UNIVERSITÀ DEGLI STUDI DI MESSINA

DEPARTMENT OF ENGINEERING

DOCTORAL PROGRAMME IN CYBER PHYSICAL SYSTEMS XXXIV CYCLE

DEEP LEARNING FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Student:

MUHAMMAD AHMAD

Advisor:

PROF. DR. SALVATORE DISTEFANO

Co-Advisors:

PROF. DR. MANUEL MAZZARA

PROF. DR. ADIL MEHMOOD KHAN

ACADEMIC YEAR 2020 - 2021

Dedicated to My Beloved Family…

Advisor:

Prof. Dr. Salvatore Distefano

Co-Advisors:

Prof. Dr. Manuel Mazzara
Prof. Dr. Adil Mehmood Khan

Declaration of Authorship

I, MUHAMMAD AHMAD, declare that this thesis titled, "Deep Learning for Hyperspectral Image Classification" and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Doctorate degree at Università degli Studi di Messina.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Università degli Studi di Messina or any other institution, this has been clearly stated, credited and properly cited.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, the proposed scheme is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Acknowledgements

In the Name of Allah (Subhanahu Wa Ta'ala), the Most Beneficent and the Most Merciful. All praise and glory go to Allah Almighty (Subhanahu Wa Ta'ala) Who gave me the courage and patience to carry out this research work. Peace and blessings of Allah be upon His last Prophet Muhammad (Sallulah-o-Alaihihe-Wassalam) and all his Sahaba (Razi-Allah-o-Anhum) who devoted their lives for the prosperity and spread of Islam. By the grace of Allah Almighty (Subhanahu Wa Ta'ala), I would like to express my admiration for the assistance provided during the groundwork of this thesis.

My sincere thanks go to **Università degli Studi di Messina** for offering me the Ph.D. position and Innopolis University for providing me the Scholarship and Teaching opportunities to complete this milestone.

Foremost, I would also like to express the deepest appreciation to my Advisor **Prof. Dr. Salvatore Distefano** , my co-Advisors **Prof. Dr. Manuel Mazzara** **Prof. Dr. Adil Mehmood Khan** who have shown the attitude and the substance of a genius: they continually and persuasively conveyed a spirit of adventure regarding research and excitement in regards to teaching. Without their supervision and constant help, this dissertation would not have been possible. I could not have imagined having a better advisors and mentors for my Ph.D. study. Besides my advisor, I would like to thank the rest of my thesis committee for their encouragement, insightful comments, and hard questions.

I extend my gratitude to all my close friends and fellows who helped me a lot during my research and completion of this thesis. I am also thankful to all my fellows for their materialistic support and prayers.

Last but not the least, I would like to thanks My Parents, Siblings, wife, and son. Their prayers and encouragement have always helped me to take the right steps in my life. There is no way, no words, to express my love and gratitude. May Allah (Subhanahu Wa Ta'ala) help us in following the true spirit and principles of ISLAM write down in the Holy Quran and Sunnah! (Ameen).

Abstract

Deep Learning for Hyperspectral Image Classification

Hyperspectral Imaging (HSI) has been extensively utilized in many real-life applications because it benefits from the detailed spectral information contained in each pixel. Notably, the complex characteristics i.e., the nonlinear relation among the captured spectral information and the corresponding object of HSI data make accurate classification challenging for traditional methods. In the last few years, Deep Learning (DL) has been substantiated as a powerful feature extractor that effectively addresses the nonlinear problems that appeared in a number of computer vision tasks. This prompts the deployment of DL for HSI Classification (HSIC) which revealed good performance.

Keeping in mind the aforementioned issues, this thesis first enlists a systematic overview of DL for HSIC and compared state-of-the-art strategies of the said topic. Primarily, this thesis encapsulates the main challenges of traditional machine learning for HSIC and then acquaint the superiority of DL to address these problems. The literature is breakdown the state-of-the-art DL frameworks into spectral features, spatial features, and together spatial-spectral features to systematically analyze the achievements and future directions. This thesis also investigates the behavior and performance in terms of computational cost and classification accuracy, of the most commonly and widely used classification algorithms under different experimental setups. In a nutshell, the following specific contributions are made in this thesis:

1. A Fast and Compact 3D CNN that utilizes both spatial-spectral feature maps to improve the performance of HSIC.

2. 3D CNNs are computationally expensive and 2D CNN alone cannot efficiently extract discriminating spectral-spatial features. Therefore, to overcome these challenges, this part presents a compact hybrid CNN model which overcomes the aforementioned challenges by distributing spatial-spectral feature extraction across 3D and 2D layers.

