for data pre-processing, image compression, data reduction etc. It develops an orthogonal transformation that linearly projects the patterns from high dimensional, R_H to low dimensional, R_L using the formula $y = W^T x$, where the matrix W is composed by d principal components, which are vectors defining the directions with the maximum variability in R_H. The subset of the principal components associated to the d largest eigenvalues define the low dimensional space R_L. PCA computes the nxn order covariance matrix, $C = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$ and solves the following eigen problem:

$$Cv = \lambda v \qquad (4.1)$$

4.3 NON-LINEAR DR TECHNIQUES

The current section groups nonlinear methods which have been traditionally used for DR: global and local techniques based on the type of information they preserve.

4.3.1 Global nonlinear methods

These methods attempt to preserve the global properties of the pattern set, usually the pairwise distance between patterns, however allowing non-linear mappings between R_H and R_L.

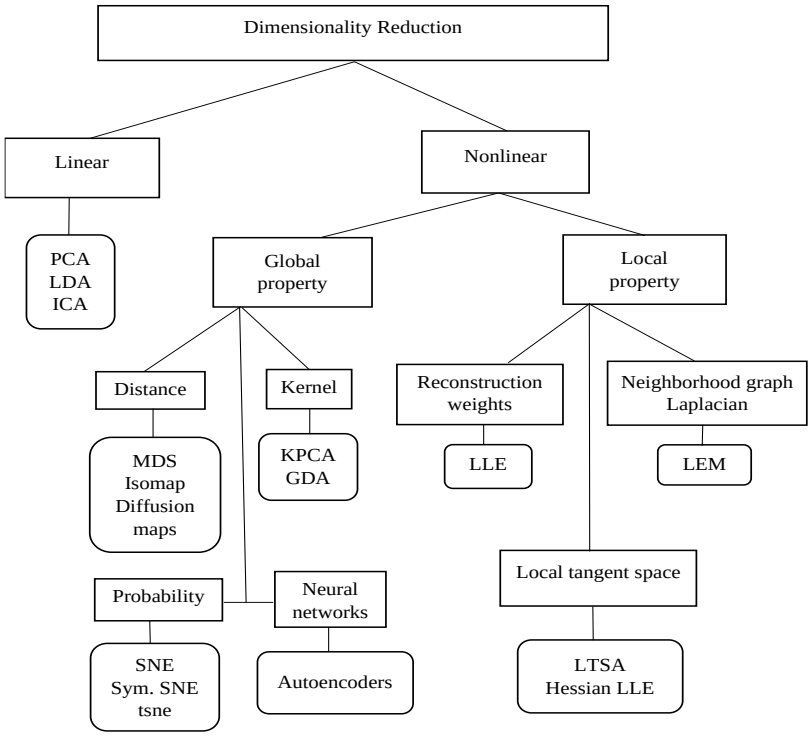


Figure 4.1 Taxonomy of DR techniques

A) Multidimensional data scaling (MDS): The MDS (Borg et al. 1997) maps pairwise patterns from R_H onto R_L while retaining the inter distance between patterns as much as possible in both spaces, using a dissimilarity matrix that describes the geometric structure of the pattern set. Hence, it only requires a matrix with the pairwise distances between patterns, and not the whole patterns themselves, which is an advantage for computing and storing requirements. The loss function (eq.4.2) of the distances is called stress (φ), and it measures the difference between the pairwise distances in R_H and R_L.

$$\varphi(Y) = \sum_{i<j}(|\delta x_{ij}| - |\delta y_{ij}|)^2 \qquad (4.2)$$

B) Isomap: The approach is a combination of MDS and PCA which preserves the intrinsic geometry of the data (Bachmann et al. 2005), and specifically the pairwise geodesic distance between patterns, i.e., the distance alongside the curvilinear manifold which best describes the pattern set. The geodesic distances are calculated by creating a neighbourhood graph G which connects each original pattern x_i to its k nearest neighbours. The shortest path between two points in the graph forms an estimate of the geodesic distance between these two points and

can easily be computed using Dijkstra's or Floyd's shortest-path algorithm. The geodesic distances between all datapoints in X are computed, thereby forming a pairwise geodesic distance matrix. The low-dimensional representations y_i of the datapoints x_i in the low-dimensional space R_L are computed by applying PCA on the resulting pairwise geodesic distance matrix.

C) Diffusion maps: A diffusion map (Coifman et al. 2006) creates a graph of the patterns using Markov random walks as the first step. All the nodes in the graph are connected together and the weight of the edges in the graph between two patterns (x_i,x_j) is computed using the Gaussian (or diffusion) kernel and diffusion matrix is also calculated. Further, eigenvectors and eigenvalues of the diffusion matrix are calculated to represent the patterns y_i in R_L by selecting directions in the diffusion space associated with the largest eigenvalues. The diffusion map attempts to retain the diffusion distance, being robust to noise perturbation, and its computational cost is relatively low.

D) Kernel PCA: KPCA is the reformulation of traditional linear PCA in high dimensional space constructed using a kernel function (Fauvel et al. 2009). KPCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. It projects $x_i \in R_H$ to a feature space of dimension F which may be infinite $\overrightarrow{\varphi} : R_H \rightarrow F$, defined implicitly using kernel trick.

E) Generalized discriminant analysis (GDA): GDA (Ye et al. 2004) is a derivation of LDA which uses the kernel trick to solve non-linearly separable classification problems, which cannot be appropriately separated by LDA. The technique uses a non-linear kernel function to project the patterns in the input space into a feature (or Hilbert) space F, where the dot product is calculated in order to learn nonlinear classification functions. The GDA maximizes the scatter between classes and minimizes the scatter within a class, similarly to the LDA, but in the high-dimensional space F, while the LDA maximizes the Fisher criterion in the original R_H space.

F) Stochastic Neighbour Embedding (SNE): It is a nonlinear data reduction technique which maps patterns from R_H to R_L preserving a neighbourhood identity while retaining the pairwise Euclidean distance between the original patterns $x_i \in R_H$ as much as possible (Hinton et al.

2003). The SNE minimizes a cost function given by the sum of KL divergences of $p_{ij} \epsilon R_H$ and $q_{ij} \epsilon R_L$ between all neighbours:

$$\emptyset(Y) = \sum_i \sum_j p_{j/i} \; log \; \frac{p_{j/i}}{q_{j/i}} \qquad (4.3)$$

G) Symmetric SNE (SSNE): It is a variant of SNE that maps the data from R_H to R_L, based on the pairwise similarity matrix attempting to preserve the neighbor identity (Lavander Maaten et al. 2008). The use of a symmetric cost function, whose gradient is simpler than SNE and the addition of momentum terms speed up the optimization with respect to SNE.

H) t-distribution SNE (t-SNE): The t-SNE is a variation of SSNE (Lavander Maaten et al. 2008) which uses a student t-distribution with a single degree of freedom as a distribution in R_L with a tail heavier than a Gaussian distribution in order to avoid the crowding problem of SSNE.

I) Probabilistic PCA (ProbPCA): In PCA, the size of the covariance matrix is proportional to the dimensionality of the datapoints. As a result, the computation of the eigenvectors might be infeasible for very high-dimensional data. Alternatively, iterative techniques such as probabilistic PCA may be employed. The ProbPCA (Smola et al. 2004) is an iterative extension of PCA that uses a probabilistic Gaussian latent variable model to solve the limitations inherent in the regular PCA, such as dealing with missing data patterns and lack of an explicit generative model. ProbPCA uses an iterative algorithm called Expectation Maximization (EM) to accomplish the task.

J) Multi-layer autoencoders: An autoencoder is a feed forward neural network with encoder and decoder sub-models (Hinton et al. 2006). The encoder compresses the input via bottleneck layer and decoder tries to reconstruct the original input. The autoencoder is trained using the original patterns $x \epsilon R_H$, to minimize the mean squared error between the actual and predicted output (Behnood Rasti et al. 2020). A deep autoencoder is composed of two symmetrical deep-belief networks with four or five shallow layers to learn more complex features. The layers are restricted Boltzmann machines which are the building blocks of deep-belief networks.

4.3.2 Local nonlinear methods

The methods presented in this section are non-linear mappings oriented to preserve properties

which are valid only in small neighborhood around the patterns. Hence, they are called local nonlinear mappings.

A) Local linear embedding (LLE): The LLE (Kim et al. 2003) creates a representation graph of patterns, however preserving the local structure by describing the original patterns x_i as a linear combination of their k nearest neighbors x_{ij}, with j = 1,…..,k. LLE maps the patterns to R_L in such a way that it preserves the local geometry of the original patterns x_i in the manifold, i.e., to keep the same weights as in R_H, as much as possible. Specifically, LLE calculates the low dimensional mapped patterns $y_i \in R_L$ to minimize the following cost function:

$$\phi(Y) = \sum_{i=1}^{N} |y_i - \sum_{j=1}^{N} w_{ij} y_{ij}|^2 \quad (4.4)$$

B) Local tangent space alignment (LTSA): LTSA (Zhang et al. 2007) describes the local properties of patterns in R_H using the local tangent space Θ_i of each pattern x_i. Assuming that patterns lie on a manifold which is locally linear; there are two linear mappings, one from x_i and other from its low-dimensional version $y_i \in R_L$, both into Θ_i. The LTSA proceeds by aligning both mappings to construct the LTS from R_L searching simultaneously for y_i and for the mappings L_i from y_i to Θ_i. Initially, LTSA applies PCA on the set of k neighbors of x_i, creating a mapping from this set to Θ_i. Then, it minimizes:

$$\phi(Y) = \sum_{i=1}^{N} |J_k y_i - L_i \Theta_i|^2 \quad (4.5)$$

where J_k is the centring matrix of size k, so LTSA minimizes the sum, over all the patterns, of the squared norms of the differences between the local tangent space in R_H, given by $L_i\Theta_i$, and the centred low dimensional pattern y_i.

C) Laplacian Eigenmaps (Laplacian EM): LEM (Belkin et al. 2001) maps the patterns from R_H to R_L by preserving the local properties, the pairwise Euclidean distances between neighbors of the manifold where the patterns lie. The mapped patterns y_i are calculated in such a way that they minimize the distance between y_i and its nearest neighbours. In the cost function, distances are weighted decreasingly with the neighbour order. Initially, it creates a graph where each pattern x_i is connected to its k nearest neighbours. The cost function which is minimized to calculate $y_i \in R_L$ is given by eq. 4.6,

$$\phi(Y) = \sum_{i,j} w_{ij} |\delta y_{ij}|^2 \quad (4.6)$$

In addition to the algorithm details explained above, it is crucial to reveal few interesting facts about DR algorithms to facilitate proper selection of a DR technique for a particular application. PCA fails when mean and covariance alone are insufficient to describe datasets. MDS aims to maintain the order of the distances and, as a result pursues a monotonic relationship between the embedding space distances and similarities/differences. Isomap is an extension of MDS and KPCA. Diffusion maps are similar to Isomap but less sensitive to short circuiting since they integrate the overall paths through the graph rather than shortest paths. However, it performs well solely on noise free and densely sampled data. Isomap, LLE and their variants work best to unfold a single continuous low-dimensional manifold, whereas t-SNE concentrates on the local structure of the data and tends to recover clustered local groupings of samples. It may be possible to visually detangle a dataset made up of multiple manifolds by using this ability to group samples depending on the local structure. LLE aims to project the data in a lower dimension while maintaining local distances. To determine the optimal non-linear embedding, it can be compared globally as a collection of local PCAs. LTSA and LLE is sufficiently comparable. LTSA aims to characterise the local geometry in each neighbourhood via its tangent space rather than preserving neighbourhood distances as in LLE. Additionally, performs a global optimization to align these local tangent spaces to learn the embedding. t-SNE aids in minimising the inclination to data points near the centre to avoid crowding problem inherent in SNE. Neural networks, such as autoencoders are effective at processing specific categories of data, including audio and image data. However, as autoencoders are neural networks, they require more data to be trained.

Table 4.1 depicts useful properties of few DR techniques (Espadato et al.2019). In the table, eight DR techniques are listed by four general properties: the main free parameters that has to be optimized, the computational complexity of the technique, the memory complexity of the technique and the corresponding features. The four general properties are discussed below. The objective functions of nonlinear techniques for DR have free parameters that needs to be optimized. Free parameters refer to the parameters which directly influences the cost function to be optimized. Non-convex technique for DR, autoencoders have additional free parameters namely, the learning rate and the permitted maximum number of iterations. Moreover, LLE uses a regularization parameter in the computation of the reconstruction weights. The presence of free parameters has both advantages and disadvantages. The main advantage of the presence of free parameters is that they provide more flexibility to the technique, whereas the main disadvantage is the necessity to be tuned to optimize the performance of the DR techniques.

Table 4.1 Properties of few DR techniques.

Sl.No.	Technique	Parameters	Computation	Memory	Features
1.	PCA	-	$O(D^3)$	$O(D^2)$	Linear, unsupervised
2.	Isomap	k	$O(n^3)$	$O(n^2)$	Preserves pairwise geodesic distance, neighbourhood graph
3.	KPCA	$\kappa(\cdot,\cdot)$	$O(n^3)$	$O(n^2)$	Unsupervised, nonlinear, kernel
4.	Diffusion maps	σ, t	$O(n^3)$	$O(n^2)$	Preserves distance, diffusion kernel, neighbourhood graph, not explicit
5.	LLE	k	$O(pn^2)$	$O(pn^2)$	Preserves neighbour weights, neighbourhood graph
6.	Laplacian Eigenmaps	k,σ	$O(pn^2)$	$O(pn^2)$	Preserves pairwise distance, neighbourhood graph
7.	LTSA	k	$O(pn^2)$	$O(pn^2)$	Local tangent space, nearest neighbours, not explicit
8.	Autoencoders	network size	O(inw)	O(w)	Neural network, unsupervised

For properties 2 and 3, Table 4.1 provides insight into the computational and memory complexities of the computationally most expensive algorithmic components of the techniques. The computational complexity of a DR technique is of importance to its practical applicability. If the memory or computational resources needed are too large, application becomes infeasible. The computational complexity of a dimensionality reduction technique is determined by: (i) properties of the dataset such as the number of datapoints n and their dimensionality D, and (ii) by parameters of the techniques, such as the target dimensionality d, the number of nearest neighbours k (for techniques based on neighbourhood graphs) and the number of iterations i (for iterative techniques). In Table 4.1, p denotes the ratio of nonzero elements in a sparse matrix to the total number of elements, and w is the number of weights in a neural network.

From the discussion of general properties of the techniques for DR above it can be observed that,

1) Most of the nonlinear techniques for DR do not provide a parametric mapping between the high dimensional and the low dimensional space.

(2) All nonlinear techniques require the optimization of one or more free parameters.

(3) When $D < n$ (which is true in most cases), nonlinear techniques have computational disadvantages compared to PCA.

(4) Majority of nonlinear techniques suffer from a memory complexity that is square or cube with the number of datapoints n.

From these observations, it is clear that nonlinear techniques impose considerable demands on computational resources, as compared to PCA.

4.4 CRITICAL FACTORS AFFECTING THE CHOICE OF AN APPROPRIATE DR TECHNIQUE

The characteristics of DR techniques to be selected for a particular application depends on the following factors.

- Linearity: A transformation can be linear or nonlinear. Linear projections are easy to understand and use but cannot capture well sample distributions spread on complex manifolds. Nonlinear projections are better for such datasets, however harder to handle with respect to parameters.
- Input type: A DR technique can accept either a distance matrix or high dimensional samples themselves as input. When samples are available, one can always derive a distance matrix from them, but not conversely.
- Neighbourhood: A DR technique claims to preserve local or global neighbourhoods. Local neighbourhood methods try to preserve distances between a point and its neighbours leading to better cluster separation, distances between clusters in the projected space is not retained. Global methods try to preserve pair-wise distances, which may result in more faithful projections of the high dimensional space but exhibits less cluster separation.
- Ease of use: Number of free parameters (hyperparameters) that a DR technique exposes to the end user. Additional parameters give more flexibility, however finding optimal settings is harder.
- Computational complexity: Low-complexity methods are best for interactive visual exploration. The final results may not be accurate.
- Out-of-sample: Ability to project new data based on earlier training which can be extremely useful when one wants to study dynamic datasets which adds new samples over time.
- Inverse transform: Ability to map low-dimensional data to the original space, particularly beneficial for explaining patterns in the projection.

- Determinism: Ability to reproduce the results regardless of random seed initialization, suitable when reproducible results are expected.

The end user has to consider all these factors while selecting a DR method for a particular application.

4.5 METHODOLOGY

Classical mineral exploration and geologic mapping techniques utilize physical characteristics of soils and rocks namely temperature, pH, fluid/rock ratio, weathering, geochemical signatures, landforms, etc., to identify minerals (Clark et al.1995). However, due to subtle mineralogical variations, it can be challenging to distinguish minerals that share similar traits and are frequently challenging to map in the field. For this purpose, indirect data such as HRS can be employed for mineral mapping. Numerous studies have been reported in the literature regarding application of HRS for mineral exploration (Kruse et al. 2003; Adep et al. 2016). To identify minerals, absorption features are considered as diagnostic characteristics. Each mineral has unique reflectance and absorption pattern across different wavelength region which helps to identify them uniquely. In the current study, Cuprite dataset with 188 bands has been employed.

4.5.1 Quality measures

It is often too hard for researchers to judge the quality of the resulting embedding by visual inspection. In addition, it cannot be compared against ground truth due to high dimensional nature of HSI data. Therefore, formal measures play a vital role in judging the quality of a given data embedding. Recent research focuses on assessing DR methods from the geometric point of view. The assumption implies that the neighboring points in the input space must be mapped to neighbors in the output space, and vice versa for the inverse mapping. The phenomenon is called “topology preservation”. Recently, few rank-based measures have been proposed with broader applicability which includes mean relative rank errors, trustworthiness and continuity, local continuity meta criterion and the agreement rate metric (Gracia et al. 2014).

In this study, a quality measure based on the co-ranking matrix which is a combination of the above-mentioned metrics is used to evaluate the performance of different DR techniques on hyperspectral data (Lee et al. 2009). The co-ranking matrix is an effective way to capture the changes in ordinal distance. The column wise distances in a distance matrix are replaced by their ranks. The comparison of the ranks in the high and low dimensional spaces is carried out in a systematic way. In a perfect DR, the matrix will only have non-zero entries in the diagonal, if most of the non-zero entries are in the lower triangle, then the process of DR collapsed far away points onto each other; if most of the non-zero entries are in the upper triangle, then it is understood that close points are torn apart. Rank errors and concepts such as neighborhood intrusions and extrusions can be associated with different blocks of the co-ranking matrix.

The high dimensional hyperspectral dataset is represented by, $X = \{x_1, x_2,\ldots\ldots,x_N\} \epsilon R_H$ and low dimensional dataset $Y = \{y_1, y_2,\ldots\ldots,y_N\} \epsilon R_L$. Let δ_{ij} be the distance from x_i to x_j in R_H and d_{ij} be the distance from y_i to y_j in R_L. The rank of x_j with respect to x_i in R_H is given by,

$$\rho_{ij} = |\{k \mid \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}| \quad (4.7)$$

Similarly, the rank of y_j with respect to y_i in low dimensional space is,

$$r_{ij} = |\{k \mid d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}| \quad (4.8)$$

The differences $R_{ij} = r_{ij} - \rho_{ij}$ are the rank errors. The co-ranking matrix C is the histogram of all rank errors and is given by,

$$C_{kl} = |\{(i,j)| \rho_{ij} = k \text{ and } r_{ij} = l\}| \quad (4.9)$$

Pairs of points which change their rank between the original data and its projection are considered as errors which results in non-zero off-diagonal entries in the co-ranking matrix. A point y_j with $\rho_{ij} > r_{ij}$ is termed as intrusion and $\rho_{ij} < r_{ij}$ is termed as extrusion.

The un-weighted sum of C is expressed as a quality (Mokbel et al. 2018),

$$Q_{NX}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} C_{kl} \quad (4.10)$$

where K defines the neighbourhood points.

4.5.2 Similarity metrics

The ranking matrix of high dimensional data points (input ranking matrix) transforms into the ranking matrix of low dimensional points (output ranking matrix) as a result of the change in distances between data points in the DR process. Since the ranking matrices may be considered as 2D images, the degree of similarity between the input and output ranking matrices can be measured using image similarity techniques. The entropy and mutual information of the probability distribution defined over the joint histogram of ranking matrices are extensively used measurements of image similarity (Babaee et al. 2013). A joint probability distribution $P(i,j)$ is defined over the co-ranking matrix C by,

$$P(i,j) = \frac{1}{N-1} C \tag{4.11}$$

Therefore, the entropy H and mutual information (MI) is given by,

$$H = -\sum_i \sum_j P(i,j) log P(i,j) \tag{4.12}$$

$$MI = \sum_i \sum_j P(i,j) log \frac{P(i,j)}{P(i)P(j)} \tag{4.13}$$

If a particular DR technique preserves more structural information, then the value of MI is maximum.

4.5.3 Loss of Quality

The loss of quality concept could be seen as a simple way of referring to the process of losing the original data geometry associated with a reduction in the data dimensionality, when using a DR algorithm. The rationale for using the concept is to emphasise the loss of quality that occurs in a DR process, rather than the value itself obtained by a quality measure (Gracia et al. 2014).

To achieve this, the loss of quality is quantified when reducing the dimensionality of the data over a pre-specified dimensional range. The loss of quality concept is defined as,

Quality Loss = (1 - quality value) (4.14)

where 1 represents a perfect preservation of geometry, and the quality value is the value obtained by a particular quality measure. The domain for quality value is [0,1], where 0 means the worst preservation of geometry and 1 is the best possible result. The loss of quality is the achieved quality value subtracted from 1. Therefore, the smaller loss of quality value, the better preservation of geometry. Figure 4.2 depicts the methodology involved in DR of HSI data.

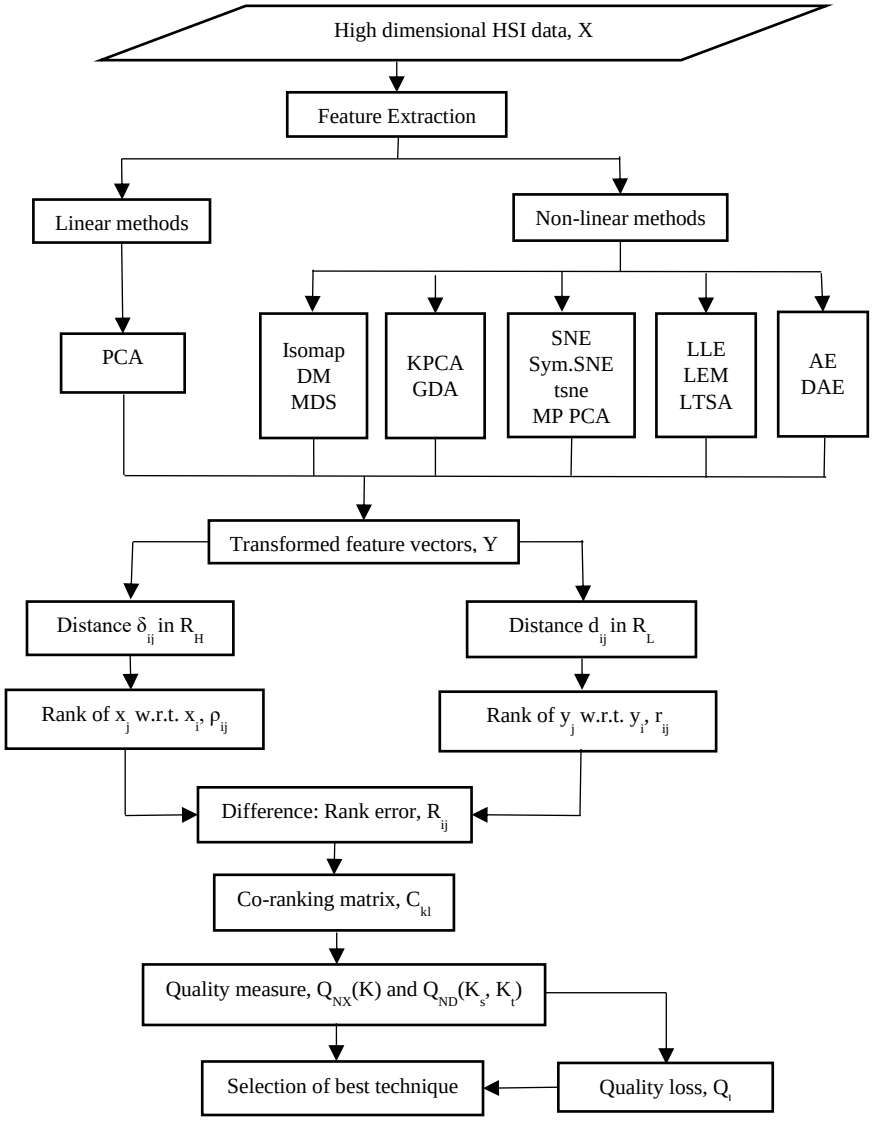


Figure 4.2 Flowchart for DR of HSI data

The primary step includes collection of pre-processed hyperspectral data, X. Various linear and nonlinear DR techniques are applied on the data to address the problem of curse of dimensionality. In the next step, the transformed feature vectors, Y are obtained for each technique. The distance between two points i,j is calculated in high dimensional space R_H as δ_{ij} and in low dimensional space R_L as d_{ij}. Based on this, the ranks are calculated and denoted as ρ_{ij} and r_{ij} respectively. The difference between these two ranks is calculated, which indicates the rank error, R_{ij}. The histogram of all these rank errors is used to obtain co-ranking matrix, C_{kl}. The unweighted sum of the matrix is expressed as a quality, Q_{NX} (K). For better visualisation, this quality metric is replaced by Q_{ND} (K_s,K_t). Based on the results, quality loss

Q_l is calculated which plays a major role in selecting the best technique for a particular application.

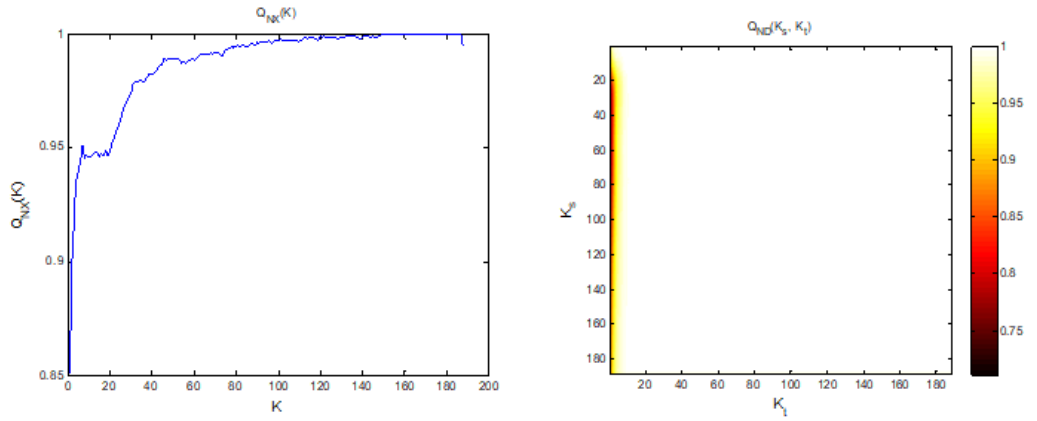
4.6 RESULTS AND DISCUSSION

The qualitative and comparative assessment of 15 DR techniques has been carried out. All the experiments are carried out in MATLAB using the DR toolbox proposed by Laurens Van Der Maaten et al. 2009. The spectral dimensions have been reduced to 10. The results are displayed in terms of quality of embedding. A curve of $Q_{NX}(K)$ is plotted for fixed range of K. For better visualisation, a single parameter K is replaced by the pair (K_s, K_t), where K_s determines the region of interest and K_t is the size of tolerated rank errors which results in a new quality measure $Q_{ND}(K_s, K_t)$ (Mokbel B et al. 2018). Hence, the results can now be characterized by a surface rather than a single curve. The output is displayed as a coloured matrix and (K_s, K_t) is assigned a colour value based on $Q_{ND}(K_s, K_t)$.