3. CNN's are known to be effective in exploiting joint spatial-spectral information with the expense of lower generalization performance and learning speed due to the hard labels and non-uniform distribution over labels. Several regularization techniques such as dropout, L1, L2, etc., have been used to overcome the aforesaid issues. However, sometimes models learn to predict the samples extremely confidently which is not good from a generalization point of view. Therefore, this thesis proposed an idea to enhance the generalization performance of a hybrid CNN for HSIC using soft labels that are a weighted average of the hard labels and uniform distribution over ground labels. The proposed method helps to prevent CNN from becoming over-confident.

4. DL usually required a large amount of labeled training samples which is not a real scenario. Thus, a fully automatic Spatial-Spectral approach has been proposed for the selection of most informative and heterogeneous samples for training using a novel Spectral Angle Mapper (SAM) based objective function for the computation of attribute profiles in a computationally efficient fashion.

Contents

List of Figures

List of Tables

Chapter 1

Preface

1.1 Overview of Remote Sensing

Remote sensing is a field of science that includes all those activities necessary for the observation, acquisition, and interpretation of information related to objects, events, phenomena, or any other item under investigation, without making physical contact with the object, event, or phenomenon under investigation.

Remote sensing systems (space-borne and airborne) used for earth observation collect data by detecting the energy that is reflected from an object or area under investigation on the earth's surface. Considering electromagnetic radiation as the principal physical carrier of information, the main differentiation of such systems is based on the type of source of energy exploited. Depending on whether these systems measure the radiation that is naturally available or omitted by the sensor, they can be defined as passive or active sensors.

Passive sensors rely on the energy provided by the Sun, which is either reflected or absorbed and then re-emitted from the Earth's surface. The reflected energy (e.g., visible radiation) is available only when the Sun illuminates the Earth. The emitted energy can be detected as long as the amount of energy is large enough to be recorded. Examples of the most popular passive sensors include cameras, scanning sensors, and microwave radiometers.

Active sensors, on the other hand, emit the energy required to illuminate the target under investigation and then detect the back-scattered radiation. Examples of broadly used active systems are Radio Detection and Ranging (RADAR) and Light Detection and Ranging (LiDAR). In this case, being the sensor, the source of radiation, the data acquisition can be performed at any time. The vast variety of available sensors, which provide data either in image or signal formats, allows tackling a large number of applications with remarkable advantages. In general, each family of sensors is characterized by properties such as spatial, spectral, radio-metrical, and temporal resolutions, which are strictly related to their physical implementation resulting in more or less suitable for a precise application. This entails the

development of advanced techniques for data processing and interpretation that are sensor and application-dependent.

1.2 Hyperspectral Imaging (HSI)

The human eye sees the color of visible light mostly in red, green, and blue (RGB) channels (bands). In contrast, spectral imaging divides the spectrum into many more channels. Therefore, in comparison to a traditional camera, an HSI camera (sensor) does not record images in RGB channels only, but in hundreds of channels. For each of these channels, an image is created and coded with grayscale levels. When we combine these images or channels we form an HSI cube for processing and analysis.

In other words, HSI combines digital imaging with spectroscopy - and provides high spatial and spectral information in each image pixel. Each pixel can then be associated with the spectral signature of the target. This result can be then used to identify, measure, and locate different materials and their chemical and physical properties [1]. It is for these reasons, that HSI has attracted the formidable interest of the scientific community in recent years and has been applied to an increasing number of applications in different fields. These applications include biomedical imaging, geosciences, surveillance, object detection, and recognition, change detection, human-made material identification, semantic annotation, unmixing, and remote sensing [2, 3].

1.3 Problem Statement

HSIC is a challenging task due to high inter-class similarity, high intra-class variability, overlapping, and nested regions. 2D CNN is a viable classification approach since HSIC depends on both Spectral-Spatial information. 3D CNN is a good alternative for improving the accuracy of HSIC, but it can be computationally intensive due to the volume and spectral dimensions of HSI. Furthermore, these models may fail to extract quality feature maps and underperform over the regions having similar textures.