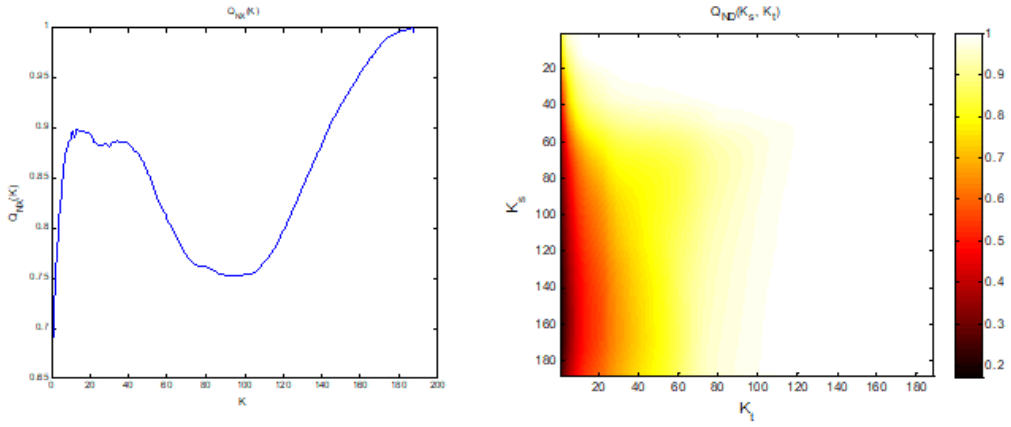
Since similar results are obtained, for simplicity purposes only 6 out of 15 results are displayed. Figure 4.3 (a) – (f) depicts the quality measure $Q_{NX}(K)$ and $Q_{ND}(K_s, K_t)$ for PCA, Isomap, t-SNE, KPCA, deep autoencoders and LEM respectively. As shown in the figure, for $Q_{NX}(K)$ the curve steadily rises to the maximum. A perfect embedding results in a Q value of 1 which indicates global errors are minimum. Furthermore, $Q_{ND}(K_s,K_t)$ provides better visual interpretation compared to $Q_{NX}(K)$. The perfect embedding with quality index value 1 is represented by a white surface. The off-diagonal entries of the co-ranking matrix have to be strictly zero in ideal case. However, it is not true due to intrusions and extrusions induced by rank errors. For a smaller region, the errors are minimum.

The best algorithm for the given data is deep autoencoders followed by PCA, MDS, Autoencoders. The approach with better quality, less sensitivity, and faster processing time is a good global DR method. It is interesting to note that, despite the recent advancement of other DR techniques, PCA and deep autoencoders are still regarded as state-of-the-art when weighing accuracy, speed, and robustness. LLE, LTSA also provides similar results. KPCA and GDA are computationally complex compared to LLE and LTSA. Diffusion maps do not provide better results on this dataset as it works well only on well sampled and clean data. Hence, it is not a better choice when data lies on multiple sub-manifolds. The computation cost of t-SNE is high. Global structure is not specifically maintained. Initializing data points with PCA can assist to solve the issue. LLE and LTSA are computationally efficient than Isomap,

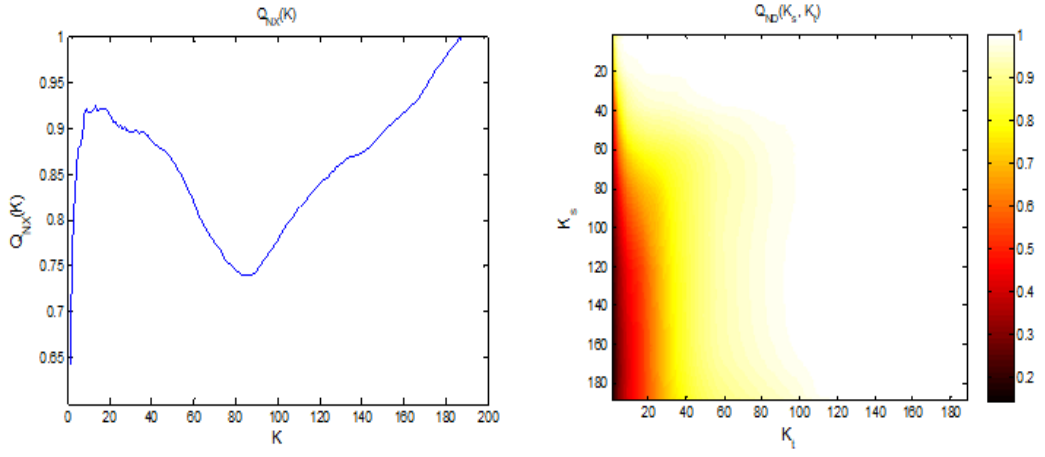
Diffusion maps. t-SNE is a widely used method for data visualisation; if t-SNE fails other algorithms for instance Isomap, LEM, etc. can be employed. Unless the data is significantly non-linear, it can be inefficient for large data and is definitely not an optimal solution. Autoencoders are capable of extracting features at various levels and can be trained to generate data or denoise data.



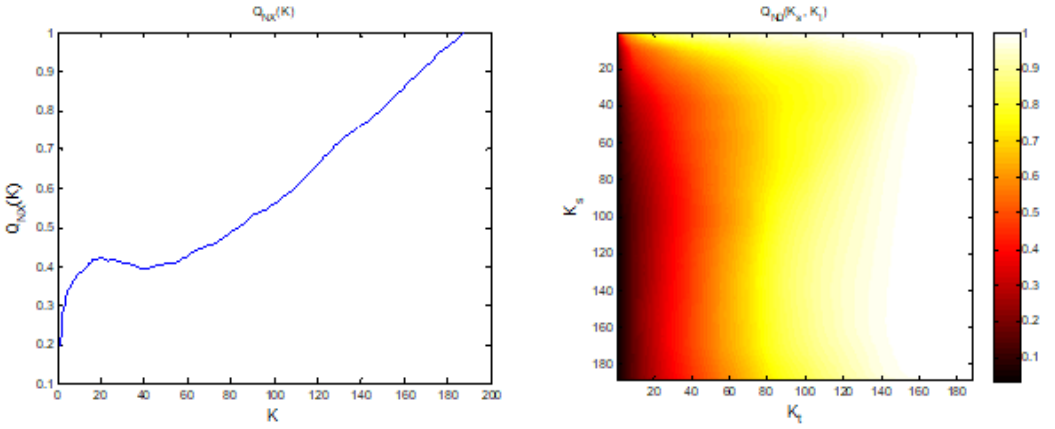
(a) PCA



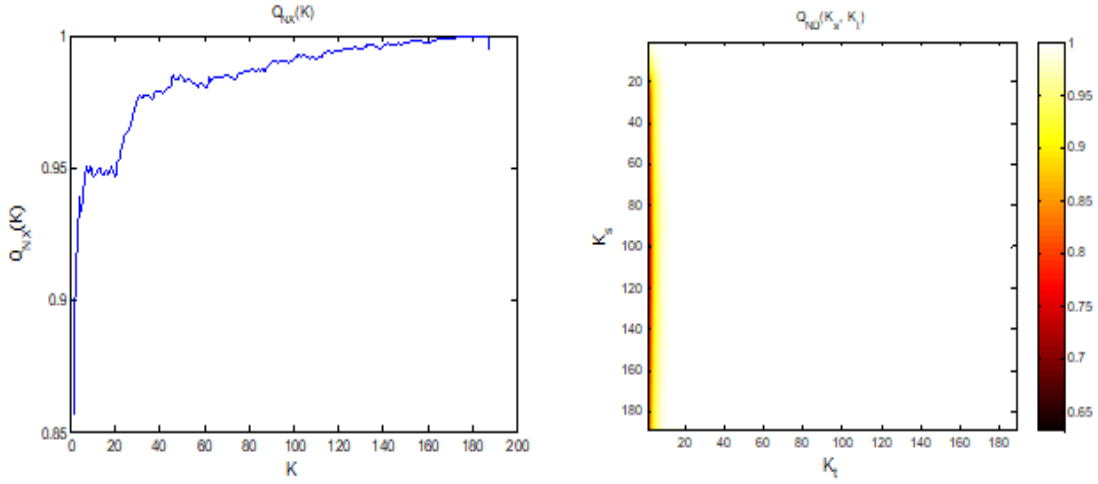
(b) Isomap



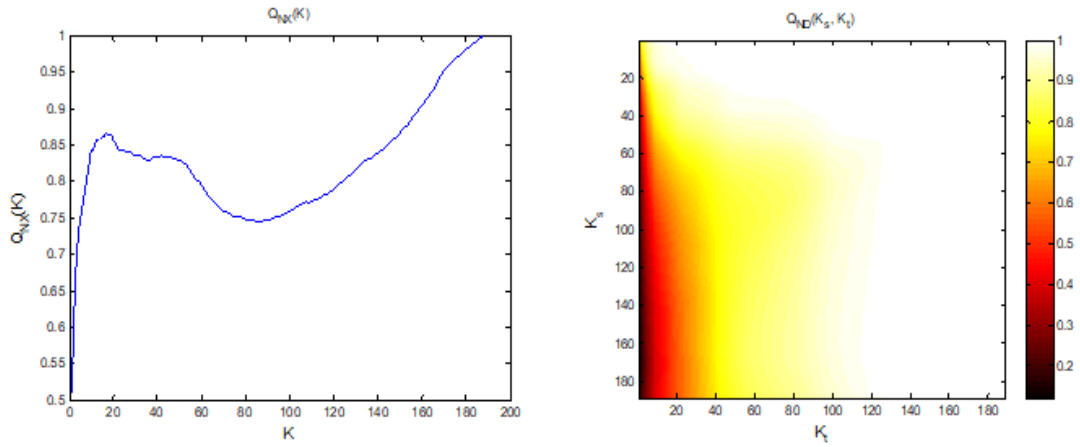
(c) t-SNE



(d) KPCA



(e) Deep autoencoders



(f) Laplacian Eigenmaps

Figure 4.3 $Q_{NX}(K)$ and $Q_{ND}(K_s,K_t)$ of few DR techniques

To obtain a better insight of these results, Table 4.2 provides the quality values $Q_{NX}(K)$ for different values of K and Table 4.3 provides quality loss Q_l values for K = 120. For each of the 15 DR approaches, the Normalized Mutual Information (NMI) and entropy of the co-ranking matrix are calculated. In addition, clustering is performed using K-Means on the original dataset with 10 clusters, accuracy is calculated and the results are included in Table 4.3.

For better visualization, Figure 4.4 represents quality index values of 6 DR techniques for different values of K and Figure 4.5 represents quality loss of 15 DR techniques for K = 120. From the experimental results it is clear that, deep autoencoders provides good embedding from $K \geq 40$. The value of $Q_{NX}(K)$ is 0.9977 for K = 160 which is almost close to 1. The loss of quality is also less for deep autoencoders, 0.0062 for K = 120 which is almost close to 0. Since there are more hidden layers, the network can learn features more complicated features inherent in the data which in turn results in a better embedding. Though PCA provides better results, it cannot be accepted as the best technique for this dataset since it is incapable of capturing nonlinear relationships inherent in hyperspectral data. It is noteworthy that PCA is still regarded as state-of-the-art despite the recent advancement of numerous DR approaches. However, if most accurate technique with non-linearity and denoising is required it is better to employ deep autoencoders. The user can choose the method depending on the application, since a single DR method could not provide better embedding for different datasets. From the experimental results, it can be concluded that larger values of MI lead to better clustering performance as DR and clustering are directly related to each other. The visualization of relationship between DR quality and clustering is depicted in Figure 4.6 where NMI provides an R^2 value of 0.7815 and clustering accuracy provides an R^2 value of 0.7691. Hence clustering and DR is positively correlated.

Statistical hypothesis test (t-test) has been conducted on the two pairs of data: $Q_{NX}(K)$-NMI and $Q_{NX}(K)$-Clustering accuracy for two-dimensional data representations. The hypothesis mean difference is equal to 0. The p-value obtained in both the cases is less than 0.05 (significance level). Hence, it can be concluded that there is no significant difference between the two groups of data and they are positively correlated.

Table 4.2 $Q_{NX}(K)$ for different values of K

Sl.No.	Technique	$Q_{NX}(K)$ for different values of K							
		K=20	K=40	K=60	K=80	K=100	K=120	K=140	K=160
1.	Diffusion maps	0.1013	0.1791	0.2719	0.4203	0.5173	0.625	0.7111	0.8503
2.	GDA	0.1891	0.2787	0.3715	0.4658	0.5500	0.6629	0.7769	0.8708
3.	Isomap	0.8936	0.8836	0.8090	0.7614	0.7541	0.8006	0.8832	0.9547
4.	Kernel PCA	0.4226	0.3964	0.4227	0.4883	0.5621	0.6645	0.7623	0.8613
5.	Laplacian EM	0.8566	0.8331	0.7941	0.7485	0.7600	0.7889	0.8338	0.9045
6.	LLE	0.8989	0.8473	0.7684	0.7094	0.6862	0.7412	0.8481	0.9371
7.	LTSA	0.9040	0.8860	0.8021	0.7436	0.7187	0.7625	0.8550	0.9312
8.	MDS	0.9479	0.9820	0.9885	0.9945	0.9971	0.9987	0.9983	0.9995
9.	MP_PCA	0.5947	0.7279	0.8267	0.9028	0.9449	0.9691	0.9824	0.9910
10.	PCA	0.9479	0.9820	0.9885	0.9945	0.9971	0.9987	0.9983	0.9995
11.	SNE	0.9303	0.8891	0.8168	0.8187	0.8548	0.8820	0.8867	0.8989
12.	Sym. SNE	0.1146	0.2152	0.3211	0.4184	0.5245	0.6312	0.7375	0.8534
13.	Autoencoders	0.8965	0.8887	0.8384	0.8463	0.8780	0.9113	0.9364	0.9765
14.	Deep AE	0.9463	0.9793	0.9808	0.9870	0.9919	0.9938	0.9957	0.9977
15.	t-SNE	0.9141	0.8878	0.8191	0.7455	0.7792	0.8389	0.8733	0.9171

Table 4.3 Loss of Quality Q_l for K=120 along with NMI and Clustering accuracy values

Sl.No.	Technique	Q_l for K=120	NMI	Accuracy
1.	Diffusion maps	0.3750	0.4018	0.40015
2.	GDA	0.3371	0.1553	0.45183
3.	Isomap	0.1994	0.4357	0.51342
4.	Kernel PCA	0.3355	0.3658	0.44176
5.	Laplacian EM	0.2111	0.3247	0.49382
6.	LLE	0.2588	0.4257	0.42176
7.	LTSA	0.2375	0.3864	0.46153
8.	MDS	0.0013	0.7585	0.71236
9.	MP PCA	0.0309	0.4125	0.68512
10.	PCA	0.0013	0.7585	0.72253
11.	SNE	0.118	0.3321	0.5743
12.	Sym. SNE	0.3688	0.3654	0.41376
13.	Autoencoders	0.0887	0.5615	0.61382
14.	Deep AE	**0.0062**	**0.7655**	**0.74921**
15.	t-SNE	0.1611	0.3147	0.54721

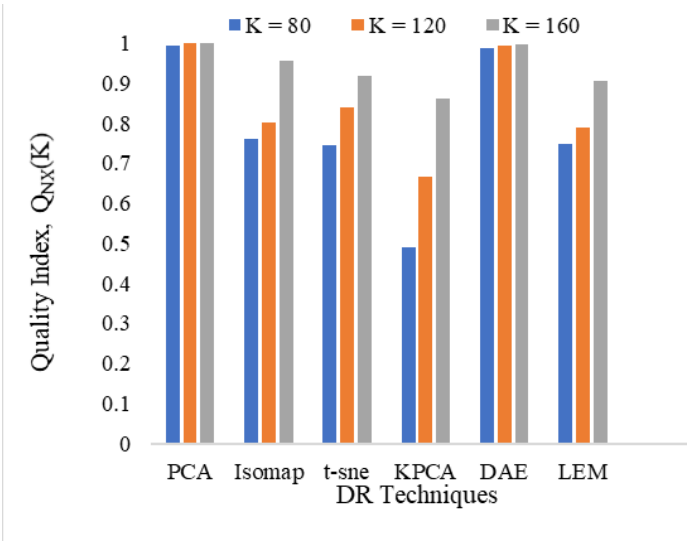


Figure 4.4 $Q_{NX}(K)$ of few DR techniques for different values of K

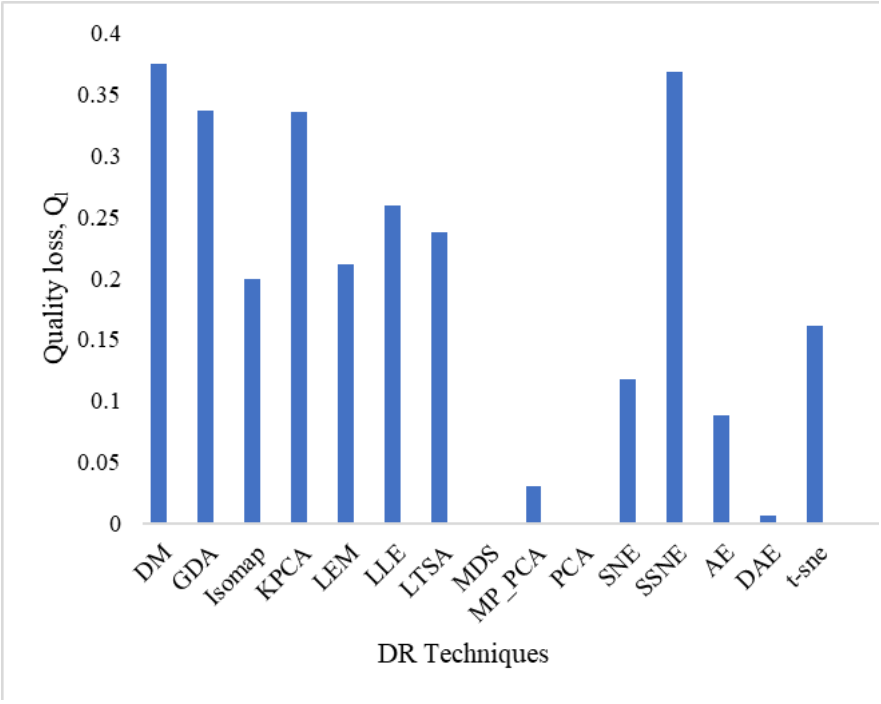


Figure 4.5 Quality loss, Q_l for K = 120

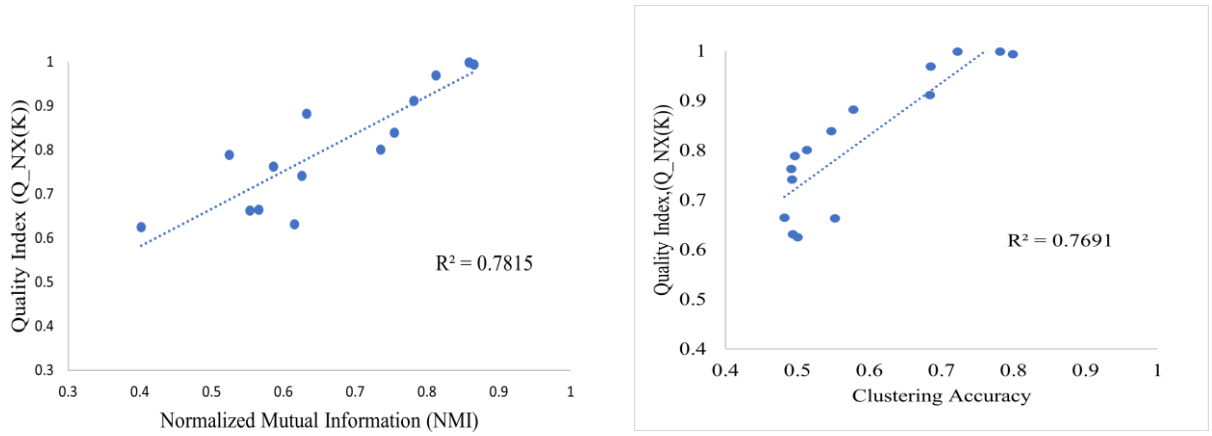


Figure 4.6 Quality index vs clustering accuracy and NMI

In the next chapter, a hybrid CNN model with different FE strategies and dimensions has been proposed to analyze the impact of DR on hyperspectral data.

CHAPTER 5
HYBRID DR TECHNIQUES

5.1 INTRODUCTION

As discussed in the previous chapter, though several techniques exist for DR, most of them concentrates on raw spectral information ignoring the spatial and spectral relationships across the bands. Hence, fails to extract discriminative features for classification. In order to obtain significant discriminating spectral-spatial features for HSIC, it is intended to combine the strengths of 3D and 2D CNN. To be more precise, the additional 2D convolution can combine the 3D convolution generated feature maps, providing richer information while also reducing the size of the spectral bands (Roy et al. 2019). In addition, the majority of the studies utilize PCA as a standard pre-processing technique for DL models without exploring the capabilities of other well-known DR techniques in HSIC. In the current study, DR is used as a pre-processing step before FE and the effects of various cutting-edge DR algorithms on the performance of the hybrid CNN model is investigated in detail.

5.2 METHODOLOGY

This section provides a detailed description of DR techniques and hybrid CNN used in the study and flow diagram of the proposed scheme.

5.2.1 DR techniques

Dealing with high dimensional feature space is challenging due to problems associated with the analysis, visualization, and training of the model (Deepa et al. 2020). Consequently, DR is a crucial stage in the classification pipeline since hyperspectral images have a stronger spectral redundancy than spatial ones. The process saves computational/storage resources while analyzing features and also avoids overfitting. Hyperspectral image X is represented as a 3-D cube, $X = [x_1, x_2, x_3, \ldots, x_D]^T$ of size $(M \times N) \times D$, where M and N indicates the spatial width and height of the image and D denotes the number of spectral bands (Mohan et al. 2020). The ground truth of the input image G is coded using one-hot encoding and represented as $G = [g_1, g_2, \ldots g_C]$, where C denotes the number of ground truth classes. Suppose $(x_i, g_i) \in (\mathbf{R}^{(M \times N) \times D}, \mathbf{R}^{(G)})$, where $x_i = [x_{1i}, x_{2i}, x_{3i}, \ldots\ldots, x_{Di}]^T$ is the i^{th} sample in X and g_i is the class label of the i^{th} sample in G. Different DR algorithms are applied on X to reduce spectral redundancy to Y =

$[y_{1i}, y_{2i}, y_{3i}, \ldots, y_{Bi}]^T$ and the dimensions vary to (M×N)×B where B << D. The spatial dimensions of the input image are retained after DR while the spectral dimension reduced from D to B, the desired number of bands. In this work, the performance of the following DR approaches for the hybrid CNN model is evaluated.

i) Principal Component Analysis (PCA): PCA is one of the most extensively used methods for data processing, compression, and visualization. Principal components are generated by linearly combining the input variables. The combinations are made such that the new variables (i.e., principal components) are uncorrelated and the majority of the data from the input variables are crammed into the first few components. Principal components represent the directions of the data that explain a maximal amount of variance. The transformation is given by, $y_i = W^T x_i$, where W is the transformation matrix (Jiang et al. 2018). The linear transformation matrix $\widehat{W}$ is obtained by solving the following objective function as follows,

$$\widehat{W} = argmax_{W^T W = I} Tr(W^T Cov(X) W) \quad (5.1)$$

where Cov(X) is the covariance matrix of X and Tr(X) is the trace of matrix X.

ii) Fast Independent Component Analysis (FastICA): FastICA is an efficient and quickest algorithm for ICA which is capable of disentangling every single individual signal from a mixed signal. The goal is to formulate a linear combination of independent source signals from a multivariate random measurement signal (Ibarrola Ulzurrun et al. 2017). Unlike PCA, ICA reduces higher order statistical errors as well as decorrelates second order statistical dependencies, maximizing the degree of signal independence (Boukhechba et al. 2018). Consider an observation vector is given by, $x = [x_0, x_1, x_2, \ldots, x_{q-1}]$ as a linear mixture of Q independent elements of random source, $s = [s_0, s_1, s_2, \ldots\ldots, s_{q-1}]$. In matrix form, the model is given by, $X = A.S$ where, A is the mixing matrix. ICA estimates the unmixing matrix W, inverse of A to get best approximation of S as, $Y = W.X \approx S$.

iii) Sparse Random Projection (SRP): A sparse random matrix whose column vector is unit length is employed to project the original input space to reduce the dimensionality. Dense Gaussian random projection matrices can be replaced by sparse random matrices, which ensure comparable embedding quality while using significantly less memory and processing projected data more quickly. The Johnson-Lindenstrauss (JL) lemma is satisfied by the SRP algorithm (Jia et al. 2022). Prior knowledge of the original data is not required to implement the

algorithm. Let spatial dimension (M×N) be represented by n, then $R^{D\times B}$ is the random matrix used for projection as, $Y^{n\times B} = X^{n\times D}R^{D\times B}$.

iv) Factor Analysis (FA): FA is a multivariate statistical method similar to PCA, developed to extract potential factors from observed variables for data reduction (Tarnas et al. 2021). It is a very useful method for high-dimensional data generation models since it allows different regions in the input space to build a model of local factor data. FA has several advantages over PCA, the ability to individually model variance in each direction of the input space.

v) Non-negative Matrix Factorization (NMF): An alternate method of decomposition, NMF assumes that the data and the components are both non-negative (Menon et al. 2018). When the data matrix does not contain negative values, NMF can be substituted for PCA or its variants. In contrast to PCA, the representation of a vector is created by superimposing the components rather than by deleting any of them. These additive approaches for representing images are effective.

5.2.2 Hybrid CNN

Initially, multiple overlapping 3D patches are created from the input hyperspectral image X of size M×N×B. The class labels of these pixels are labelled based on the label of the central pixel. The 3D patches, $A \in \mathrm{R}^{(W\times W)\times B}$ that are centred at spatial location (p, q) and cover the $W \times W$ windows are generated. (M−W+1) (N− W + 1) is the total number of 3D spatial patches formed from X. The 3D patches represented by A(p,q), span the width from p−(W− 1)/2 to p+(W−1)/2 and the height from q−(W−1)/2 to q+(W−1)/2 as well as B spectral bands obtained after the DR.