Precisely, one of the main challenges in the field of HSI is connected to the characteristics of the data. More specifically, Hyperspectral data yields hundreds of contiguous and narrow spectral bands with very high spatial resolution throughout the electromagnetic spectrum [4]. When combined with the limited availability of labeled training data, it can lead to Hughes Phenomenon [5], which is also known as the curse of dimensionality. It occurs whenever the number of available labeled training samples is considerably lower than the number of spectral bands present in the data [6]. This aspect results in a relatively poor

predictive performance of supervised [7] and semi-supervised learning methods [8] for HSI classification (HSIC).

For example, some of the supervised classification methods used for HSIC include Multinomial Logistic Regression [9], Random Forests [10], Ensemble Learning [11], Deep Learning [12], Support Vector Machine [13], and K-Nearest Neighbors [6]. Figure 1.1 illustrates the loss in the predictive performance of such classification methods for a particular ground image when using two different sample sizes.

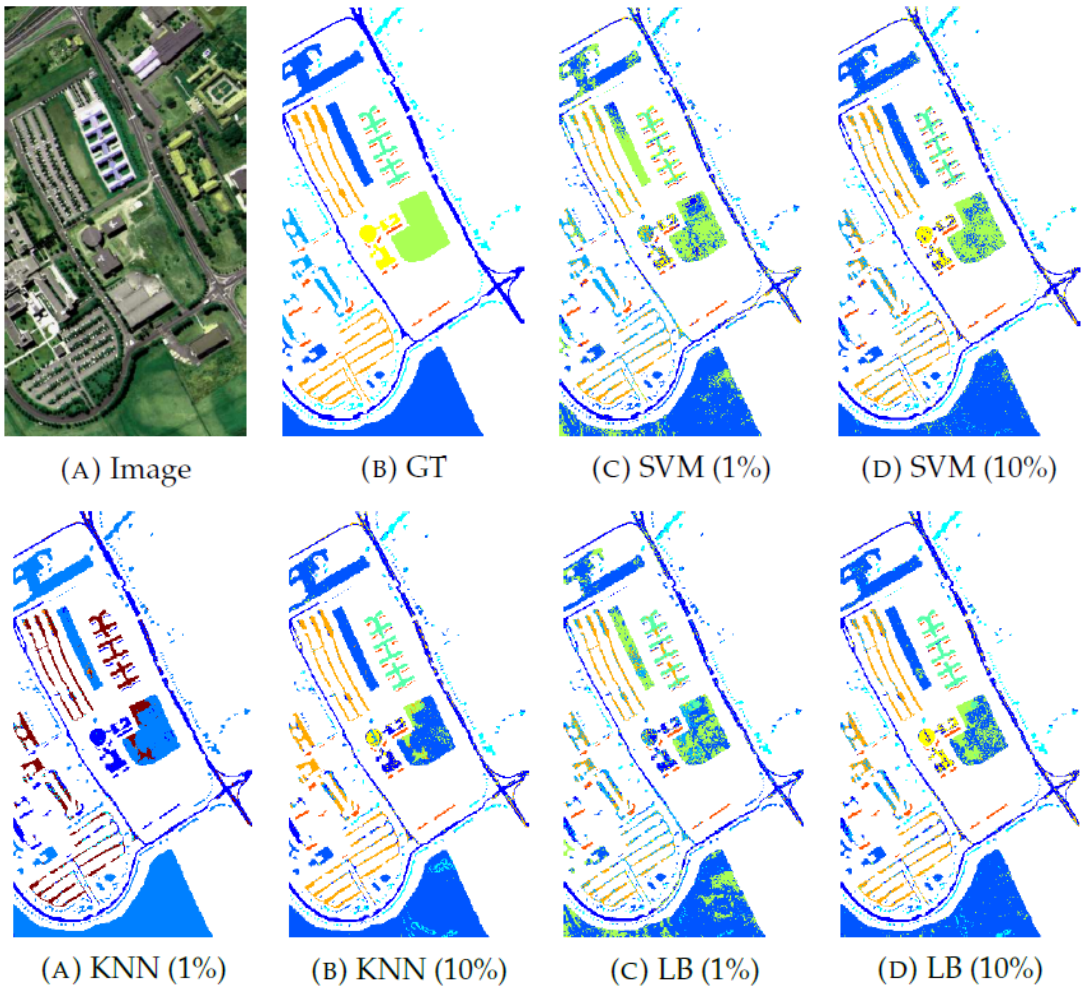


FIGURE 1.1: (A): Pavia University ground image, (B): True ground truths differentiate nine (9) classes, (C): SVM trained with 1% randomly selected training samples with overall accuracy 68.18% and kappa 58.53%, (D): SVM trained with 10% randomly selected training samples with overall accuracy 83.65% and kappa 78.04%, (E): KNN trained with 1% randomly selected training samples with overall accuracy 64.88% and kappa 48.97%, (F): KNN trained with 10% randomly selected training samples with overall accuracy 76.91% and kappa 67.91%, (G): LB trained with 1% randomly selected training samples with overall accuracy 64.88% and kappa 52.16%, (H): LB trained with 10% randomly selected training samples with overall accuracy 83.65% and kappa 75.31%.

One solution to overcome this problem is to collect large amounts of labeled training samples or to reduce the dimensions, but labeled samples collection is expensive, difficult,

and time-intensive in real-life scenarios because of the unavailability of field experts. Moreover, dimensionality reduction may lead to the loss of important geographical information associated with the HIS. Therefore, this dissertation aims to address the aforementioned HSIC problem without collecting a large number of labeled training samples and without losing the important information for HISC in a computationally efficient fashion. In this regard, this dissertation proposed several integrated design choices as listed in section 1.4.

1.4 Overview of the Dissertation

HSI has been extensively utilized in many real-life applications because it benefits from the detailed spectral information contained in each pixel. Notably, the complex characteristics i.e., the nonlinear relation among the captured spectral information and the corresponding object of HSI data make accurate classification challenging for traditional methods. In the last few years, deep learning (DL) has been substantiated as a powerful feature extractor that effectively addresses the nonlinear problems that appeared in a number of computer vision tasks. This prompts the deployment of DL for HSI classification (HSIC) which revealed good performance.

Keeping in mind the aforementioned issues and the conditions, this thesis makes the following contributions. This thesis first enlists a systematic overview of Deep Learning (DL) for HSIC and compared state-of-the-art strategies of the said topic. Primarily, this thesis encapsulates the main challenges of traditional machine learning for HSIC and then acquaint the superiority of DL to address these problems. The literature is breakdown the state-of-the-art DL frameworks into spectral features, spatial features, and together spatial-spectral features to systematically analyze the achievements (future directions as well) of these frameworks for HSIC. Moreover, we will consider the fact that DL requires a large number of labeled training examples whereas acquiring such a number for HSIC is challenging in terms of time and cost. Therefore, this thesis discusses some strategies to improve the generalization performance of DL strategies which can provide some future guidelines. In a nutshell, the following specific contributions are made in this thesis. A flow-graph is also added (Figure 1.2) to show the hierarchy among the proposed methodologies.

1. Investigates the behavior and performance, in terms of computational cost and classification accuracy, of the most common and widely used classification algorithms in the HSI domain under different experimental setups.