CNN is a multilayer neural network that consists of a convolution layer, a pooling layer, and a fully connected layer. The essential component of CNN is the convolution layer, which performs convolution operations on the receptive field (input) and kernel (learnable parameters). The output of individual neuron y, for input x is given by, $y = f(w * x + b)$ where, filter weight is w and the bias is b. The nonlinear activation applied to a weighted sum of input is represented by f (.). Before activation, the 2D CNN model convolves the input data using a two-dimensional kernel to extract spatial features from the input. Reframing the above equation for each neuron for 2D convolution, the activation value at spatial position (x,y) in the jth feature map of ith layer $v_{i,j}^{x,y}$ is given by eq. (5.2),

$$v_{i,j}^{x,y} = F\left(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho,\lambda} \times v_{i-1,\tau}^{x+\sigma,y+\rho}\right) \quad (5.2)$$

where F represents activation function, $b_{i,j}$ is the bias parameter, d_{l-1} is the number of feature maps in $(l-1)th$ layer and depth of kernel $w_{i,j}$, $2\gamma + 1$ and $2\sigma + 1$ are the width and height of the kernel respectively. In contrast, the 3D convolutional method first computes the sum of the dot product between the input patches and the 3D kernel function, implying that the 3D input patches are convolved with the 3D kernel function. The feature maps generated are passed through activation function to induce non-linearity. The hybrid model builds the 3D convolutional layer feature maps in the input layer by employing the 3D kernel function over B spectral bands retrieved after DR. The activation value at spatial location (x, y, z) at the ith layer and jth feature map in a 3D convolutional process can be expressed by eq. (5.3),

$$v_{i,j}^{x,y} = F\left(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\lambda=-v}^{v} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho,\lambda} \times v_{i-1,\tau}^{x+\sigma,y+\rho,z+\lambda}\right) \quad (5.3)$$

where all the parameters are same as explained above, except $2v + 1$ is the depth of the kernel along spectral dimension. Figure 5.1 shows the flow diagram of the proposed hybrid CNN. The architecture includes three 3D convolutional layers and a single 2D convolutional layer. The specifications of 3D convolutional kernels used for joint spectral-spatial features are 3D_Conv_layer1 = 8×3×3×7×1, 3D_Conv_layer2 = 16×3×3×5×8 and 3D_Conv_layer3 = 32×3×3×3×16 which implies 32 3D kernels of dimensions (3×3×3) both spatial and spectral for all16 3D input feature maps. 2D_Conv_layer1 = 64×3×3×96 is used to obtain discriminative features based on spatial information only. The proposed model incorporates DR stage and 3D, 2D convolutional layers in a systematic way to achieve maximum accuracy. ReLU is faster compared to other activation functions and its equation is given by $ReLU(x) = \max(0, x)$. Hence ReLU with a learning rate of 0.001 is used in all layers except softmax on the final layer. The network includes two fully connected layers with 256 and 128 nodes which is responsible for mapping features to the output. The base model proposed by Roy et al. 2019 is investigated in this work using numerous DR approaches to further explore DR method that works best in different scenarios. The weights are initially randomized before being optimized using back-propagation with the Adam optimizer and the Softmax loss function. The optimizers need the model hyper-parameters, including the learning rate utilized in optimization to be carefully initialized and adjusted. The learning rate hyper-parameter regulates network weights tuning in relation to the loss gradient. The Adam optimizer used in

the study provides a number of benefits including working with sparse gradients, naturally carrying out a type of step size annealing and invariant parameter updates to a rescaling of the gradient. The spatial dimensions of 3D input patches for all datasets are set to 9×9×15, 11×11×15, 9×9×27, 11×11×27, 9×9×33, 11×11×33, and 9×9×39, 11 ×11×39, where 15, 21, 27, 33, and 39 are the number of most informative bands extracted by PCA, FastICA, SRP, FA, and NMF, respectively. The network is trained for 50 epochs using a 256 mini-batch size without batch normalization and data augmentation. The model summary of the proposed scheme for the IP dataset with PCA as DR technique and window size 9×9, 15 bands is presented below in Table 5.1. The number of trainable parameters is less (127,104) compared to the base model (5,122,176).

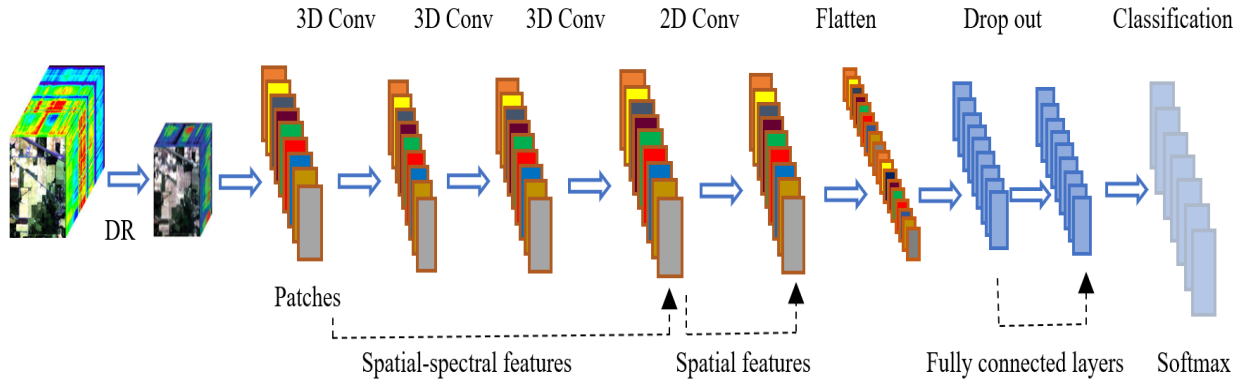


Figure 5.1 Flow diagram of the proposed hybrid CNN including DR stage and 3D-2D convolutions

Table 5.1 Model summary of IP dataset using PCA with 9×9 and 15 bands

Layer (type)	Output Shape	# Parameters
input_1 (InputLayer)	(9, 9, 15, 1)	0
conv3d (Conv3D)	(7, 7, 9, 8)	512
conv3d_1 (Conv3D)	(5, 5, 5, 16)	5776
conv3d_2 (Conv3D)	(3, 3, 3, 32)	13856
reshape (Reshape)	(3, 3, 96)	0
conv2d (Conv2D)	(1, 1, 64)	55360
flatten (Flatten)	(64)	0
dense (Dense)	(256)	16640
dropout (Dropout)	(256)	0

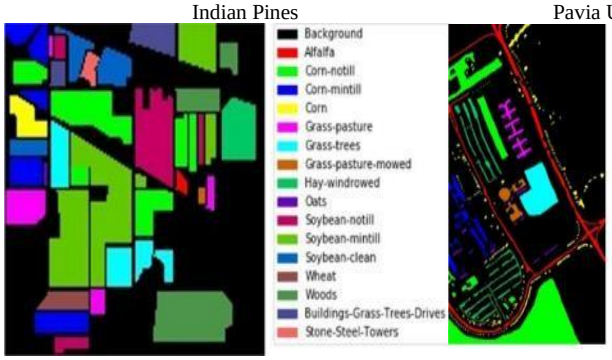
dense_1 (Dense)	(128)	32896
dropout_1 (Dropout)	(128)	0
dense_2 (Dense)	(16)	2064
Total Trainable parameters: 127,104		

5.2.3 Experimental setup

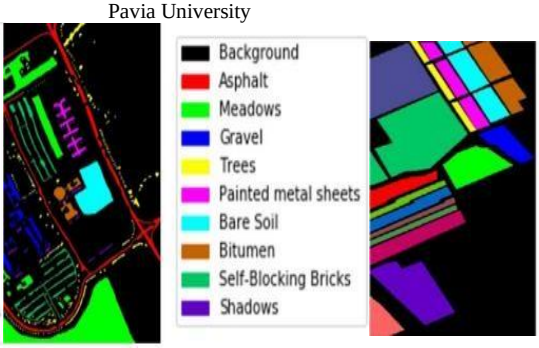
All experiments were carried out in Google Colab which is a cloud-based Jupyter notebook environment. The ground truth image with labels, number of training, validation and testing samples for various datasets is given below in Table 5.2. Each dataset is separated into a training set and a test set in a 50/50 ratio for all the experiments and the training set is then further divided into training samples and validation samples in a 50/50 ratio.

Table 5.2 Ground truth image with labels and number of samples for various datasets

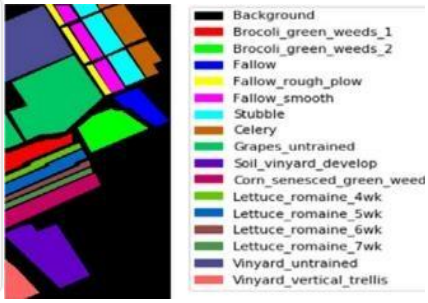
Indian Pines



Pavia University



Salinas



	Class	Tr, Val, Te	Class	Tr, Val, Te	Class	Tr, Val, Te
1.	Alfalfa	11, 12, 23	Asphalt	1657,1658,3316	Brocoligreenweeds1	502,503,1004
2.	Cornnotill	357, 357, 714	Meadows	4662,4663,9324	Brocoligreenweeds2	932,931,1863
3.	Cornmintill	208, 207, 415	Gravel	525,525,1049	Fallow	494,494,988
4.	Corn	59, 60, 118	Trees	766,766,1532	Fallowroughplow	348,349,697
5.	Grasspasture	121, 120, 242	Painted metal sheets	336,336,673	Fallowsmooth	669,670,1339
6.	Grasstrees	183, 182, 365	Bare soil	1258,1257,2514	Stubble	990,989,1980
7.	Grasspasturemowed	7, 7, 14	Bitumen	333,332,665	Celery	895,894,1790
8.	Haywindrowed	120, 119, 239	Self-Blocking bricks	921,920,1841	Grapesuntrained	2818,2817,5636
9.	Oats	5, 5, 10	Shadows	236,237,474	Soilvinyarddevelop	1550,1551,3102
10.	Soybeannotill	243, 243, 486	-	-	Cornsenescedgreenweed	820,819,1639
11.	Soybeanmintill	614, 613, 1228	-	-	Lettuceromaine4wk	267,267,534
12.	Soybeanclean	148, 148, 297	-	-	Lettuceromaine5wk	482,482,963
13.	Wheat	51, 52, 102	-	-	Lettuceromaine6wk	229,229,458
14.	Woods	316, 316, 633	-	-	Lettuceromaine7wk	267,268,535
15.	BuilGraTreDrives	96, 97, 193	-	-	Vinyarduntrained	1817,1817,3634
16.	StoneSteelTowers	23, 24, 46	-	-	Vinyardverticaltrellis	452,452,903

5.2.4 Evaluation parameters

To evaluate the performance of the HSI classification, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ) measures are used. OA measures the proportion of correctly classified samples to all tested samples. The confusion matrix generated for the classification, a C×C square matrix where C denotes the number of ground truth classes is used to calculate OA using eq. (5.4),

$$OA = \sum_{i=1}^{C} \frac{t_{ii}}{t} \tag{5.4}$$

where, t_{ii} is the proportion of test samples in class i that are correctly classified as belonging to the same class which represents the diagonal elements of the confusion matrix. The total number of test samples are denoted by t. The average of class-wise accuracy (CA) is referred to as AA. From the confusion matrix, CA is determined by eq.(5.5),

$$CA_i = \frac{t_{ii}}{\sum_{j=1}^{C} t_{ij}} \tag{5.5}$$

The class-wise accuracy of class i is indicated by CA_i and the number of samples in class i that were correctly classified into class j is indicated by t_{ij}. Therefore, CA is the proportion of correctly classified samples in class i to the total number of test samples in the same class. Hence AA can be defined as shown below eq. (5.6),

$$AA = \frac{\sum_{i=1}^{C} CA_i}{C} \tag{5.6}$$

The Kappa statistic (κ) is the difference between the row-wise and column-wise sums of the confusion matrix, which is the chance of agreement of the classified result and the chance of agreement of the actual result. κ value ranges between -1 and +1, and a more accurate classification model moves κ value towards 1. Let t_{i+} represent the total of the row elements in the confusion matrix for class i (where $\sum_{j=1}^{C} t_{ij}$) and t_{+i} represents the sum of the column elements($\sum_{i=1}^{C} t_{ij}$). The calculation of the Kappa statistic value is given by,

$$\kappa = \frac{t\sum_{i=1}^{C} t_{ii} - \sum_{i=1}^{C} t_{i+}t_{+i}}{n^2 - \sum_{i=1}^{C} t_{i+}t_{+i}} \tag{5.7}$$

In addition, to evaluate the statistical significance the following measures are used.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{5.8}$$

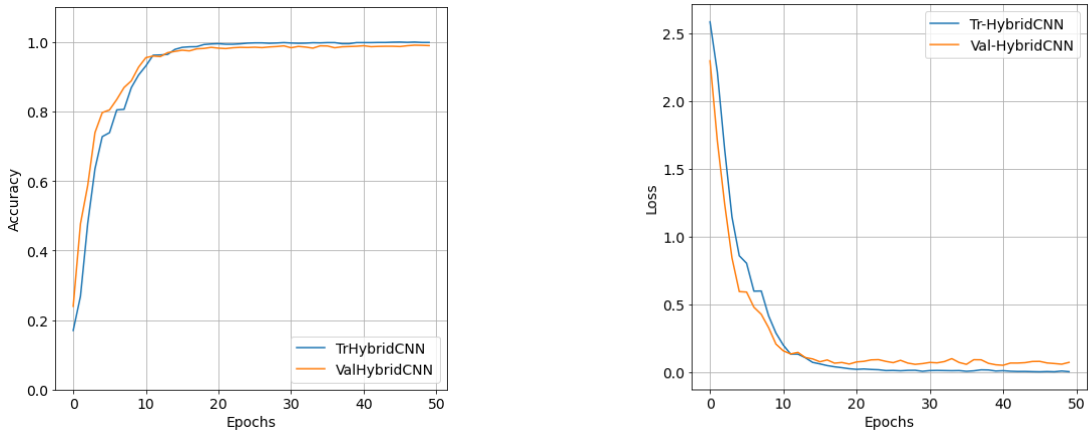
$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5.9}$$

$$F1 = 2 * \frac{precision \times recall}{precision + recall} \tag{5.10}$$

5.3 RESULTS AND DISCUSSION

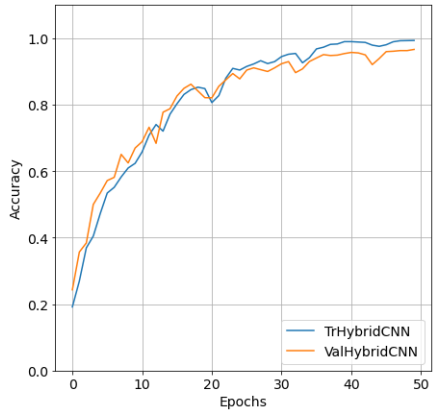
To validate the performance of the proposed hybrid CNN based land cover classification framework and to explore the impact of DR, experiments were conducted on benchmark hyperspectral datasets. Three publicly available datasets collected from Air-borne Visible/Infrared Imaging Spectrometer (AVIRIS) and Reflective Optics System (ROSIS) sensors are used in the study namely Indian Pines, Pavia University and Salinas Full scene.

Indian Pines (IP): For analyzing the convergence characteristics of the proposed method, a separate set of experiments are performed by fixing the dimensions as (9,15), (11,15) with PCA and termination condition as 50 epochs, since the model was observed to converge within 40 epochs for all the test images. The accuracy and loss curves included for IP dataset in Figure 5.2-5.3 indicate the efficiency of the proposed model to converge with a minimum number of epochs. Table 5.3 displays the results of the IP dataset for several approaches in terms of OA, AA, and κ while classification maps are depicted in Figure 5.4. In addition, the statistical significance of each class is shown in Table 5.4. To further validate the proposed model for its generic properties, additional experiments were conducted on other data sets with different dimensions and patch size.

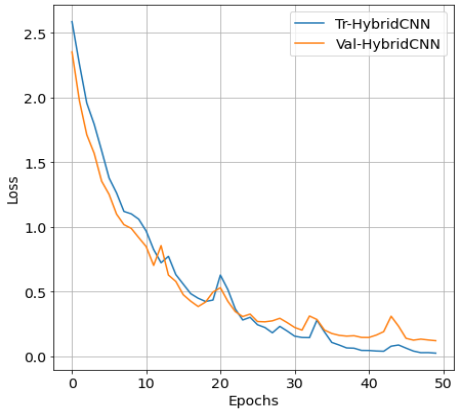


(a)Accuracy, (9,15) (b) Loss, (9,15)

Figure 5.2 Accuracy and loss curves for (9,15)



(a) Accuracy, (11,15)

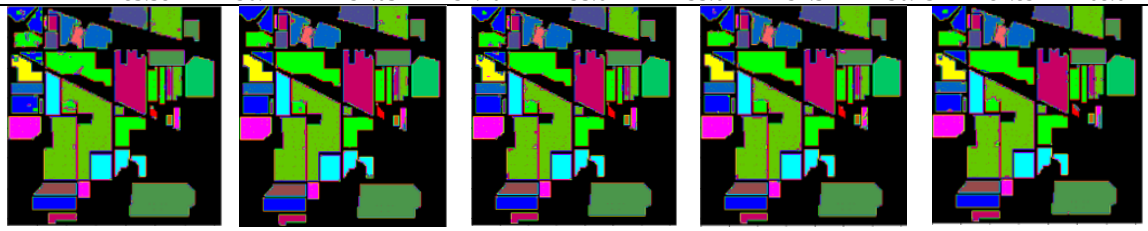


(b) Loss, (11,15)

Figure 5.3 Accuracy and loss curves for (11,15)

Table 5.3 Classification results indicating OA, AA and κ respectively of IP dataset with different DR, dimensions and window size

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	96.93	98.18	97.56	98.36	97.71	98.10	97.32	98.49	97.93	98.51
	96.50	97.93	97.22	98.13	97.39	97.83	96.95	98.28	97.51	98.30
	96.44	97.81	97.87	98.70	96.05	95.61	93.11	96.52	97.04	97.02
Fast ICA	75.31	87.43	74.45	91.02	84.15	93.96	88.95	92.85	92.17	94.96
	71.44	85.59	70.38	89.73	81.84	92.72	87.25	91.32	91.77	94.25
	64.15	72.27	56.59	84.87	68.89	91.25	87.64	91.86	91.88	93.92
SRP	23.96	23.96	23.96	23.96	55.29	24.87	24.0	23.96	23.96	23.96
	0.0	0.0	0.0	0.0	47.90	1.68	0.05	0.0	0.0	0.0
	6.25	6.25	6.25	6.25	41.06	7.96	6.28	6.25	6.25	6.25
FA	98.06	98.84	98.67	98.20	98.45	98.88	98.47	98.96	98.45	99.04
	97.79	98.68	98.48	97.95	98.24	98.73	98.26	98.82	98.24	98.90
	97.00	98.03	95.67	96.97	98.34	97.34	97.35	98.96	97.94	98.00
NMF	87.57	93.67	93.71	98.06	96.42	95.60	95.53	95.21	95.51	89.54
	85.81	92.78	92.83	97.79	95.92	94.99	94.90	94.54	94.88	88.07
	83.58	90.11	94.05	97.10	95.62	95.62	94.51	90.43	84.89	68.61



(a) (15,9) (b) (15,11) (c) (21,9) (d) (21,11) (e) (27,9)

(f) (27,11) (g) (33,9) (h) (33,11) (i) (39,9) (j) (39,11)

Figure 5.4 Classification results of IP for different number of dimensions and patch size

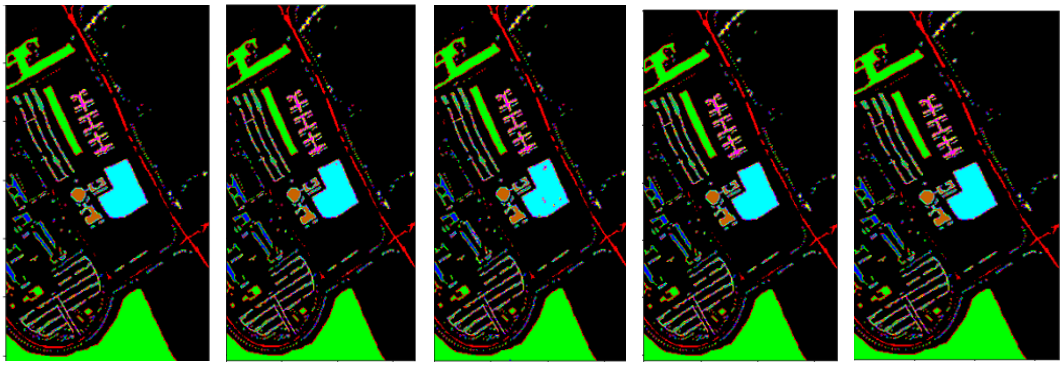
Table 5.4 Statistical significance of IP dataset with PCA

Class	15 Bands			21 Bands			27 Bands			33 bands			39 Bands		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2		
1.	1.00/ 1.00	0.87/ 0.96	0.93/ 0.98	1.00/ 1.00	0.96/ 0.96	0.98/ 0.98	1.00/ 1.00	0.83/ 1.00	0.90/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
2.	0.97/ 0.99	0.92/ 0.97	0.95/ 0.98	0.97/ 0.98	0.97/ 0.99	0.97/ 0.98	0.98/ 0.99	0.97/ 0.98	0.97/ 0.99	0.98/ 0.98	0.96/ 0.99	0.97/ 0.99	0.98/ 0.99	0.97/ 0.98	0.97/ 0.98
3.	0.98/ 0.96	0.96/ 0.99	0.97/ 0.98	0.98/ 0.97	0.98/ 0.99	0.98/ 0.98	0.94/ 0.95	0.98/ 098	0.96/ 0.97	0.92/ 0.98	0.98/ 0.98	0.95/ 0.98	0.98/ 0.99	1.00/ 0.98	0.97/ 0.97
4.	0.97/ 0.93	0.97/ 0.99	0.97/ 0.96	0.96/ 0.99	0.93/ 1.00	0.94/ 1.00	0.97/ 0.97	0.96/ 094	0.97/ 0.96	0.97/ 1.00	0.97/ 0.95	0.97/ 0.97	0.96/ 0.98	0.99/ 0.98	0.97/ 0.96
5.	0.99/ 1.00	0.97/ 0.95	0.98/ 0.97	1.00/ 0.97	0.97/ 0.95	0.98/ 0.96	1.00/ 1.00	0.95/ 0.97	0.97/ 0.98	1.00/ 1.00	0.96/ 0.97	0.98/ 0.98	0.98/ 0.99	0.97/ 0.98	0.96/ 0.97
6.	1.00/ 1.00	0.99/ 0.98	1.00/ 0.99	0.99/ 1.00	0.99/ 0.99	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
7.	0.93/ 0.88	1.00/ 1.00	0.97/ 0.93	1.00/ 0.93	1.00/ 1.00	1.00/ 0.97	1.00/ 0.88	1.00/ 1.00	1.00/ 0.93	0.88/ 1.00	1.00/ 0.86	0.93/ 0.92	0.95/ 0.97	0.96/ 0.98	0.97/ 0.98
8.	0.99/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
9.	0.82/ 1.00	0.90/ 0.90	086/ 0.95	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.80/ 0.90	0.89/ 0.95	1.00/ 1.00	0.90/ 0.90	0.93/ 0.96	1.00/ 1.00	0.97/ 0.96	0.96/ 0.98
10.	0.98/ 0.98	0.93/ 0.95	0.95/ 0.97	0.97/ 0.98	0.94/ 0.97	0.95/ 0.97	0.96/ 0.99	0.95/ 0.96	0.95/ 0.97	0.95/ 0.97	0.96/ 0.98	0.95/ 0.97	0.96/ 0.97	0.97/ 0.99	0.97/ 0.98
11.	0.94/ 0.98	0.99/ 0.99	0.97/ 0.99	0.98/ 0.99	0.97/ 0.98	0.97/ 0.99	0.99/ 0.97	0.98/ 0.99	0.98/ 0.98	0.98/ 0.99	0.97/ 0.99	0.98/ 0.99	0.98/ 0.98	0.97/ 0.97	0.95/ 0.96
12.	0.94/ 0.96	0.97/ 0.97	0.96/ 0.96	0.92/ 0.97	0.99/ 0.98	0.95/ 0.97	0.95/ 098	0.99/ 0.96	0.97/ 0.97	0.94/ 0.97	0.95/ 0.96	0.94/ 0.96	0.98/ 0.99	0.99/ 0.98	0.97/ 0.98
13.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	0.98/ 0.98	0.99/ 0.98	0.99/ 0.98	0.98/ 1.00	0.97/ 0.98	0.98/ 0.99	1.00/ 1.00	0.98/ 0.99	0.99/ 1.00
14.	0.99/ 0.99	0.99/ 1.00	0.99/ 0.99	1.00/ 1.00	0.99/ 0.99	0.99/ 0.99	1.00/ 0.98	1.00/ 100	1.00/ 0.99	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
15.	0.95/ 0.96	0.97/ 0.99	0.96/ 0.98	0.92/ 0.95	0.97/ 1.00	0.95/ 0.97	0.91/ 0.97	0.99/ 0.94	0.95/ 0.95	0.96/ 0.95	0.99/ 1.00	0.97/ 097	0.96/ 0.98	0.98/ 0.99	0.97/ 0.97
16.	0.92/ 0.96	1.00/ 1.00	0.96/ 0.98	0.98/ 0.92	1.00/ 1.00	0.99/ 0.96	0.98/ 0.98	1.00/ 1.00	0.99/ 0.99	0.96/ 0.90	1.00/ 1.00	0.98/ 0.95	0.95/ 0.97	0.93/ 0.95	0.95/ 0.96

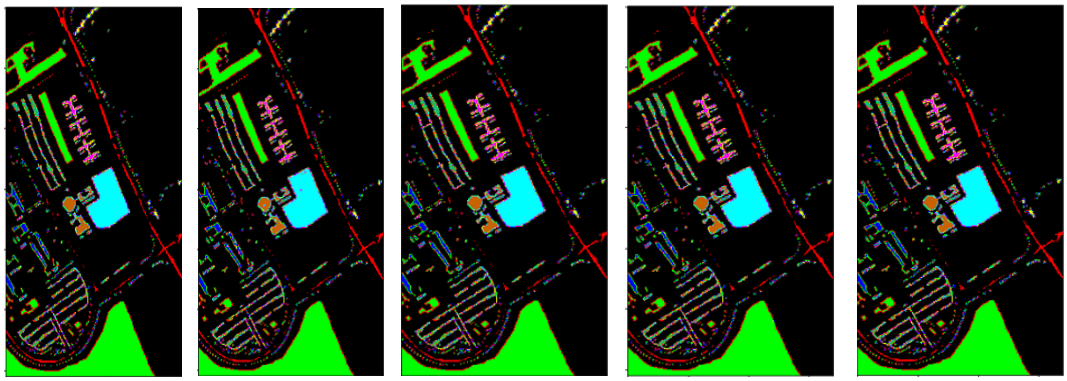
Pavia University (PU): Table 5.5 displays the results of the PU dataset for several approaches in terms of OA, AA, and κ while classification maps are depicted in Figure 5.5. In addition, the statistical significance of each class is shown in Table 5.6.