2. This thesis develops the following novel strategies to overcome the aforementioned problems.

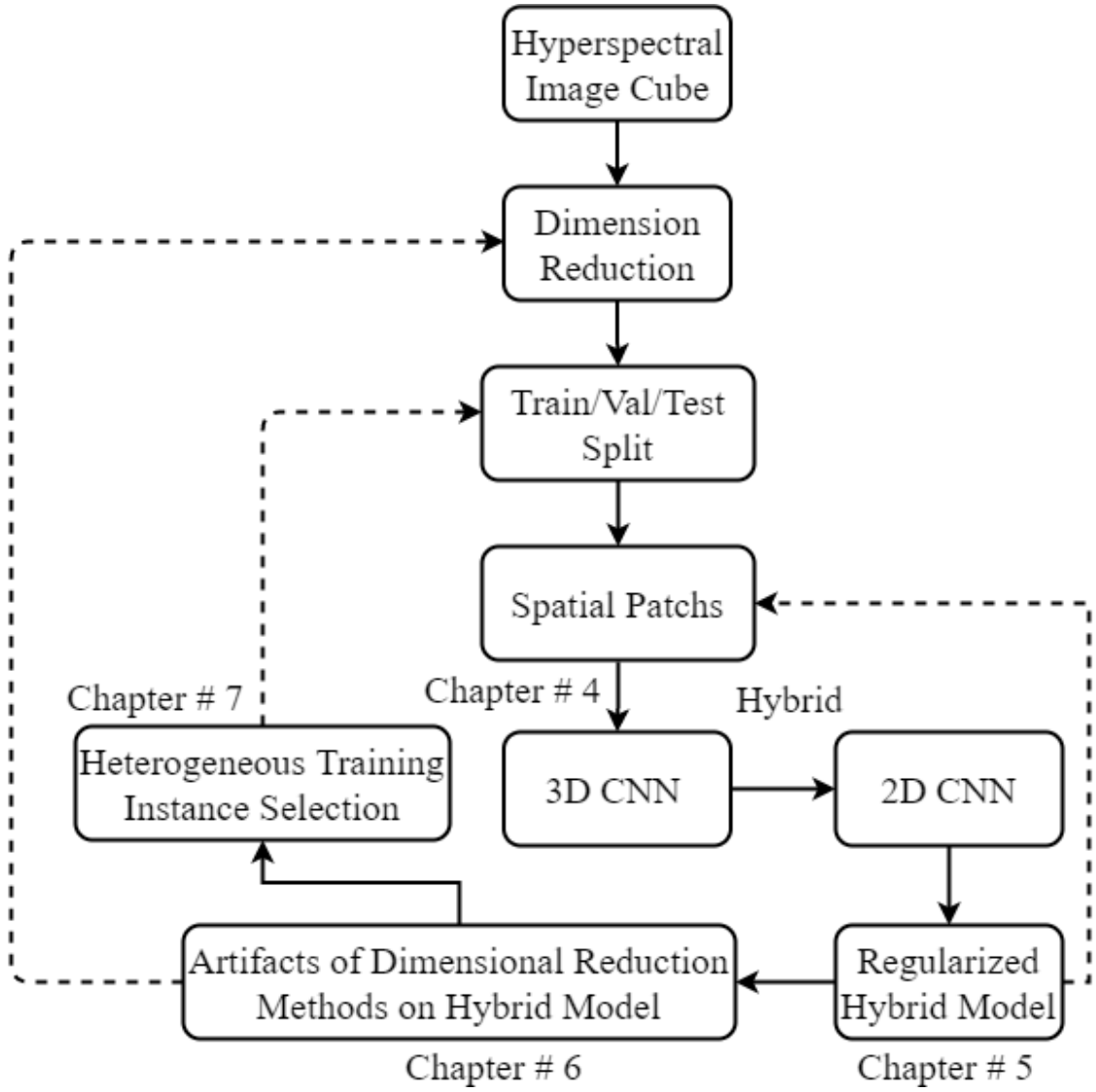


FIGURE 1.2: Workflow of proposed methodologies.

(a) A Fast and Compact 3D CNN that utilizes both spatial-spectral feature maps to improve the performance of HSIC. For this purpose, the HSI cube is first divided into small overlapping 3D patches, which are processed to generate 3D feature maps using a 3D kernel function over multiple contiguous bands of the spectral information in a computationally efficient way. In brief, the end-to-end trained model requires fewer parameters to significantly reduce the convergence time while providing better accuracy than existing models.

(b) 3D CNNs are computationally expensive and 2D CNN alone cannot efficiently extract discriminating spectral-spatial features. Therefore, to overcome these challenges, this part presents a compact hybrid CNN model which overcomes the aforementioned challenges by distributing spatial-spectral feature extraction across

3D and 2D layers. The experimental results show that the proposed pipeline outperformed in terms of generalization performance and statistical significance as compared to the state-of-the-art CNN models except commonly used computationally expensive design choices.

(c) CNN's are known to be effective in exploiting joint spatial-spectral information with the expense of lower generalization performance and learning speed due to the hard labels and non-uniform distribution over labels. Several regularization techniques such as dropout, L1, L2, etc., have been used to overcome the aforesaid issues. However, sometimes models learn to predict the samples extremely confidently which is not good from a generalization point of view. Therefore, this paper proposed an idea to enhance the generalization performance of a hybrid CNN for HSIC using soft labels that are a weighted average of the hard labels and uniform distribution over ground labels. The proposed method helps to prevent CNN from becoming over-confident. We empirically show that in improving generalization performance, label smoothing also improves model calibration which significantly improves beam-search.

(d) A fully automatic approach for the selection of most informative and heterogeneous samples for training using a novel Spectral Angle Mapper (SAM) based objective function for the computation of attribute profiles in a computationally efficient fashion.

1.5 Dissertation Organization

This dissertation is organized as follows: Part I, which consists of chapters 1 and 2, provides an introduction to the HSI field and the context in which the dissertation is developed. More specifically, chapter 2 introduces the HSI field, describing both the challenges and the objectives addressed in this dissertation. Whereas, chapter 3 presents an overview of the several frameworks proposed for HISC. Moreover, it provides the theoretical background of the proposed methodologies.

Part II consisting of chapters 4-7 (Chapter 4, 5, 6, and 7 presents the contributions made in this dissertation. Part III consisting of Chapter 8 presents the experimental evaluation of the proposed methodologies. Finally, Chapter 9 concludes this dissertation remarking its most important findings and discussing the most prominent future research directions.

Chapter 2

Introduction to Hyperspectral Imaging

This chapter introduces HSI and its scope in real-world applications. This chapter also discusses the issues related to the HSIC in real-life scenarios along with the detailed description of the problem statement and their solutions proposed in this dissertation.

2.1 Hyperspectral Imaging Technology

The concept of HSI was first introduced by A. F. H. Goetz and his colleagues at the National Aeronautics and Space Administration (NASA's) Jet Propulsion Laboratory (JPL) in the 1980s, where a system called Airborne Imaging Spectrometer (AIS) was built to demonstrate HI technology [2, 14].

Nowadays, NASA is continuously collecting high-dimensional HSI datasets with instruments such as Airborne Visible Infrared Imaging Spectrometer (AVIRIS), an example is shown in Figure 2.1. The advanced AVIRIS sensor for earth observation records the visible to near and mid-infrared spectrum of the reflected light using more than 200 spectral bands, thus producing a stack of images in which each pixel vector is represented by a spectral signal that uniquely characterizes the underlying objects [4, 15].

According to the characteristics of the scanner, sensor systems are distinguished by their different resolutions, which also define the characteristics of the acquired images. The minimum size of an object that the sensor can distinguish from the ground represents the spatial resolution and depends on the altitude of the sensor and its angle of view (i.e., the angle subtended by the sensor), which is defined in terms of Instantaneous Field of View (IFOV).

The spectral resolution is the minimum wavelength at which the instrument is sensitive while the radiometric resolution is defined as the minimum energy able to be detected by the sensing system. The intrinsic radiometric resolution of a sensor depends on the detector's signal-to-noise ratio (SNR). In a digital image, the radiometric resolution is limited by the number of discrete quantization levels used to digitize the continuous intensity value.

The spectral resolution is the minimum bandwidth on which the measured radiation is integrated. Although the acquisition system could detect signals with high resolutions, it