Table 5.5 Classification results indicating OA, AA and κ respectively of PU dataset with different DR, dimensions and window size

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	99.79 99.72 99.66	99.72 99.64 99.49	98.68 98.25 97.56	99.76 99.45 99.28	99.72 99.63 99.41	99.71 99.62 99.46	99.02 98.71 98.20	99.82 99.76 99.71	99.60 99.47 99.27	99.84 99.78 99.76
Fast ICA	89.40 89.21 88.87	89.32 89.69 89.33	91.59 90.92 90.36	90.04 90.35 90.15	93.21 92.85 92.36	92.88 92.36 92.17	95.36 94.23 94.11	96.85 96.32 96.15	96.55 96.21 96.04	99.40 99.21 98.87
SRP	43.59 0 11.11	43.59 0 11.11	43.59 0 11.11	43.59 0 11.11	82.74 76.50 60.93	83.59 84.22 84.36	70.38 58.93 46.34	91.29 88.45 85.12	89.40 86.05 88.75	96.22 94.98 94.29
FA	99.45 99.28 99.13	99.14 98.86 99.01	99.43 99.25 99.13	99.78 99.71 99.51	99.11 98.82 98.49	99.32 99.10 98.95	98.20 97.61 97.41	99.59 99.46 99.28	99.47 99.30 98.99	99.32 99.10 99.02
NMF	99.28 99.05 98.70	99.34 99.13 98.98	99.47 99.30 99.24	99.66 99.55 99.52	99.50 99.34 99.35	99.31 99.09 99.12	99.29 99.07 98.76	99.85 99.80 99.78	99.70 99.60 99.46	99.78 99.71 99.67



(a) (15,9) (b) (15,11) (c) (21,9) (d) (21,11) (e) (27,9)



(f) (27,11) (g) (33,9) (h) (33,11) (i) (39,9) (j) (39,11)

Figure 5.5 Classification results of PU for different number of dimensions and patch size

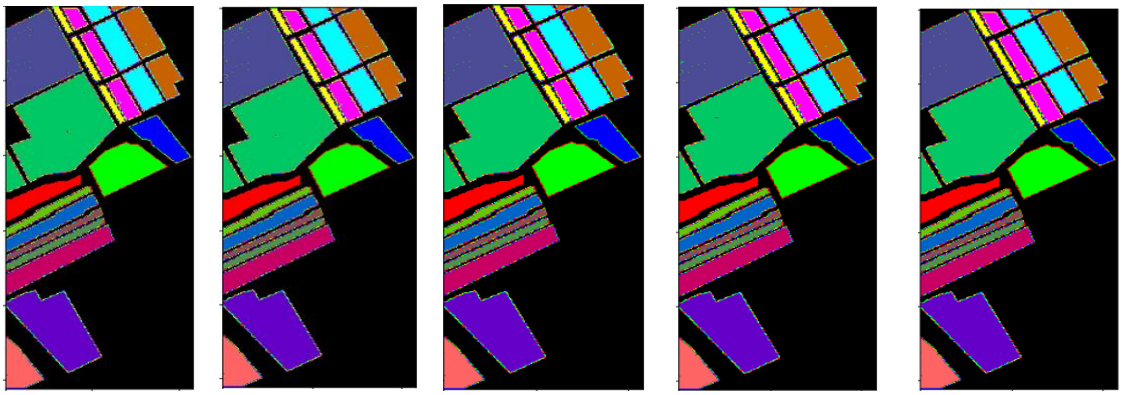
Table 5.6 Statistical significance of PU dataset with PCA

Class	15 Bands			21 Bands			27 Bands			33 Bands			39 Bands		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	W1/W2			W1/W2			W1/W2			W1/W2			W1/W2		
1.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
2.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00/	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
3.	0.99/ 0.99	0.99/ 0.98	0.99/ 0.98	0.91/ 0.99	0.93/ 0.99	0.92/ 0.98	0.99/ 0.99	0.98/ 0.97	0.98/ 0.98	0.94/ 0.99	0.94/ 0.99	0.94/ 0.99	0.98/ 0.99	0.98/ 0.99	0.98/ 0.99
4.	1.00/ 1.00	0.99/ 1.00	0.99/ 1.00	1.00/ 1.00	0.99/ 0.99	0.99/ 0.99	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	0.99/ 1.00	0.98/ 0.99	0.99/ 1.00	1.00/ 1.00	0.99/ 1.00	0.99/ 1.00
5.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
6.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 0.99	0.99/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
7.	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	0.99/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	100/ 1.00	0.99/ 1.00	0.97/ 1.00	0.98/ 1.00	0.99/ 0.99	0.99/ 1.00	0.99/ 1.00
8.	0.99/ 0.99	1.00/ 0.99	1.00/ 0.99	0.94/ 1.00	0.94/ 0.99	0.94/ 0.98	0.99/ 0.98	0.99/ 0.99	0.99/ 0.99	0.96/ 0.99	0.96/ 0.99	0.96/ 0.99	0.98/ 0.99	0.99/ 0.99	0.99/ 0.99
9.	1.00/ 1.00	1.00/ 0.99	1.00/ 0.99	1.00/ 1.00	0.95/ 0.99	0.97/ 0.99	0.99/ 0.99	0.99/ 0.99	0.99/ 0.99	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	0.99/ 1.00

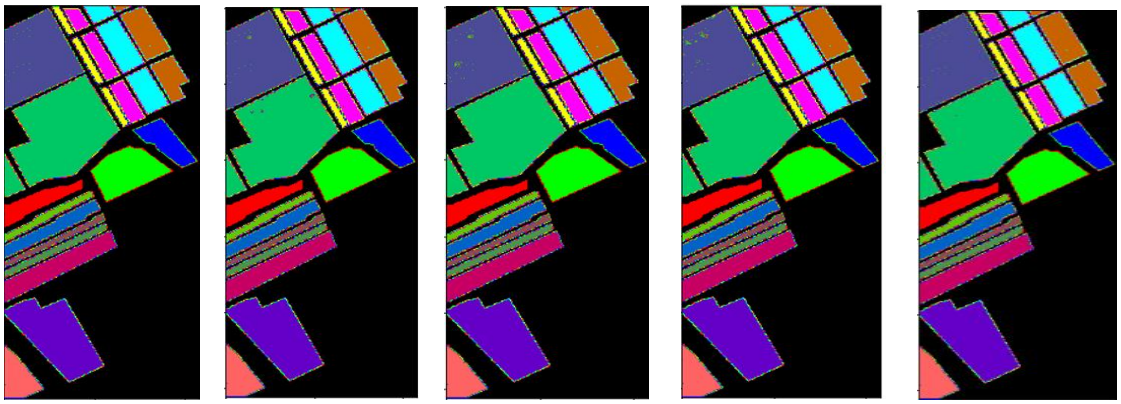
Salinas Full scene (SA): Table 5.7 displays the results of SA dataset for several approaches in terms of OA, AA, and Kappa while classification maps are depicted in Figure 5.6. In addition, the statistical significance of each class is shown in Table 5.8.

Table 5.7 Classification results indicating OA, AA and κ respectively of SA dataset with different DR, dimensions and window size

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	99.63	99.82	99.83	99.94	99.98	99.97	99.85	99.94	99.86	99.97
	99.59	99.80	99.81	99.93	99.98	99.97	99.83	99.93	99.84	99.97
	99.73	99.84	99.89	99.95	99.98	99.96	99.92	99.95	99.91	99.98
Fast ICA	99.88	99.92	97.98	99.65	99.95	100	99.97	100	99.96	99.98
	99.87	99.92	97.75	99.61	99.95	100	99.97	100	99.96	99.98
	99.91	99.95	95.91	99.69	99.97	100	99.96	100	99.95	99.97
SRP	41.52	20.89	20.82	20.82	20.82	20.82	20.82	20.82	38.82	20.82
	32.43	0.10	0	0	0	0	0	0	25.58	0
	22.31	6.31	6.25	6.25	6.25	6.25	6.25	6.25	27.81	6.25
FA	99.92	99.80	99.89	99.97	99.65	99.95	99.91	99.93	99.95	99.85
	99.91	99.78	99.88	99.97	99.61	99.95	99.90	99.93	99.94	99.83
	99.96	99.90	99.93	99.96	99.79	99.94	99.94	99.96	99.97	99.82
NMF	96.35	97.63	96.78	97.23	96.68	98.14	97.55	99.25	98.63	99.38
	96.21	97.58	96.24	97.15	96.39	97.96	97.29	99.05	98.41	99.14
	96.76	97.84	96.81	97.36	96.72	98.36	97.68	99.38	98.92	99.45



(a) (15,9) (b) (15,11) (c) (21,9) (d) (21,11) (e) (27,9)



(f) (27,11) (g) (33,9) (h) (33,11) (i) (39,9) (j) (39,11)

Figure 5.6 Classification results of SA for different number of dimensions and patch size

Table 5.8 Statistical significance of SA dataset with PCA

Class	15 Bands			21 Bands			27 Bands			33 bands			39 Bands		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
	W1/W2			W1/W2			W1/W2			W1/W2			W1/W2		
1.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
2.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
3.	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
4.	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
5.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
6.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
7.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
8.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
9.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
10.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
11.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
12.	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
13.	1.00/ 0.99	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
14.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00
15.	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	0.99/ 1.00	1.00/ 1.00
16.	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 0.99	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00	1.00/ 1.00

In all the experiments mentioned above, the performance of the proposed model is analyzed using a variety of DR approaches for hybrid CNN and evaluated for different number of spectral bands. Furthermore, by using two distinct patch sizes (9×9 and 11×11), the impact of input window size on the classification performance is investigated. For IP dataset FA and PCA provides better results compared to other DR techniques with increase in bands. FastICA and NMF also shows improvement in accuracy with more bands. However, the performance of SRP is not acceptable. PCA performs better with a marginal improvement in accuracy as compared to FA for PU and SA datasets. The accuracy of FastICA and NMF algorithm is improved with PU and SA datasets. According to the experimental results, PCA in combination with a hybrid CNN model offers excellent classification accuracy for all three datasets. The patch size selected also has a significant impact on the final results. When the patch size is set too small, it reduces the inter-class diversity in samples, and when the patch size is set too large, it may include pixels from different classes, leading to misclassification in both situations. In most of the cases, a patch size of (11×11) provides better results which is more pre-dominant in IP dataset compared to PU and SA datasets. According to the results, the

classification outcomes for both the PU and SA datasets exhibit a marginal improvement with larger window size. Compared to IP and PU, SA dataset provides the highest accuracy since field area is not overlapped and each crop area has a precise boundary. The statistical significance test performed with PCA for all three datasets provides detailed information regarding accurate prediction of each class and the variations with spectral bands and patch size. However, it is clear from the experimental results that, OA, AA and κ values nearly remain constant as more spectral bands are extracted using different DR methods. Hence a smaller patch size with few bands is adequate for accurate classification.

5.3.1 Computational cost analysis

The reduction of computational complexity is a key factor in the design of DL models. Hybrid CNN models reduce the number of computations and thus reduce the training time for feature extraction and classification. In this regard, the major goal of the proposed hybrid model is to reduce the computational complexity of existing CNN models. Consequently, it is crucial to evaluate the computational costs of the proposed model using various DR approaches and patch sizes. Training time is influenced by critical factors such as training data size, device processing speed, and the availability of Graphical Processing Units (GPU) and RAM. As a result, all the DR techniques are evaluated on hybrid model on the same device to compare and determine the computing load required for successful model training. Table 5.9-5.11 describes the time needed in seconds for training, testing and DR phases respectively for all three datasets. A lower value in the execution time results in a simpler model, which reduces the computing cost, processing power and memory requirements. For IP dataset, the training time difference between each DR technique is less since the spatial size is small. However, when datasets with larger spatial size like PU and SA is considered, the training time has a significant impact on final results. With the increase in spectral bands and patch size, overall execution time also increases. However, it is worth to be noted that there is an increase in time taken for DR compared to training and testing time. The hybrid model with PCA converges quickly compared to other DR techniques. The architectures of the chosen model yield a relatively similar number of trainable weights. From the analysis of the execution time, it is clear that the proposed hybrid model with PCA delivers highest classification accuracy in less time compared to other DR approaches for all datasets. Furthermore, the proposed model can be implemented with larger spatial patch size to increase accuracy.

Table 5.9 Execution time (Tr, Te, DR) of IP dataset

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	10.85	11.14	10.88	15.58	21.85	21.96	21.29	26.59	28.15	29.26
	0.52	0.57	0.74	0.56	0.58	0.96	0.63	0.78	0.86	0.94
	0.36	0.37	0.23	0.25	0.28	0.83	0.36	0.32	0.97	0.42
Fast ICA	21.11	10.88	10.91	14.31	12.18	20.14	21.11	24.53	21.13	41.64
	0.52	0.57	0.76	0.76	0.76	0.55	0.76	0.58	0.66	0.80
	1.59	1.65	1.65	1.67	1.67	3.53	1.72	1.76	8.66	10.12
SRP	10.89	9.95	11.01	14.58	21.08	21.29	21.15	41.63	19.20	41.64
	0.75	0.58	0.75	0.61	0.61	0.77	0.77	0.78	0.78	1.44
	0.05	0.2	0.12	0.24	0.34	0.38	0.47	0.46	0.74	0.62
FA	8.23	9.65	10.86	13.92	21.08	41.56	21.12	25.13	21.11	29.65
	0.74	0.51	0.79	0.55	0.75	0.77	0.57	0.77	0.63	0.61
	1.75	1.84	4.44	5.43	3.52	3.43	5.52	5.10	3.06	2.93
NMF	10.89	9.96	10.92	21.11	21.13	20.38	21.14	41.60	21.12	28.97
	0.75	0.75	0.75	0.58	0.76	0.57	0.77	0.79	0.78	1.50
	7.90	8.19	10.90	9.46	11.30	11.13	17.53	15.12	19.56	19.34

Table 5.10 Execution time (Tr, Te, DR) of PU dataset

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	23.71	36.03	34.90	51.69	54.23	76.23	82.60	94.42	61.69	113.64
	2.74	1.94	2.76	2.82	2.94	2.86	2.83	2.96	2.88	3.12
	1.68	2.29	2.07	1.98	2.02	2.23	3.45	3.04	3.63	7.93
Fast ICA	26.20	41.60	41.61	83.50	43.56	71.05	82.59	95.82	65.45	142.84
	1.61	2.95	2.75	1.83	1.70	2.88	2.82	2.02	2.90	2.66
	61.98	46.59	35.94	36.16	67.23	66.33	34.62	26.89	32.77	33.83
SRP	41.57	34.89	33.30	82.61	44.10	71.29	82.61	99.20	65.44	142.89
	1.64	2.75	2.75	1.88	2.79	3.23	1.89	2.25	1.89	2.50
	0.29	0.25	0.26	0.23	0.33	0.30	0.33	0.47	0.39	0.66
FA	41.58	34.28	35.05	82.61	82.58	82.66	53.81	100.05	65.85	119.34
	2.75	1.74	2.96	2.83	2.78	2.03	2.89	3.00	2.89	2.44
	36.09	43.46	41.48	46.03	45.63	55.14	131.91	130.98	234.09	230.50
NMF	22.67	41.61	34.40	53.12	82.60	76.14	57.12	142.70	70.51	121.50
	2.76	1.75	2.77	1.92	1.83	2.01	2.85	2.41	1.94	2.53
	50.18	52.62	79.46	80.49	116.01	109.10	136.98	141.60	166.89	182.55

Table 5.11 Execution time (Tr, Te, DR) of SA dataset

DR Method	15 bands		21 bands		27 bands		33 bands		39 bands	
	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11	9 × 9	11 × 11
PCA	31.76	40.69	40.89	67.41	52.45	142.75	82.74	142.68	82.92	162.72
	2.76	2.10	2.05	2.38	2.80	3.09	2.99	2.47	2.22	3.07
	4.86	4.35	3.35	2.42	2.30	3.94	2.73	3.83	2.74	2.61
Fast ICA	30.33	82.62	41.61	67.07	54.23	142.60	82.61	115.33	82.62	138.09
	2.81	2.77	2.77	2.53	2.82	3.05	2.18	3.16	2.34	3.23
	9.69	10.13	12.61	13.40	42.46	46.55	45.16	52.72	31.58	41.59
SRP	32.28	41.58	82.60	82.60	83.19	142.62	67.77	117.95	82.61	143.81
	1.93	2.77	2.77	2.19	2.29	2.88	2.32	2.59	3.04	3.46
	0.29	0.24	0.31	0.30	0.32	0.28	0.36	0.27	0.36	0.55
FA	29.29	41.68	41.13	82.63	82.61	93.99	69.47	121.63	142.65	142.08
	1.98	2.20	2.88	2.28	2.82	3.06	3.00	3.39	2.87	2.76
	81.92	77.56	37.28	36.32	28.50	24.74	25.41	25.62	30.01	29.18
NMF	32.08	43.52	43.02	69.25	55.36	138.27	85.96	169.63	88.36	180.93
	2.75	2.34	1.96	2.05	1.99	2.32	2.08	2.89	2.68	3.04
	5.12	10.25	19.36	36.25	41.58	50.02	52.37	59.10	68.36	96.32

5.3.2 Comparison with state-of-the-art methods

In order to evaluate the efficacy of the proposed hybrid CNN model, it is compared with a number of cutting-edge frameworks that have recently been published for HSIC using publicly available codes through website https://github.com/eecn/Hyperspectral-Classification.Support Vector Machine (SVM) (Melgani et al. 2004) with radial basis function is implemented as it is less sensitive to the number of training samples and provides high classification accuracy. Multi-layer Perceptron (MLP) with four fully connected layers and dropout is used as a baseline model for comparison. 2D CNN uses PCA as a pre-processing technique and 2D convolutions to extract local spatial features. The extracted features are classified by the linear regression model. To extract joint spectral-spatial features 3D CNN incorporates 3D convolutions. Both 2D and 3D models (Chen et al. 2016) use ReLU as the activation function and mini-batch to update weights. Dropout has been introduced to avoid overfitting. Spectral-Spatial Unified Networks (SSUN) (Xu et al. 2018) integrate spectral and spatial feature extraction modules with a classifier into a unified framework. Spectral features are extracted using Long Short-term Memory (LSTM) networks while spatial features are extracted by multiscale CNN. The classification layer is a fully connected layer with 128 neurons. Spectral-Spatial Residual Networks (SSRN) (Zhong et al. 2017) is similar to 3D CNN with batch normalization and dropout of 0.5 as a regularizer during the training process. The model is trained up to 200 epochs with batch size 16 and residual blocks are introduced to avoid a decrease in classification accuracy as the number of convolutional layers is increased. Synergistic 2D/3D CNN with attention module (SynCNN-ATT) (Yang et al. 2020) consists of a hybrid module for 2D-3D convolutions with a data interaction module for spectral-spatial feature fusion. Additionally, a 3D attention mechanism is introduced before fully connected layers to filter out interfering features and information effectively. Multi-scale 3D CNN (MS 3D CNN) (He et al. 2017) jointly learns both 2D multi-scale spatial features and 1D spectral features without pre-processing hyperspectral data. Fast and Compact 3D CNN (FC 3D CNN) (Ahmad et al. 2020) has been proposed to generate 3D feature maps using 3D kernels over multiple contiguous spectral bands in a computationally efficient way. All the experiments are implemented on IP and SA datasets. To compute the results, publicly accessible codes are used for the methods that were being compared. With the exception of the number of dimensions and patch size (i.e., 15 dimensions selected using PCA, and 11×11 patch size is employed for experimental purposes), all the comparison models are being trained in accordance with the settings mentioned in their respective articles. From the experimental results shown in Table

5.12, the proposed hybrid CNN model outperforms other baseline models to a certain extent and produced results that are equivalent to those of state-of-the-art techniques with reduced complexity. As compared to hybrid CNN, 3D or 2D CNNs alone are unable to extract highly discriminative features. In addition, the accuracy and loss curve between the 3D and hybrid CNN model for the IP dataset with PCA for (15×11) is depicted in Figure 5.7. The convergence of hybrid CNN model is fast compared to 3D CNN model.

Table 5.12 Comparison with state-of-the-art approaches

Methods	Indian Pines			Salinas		
	OA	AA	Kappa	OA	AA	Kappa
SVM	85.30	79.03	83.10	92.95	94.60	92.11
MLP	87.57	89.07	85.80	79.79	67.37	77.40
2D CNN	80.27	68.32	75.26	96.34	94.36	95.93
3D CNN	82.62	76.51	79.25	85.00	89.63	83.20
SSUN	95.27	94.29	95.14	98.39	98.58	98.72
SSRN	97.43	96.38	98.17	98.27	98.43	98.12
SynCNN-ATT	97.31	97.43	96.90	98.92	99.35	98.80
MS 3D CNN	91.87	92.21	90.80	94.69	94.03	94.10
FC 3D CNN	98.10	96.46	97. 59	98.06	98.80	97.85
Proposed	**98.18**	**97.93**	**97.81**	**99.82**	**99.80**	**99.84**

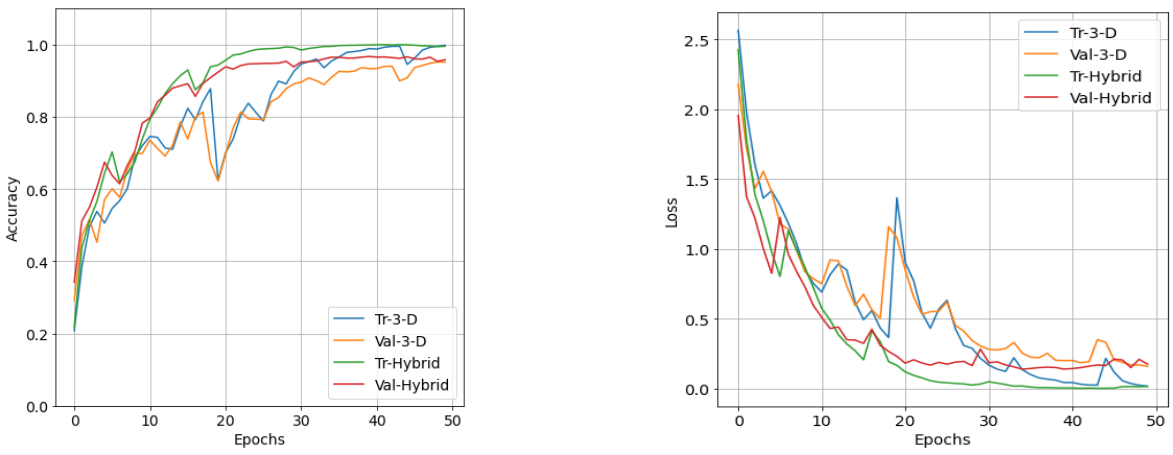


Figure 5.7 Accuracy and loss curves for IP dataset with PCA (11,15)

The next chapter proposes a deep feature selection technique inspired by KD for hyperspectral data.

CHAPTER 6
DEEP FEATURE SELECTION BASED ON KNOWLEDGE DISTILLATION

6.1 INTRODUCTION

FE approaches alter the physical characteristics of the original data when transformed to low dimensional representation (Rodarmel and Shan 2002; Mou et al. 2017; Hao et al. 2018). However, few applications demand the original semantics of data. Band selection can be effectively utilized in such scenarios to extract highly informative and discriminative bands while ignoring redundant information. In BS, the classification accuracy depends on:

i. Pre-processing technique employed in the study.
ii. The selection and search criteria adopted in the overall analysis.

Conventional BS techniques evaluate the relevance of each band individually. Since the adjacent bands in hyperspectral data are highly correlated, the influence of neighboring bands is ignored which leads to sub-optimal band subset. The exhaustive search techniques lead to high computational complexity. In addition, due to inadequate training samples and ground truth, unsupervised learning-based BS is more appropriate. To address these issues, this chapter provides an overview of Knowledge Distillation (KD) and autoencoders for deep feature selection for the first time in HRS. The goal is to explain the difference between autoencoders based on principles of representation learning rather than focusing on the mathematical explanation of parameters introduced. The potential of convolutional autoencoders (CAE) and sparsity regularization is well explored for deep feature selection of hyperspectral data.

6.1.1 Model compression

Since hyperspectral datasets are voluminous, large weights need to be updated in the training process which is computationally expensive and time-consuming. In recent years, portable devices like mobile phones, and cameras are more popular. However, massive deep neural networks cannot be deployed in such devices with limited memory which necessitates the use of deep neural network model compression in embedded systems and mobile devices (Li H T et al. 2019). The current section provides an overview of model compression.

Operator factorization techniques break down complex operations like dense layer matrix multiplication into simpler ones. Singular value decomposition (Sainath et al. 2013) can be used for matrices, and tensor train decomposition can be used for tensors (Kanjilal et al. 1993, Zhao et al. 2019). Value quantization techniques find a superior low-precision encoding for

network values like weights, gradients, and activations. Various integer and floating-point formats can be employed to encode data effectively instead of typical 32 or 64-bit datatypes. Model distillation (Hinton G et al. 2015) and Neural Network Architecture search (Elsken Thomas et al. 2019) are used to downsize models, resulting in less dense networks that handle the same task with less complexity. Value compression techniques based on entropy methods (Han Song et al. 2016) and loss bounded type-specific methods employing correlation can be adopted for model compression (Jin Sian et al. 2019). KD techniques extract the knowledge of a complicated neural network (teacher) to a smaller network (student). The small network mimics the functionality of the large neural network leading to model compression. The sparsification process can also lead to efficient models even though it works in high dimensional feature space. It can be achieved by employing only a subset of dimensions at a time. Parameter sharing techniques try to exploit redundancy in the parameter space during the training process and compress the model (Plummer et al. 2020). Network pruning strategies eliminates nodes, filters, and layers from the model that are redundant to the task at hand (Cheng et al. 2018, Molchanov et al. 2019) and supports training as well as transfer learning from pre-trained models. All of the above approaches lead to low memory, storage and computational costs. The current study investigates the most complex and efficient of these techniques, KD. Since the number of trainable parameters in the student network decreases, it can learn the knowledge from teacher network without much effort, leading to a simpler design and faster training. The concept was first introduced for speech recognition and classification of the MNIST dataset (Hinton G et al. 2015). The methodology adopted in the study is described in the next section as follows.

6.2 METHODOLOGY

6.2.1 Deep Teacher Student Feature Selection (TSFS)

Existing FS approaches, such as AEFS and GAFS, employ simple autoencoders for both FS and reconstruction. A simple network is beneficial for FS since error can be easily backpropagated and top features can be chosen with less complexity. However, for reconstruction a simple autoencoder is insufficient. A novel two-stage deep FS technique inspired by KD is presented to solve the contrasting criteria. The typical procedure of the proposed scheme is represented in Figure 6.1. The proposed strategy can be divided into two phases. FE is performed in the first stage by mapping input data to a low-dimensional subspace.

In the second stage, FS is used to select only significant features from a low-dimensional subspace. Since the two steps are now distinct, a network architecture that is both complicated and simple can be used. The following are the two stages of the proposed approach.

Teacher stage: A complex FE method is utilized that best represents the given data which can be linear or nonlinear, supervised or unsupervised depending on the nature of the dataset and its intended application. A teacher capable of obtaining better latent space can use a variety of manifold learning or DL techniques.

Student stage: A simple FS approach is utilized, usually with a single layer autoencoder. The weights of the first layer of the network are subjected to row sparsity constraints. Finally, the features are ranked and the features corresponding to the highest scores are then selected. The selected features can then be used as input to other subsequent applications.

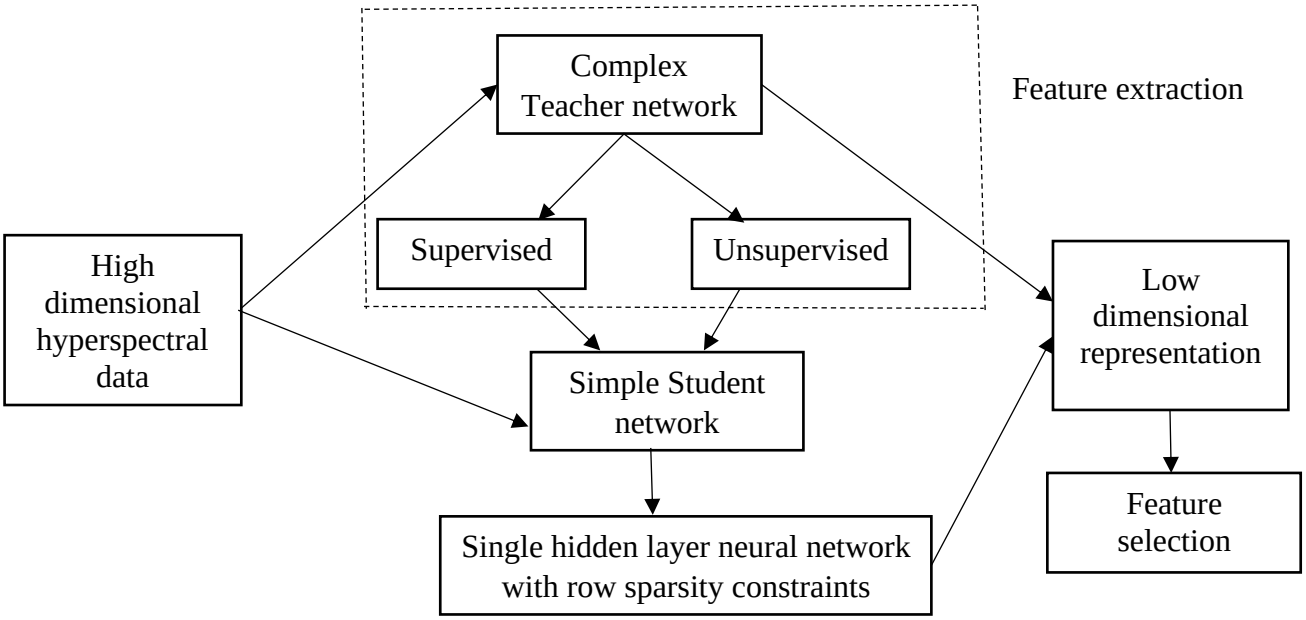


Figure 6.1 Schematic of the proposed D-TSFS technique

6.2.2 Autoencoder framework

If the vector, $\boldsymbol{X} \in \mathbb{R}_{n\times b}$ represents the hyperspectral data with *n* samples and *b* bands, the autoencoder first maps the input to a code through a deterministic mapping called the encoder (Ali Mirzaei et al. 2020) represented as,

$$h = f_\theta(x) = \sigma(xW + b) \tag{6.1}$$

where $\theta = \{W, b\}$; $W \epsilon \mathbb{R}_{m \times n}$ is a weight matrix; $b \epsilon \mathbb{R}_m$ is a bias vector; $f_\theta(x)$ is the encoder and σ is an activation function. The code is mapped back to the original input dimension reproducing the output layer $\hat{x}$ through decoder function represented as,

$$\hat{x} = g_{\hat{\theta}}(h) = \sigma(h\widehat{W} + \hat{b}) \quad (6.2)$$

where, $\hat{\theta} = (\widehat{W}, \hat{b})$ and $g_{\hat{\theta}}(h)$ is the decoder.

6.2.3 Algorithm

Assume $\boldsymbol{X} \epsilon \mathbb{R}_{n \times b}$ represents the original high dimensional hyperspectral data with *n* samples and *b* bands. The top *p* features are selected based on the following steps:

1. Teacher step: In this step, a DNN model based on autoencoders is used for FE to obtain the best possible low dimensional representation, *Y*.

   $$Y_{n \times k} = F(X_{n \times b}), \quad (6.3)$$

   where *k* is the latent space dimension and in general, $k \ll b$ and *F* is a complicated nonlinear function. After *F* is trained, all training data is fed to *F*, and the corresponding low dimensional representation, *Y* is obtained.

2. Student step: A single hidden layer autoencoder network is trained to reproduce the latent code from the teacher network during the feature selection stage. For better training, the latent codes are normalized between 0 and 1 using,

   $$Y_n = \frac{Y - min(Y)}{max(Y) - min(Y)} \quad (6.4)$$

   The neural network output can be written as,

   $$Y_p = (Relu(XW^1 + b_1))W^2 + b_2\,, \quad (6.5)$$

   where $W^1(b \times h)$ and $W^2(h \times k)$ are the weight matrices of hidden and output layers of the network respectively. b_1 and b_2 are the bias vectors for network layers. To perform feature selection, a row sparsity constraint is applied as follows,

   $$\|W^1\|_{2,1} = \sum_i^b \sqrt{\sum_j^h (W_{ij}^1)^2}, \quad (6.6)$$

   where W^1 is the weight matrix of the first layer with *b* rows and *h* columns.

   The loss function of the neural network is defined as,

$$J(\Theta) = \frac{1}{2n}\left\|Y_n - Y_p\right\| + \lambda\|W^1\|_{2,1} \quad (6.7)$$

where λ is the trade-off parameter between the reconstruction loss and regularization term.

3. Feature selection step: Once the student network is trained, the feature scores are calculated as:

$$s = diag(W^1(W^1)^T \quad (6.8)$$

where *s* is a *d* dimensional vector containing weights of significant features. The top p percent of s is chosen to retain the p percent of features. The complete algorithm is summarized in Algorithm 1.

Algorithm 1: Teacher Student Deep Feature Selection Algorithm

---

Step 1: **Input:** Data: $X_{n\times b}$, labels(optional),$T_{n\times c}$, percent of features $p \in (0,100)$

Step 2: *X*, *T* are used to train the DNN.

Step 3: To obtain latent space *Y*, feed *X* to the teacher network (autoencoder).

Step 4: *X* and *Y* are used to train the student network.

Step 5: Calculate the feature scores based on the weights of the first layer of the student network, $s = diag(W^1(W^1)^T$.

Step 6: **return** the first m values of $argsort_{descending}(s)$.

Output: Selected features: $\{f_1, f_2, \ldots\ldots, f_m\}, f_i \in \{1, \ldots\ldots, \text{b}\}$, where $m = p * b/100$.

The flowchart of the proposed scheme is as shown in Figure 6.2(a).

6.2.4 Models

Teacher model: Two teacher models have been adopted in the study; the first model is a 1D autoencoder with dense layers for compression of data in the spectral domain. Since the network cannot effectively capture the spatial relationships inherent in the data, a 2D

convolutional autoencoder (CAE) is introduced in the next stage for better analysis (Pintelas et al. 2021) and to compress the data in the spatial domain.

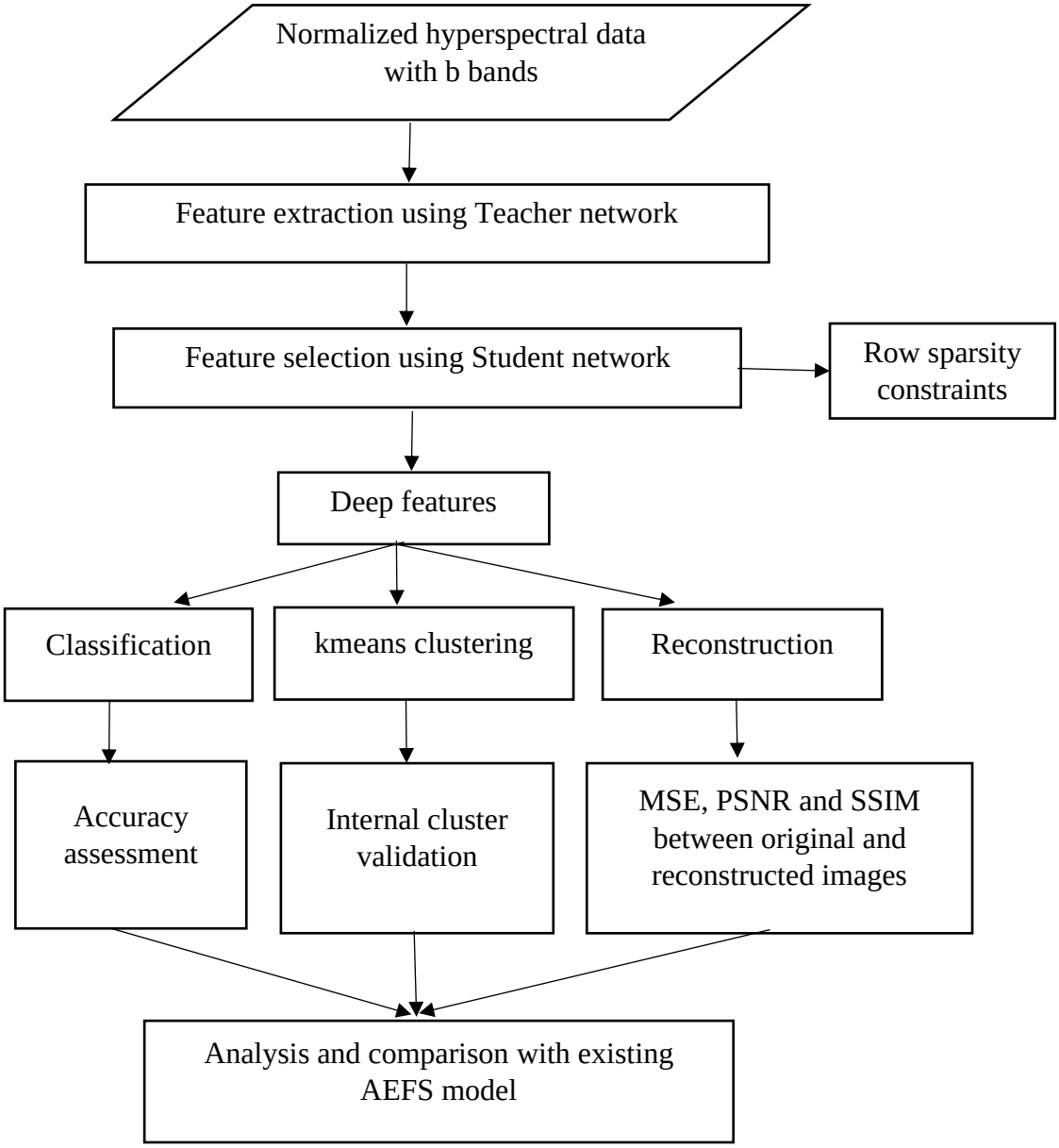


Figure 6.2(a) Flowchart of deep feature selection using autoencoders

Student model: A single hidden layer autoencoder is used for FS, i.e., only top-ranked features are selected from the total number of features. Hence it is mandatory to make sure that, the relevance of the required features rarely spread out and instead remains concentrated on a few features. It has been achieved by adding a sparsity constraint on the hidden layer. By imposing the regularization constraint, the hidden layer is forced to set its weights close to zero. Hence results in only a few features possessing a larger magnitude of weight.

6.2.5 Architecture

The encoder of the proposed 2D CAE stacks 2D convolutional layers after the input with max-pooling layers of size (2,2) to learn spatial features. To train the encoder for feature extraction, a companion 2D convolutional decoder is designed to reconstruct hyperspectral data from features extracted by the encoder. The size of feature maps is 64,16 and 4 with a kernel size of 3x3. The labelled samples are not required for the proposed training scheme. The architecture of the proposed technique with 2D CAE is shown in Figure 6.2(b).

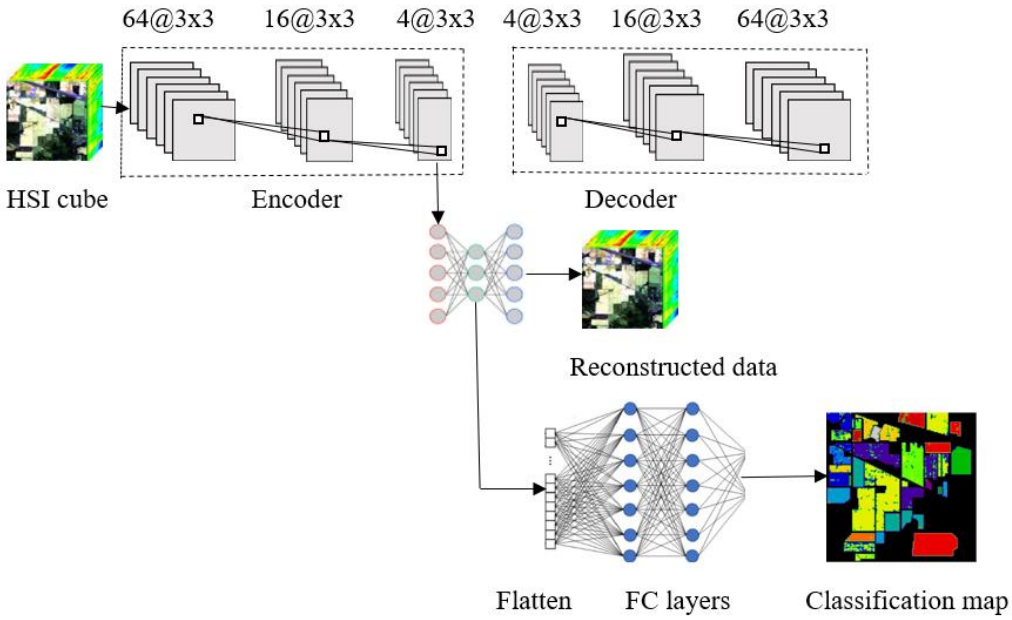


Figure 6.2(b) Architecture of the proposed 2D-TSFS model

The output of the encoder is fed to a student network composed of a single hidden layer with 90 nodes. For unsupervised analysis, clustering and reconstruction task is carried out. For supervised classification, the output from the student encoder is flattened and fully connected layers are added on top of the student encoder to generate the corresponding classification maps.

6.3 EVALUATION METRICS

The effectiveness of the proposed technique on different DL models is evaluated using both supervised (classification) and unsupervised (clustering and reconstruction) evaluations on hyperspectral datasets for different percentages of selected features.

6.3.1 Classification: SVM is used due to its simple and efficient implementation on large datasets. Classification accuracy and F1 scores are used as the evaluation metrics (Windrim et al. 2019) on IP and PU datasets where,

$$Accuracy = \frac{Correct\ predictions}{total\ predictions} \tag{6.9}$$

$$F1 = 2 * \frac{precision \times recall}{precision + recall} \tag{6.10}$$

$$where,\ Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{6.11}$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{6.12}$$

6.3.2 Clustering: K-means clustering is performed on the selected features. Since Cuprite and Samson do not have the corresponding label information, unsupervised analysis is carried out. To evaluate clustering performance, three different internal clustering validation measures (Hassani and Seidl, 2017) are widely used and can be described as follows.

Silhouette Index, S computes compactness a(x) using the pairwise distance between all the items in a cluster and b(x) measures separation using the average distance of objects to the second closest cluster.

$$S = \frac{1}{NC} \sum_i \left(\frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{max[b(x), a(x)]} \right) \tag{6.13}$$

$$\text{where, } a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i, y \neq x} d(x, y) \tag{6.14}$$

$$b(x) = min_{j \neq i} \left[\frac{1}{n_j} \sum_{y \in C_j} d(x, y) \right] \tag{6.15}$$

Calinski Harabasz index, CH uses the average between and inside the cluster sum of squares to quantify both compactness and separation requirements simultaneously.

$$CH = \frac{\sum_i d^2(c_i, g)/(NC - 1))}{\sum_i \sum_{x \in C_i} d^2(x, c_i)/(n - NC)} \tag{6.16}$$

The numerator represents the degree of separation by exploring the spread of cluster centers. The denominator is compactness, which shows how close objects in the same cluster are grouped around the cluster center.

Davies Bouldin index, DB includes intra-cluster variance and inter-cluster center distance to find the closest most scattered one for each cluster. The optimum number of clusters is obtained by minimizing DB.

$$DB = \frac{1}{NC}\sum_i max_{j\neq i}\frac{\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)}{d(c_i,c_j)} \quad (6.17)$$

All the measures have been designed to represent both compactness and separation simultaneously to evaluate complex clustering.

6.3.3 Reconstruction: Additionally, an attempt has been done to reconstruct the original data using the selected features to analyze the quality of selected features. The quantitative metrics based on reconstruction used for comparison of performance between different models are Structural Similarity Index Metric (SSIM), Mean Squared Error (MSE), and Peak Signal to Noise Ratio (PSNR) between the original and reconstructed images (Samajdar and Quraishi, 2015).

MSE is relatively simple with fewer computations which satisfies non-negativity, symmetry, identity and triangular inequality constraints. Lower MSE indicates a higher similarity between original and reconstructed images f(x,y) and g(x,y) respectively. If MN represents the dimensions of the image, then

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[f(x,y) - g(x,y)]^2 \quad (6.18)$$

The ratio of maximum signal power to maximum noise power is termed as PSNR. A greater PSNR score suggests that the images are more comparable.

If L is the dynamic range of pixels in the image, then

$$PSNR = 20log_{10}\frac{L^2MN}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[f(x,y)-g(x,y)]^2} \quad (6.19)$$

SSIM is a feature-based Human Visual System (HVS) metric that outperforms MSE and PSNR measures. The similarity between two images is calculated over several windows of an image as,

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \quad (6.20)$$

where μ_x and μ_y are the average of x and y respectively. σ_x^2 and σ_y^2 are the variances of x and y respectively. σ_{xy} is the covariance of x and y, and c_1, c_2 are constants.

6.4 RESULTS AND DISCUSSION

Autoencoders are trained to reconstruct the input data. Hence, they learn the complex features of the data. With the presence of the bottleneck layer, which usually contains a lesser number of nodes than the input, the autoencoder effectively learns a lower dimensional representation of the input. FE is performed using 1D deep autoencoder and 2D CAE termed as Teacher network. FS is performed using the student model which utilizes the low dimensional data obtained from the Teacher network to further enhance compression. For FS, an autoencoder with just one hidden layer is used to compress the data by another factor. The squares of the weights of the hidden layer are extracted and sorted in descending order. The list of weights is the score for the corresponding features, based on which a certain percentage of features are finally selected. All the models have been compiled using Adam optimizer, ReLU activation function with loss function as MSE. In addition, one of the existing feature selection methods, AEFS is implemented for comparison with the proposed models. Classification datasets IP and PU are subjected to supervised analysis in order to evaluate the effectiveness of the proposed scheme. However, as there is no ground truth, Spectral Unmixing datasets Cuprite and Samson are utilised for unsupervised analysis (clustering). The classification accuracies of IP and PU datasets for different percentages of features are reported in Table 6.1 and has been depicted in Figure 6.3 for better visualization.

Table 6.1 Classification accuracies with different percentages of features

Model	Dataset	100%	75%	50%	25%
AEFS	IP	87.60	87.51	85.70	80.14
	PU	92.28	88.86	88.32	86.18
1D-TSFS	IP	88.24	82.43	88.58	82.78
	PU	93.25	91.94	89.15	90.67
2D-TSFS	IP	96.15	95.32	95.05	94.82
	PU	97.82	96.86	96.34	94.65

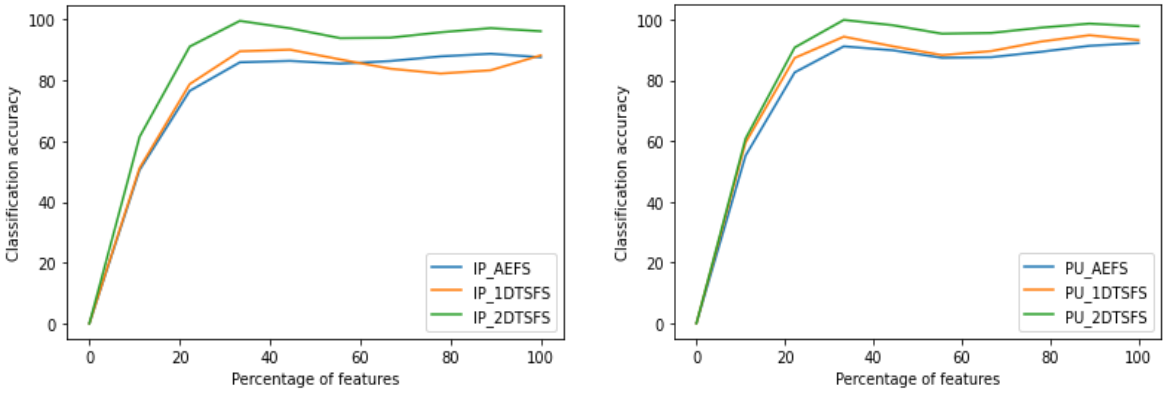


Figure 6.3 Classification accuracies of IP and PU datasets with different percentages of features

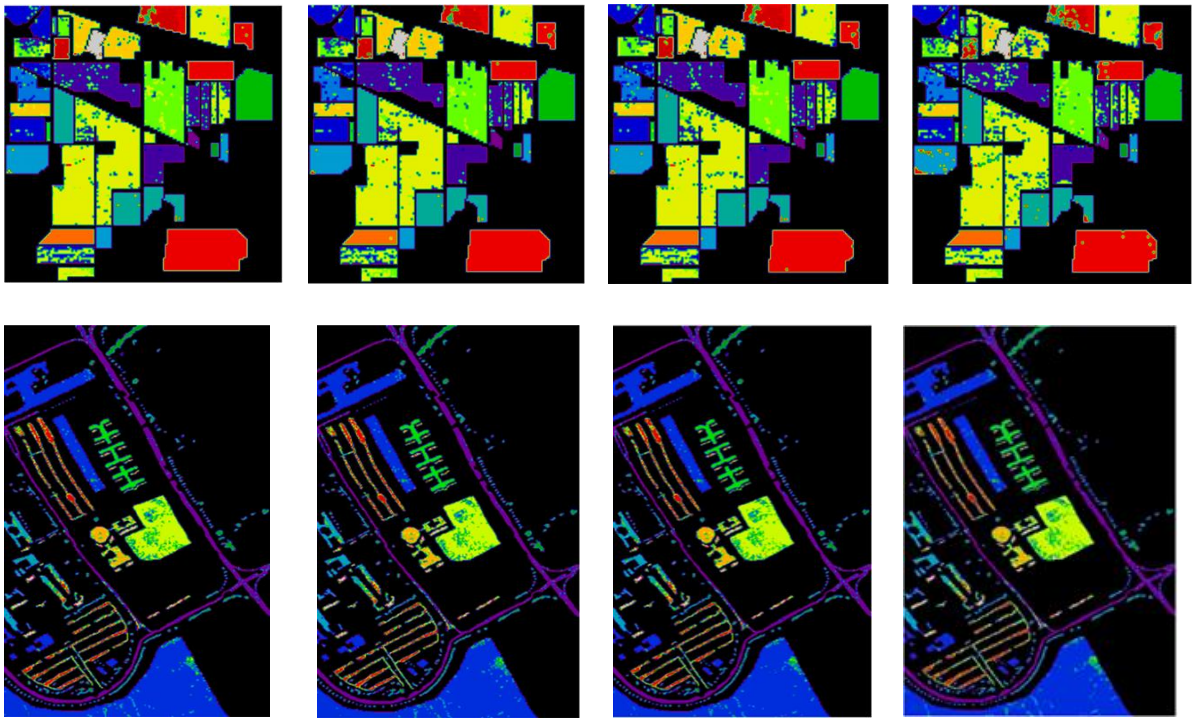
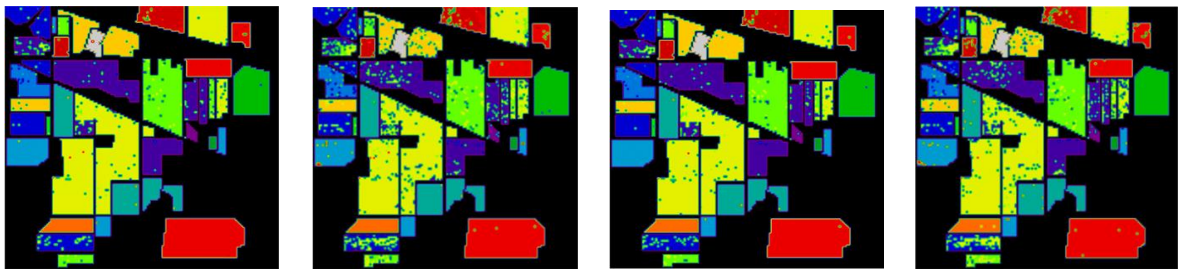


Figure 6.4 Classification maps of IP and PU datasets for AEFS model with 100,75,50 and 25% of features respectively.



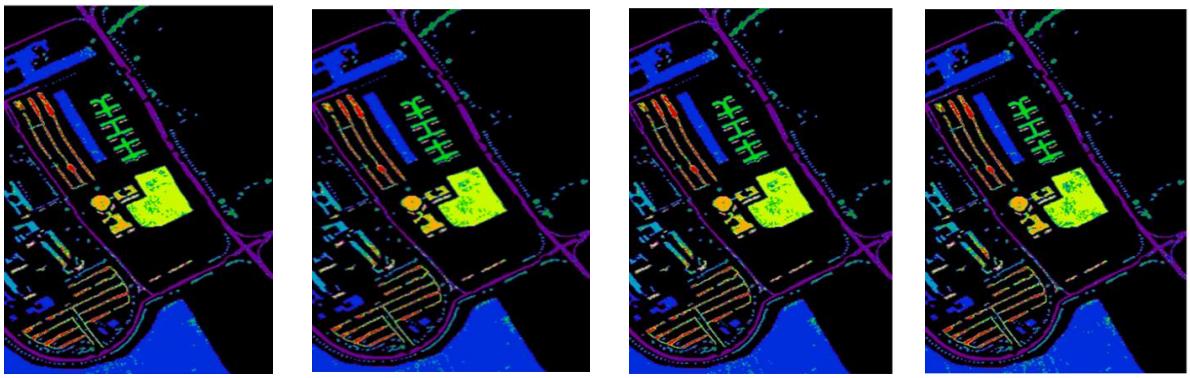


Figure 6.5 Classification maps of IP and PU datasets for the 1D-TSFS model with 100,75,50 and 25% of features respectively.

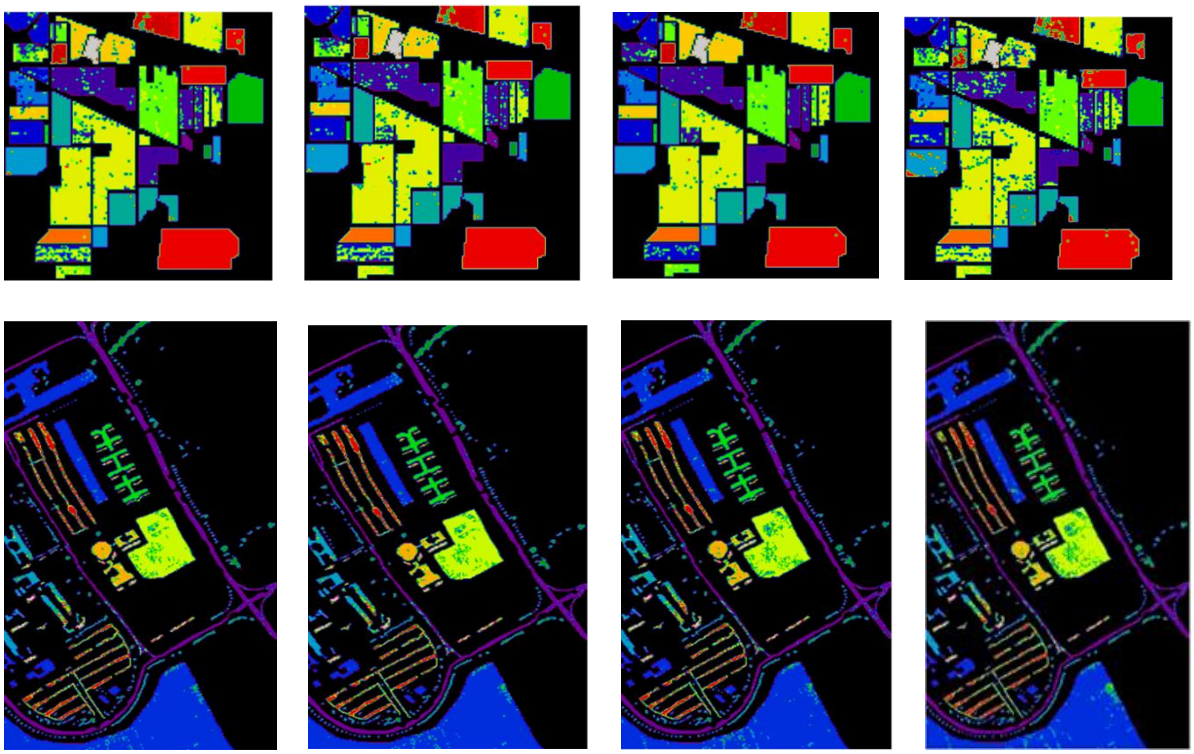


Figure 6.6 Classification maps of IP and PU datasets for 2D-TSFS model with 100,75,50 and 25% of features respectively.

From the above results, it can be observed that the 2D-TSFS model provides better classification accuracies compared to the other two models for different percentages of features with the highest accuracy value of 96.15% for the IP dataset and 97.82% for the PU dataset. The classification maps of these two datasets with different percentages of features for different models are shown in Figure 6.4, 6.5, and 6.6. The corresponding F1 scores are depicted in Figure 6.7. The maximum F1 score is reported by the 2D-TSFS model for both the datasets.

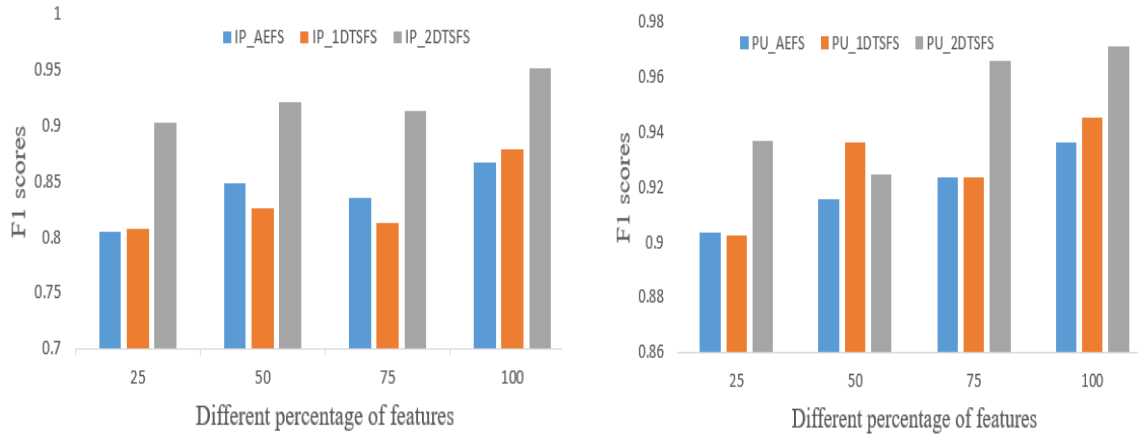


Figure 6.7 F1 scores of IP and PU datasets with different percentage of features for different models AEFS, 1D-TSFS and 2D-TSFS respectively

The clustering results are reported in Table 6.2. Clustering is performed on Cuprite and Samson datasets to explore the benefits of DR on datasets without ground truth.

Table 6.2 Clustering results

Model	Dataset	Silhouette	CH	DB index
AEFS	Cuprite	0.6792	6520.40	1.4576
	Samson	0.5225	5457.87	0.7944
1D-TSFS	Cuprite	0.6679	7810.864	1.0891
	Samson	0.5091	6282.06	0.5944
2D-TSFS	Cuprite	0.8038	10173.90	1.0032
	Samson	0.7312	7771.57	0.4673

In general, Silhouette and CH scores must be maximum and DB score has to be minimum for optimum performance. AEFS model has the highest DB index value and lowest scores for CH and Silhouette index. 2D-TSFS model reports maximum scores for CH and Silhouette index, with a minimum DB index score of 1.0032 compared to the other two models for both datasets. From the above results, it can be concluded that the 2D-TSFS model provides compact clusters for both datasets.

The MSE for four datasets with different percentages of features is presented in Table 6.3. Similarly, Table 6.4 represents PSNR, and Table 6.5 represents SSIM between original and reconstructed images.

Table 6.3 MSE between the original and reconstructed image

Model	Dataset	100%	75%	50%	25%
AEFS	IP	0.00016	0.00598	0.05462	0.1666
	PU	2.916e-06	0.00036	0.00277	0.0054
	Cuprite	2.39e-05	0.02030	0.03063	0.0637
	Samson	3.583e-05	0.00536	0.05367	0.1233
1D-TSFS	IP	0.03272	0.03281	0.03701	0.0473
	PU	0.00849	0.00851	0.00914	0.0131
	Cuprite	0.0098	0.01019	0.01115	0.0132
	Samson	0.01658	0.01748	0.02503	0.0349
2D-TSFS	IP	0.0285	0.04292	0.03950	0.03177
	PU	0.00077	0.00081	0.00086	0.0011
	Cuprite	0.00069	0.00067	0.00121	0.0122
	Samson	0.011963	0.01019	0.00923	0.0021

Table 6.4 PSNR between the original and reconstructed image

Model	Dataset	100%	75%	50%	25%
AEFS	IP	85.879	70.363	60.755	55.912
	PU	103.4814	82.5075	73.7038	70.7828
	Cuprite	93.1052	77.2129	68.2280	71.1833
	Samson	44.4574	22.7038	12.7024	9.0890
1D-TSFS	IP	62.981	62.970	62.447	61.3805
	PU	68.8395	68.82952	68.5213	66.9477
	Cuprite	67.4381	67.3697	67.0520	66.1372
	Samson	17.8021	17.5725	16.8749	14.5663
2D-TSFS	IP	63.567	61.799	62.1615	63.1102
	PU	31.0835	30.8747	30.6167	29.5596
	Cuprite	72.9236	71.9160	71.2678	65.5772
	Samson	67.3522	68.0477	68.4750	74.8856

Table 6.5 Structural similarity between the original and reconstructed image

Model	Dataset	100%	75%	50%	25%
AEFS	IP	0.9898	0.9611	0.7394	0.2898
	PU	0.9995	0.9864	0.8337	0.8112
	Cuprite	0.997	0.661	0.477	0.0392
	Samson	0.9955	0.8110	0.4306	0.0939
1D-TSFS	IP	0.6517	0.6510	0.6204	0.5347
	PU	0.6791	0.6783	0.6461	0.4707
	Cuprite	0.856	0.851	0.834	0.7981
	Samson	0.8934	0.8874	0.8632	0.7781
2D-TSFS	IP	0.7848	0.6720	0.6778	0.6298
	PU	0.9297	0.9267	0.9248	0.9180
	Cuprite	0.9188	0.9182	0.9111	0.7605
	Samson	0.9086	0.9184	0.8934	0.9246

From the obtained results, it can be observed that MSE, SSIM, and PSNR of AEFS models are sometimes better than 2D-TSFS for 100% features. However, as the percentage of features is decreased the performance of the AEFS model reduces drastically while the 2D-TSFS model performance remains almost stable. Hence proves the superiority of 2D-TSFS over AEFS for effectively generalizing and capturing only the most important features. Even though the MSE and SSIM of 1D-TSFS are also almost stable for different percentages of features, the 2D-TSFS model manages to capture the features better than 1D-TSFS with less MSE and high PSNR and SSIM values. Hence, the proposed models exhibit better generalization and feature selection capability.

For better visual comparison, the reconstructed images of AEFS, 1D-TSFS, and 2D-TSFS models with various selections of features (100, 75, 50, and 25 respectively) have been employed. The original images of IP, PU, Cuprite, and Samson datasets respectively are depicted in Figure 6.8. The reconstructed images for the AEFS model are shown in Figure 6.9. Similarly, Figure 6.10 represents 1D-TSFS model results and Figure 6.11 represents 2D-TSFS model results.

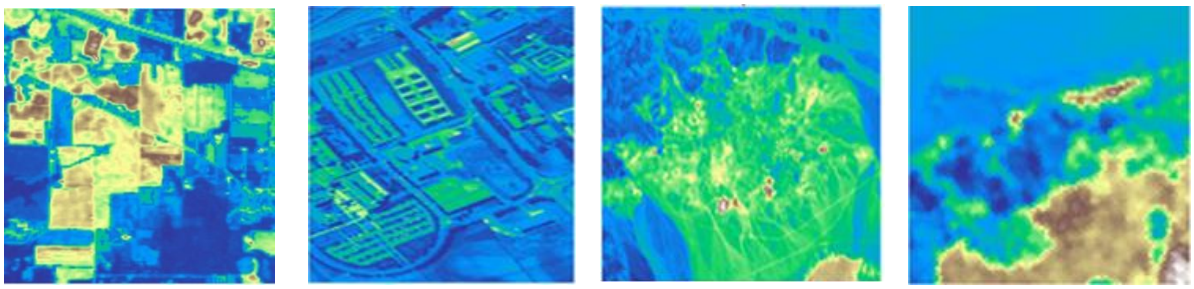


Figure 6.8 Original images

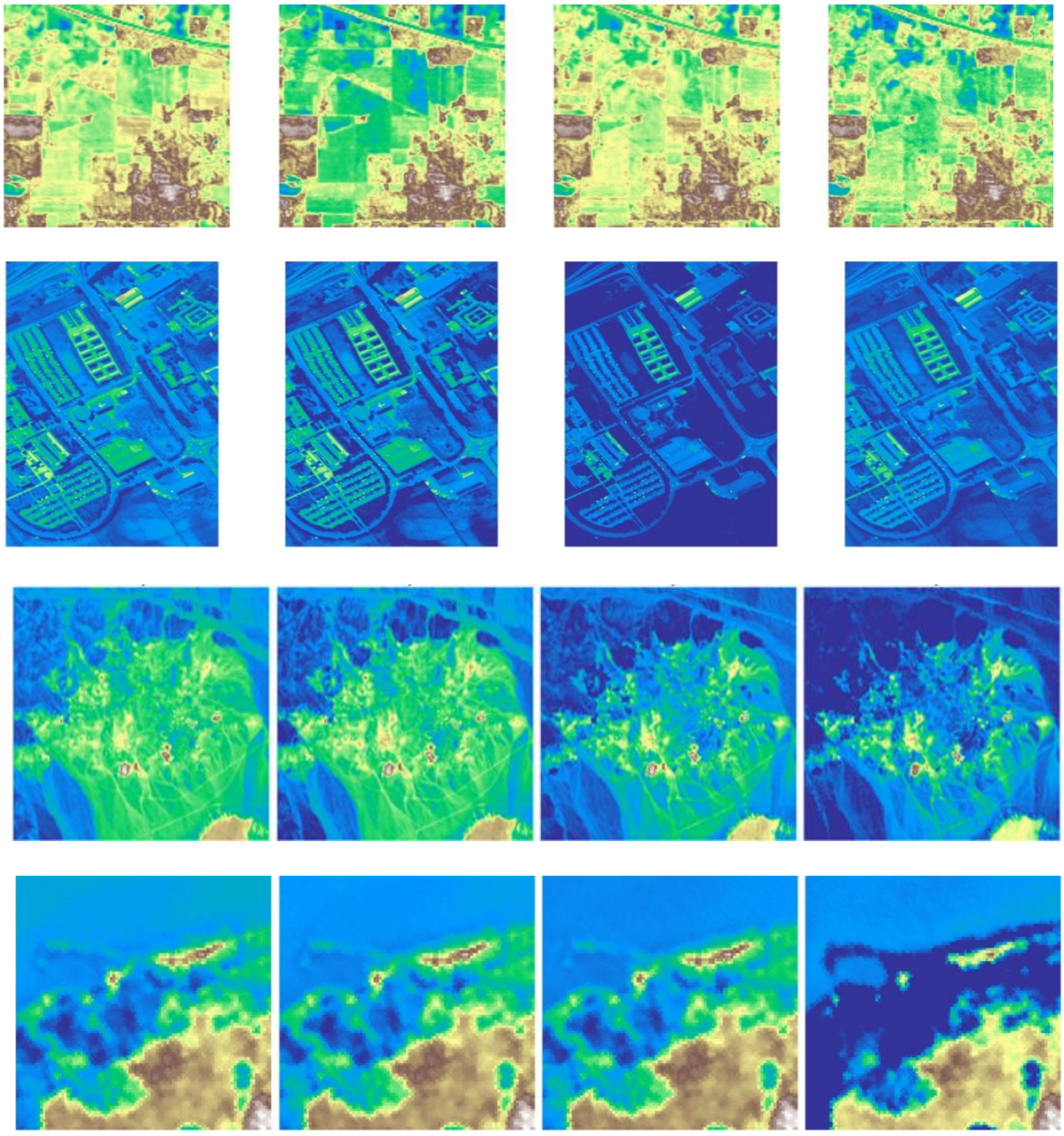


Figure 6.9 Reconstructed images for AEFS model for IP, PU, Cuprite, Samson datasets

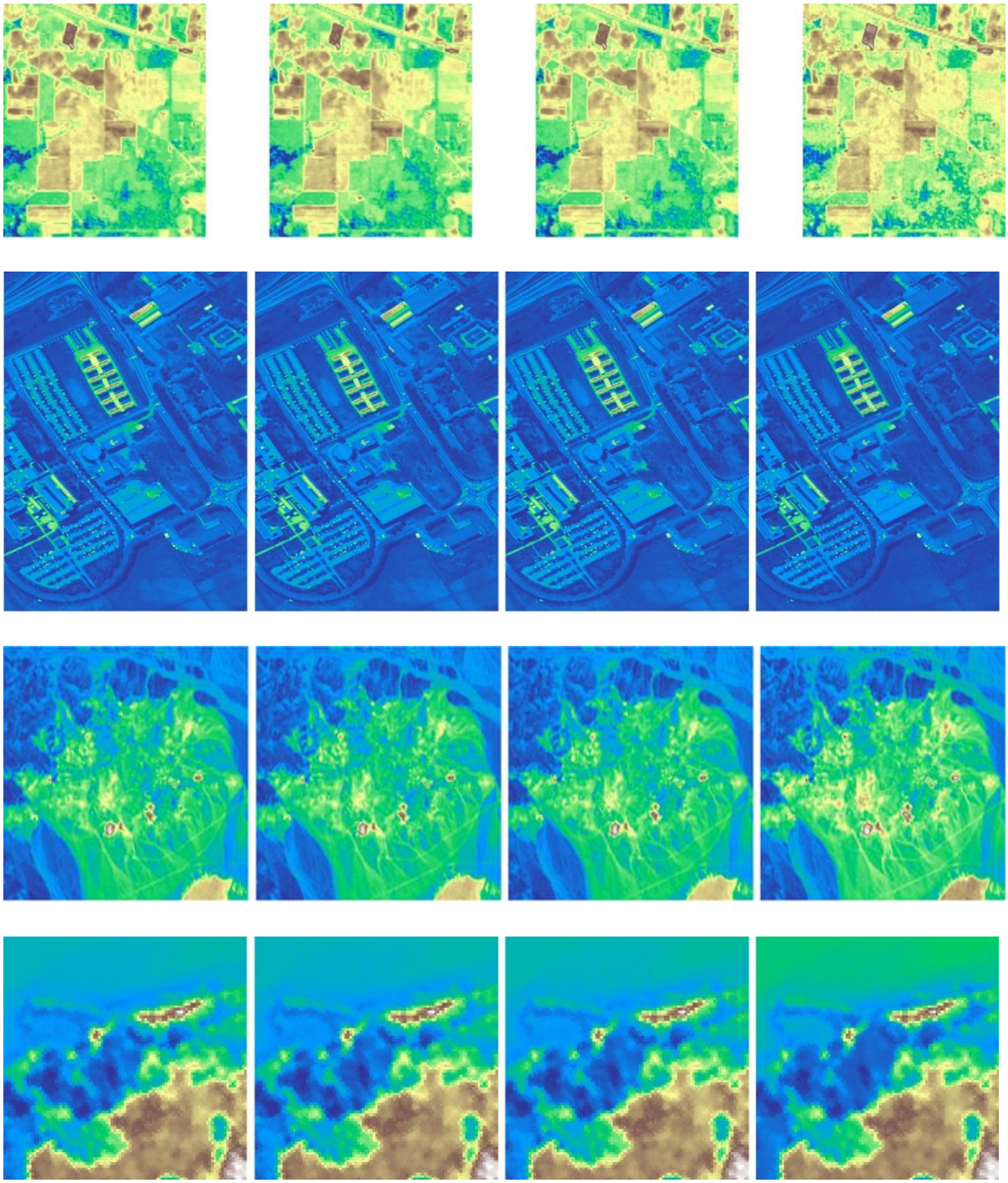
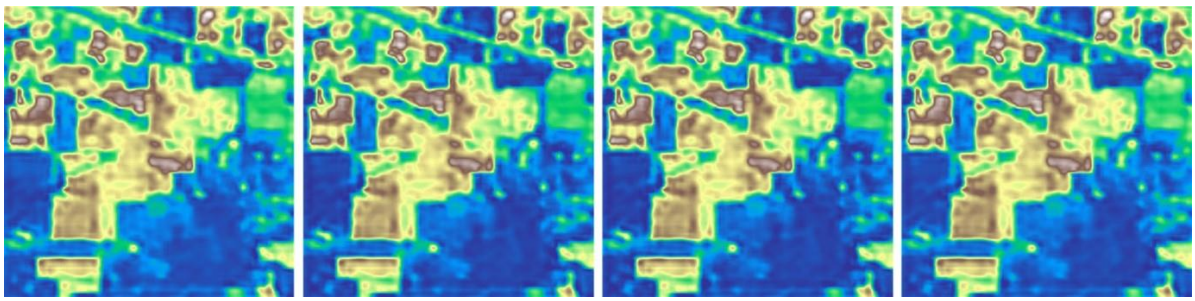


Figure 6.10 Reconstructed images for 1D-TSFS model for IP, PU, Cuprite, Samson datasets



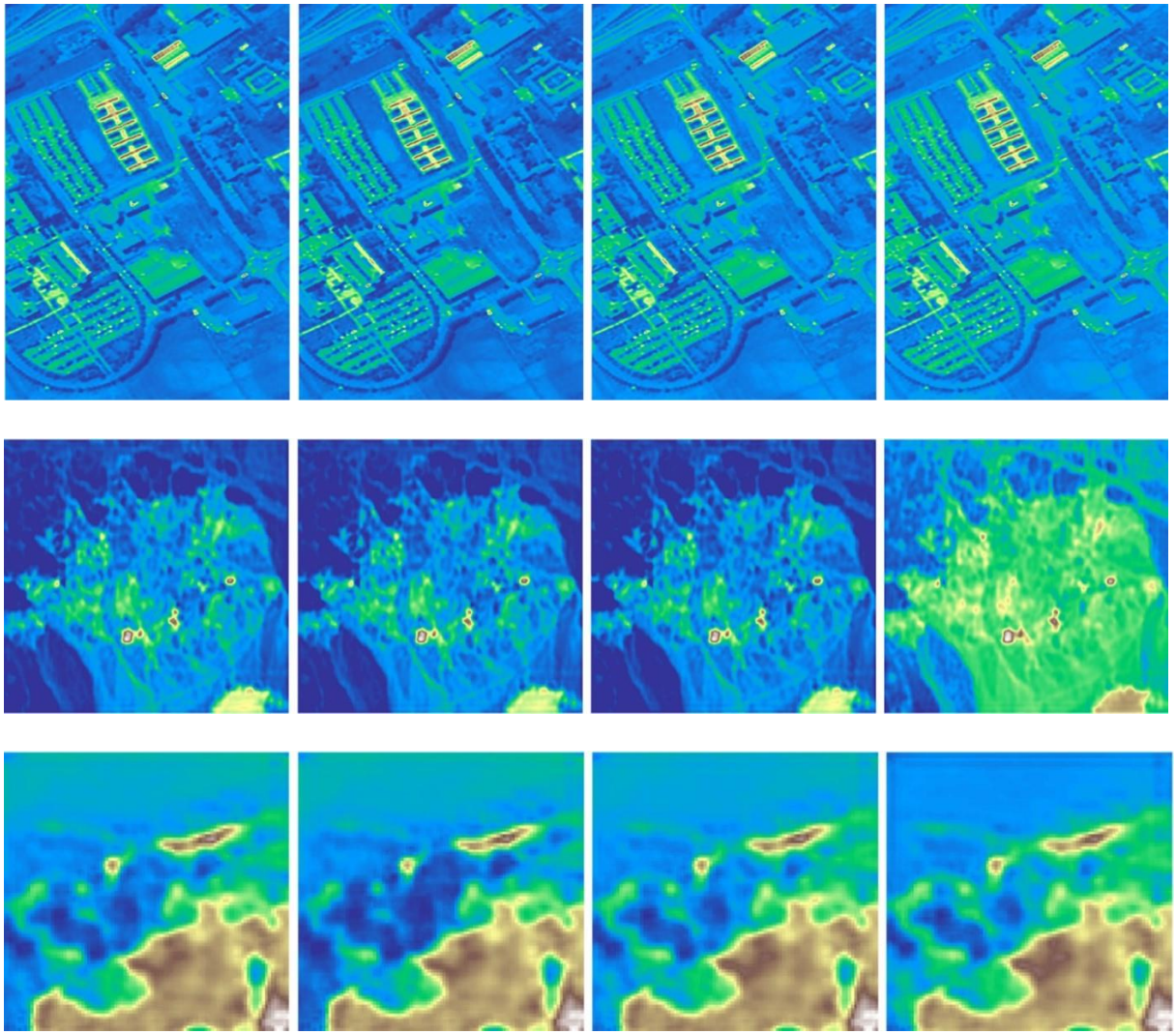


Figure 6.11 Reconstructed images for 2D-TSFS model for IP, PU, Cuprite, Samson datasets

From the above results, it can be concluded that 2D-TSFS models have better reconstruction capability compared to AEFS and 1D-TSFS models. From the overall analysis, it can be observed that the 2D-TSFS model plays a major role in the overall workflow. Compared to a fully connected Autoencoder, CAE can be used to reduce the number of trainable parameters and to preserve spatial information thus improving the overall analysis of the proposed scheme.

6.4.1 Sensitivity Analysis

Two parameters have to be fine-tuned in the proposed approach. The number of neurons in the hidden layer of the student network and the learning rate. It is tested on the Cuprite dataset (2D-TSFS) for the different number of nodes, batch size, epochs, and λ values. By trial and error, the number of neurons in the student network is set to be 90. Since the loss does not vary

much for different values of λ, it is fixed to be 0.1 for all the experiments. The proposed algorithm is unaffected by the value of λ. The fluctuation in loss with parameters like, number of nodes, and batch size were noted for the student network trained on the output of the 2D Teacher network. Table 6.6 summarizes the results of the experiment.

Table 6.6 Sensitivity Analysis

Nodes	Batch size	λ	Loss
90	100	10e-4	3.1493e-05
90	500	10e-4	6.4446e-05
90	1000	10e-4	2.8971e-05
90	500	10e-5	2.9908e-04
90	512	10e-5	2.0521e-05
90	1000	10e-5	6.6340e-05
100	500	10e-4	1.0020e-04
100	512	10e-4	2.2787e-04
100	1024	10e-4	2.7672e-05
100	1024	10e-5	1.1133e-04
110	500	10e-3	2.5471e-04
110	500	10e-4	9.5435e-05
110	500	10e-5	1.8665e-04

The next chapter provides a detailed description of GAN model for compact representation and virtual sample generation.

CHAPTER 7
COMPACT REPRESENTATION OF DATA USING GAN

7.1 INTRODUCTION

Recently, DL based models incorporating spatial information have received tremendous attention in HRS community. However, there are complications associated with hyperspectral image classification especially using deep neural networks. The large spectral size of the database introduces a lot of tunable parameters to the model and the unavailability of adequate training samples to train the model effectively. In particular, when CNNs are used overfitting phenomena occurs (Zhang et al. 2018). Recently, the advent of Generative Adversarial Networks (GAN) has received considerable interest in HRS since it can be used to generate virtual samples for training (Hennessy et al. 2021). Therefore, in this chapter, a supervised spatial-spectral feature learning strategy is proposed for hyperspectral data using GAN. The proposed technique seems to be extremely beneficial, since there is lack of training samples and ground truth information in HRS.

7.1.1 Data Augmentation

In order to artificially increase the size of a dataset for training machine learning models, a set of techniques known as Data Augmentation (DA) is used (Abdollahi et al. 2020). Expanding image datasets for use in DL models for computer vision tasks has been the main focus in recent years. In addition to enhancing the model ability to generalize, the most challenging problems can be addressed. The methods which alter the images contained in the initial training set are utilized in the idea that DA techniques can add extra information to the original dataset (Zhu et al. 2018). By data warping or oversampling, the augmentations artificially increase the size of the training dataset (Shorten and Khoshgoftaar 2019).

The conventional methods improved the performance of models to a limited extent since they modify the original images using different angles, sizes, and filters only. Consequently, efforts over the past few years have focused on creating new and improved strategies to assist model training with limited datasets and unbalanced classes (Buda et al. 2018). One of the main areas of research has been the use of generative models for the synthesis of images that replicate the diversity and quality of the original datasets and thereby enrich the amount of data available. A category of DL based models that has gained tremendous attention in the recent years are the GANs (Shin et al. 2018; Alipour et al 2020).

7.1.2 Generative Adversarial Networks

In essence, GANs are made up of two networks that compete with each other. One network creates fake data, and the other network classifies the data as real or fake. The two networks that are in competition are Generator and Discriminator. The discriminator network D tries to separate the fake images created by the generator from real images provided by a training dataset while the generator network G creates fake images (Hi et al. 2017; Li et al. 2021). Both networks are competitively trained together. The objective is to train the generator network to output images y that are fake image samples from random noise z. In other words, the objective of G is to create samples from the learning distribution and to predict the data distribution of a training dataset as accurately as feasible. The discriminator network D is used to optimize G during training by differentiating between actual images y and fake images y = G(z) produced by G. Formally, a noise vector z in the latent space is mapped to an image y by the generator.

$$G(z) \rightarrow y, \tag{7.1}$$

whereas the discriminator is defined as:

$$D(y) \rightarrow [0,1] \tag{7.2}$$

and classifies an image y as real (close to 1) or as fake (close to 0). The two networks are trained in a competitive fashion with backpropagation. The loss function is generally formulated as (Gao et al. 2019),

$$\min_G \max_D \mathcal{L}_{GAN}(G,D) = \min_G \max_D \mathbb{E}_{y \in Y}[\log D(y)] + \mathbb{E}_{z \in Z}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \tag{7.3}$$

where $\mathbb{E}$ denotes the expected value, Y the set of real images and Z denotes the latent space. The loss function is the adversarial loss. The term $\mathbb{E}_{y \in Y}[\log D(y)]$ represents the predicted log probability of D that y is real and the term $\mathbb{E}_{z \in Z}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$ represents the predicted log probability of D that G(z) is fake. D is a classification network that typically calculates the likelihood that an image belongs to class 1 (real) or class 0 (fake). The generator network G is used to create false data after the training operation is complete, while D is only used during the training phase to enhance the generato. The networks are specifically optimized by switching between training D and G, maximizing the GAN loss with respect to the parameters of the discriminator network D θ_D, and then minimizing the loss with respect to the parameters of the generator network G θ_G. As a result, D attempts to convey the term D(G(z)) from Equation 7.3 as near to 0 as possible, or when all (false) images produced by G are identified

and correctly labelled as fake by D. On the other hand, G aims to get the term D(G(z)) close to 1, which occurs when all (false) images are not recognized by D and are incorrectly classified as real. G will subsequently learn an estimation of the real data distribution using this method from the training dataset. The discriminator network is trained using real samples from the training set as well as fake samples that G generates. Figure 7.1 depicts the schematic of a GAN.

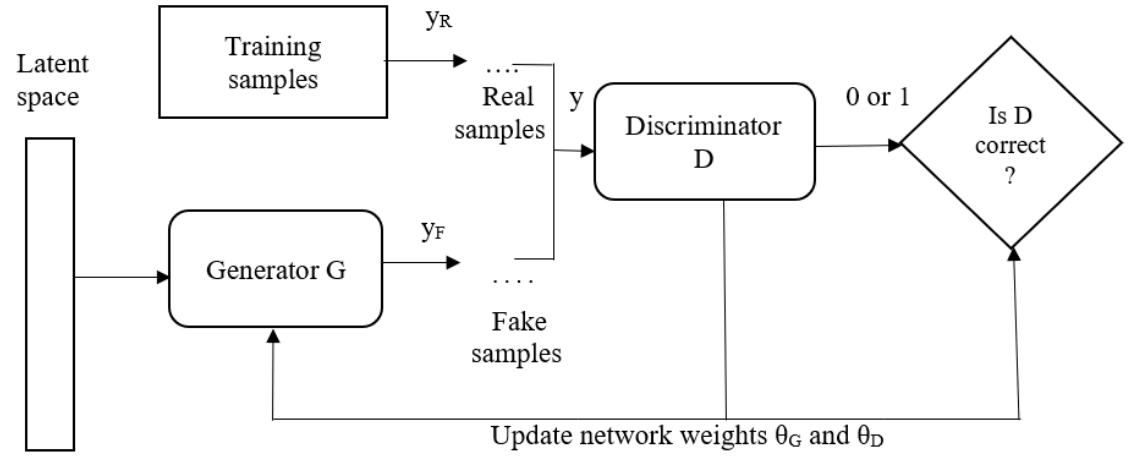


Figure 7.1. Schematic of GAN

Despite the outstanding results obtained by GANs in computer vision applications, (Yi et al. 2019, Wang et al. 2019) there are few issues that impairs training stability. These challenges arise regularly and pose a significant barrier in obtaining suitable generators for a variety of applications, resulting in a loss in the quality and diversity of the generated images (Neyshabur et al. 2017). The prevalent issues related to training instability (Li et al. 2020; Costa et al. 2020) of GANs are:

• Mode Collapse: A circumstance in which the generator can only synthesize a small subset of images from the whole distribution since the training did not allow for generalization of the richness of variants of the original images.

• Vanishing Gradient: The condition occurs when the discriminator or generator gets powerful enough to induce an irreversible imbalance in training and when suitable cost functions are not used to obtain adequate learning gradients. This stops the other network from improving its performance, leading to a standoff.

It is crucial to select proper hyperparameters, architectures, and training technique in order to generate an effective generator due to the aforementioned issues.

7.2 DEEP CONVOLUTIONAL GAN (DCGAN)

DCGANs were proposed to improve the training instability of conventional GAN (Vanilla GAN) where Multi-Layer Perceptron (MLP) are used in both generator and discriminator (Radford et al. 2015). The training procedure remains the same, except the design of generator and discriminator using CNN architecture. The discriminator employs a standard CNN since the discriminator classification task is the same as the conventional classification tasks for which CNNs are frequently employed. Simultaneously, a deconvolutional network is necessary for the generator. The network is similar to traditional CNN, with the main difference being that, as a generative model, it is responsible for synthesising images from high level information provided by random noise. The CNN convolutional layers are replaced by deconvolutional layers in the deconvolutional network. Additionally, the pooling layers are not used in the architecture as the goal is to extend the feature maps (upsampling), not to reduce them (downsampling) as in CNN.

The activation layers are another difference compared to CNNs. Sigmoid or softmax functions are typically employed after the fully connected layer that performs classification in CNNs, while ReLU is frequently utilized as the activation function in each convolutional layer. ReLU is also employed in the generator of a DCGAN, however the tanh function is used in the final layer since the training images are normalised in the range [-1,1]. The last layer which classifies between false and real images, employs the sigmoid function to obtain a binomial probability, while all the other layers use Leaky ReLU activation for the discriminator. Since its inception, DCGAN has full-fledged to serve as a standard model for image synthesis in numerous state-of-the-art works.

7.3 Methodology

In the current study, two DCGAN architectures are implemented on IP dataset with spatial dimensions 145×145 and 200 spectral bands. 1D GAN uses only spectral features and 3D GAN uses both spatial and spectral features for processing. PCA is employed for reducing the spectral bands from 200 to 10 for 1D GAN and 3 for 3D GAN. Batch normalization is used to stabilize training. The methodology adopted for DCGAN is presented in Figure 7.2.

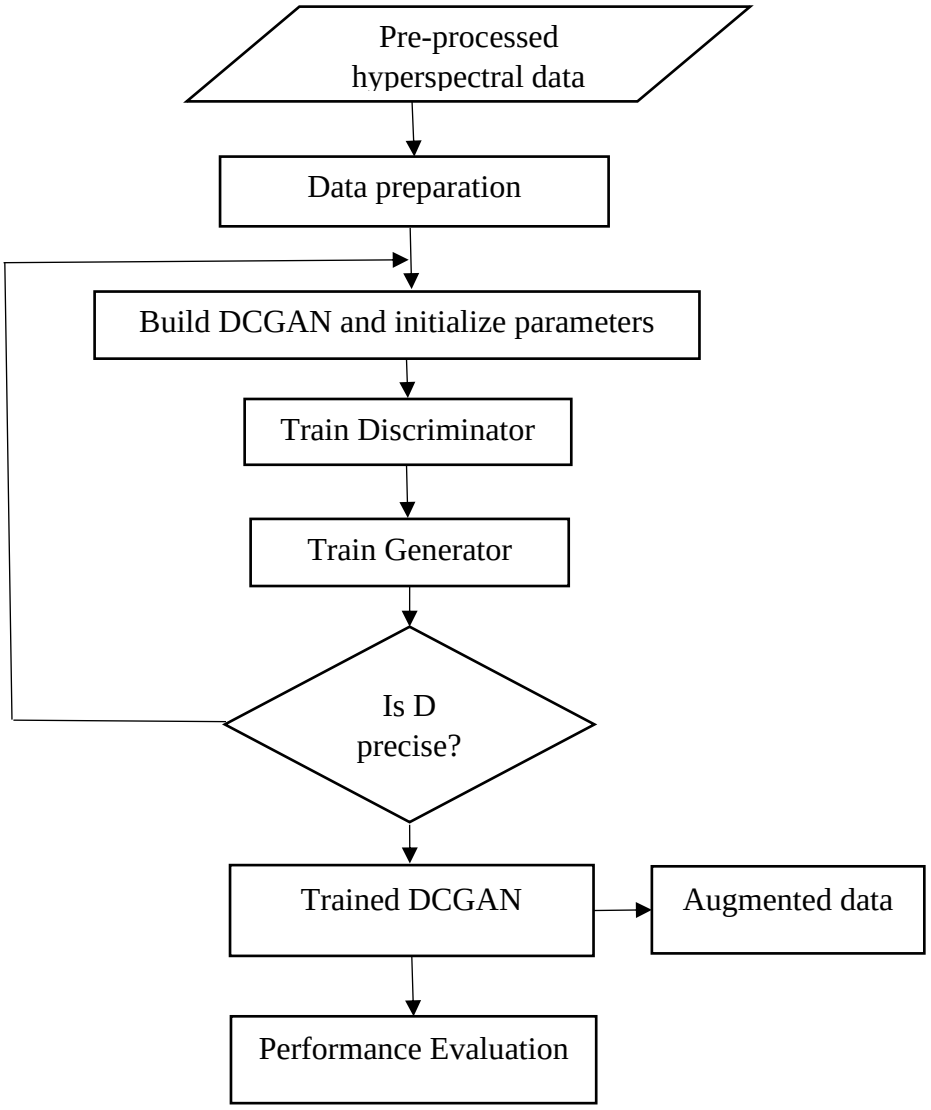


Figure 7.2. Flowchart of DCGAN

The architecture design criteria based on the original DCGAN study (Radford et al. 2015) is used. The number of convolution filters used are 32, 64, 128 and 256 in different convolutional layers. The specifications related to activation functions and optimizers are as follows:

Activation Functions:

Discriminator: LeakyReLU with alpha = 0.2, sigmoid, softmax

Generator: ReLU, tanh

Optimizers:

Discriminator: Adam with learning rate 0.001, beta1 0.9

Generator: Adam with learning rate 0.002, beta1 0.5

7.4 RESULTS AND DISCUSSION

The visualization of the obtained results for 3D GAN is done by comparing the real and synthetic spectrum for all the classes. The spectrum graphs show the mean and standard deviation for classes 1 and 12 respectively. All the spectral features, including the mean and standard deviation are reproduced accurately by the GAN model as shown in Figure 7.3. The DCGAN employed in the current study improves the performance of the model to a considerable extent compared to shallow layers in MLP with limited functionality. classification results and comparative results with other state-of-the-art methods 3D GAN (Zhu et al. 2018) for spectral-spatial FE, 3D bilateral filter (3DBF) (He et al. 2017) are depicted in Table 7.1 and visualized in Figure 7.4. From the obtained results, it is evident that, the proposed 3D GAN model outperforms other methods.

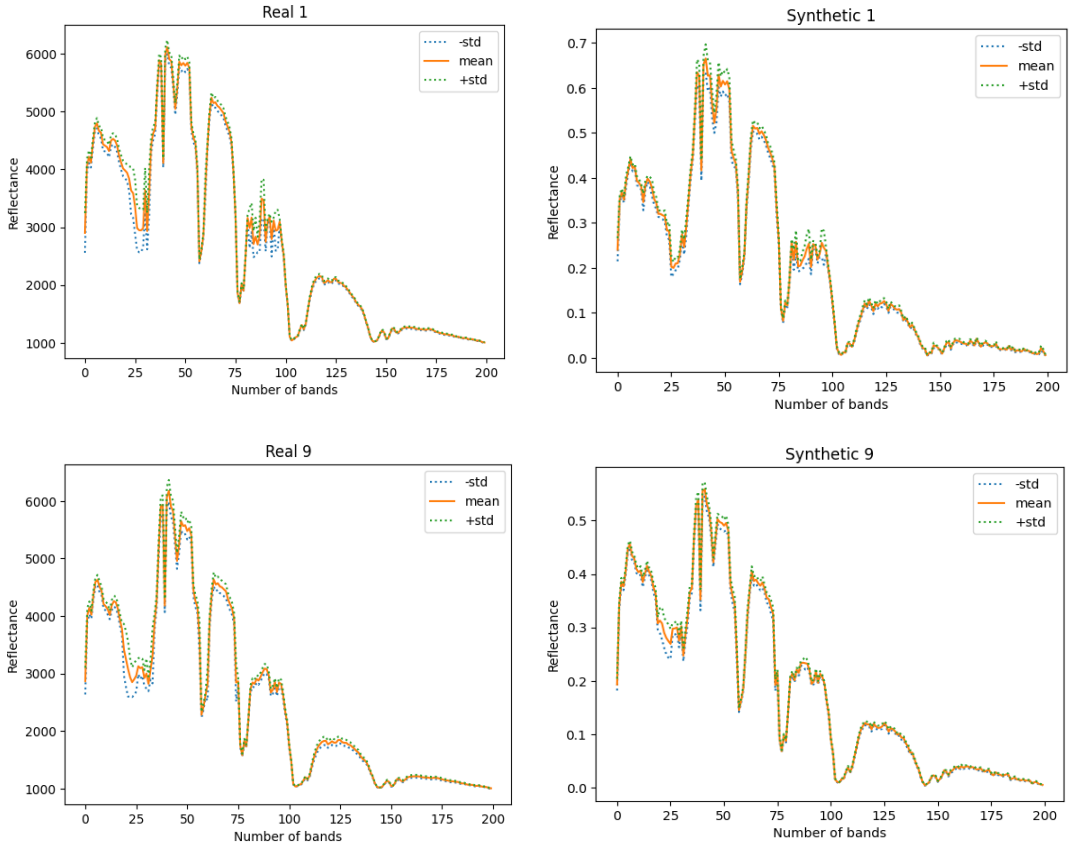


Figure 7.3 Spectrum of various classes generated by GAN model

Table 7.1 Classification results on GAN model

Model	Classification accuracy
1D GAN	69.13
3D GAN (proposed)	**90.93**
3DBF GAN	75.62
3D GAN	89.09

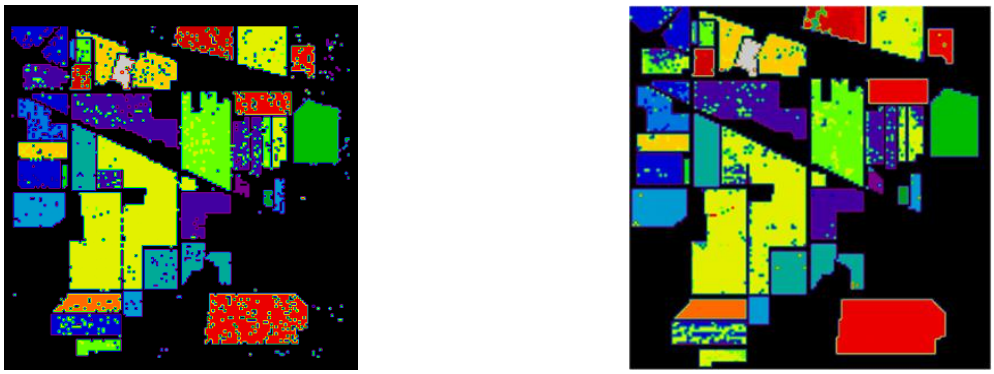


Figure 7.4 Classification results of GAN on IP dataset

The deep CNN obtains greater classification accuracy with the help of GANs, and the overfitting problem raised by CNNs is significantly minimized. For HSI classification, two frameworks, 1D GAN and 3D GAN are proposed, and the classification results produced by these two frameworks show that GANs outperform classic CNNs even with minimal training sets. In addition, GAN also creates samples that can be utilized as virtual samples. The proper application of virtual samples increases classification performance. The generated samples are employed to improve classification accuracy and the experimental results demonstrate the efficacy of generated samples. DCGAN is highly appropriate for image processing due to the benefits of CNN. Furthermore, experiments are carried out with augmented data and different input patch size to explore the efficacy of the proposed model. The augmented dataset provides classification accuracy of 92.47% i.e., there 1.54% increase in accuracy for augmented data. The results with different patch sizes are tabulated in Table 7.2.

Table 7.2 Classification accuracy of GAN model with different patch size

Patch size	Accuracy
(64,64,3)	94.42
(32,32,5)	93.07
(16,16,7)	87.77

CHAPTER 8
CONCLUSION AND FUTURE PERSPECTIVES

8.1 SUMMARY

The field of HRS, often known as image spectroscopy is interesting as well as challenging. When the whole spectral signatures of the different classes are accessible, classification issues can be addressed more efficiently compared to few broad spectral bands. Simultaneously, the tremendous increase in spectrum bands has pushed the limitations of spectral FE from remote sensing imagery. Traditional approaches that employ complete spectral bands frequently fail when applied to hyperspectral data due to the curse of dimensionality. Alternatively, index-based features do not fully exploit the existing information content. They are either too general or fail to address the complicated challenges of hyperspectral data applications. Furthermore, empirically produced application-specific indicators frequently lack robustness. Hence, the high dimensionality of the data which leads to a high intrinsic information redundancy and Hughes phenomenon is still an open issue. Consequently, addressing the above challenges requires the development of a suitable FE or BS strategy which reduces the size of hyperspectral image without compromising the classification accuracy. In this regard, the thesis deals with advanced spectral and spatial approaches for DR of hyperspectral data.

8.2 CONCLUSIONS

To address the problem of the curse of dimensionality, four DR approaches have been proposed and developed for hyperspectral image classification. DR facilitates better visualization and analysis of data. In order to analyze the performance of the proposed methods, detailed experimentation has been conducted over five widely used standard hyperspectral datasets namely, Indian Pines, Pavia University, Salinas, Cuprite and Samson datasets. The obtained results are compared with other state-of-the-art approaches. The results of the study can be summarized in terms of particular objectives.

Objective 1: To explore conventional feature extraction techniques, application on hyperspectral mineral data and its evaluation based on the co-ranking framework.

- Majority of the literature works use mean squared error or classification accuracy, using the original labels for finding the error after DR. However, the proposed approach based

on ranking criteria provides a detailed analysis on structure preserving property and works well in most of the cases. In addition, it helps to trace out the loss of information after DR process.

- Since DL is a relatively new area of research, there are limited studies which compares conventional methods with DL methodologies, but not comprehensive. Both linear and nonlinear strategies for DR have been compared in numerous papers. In particular, studies related to DR of mineral exploration is scarce.
- In the current research, the performance of 15 DR techniques has been analyzed for mineral exploration. Different DR methods are compared in terms of loss of quality which clearly brings out the performance of a DR technique. Since it is a benchmarking study, more focus towards technical differences between DR techniques is not provided. Instead, the applicability of quality assessment is well explored. This study will be extremely useful to select the best method for a particular application. In this case, deep autoencoders proved to be the best technique with a quality loss value of 0.0062 for mineral exploration.
- The relationship between the clustering quality and quality indices has been introduced for the first time in hyperspectral remote sensing literature. It clearly brings out the interdependencies between them. The clustering accuracy is positively correlated with quality indices with an R^2 value of 0.7691 and similarly NMI is also positively correlated with an R^2 value of 0.7815.

Objective 2: To analyze the impact of feature extraction strategies on hybrid CNN and design an efficient model for compression of hyperspectral imagery.

- The conventional DR techniques often do not lead to discriminative features required for efficient classification. In addition, they depend on hand-crafted features and require careful parameter tuning and domain expertise.
- Since DL techniques have achieved tremendous attention in HRS, a hybrid CNN architecture with mixed 2D and 3D CNN has been proposed to test the performance of different FE strategies with different dimensions. A detailed experiment is carried out on different parameters and the results are reported under different scenarios.
- The proposed model is compact with less trainable parameters. IP dataset provides a classification accuracy of 98.18%, PU 99.72% and SA 99.82% for patch size of 11×11 and 15 bands. Furthermore, statistical analysis tests reveal the performance of the proposed scheme for each class. From the experimental results, it can be clearly pointed

out that, an increase in dimensions will not improve the classification accuracy to a great margin as the number of bands increases, discriminative features obtained are less.

- Time complexity analysis is also carried out and PCA performs well in all scenarios when compared with other state-of-the-art methods.

Objective 3: To compress deep neural networks using knowledge distillation and develop an integrated model for deep FS of hyperspectral data.

- A novel teacher-student scheme for deep FS inspired by KD is proposed and is demonstrated that KD can improve the model generalization capability as it is less sensitive to parameter tuning. Extensive testing on four different hyperspectral datasets revealed the capability of the Deep TSFS scheme when compared to the conventional AEFS model.
- 2D-TSFS models outperform the other two models on all datasets in both supervised and unsupervised scenarios. The proposed technique uses fewer training parameters, resulting in a more compact model than standard fully connected models without compromising classification accuracy. The corresponding results of the 2D-TSFS model are 96.15% for the IP dataset and 97.82% for the PU dataset.
- Also, an attempt has been made to couple the process of DR with Spectral Unmixing as well. The clustering results prove that the proposed scheme provides optimal clustering performance for the 2D-TSFS model with maximum Silhouette index, CH index as 0.8038, 10173.90 for Cuprite, and 0.7312, 7771.57 for Samson datasets respectively. Unlabelled as well as data-less scenarios can leverage the knowledge distilled from large models for resource-constrained devices.

Objective 4: To design a DL model based on GANs for virtual sample generation and compact representation of hyperspectral data.

- Since GANs have complex structures, DCGAN incorporates specialized imaging capabilities by employing convolutions and provides improved training stability which aids to reduce mode collapse.
- By visual inspection, it can be clearly pointed out that the quality of the synthetic images generated is very close to that of the real ones. This could be an indication of the absence of mode collapse, which points to a more stable training of GANs.

- The proposed 3D GAN provides a better classification accuracy of 90.93% compared to 1D GAN with 69.13% since only spectral features are considered in the analysis. Furthermore, the augmented dataset delivers an accuracy of 92.47% which indicates 1.54% increase in the result provided by DCGAN. Hence, it can be concluded that the proper utilization of generated samples leads to improvement in the classification accuracy and solves the limited training samples issue.

8.3 LIMITATIONS AND FUTURE PERSPECTIVES

- In the current study, the proposed methodologies are restricted to only five publicly accessible hyperspectral datasets acquired by various hyperspectral sensors. It would be quite interesting to further extend and apply the concepts for real-time applications such as a) forest management, b) track pollution levels, c) food and drug administration, d) plant pathology and e) map hydrological patterns etc.
- Future study will examine possibly more efficient 3D CNN based HSI classification methods that can employ unlabelled samples. Unlabelled samples are far more accessible in HSI than labelled samples which are not fully utilized by supervised classification techniques based on 3D CNN. To more effectively handle this problem, it would be preferable to combine unsupervised and semi-supervised classification algorithms based on 3D CNN.
- Different regularization techniques and network architectures can be deployed in the future to unleash the potential of deep FS based on KD and GANs for hyperspectral data.

REFERENCES

Abdollahi, B., Tomita, N. and Hassanpour, S., (2020). "Data augmentation in training deep learning models for medical image analysis." *Deep learners and deep learner descriptors for medical applications*, 167-180.

Adep, R.N., Vijayan, A.P., Shetty, A. and Ramesh, H., (2016). "Performance evaluation of hyperspectral classification algorithms on AVIRIS mineral data." *Perspectives in Science, 8*, 722-726.

Agrafiotis, D.K., (2003). "Stochastic proximity embedding." *Journal of computational chemistry*, 24(10), 1215-1221.

Ahmad, M., Shabbir, S., Oliva, D., Mazzara, M. and Distefano, S., (2020). "Spatial-prior generalized fuzziness extreme learning machine autoencoder-based active learning for hyperspectral image classification. *Optik*, 206, 163712.

Ali Mirzaei, Vahid Pourahmadi, Mehran Soltani, Hamid Sheikhzadeh, (2020). "Deep feature selection using a teacher-student network." Neurocomputing, 383, 396-408.

Alipour Fard, T., & Arefi, H. (2020). "Structure aware generative adversarial networks for hyperspectral image classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5424-5438.

Audebert, N., Le Saux, B., & Lefèvre, S. (2018). "Generative adversarial networks for realistic synthesis of hyperspectral samples." In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 4359-4362). IEEE.

Babaee, Mohammadreza, Mihai Datcu, and Gerhard Rigoll, (2013). "Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization." *IEEE international conference on big data*, 1-6.

Bachmann C M, T. L. Ainsworth, and R. A. Fusina, (2005). "Exploiting manifold geometry in hyperspectral imagery." *IEEE transactions on Geoscience and Remote Sensing*, 43(3), 441-454.

Bachmann, C.M., Ainsworth, T.L. and Fusina, R.A., (2004). "Improvements to land-cover and invasive species mapping from hyperspectral imagery in the Virginia Coast reserve." *IEEE International Geoscience and Remote Sensing Symposium*, 6, 4180-4183.

Bai, J., Xiang, S., Shi, L., & Pan, C. (2015). "Semisupervised pair-wise band selection for hyperspectral images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 2798-2813.

Banerjee, A. and Banik, D., (2022). "Pooled hybrid-spectral for hyperspectral image classification." *Multimedia Tools and Applications*, 1-13.

Behnood Rasti, Danfeng Hong, Renlong Hang, Pedram Ghamisi, Xudong Kang, Jocelyn Chanussot, Jon Atli Benediktsson, (2020). "Feature Extraction for Hyperspectral Imagery: The

evolution from shallow to deep: Overview and toolbox." *IEEE Geoscience and Remote Sensing Magazine,* 8(4), 60-88.

Beirami, B.A. and Mokhtarzade, M., (2020). "An Automatic Method for Unsupervised Feature Selection of Hyperspectral Images Based on Fuzzy Clustering of Bands." *Traitement du Signal*, 37(2).

Belkin M and Niyogi P., (2001). "Laplacian eigenmaps and spectral techniques for embedding and clustering." *Advances in Neural Information Processing Systems,* 14, MIT Press, 585–591.

Benedetto, J., Czaja, W., Dobrosotskaya, J., Doster, T., Duke, K. and Gillis, D., (2012). "Semi-supervised learning of heterogeneous data in remote sensing imagery." In *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X* , 8401 34-45.

Borg, Groenen P, (1997). "Modern Multidimensional Scaling: Theory and Applications." *Springer Science and Business Media,* New York.

Boukhechba, K., Wu, H. and Bazine, R., (2018). "DCT-based preprocessing approach for ICA in hyperspectral data analysis." *Sensors,* 18(4), 1138.

Bruni, V., Monteverde, G. and Vitulano, D., (2022). "An Entropy-Based Speed Up For Hyperspectral Data Classification Via CNN." In *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS),* 1-5.

Buda, M., Maki, A. and Mazurowski, M.A., (2018). "A systematic study of the class imbalance problem in convolutional neural networks." *Neural networks,* 106, 249-259.

Cao, X., Wu, B., Tao, D., & Jiao, L. (2016). "Automatic band selection using spatial-structure information and classifier-based clustering." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(9), 4352-4360.

Cao, X., Xu, L., Meng, D., Zhao, Q., & Xu, Z. (2017). "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification." *Neurocomputing*, *226*, 90-100.

Chen, Y., Jiang, H., Li, C., Jia, X. and Ghamisi, P., (2016). "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks." *IEEE Transactions on Geoscience and Remote Sensing,* 54(10), 6232-6251.

Chen, Y., Lin, Z., Zhao, X., Wang, G. and Gu, Y., (2014). "Deep learning-based classification of hyperspectral data." *IEEE Journal of Selected topics in applied earth observations and remote sensing*, *7*(6), 2094-2107.

Chen, Y., Zhu, L., Ghamisi, P., Jia, X., Li, G. and Tang, L., (2017). "Hyperspectral images classification with Gabor filtering and convolutional neural network." *IEEE Geoscience and Remote Sensing Letters*, *14*(12), 2355-2359.

Cheng, Wang D, Zhou P, Zhang T, (2018). "Model compression and acceleration for deep neural networks: The principles, progress, and challenges." *IEEE Signal Processing Magazine,* 35(1), 126-136.

Clark, R.N. and Swayze, G.A., (1995). “Mapping minerals, amorphous materials, environmental materials, vegetation, water, ice and snow, and other materials: the USGS Tricorder algorithm.” In *JPL, Summaries of the Fifth Annual JPL Airborne Earth Science Workshop. Volume 1: AVIRIS Workshop*.

Coifman R R and Lafon S, (2006). “Diffusion maps.” *Applied and computational harmonic analysis,* 21(1), 5-30.

Costa, V., Lourenço, N., Correia, J. and Machado, P., (2020). “Neuroevolution of generative adversarial networks.” *Deep Neural Evolution: Deep Learning with Evolutionary Computation*, 293-322.

Deepa, C., Shetty, A. and Narasimhadhan, A.V., (2020). “Quality assessment of dimensionality reduction techniques on hyperspectral data: A neural network-based approach.” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 389-394.

Demartines, P. and Herault, J., (1997). “Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets.” *IEEE Transactions on neural networks*, 8(1), 148-154.

DeMers, D. and Cottrell, G., (1992). “Non-linear dimensionality reduction.” *Advances in neural information processing systems, 5*.

Du, Q. and Fowler, J.E., (2007). “Hyperspectral image compression using JPEG2000 and principal component analysis.” *IEEE Geoscience and Remote sensing letters*, 4(2), 201-205.

Elsken T, Jan Hendrik Metzen, Frank Hutter, (2019). “Neural Architecture Search: A Survey. Journal of Machine Learning Research.” 20(55), 1-21.

Espadoto, Mateus, Rafael M. Martins, Andreas Kerren, Nina ST Hirata, and Alexandru C. Telea, (2019). "Toward a quantitative survey of dimension reduction techniques." *IEEE transactions on visualization and computer graphics* 27(3), 2153-2173.

Fauvel M, J. Chanussot, and J. Benediktsson, (2009). “Kernel principal component analysis for the classification of hyperspectral remote sensing data of urban areas.” *EURASIP Journal on Advances in Signal Processing*, 783194, 1-14.

Feng, J., Jiao, L., Liu, F., Sun, T., & Zhang, X. (2014). “Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy.” *IEEE transactions on geoscience and remote sensing*, *53*(5), 2956-2969.

Feng, J., Jiao, L., Sun, T., Liu, H., & Zhang, X. (2016). “Multiple kernel learning based on discriminative kernel clustering for hyperspectral band selection.” *IEEE Transactions on Geoscience and Remote Sensing*, *54*(11), 6516-6530.

Feng, S., Itoh, Y., Parente, M., & Duarte, M. F. (2017). “Hyperspectral band selection from statistical wavelet models.” *IEEE Transactions on Geoscience and Remote Sensing*, *55*(4), 2111-2123.

Gao, H., Yao, D., Wang, M., Li, C., Liu, H., Hua, Z., & Wang, J. (2019). "A hyperspectral image classification method based on multi-discriminator generative adversarial networks." *Sensors, 19*(15), 3269.

Gao, P., Wang, J., Zhang, H., & Li, Z. (2018). "Boltzmann entropy-based unsupervised band selection for hyperspectral image classification." *IEEE Geoscience and Remote Sensing Letters, 16*(3), 462-466.

Geng, X., Sun, K., Ji, L., & Zhao, Y. (2014). "A fast volume-gradient-based band selection method for hyperspectral image." *IEEE Transactions on Geoscience and Remote Sensing*." *52*(11), 7111-7119.

Goodenough, D.G., Pearlman, J., Chen, H., Dyk, A., Han, T., Li, J., Miller, J. and Niemann, K.O., (2004). "Forest information from hyperspectral sensing." *IEEE International Geoscience and Remote Sensing Symposium,* 4, 2585-2589.

Gracia, A., Gonzalez, S., Robles, V. and Menasalvas, E., (2014). "A methodology to compare dimensionality reduction algorithms in terms of loss of quality." *Information Sciences*, 270, 1-27.

Green, A.A., Berman, M., Switzer, P. and Craig, M.D., (1988). "A transformation for ordering multispectral data in terms of image quality with implications for noise removal." *IEEE Transactions on geoscience and remote sensing*, 26(1), 65-74.

Green, R.O., Eastwood, M.L., Sarture, C.M., Chrien, T.G., Aronsson, M., Chippendale, B.J., Faust, J.A., Pavri, B.E., Chovit, C.J., Solis, M. and Olah, M.R., (1998). "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)." *Remote sensing of environment*, 65(3), 227-248.

Hamida, A.B., Benoit, A., Lambert, P. and Amar, C.B., (2018). "3-D deep learning approach for remote sensing image classification." *IEEE Transactions on geoscience and remote sensing*, *56*(8), 4420-4434.

Han Song, Huizi Mao, William J. Dally. (2016). "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization, and Huffman Coding." *International Conference on Learning Representations,* 1-14.

Han, K., Wang, Y., Zhang, C., Li, C. and Xu, C., (2018). "Autoencoder inspired unsupervised feature selection." *IEEE international conference on acoustics, speech and signal processing* 2941-2945.

Hao, S., Wang, W., Ye, Y., Nie, T. and Bruzzone, L. (2018). "Two-stream deep architecture for hyperspectral image classification", *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2349–2361.

Harsanyi, J.C. and Chang, C.I., (1994). "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach." *IEEE Transactions on geoscience and remote sensing*, 32(4), 779-785.

Hassani M and Seidl T, (2017). “Using internal evaluation measures to validate the quality of diverse stream clustering algorithms.” *Vietnam Journal of Computer Science,* 4, 171-183.

He, M., Li, B. and Chen, H., (2017), “Multi-scale 3D deep convolutional neural network for hyperspectral image classification.” *IEEE International Conference on Image Processing,* 3904-3908.

He, Z., Liu, H., Wang, Y., & Hu, J. (2017). “Generative adversarial networks-based semi-supervised learning for hyperspectral image classification.” *Remote Sensing*, *9*(10), 1042.

Hennessy, A., Clarke, K., & Lewis, M. (2021). “Generative adversarial network synthesis of hyperspectral vegetation data.” *Remote Sensing*, *13*(12), 2243.

Hinton G, Oriol Vinyals, Jeff Dean, (2015). “Distilling the Knowledge in a Neural Network. Neural and Evolutionary Computing.” *Deep Learning Workshop,* 1-14.

Hinton G, Oriol Vinyals, Jeff Dean. (2015). “Distilling the Knowledge in a Neural Network”. Neural and Evolutionary Computing, *Deep Learning Workshop,* 1-14.

Hinton G., and Roweis, (2003). “Stochastic neighbour embedding.” *Advances in Neural Information Processing Systems,* 15, 833–840.

Hinton, G.E., Salakhutdinov, R.R., (2006). “Reducing the dimensionality of data with neural Networks.” *Science,* 313(5786), 504–507.

Hughes, G., (1968). “On the mean accuracy of statistical pattern recognizers.” *IEEE transactions on information theory*, 14(1), 55-63.

Huilin Xu, Hongyan Zhang, Wei He, Liangpei Zhang, 2019. Superpixel-based spatial-spectral dimension reduction for hyperspectral image classification, Neurocomputing, 300, pp.138-150.

Ibarrola Ulzurrun, E., Marcello, J. and Gonzalo-Martin, C., (2017). “Assessment of component selection strategies in hyperspectral imagery.” *Entropy*, 19(12), 666.

Jia, S., Tang, G., Zhu, J., & Li, Q. (2015). “A novel ranking-based clustering approach for hyperspectral band selection.” *IEEE Transactions on Geoscience and Remote Sensing*, *54*(1), 88-102.

Jia, S., Zhao, Q. and Li, Y., (2022). Hyperspectral Remote Sensing Image Classification Based on Partitioned Random Projection Algorithm. *Remote Sensing*, 14(9), 2194.

Jia, Y., Shi, Y., Luo, J. and Sun, H., (2023). “Y–Net: Identification of Typical Diseases of Corn Leaves Using a 3D–2D Hybrid CNN Model Combined with a Hyperspectral Image Band Selection Module.” *Sensors*, *23*(3), 1494.

Jiang, J., Ma, J., Chen, C., Wang, Z., Cai, Z. and Wang, L., (2018). “SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery.” *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), 4581-4593.

Jiao, L., Feng, J., Liu, F., Sun, T., & Zhang, X. (2014). “Semisupervised affinity propagation based on normalized trivariable mutual information for hyperspectral band selection.” *IEEE*

Journal of Selected Topics in Applied Earth Observations and Remote Sensing, *8*(6), 2760-2773.

Jiao, L., Feng, J., Liu, F., Sun, T., & Zhang, X. (2014). "Semisupervised affinity propagation based on normalized trivariable mutual information for hyperspectral band selection." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(6), 2760-2773.

Jin Sian, Sheng Di, Xin Liang, Jiannan Tian, Dingwen Tao, Franck Cappello, (2019). "DeepSZ: A Novel Framework to Compress Deep Neural Networks by Using Error-Bounded Lossy Compression." *International Symposium on High-Performance Parallel and Distributed Computing,* 159-170.

Kaewpijit, S., Le Moigne, J. and El-Ghazawi, T., (2003). "Automatic reduction of hyperspectral imagery using wavelet spectral analysis." *IEEE transactions on Geoscience and Remote Sensing*, 41(4), 863-871.

Kanjilal P, Dey P K, Banerjee D N, (1993). "Reduced-size neural networks through singular value decomposition and subset selection." *Electronics Letters,* 29(17), 1516-1518.

Kim D and Finkel L, (2003). "Hyperspectral image processing using locally linear embedding." *First International IEEE EMBS Conference in Neural Engineering,* 316 -319.

Koren, Y. and Carmel, L., (2004). "Robust linear dimensionality reduction." *IEEE transactions on visualization and computer graphics*, 10(4), 459-470.

Kruse, F.A., Boardman, J.W. and Huntington, J.F., (2003). "Comparison of airborne hyperspectral data and EO-1 Hyperion for mineral mapping." *IEEE transactions on Geoscience and Remote Sensing*, 41(6), 1388-1400.

Kruse, F.A., Boardman, J.W., and Huntington, J.F, (2003). "Comparison of airborne hyperspectral data and EO-1 Hyperion for mineral mapping," *IEEE Transactions on Geoscience and Remote Sensing*, 41(6), 1388–1400.

Kumar, A. and Kumar, R., (2016). "Manifold learning using linear local tangent space alignment (LLTSA) algorithm for noise removal in wavelet filtered vibration signal." *Journal of Nondestructive Evaluation,* 35(3), 50.

Lavander Maaten and Hinton G., (2008). "Visualizing data using t-sne." *Journal of Machine Learning Research,* 9, 2579-2605.

Lee, C. and Landgrebe, D.A., (1993). "Analyzing high-dimensional multispectral data." *IEEE Transactions on Geoscience and Remote Sensing*, 31(4), 792-800.

Lee, J.A. and Verleysen, M., (2009). "Quality assessment of dimensionality reduction: Rank-based criteria." *Neurocomputing,* 72(7-9), 1431-1443.

Lennon, M., Mercier, G., Mouchot, M.C. and Hubert-Moy, L., (2001). "Independent component analysis as a tool for the dimensionality reduction and the representation of hyperspectral images." *International Geoscience and Remote Sensing Symposium,* 6, 2893-2895.

Li H T, Lin S C, Chen C Y, Chiang C K, (2019). “Layer-Level Knowledge Distillation for Deep Neural Network Learning.” *Applied Science,* 9(10), 1966.

Li, T., Zhang, J. and Zhang, Y., (2014). “Classification of hyperspectral image based on deep belief networks.” In *2014 IEEE international conference on image processing (ICIP)*, 5132-5136.

Li, W., Prasad, S., Fowler, J.E. and Bruce, L.M., (2011). “Locality-preserving dimensionality reduction and classification for hyperspectral image analysis.” *IEEE Transactions on Geoscience and Remote Sensing*, 50(4), 1185-1198.

Li, Y., & Huang, D. (2020). “Generating Hyperspectral Data Based on 3D CNN and Improved Wasserstein Generative Adversarial Network Using Homemade High-resolution Datasets.” In *Proceedings of the 2020 International Conference on Wireless Communication and Sensor Networks* (pp. 49-55).

Li, Z., Zhu, X., Xin, Z., Guo, F., Cui, X., & Wang, L. (2021). “Variational generative adversarial network with crossed spatial and spectral interactions for hyperspectral image classification.” *Remote Sensing*, *13*(16), 3131.

Liu, B., Yu, X., Zhang, P., Tan, X., Yu, A. and Xue, Z., (2017). “A semi-supervised convolutional neural network for hyperspectral image classification.” *Remote Sensing Letters*, *8*(9), 839-848.

Luo, H., Yang, L., Yuan, H. and Tang, Y.Y., (2013). “Dimension reduction with randomized anisotropic transform for hyperspectral image classification.” *IEEE International Conference on Cybernetics*, 156-161.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). “Adversarial autoencoders.” *arXiv preprint arXiv:1511.05644*.

Martinez, A.M. and Kak, A.C., (2001). “Pca versus lda.” *IEEE transactions on pattern analysis and machine intelligence*, 23(2), 228-233.

Medjahed, S. A., Saadi, T. A., Benyettou, A., & Ouali, M. (2016). “Gray wolf optimizer for hyperspectral band selection.” *Applied Soft Computing*, *40*, 178-186.

Melgani, F. and Bruzzone, L., (2004). “Classification of hyperspectral remote sensing images with support vector machines.” *IEEE Transactions on Geoscience and remote sensing*, 42(8), 1778-1790.

Menon,V. and Du,Q., (2018) “Randomized non-negative matrix factorization for hyperspectral image classification.” *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 1-5.

Mohan, A. and Venkatesan, M., 2020. “Hybrid CNN based hyperspectral image classification using multiscale spatio-spectral features.” *Infrared Physics & Technology*, 108, 103326.

Mokbel, B., Lueks, W., Gisbrecht, A. and Hammer, B., (2013). Visualizing the quality of dimensionality reduction. *Neurocomputing*, 112, 109-123.

Molchanov P, Mallya A, Tyree S, Frosio I, Kautz J, (2019). “Importance estimation for neural network pruning.” *IEEE Conference on Computer Vision and Pattern Recognition*, 11264-11272.

Mou, L., Ghamisi, P. and Zhu (2017), “Deep recurrent neural networks for hyperspectral image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3639–3655.

Mou, L., Ghamisi, P. and Zhu, X.X., (2017). “Deep recurrent neural networks for hyperspectral image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 3639-3655.

Mou, L., Lu, X., Li, X. and Zhu, X.X., (2020). “Nonlocal graph convolutional networks for hyperspectral image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, *58*(12), 8246-8257.

Mukherjee, S., Asnani, H., Lin, E., & Kannan, S. (2019). “Clustergan: Latent space clustering in generative adversarial networks.” In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4610-4617).

Neyshabur, B., Bhojanapalli, S. and Chakrabarti, A., (2017). “Stabilizing GAN training with multiple random projections.” *arXiv:1705.07831*.

Pintelas E, Livieris I E, Pintelas P E, (2021). A Convolutional Autoencoder Topology for Classification in High-Dimensional Noisy Image Datasets, *Sensors*, 21(22), 7731.

Plaza, A., Martínez, P., Plaza, J. and Pérez, R., (2005). “Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations.” *IEEE Transactions on Geoscience and remote sensing*, 43(3), 466-479.

Plummer B, Nikoli Dryden, Julius Frost, Torsten Hoeer, Kate Saenko, (2020). “Neural Parameter Allocation Search.” *International Conference on Learning Representations*, 1-16.

Prasad, S. and Bruce, L.M., (2008). “Decision fusion with confidence-based weight assignment for hyperspectral target recognition.” *IEEE Transactions on Geoscience and Remote Sensing*, 46(5), 1448-1456.

Radford, A., Metz, L. and Chintala, S., (2015). “Unsupervised representation learning with deep convolutional generative adversarial networks.” *arXiv:1511.06434*.

Rodarmel and Jie Shan, (2002). “Principal Component Analysis for Hyperspectral image Classification.” *Surveying and Land Information Systems,* 62(2), 115-123.

Romero, A., Gatta, C. and Camps-Valls, G., (2015). “Unsupervised deep feature extraction for remote sensing image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, 54(3), 1349-1362.

Roy, S.K., Krishna, G., Dubey, S.R. and Chaudhuri, B.B., (2019). “HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification.” *IEEE Geoscience and Remote Sensing Letters,* 17(2), 277-281.

Sainath N Tara, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, Bhuvana Ramabhadran, (2013). "Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets." *International Conference on Acoustics, Speech and Signal Processing,* 6655-6659.

Samajdar, T. and Quraishi, M.I., (2015). "Analysis and evaluation of image quality metrics." In *Information Systems Design and Intelligent Applications: Proceedings of Second International Conference INDIA 2015,* 2, 369-378.

Serpico, S. B., & Moser, G. (2007). "Extraction of spectral channels from hyperspectral images for classification purposes." *IEEE transactions on geoscience and remote sensing*, *45*(2), 484-495.

Sheikhpour, R., Sarram, M.A., Gharaghani, S. and Chahooki, M.A.Z., (2017). "A survey on semi-supervised feature selection methods." *Pattern Recognition*, 64, 141-158.

Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P. and Michalski, M., (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks." In *Simulation and Synthesis in Medical Imaging: Third International Workshop,* 1-11.

Shorten, C. and Khoshgoftaar, T.M., (2019). "A survey on image data augmentation for deep learning." *Journal of big data*, *6*(1), 1-48.

Shukla, U. P., & Nanda, S. J. (2018). "A binary social spider optimization algorithm for unsupervised band selection in compressed hyperspectral images." *Expert Systems with Applications*, *97*, 336-356.

Smola and Bernhard B. Scholkopf, (2004). "A tutorial on support vector regression." *Statistics and Computing,* 14(3), 199–222.

Su, H., Cai, Y., & Du, Q. (2016). "Firefly-algorithm-inspired framework with band selection and extreme learning machine for hyperspectral image classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(1), 309-320.

Su, H., Du, Q., Chen, G., & Du, P. (2014). "Optimized hyperspectral band selection using particle swarm optimization." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(6), 2659-2670.

Sui, C., Tian, Y., Xu, Y., & Xie, Y. (2014). "Unsupervised band selection by integrating the overall accuracy and redundancy." *IEEE Geoscience and Remote Sensing Letters*, *12*(1), 185-189.

Sumithra, Subu, (2015). "A Review of various linear and nonlinear dimensionality reduction techniques." *International Journal of Computer Science and Information Technologies,* 6(3), 2354-2360.

Sun, M., Wang, C., Wang, S., Zhao, Z. and Li, X., (2018). "A New Semisupervised-Entropy Framework of Hyperspectral Image Classification Based on Random Forest." *Advances in Multimedia*.

Tang, C., Liu, X., Zhu, E., Wang, L. and Zomaya, A.Y., (2021). “Hyperspectral Band Selection via Spatial-Spectral Weighted Region-wise Multiple Graph Fusion-Based Spectral Clustering.” In *IJCAI*, 3038-3044.

Tarnas, J.D., Mustard, J.F., Wu, X., Das, E., Cannon, K.M., Hundal, C.B., Pascuzzo, A.C., Kellner, J.R. and Parente, M., (2021). “Successes and challenges of factor analysis/target transformation application to visible-to-near-infrared hyperspectral data.” *Icarus*, 365, 402.

Vaddi, Prabukumar, (2017). “Comparative study of Feature extraction techniques for Hyperspectral remote sensing Classification: A survey.” *Proceedings of International Conference on Intelligent Computing and Control systems*, 543-548.

Vane, G. and Goetz, A.F., (1988). “Terrestrial imaging spectroscopy.” *Remote sensing of environment*, 24(1), 1-29.

Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G. and Koudas, N., (2002). “ Non-linear dimensionality reduction techniques for classification and visualization.” In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 645-651.

Wang, Q., Lin, J., & Yuan, Y. (2016). “Salient band selection for hyperspectral image classification via manifold ranking.” *IEEE transactions on neural networks and learning systems*, *27*(6), 1279-1289.

Wang, Z., She, Q. and Ward, T.E., (2021). “Generative adversarial networks in computer vision: A survey and taxonomy.” *ACM Computing Surveys*,54(2), 1-38.

Weinberger, K.Q. and Saul, L.K., (2006). “An introduction to nonlinear dimensionality reduction by maximum variance unfolding.” In *AAAI*, 6, 1683-1686.

Windrim L, Ramakrishnan R, Melkumyan A, Murphy R J, Chlingaryan A, (2019). “Unsupervised Feature-Learning for Hyperspectral Data with Autoencoders.” *Remote Sensing*, 11(7), 864.

Wu, Z., Li, Y., Plaza, A., Li, J., Xiao, F., & Wei, Z. (2016). “Parallel and distributed dimensionality reduction of hyperspectral data on cloud computing architectures.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(6), 2270-2278.

Xie, F., Li, F., Lei, C., Yang, J., & Zhang, Y. (2019). “Unsupervised band selection based on artificial bee colony algorithm for hyperspectral image classification.” *Applied Soft Computing*, *75*, 428-440.

Xie, L., Li, G., Peng, L., Chen, Q., Tan, Y., & Xiao, M. (2017). “Band selection algorithm based on information entropy for hyperspectral image classification.” *Journal of Applied Remote Sensing*, *11*(2), 026018-026018.

Xu, Y., Zhang, L., Du, B. and Zhang, F., (2018). “Spectral–spatial unified networks for hyperspectral image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, 56(10), 5893-5909.

Yang, L., Su, H., Zhong, C., Meng, Z., Luo, H., Li, X., Tang, Y.Y. and Lu, Y., (2019). "Hyperspectral image classification using wavelet transform-based smooth ordering." *International Journal of Wavelets, Multiresolution and Information Processing*, *17*(06), 1950-1965.

Yang, R., Su, L., Zhao, X., Wan, H., & Sun, J. (2017). "Representative band selection for hyperspectral image classification." *Journal of Visual Communication and Image Representation*, *48*, 396-403.

Yang, X., Ye, Y., Li, X., Lau, R.Y., Zhang, X. and Huang, X., (2018). "Hyperspectral image classification with deep learning models." *IEEE Transactions on Geoscience and Remote Sensing*, *56*(9), 5408-5423.

Yang, X., Zhang, X., Ye, Y., Lau, R.Y., Lu, S., Li, X. and Huang, X., (2020). "Synergistic 2D/3D convolutional neural network for hyperspectral image classification." *Remote Sensing*, *12*(12), 2033.

Ye J., Janardan R., Park, C H., Park, H., (2004). "An optimization criterion for generalized discriminant analysis on under-sampled problems." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8), 982-994.

Yi, X., Walia, E. and Babyn, P., (2019). "Generative adversarial network in medical imaging: A review." *Medical image analysis*, *58*, 552.

Yuan, Y., Lin, J., & Wang, Q. (2015). "Dual-clustering-based hyperspectral band selection by contextual analysis." *IEEE Transactions on Geoscience and Remote Sensing*, *54*(3), 1431-1445.

Zhai, H., Zhang, H., Zhang, L., & Li, P. (2018). "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection." *IEEE Transactions on Geoscience and Remote Sensing*, *57*(3), 1723-1740.

Zhan, Y., Hu, D., Xing, H., & Yu, X. (2017). "Hyperspectral band selection based on deep convolutional neural network and distance density." *IEEE Geoscience and Remote Sensing Letters*, *14*(12), 2365-2369.

Zhang T, J. Yang, D. Zhao, X. Ge, (2007). "Linear local tangent space alignment and application to face recognition." *Neurocomputing*, 70(7), 1547-1553.

Zhang, L., Zhong, Y., Huang, B., Gong, J. and Li, P., (2007). "Dimensionality reduction based on clonal selection for hyperspectral imagery." *IEEE Transactions on Geoscience and Remote Sensing*, 45(12), 4172-4186.

Zhang, M., Gong, M., Mao, Y., Li, J., & Wu, Y. (2018). "Unsupervised feature extraction in hyperspectral images based on wasserstein generative adversarial network." *IEEE Transactions on Geoscience and Remote Sensing*, *57*(5), 2669-2688.

Zhang, R., Nie, F. and Li, X., (2017). "Self-weighted supervised discriminative feature selection." *IEEE transactions on neural networks and learning systems*, 29(8), 3913-3918.

Zhang, W., Li, X., & Zhao, L. (2018). “A fast hyperspectral feature selection method based on band correlation analysis.” *IEEE Geoscience and Remote Sensing Letters*, *15*(11), 1750-1754.

Zhao Q, Masashi Sugiyama, Longhao Yuan, Andrzej Cichocki, (2019). “Learning Efficient Tensor Representations with Ring Structure Networks.” *International Conference on Acoustics, Speech and Signal Processing*, 8608-8612.

Zhao, W. and Du, S., (2016). “Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach.” *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4544-4554.

Zhao, W., Guo, Z., Yue, J., Zhang, X. and Luo, L., (2015). “On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery.” *International Journal of Remote Sensing*, 36(13), 3368-3379.

Zhao, Y. Q., Zhang, L., & Kong, S. G. (2010). “Band-subset-based clustering and fusion for hyperspectral imagery classification.” *IEEE Transactions on Geoscience and Remote Sensing*, *49*(2), 747-756.

Zhong, Z., Li, J., Luo, Z. and Chapman, M., (2017). “Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework.” *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 847-858.

Zhu, G., Huang, Y., Li, S., Tang, J., & Liang, D. (2017). “Hyperspectral band selection via rank minimization.” *IEEE Geoscience and Remote Sensing Letters*, *14*(12), 2320-2324.

Zhu, L., Chen, Y., Ghamisi, P., & Benediktsson, J. A. (2018). “Generative adversarial networks for hyperspectral image classification.” *IEEE Transactions on Geoscience and Remote Sensing*, *56*(9), 5046-5063.

PUBLICATIONS

International Journal Articles

1. Deepa C, Amba Shetty, Narasimhadhan AV (2023), “Knowledge Distillation: A novel approach for deep feature selection”, The Egyptian Journal of Remote Sensing and Space Sciences, Elsevier, 26(1), pp.63-73, https://doi.org/10.1016/j.ejrs.2022.12.006

2. Deepa C, Amba Shetty, Narasimhadhan AV (2023),“Performance evaluation of dimensionality reduction techniques on hyperspectral data for mineral exploration”, Earth Science Informatics, Springer nature, 16(1), pp. 25-36, https://doi.org/10.1007/s12145-023-00956-2

International Conferences

1. C Deepa, A Shetty, AV Narasimhadhan, “Quality assessment of Dimensionality Reduction Techniques on hyperspectral data: A neural network-based approach” International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2020,389–394, https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-389-2020, awarded ISPRS travel grant.

2. Deepa C, Amba Shetty, Narasimhadhan AV, “Semi-supervised framework for automated mineral mapping using Generative Adversarial Networks”, American Geophysical Union (AGU) Fall meeting, Abstract peer reviewed, Dec 12-16, 2022, Chicago, USA, awarded student travel grants, pp. IN22A-06.

3. Deepa C, Amba Shetty, Narasimhadhan AV, “An autoencoder framework for unsupervised learning of hyperspectral data”, American Geophysical Union (AGU) Fall meeting, Abstract peer reviewed, Dec 12-16, 2022, Chicago, USA, pp. IN32D-0400.

BIODATA

Name	: Ms. Deepa C
Date of Birth	: 10/12/1988
Address	: # 502, Type-V Apartment, NITK, Surathkal.
Telephone	: +91-9900661389
Email	: deeparamesh88@yahoo.com
Qualification	: B.E (Electronics and Communication Engineering), RL Jalappa Institute of Technology, VTU, Belgaum (2006-2010). M.Tech (Electronics), Canara Engineering College, VTU, Belgaum (2011-2013).
Publications	: International Journal: 02 International Conference: 03 National conference: 03
Awards	: ISPRS, AGU student travel grants