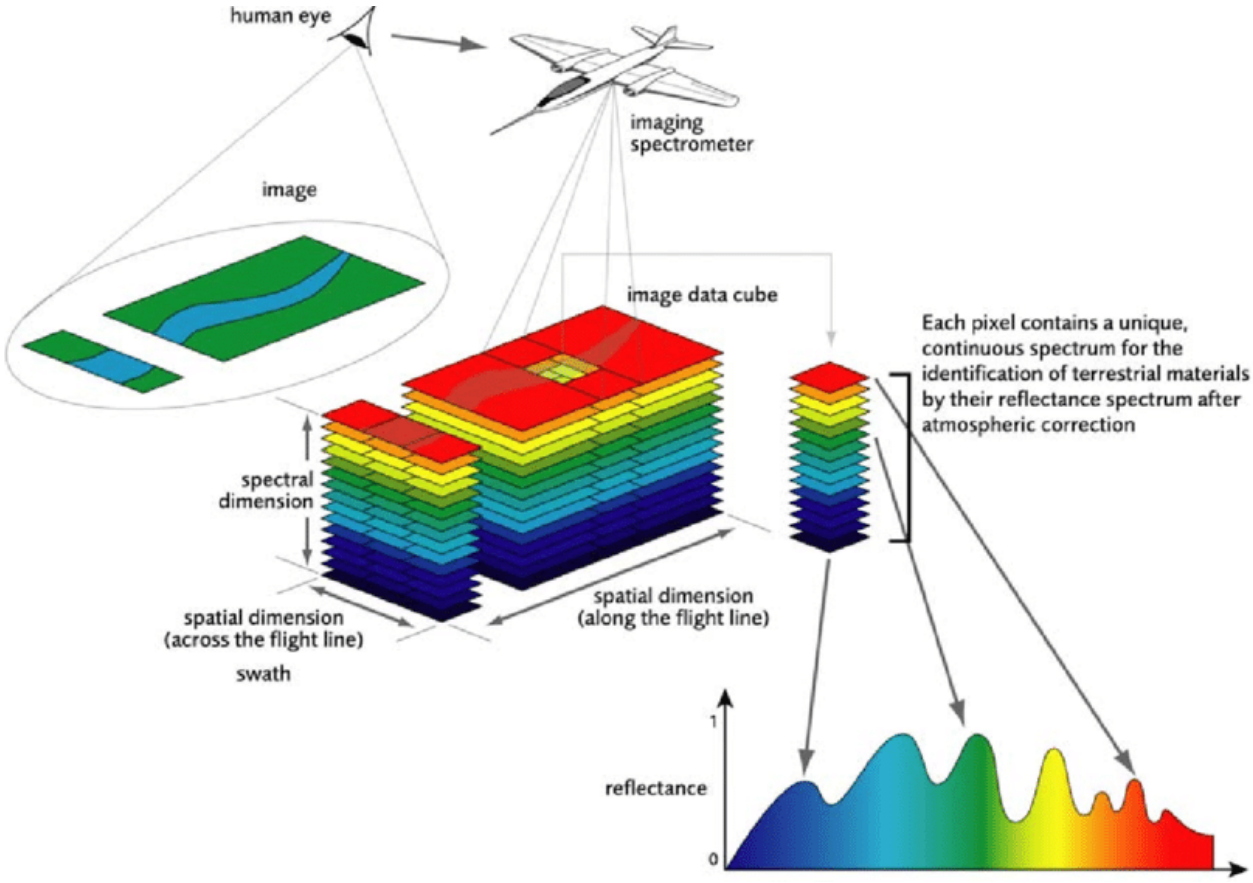


FIGURE 2.1: Hyperspectral Imaging Concept in Remote Sensing [16].

counts on various critical points due to physical constraints and instrumental limitations. Indeed, the acquisition of the images is usually affected by the sensor's noise, bad pixel location, and atmospheric contribution, requiring different levels of pre-processing to ensure the image quality in terms of spectral, spatial, and radiometric accuracy and make the data available for further analysis. Moreover, recent technological advances in sensor technology have led to the development of a new generation of Hyperspectral sensors able to provide images with improved spatial resolution.

For instance, data acquired by Hyperion sensors (mounted on EO-1 satellite) has a spatial resolution of 30m, while ROSIS-3 (Airborne Spectrometer) can provide images with a spatial resolution of 1.7m if the acquisition is taken at the altitude of 3km. CASI-1500 can provide a data cube of 144 spectral bands with a spectral resolution of 1.25m. From these few examples, one can see that contextual information becomes an important source of information that can be exploited for distinguishing different objects on the ground.

Moreover, the concept of HSI is extended to describe systems with hundreds to thousands of spectral bands with many new instruments currently in development for spaceborne operations. Table 2.1 presents a summary of several hyperspectral sensor systems which are currently in operation [17–21]. Whereas, Table 2.2 provides a summary of the

TABLE 2.1: Characteristics of Different HSI Sensors.

Sensor	**AVIRIS 62**	**HyDICE 143**	**Hymap 134**	**Probe-1 104**	**Hyperion 167**
Year	1997	1995	1996	1997	2000
Platform	Airborne	Airborne	Airborne	Airborne	Space borne
Nominal Altitude (km)	20	6	5	2.5	705
Spatial Resolution (m)	20	3	10	5	30
Spectral Resolution (nm)	10	10	17	10	10
Spectral Coverage (¯m)	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5	0.4-2.5
Number of Channels	224	210	128	128	220
Swath Width (km)	12	0.9	6	3	7.7

most commonly used sensors usually mounted on aircraft or spacecraft reporting the principal spectral characteristics. Thus, the characterization of HSI based on their spectral properties has led to the use of this type of dataset in a growing number of real-life applications.

TABLE 2.2: Technical characteristics of some HSI sensors developed over last years.

Sensor	**Manufacturer**	**Platform**	**No. of Bands**	**Spectral Resolution**	**Spectral Range**
Hyperion	NASA GSFC	Satellite	220	10 nm	$0.4 - 2.5\ \mu m$
MODIS	NASA	Satellite	36	40 nm	$0.4 - 14.3\ \mu m$
CHRIS Proba	ESA	Satellite	up to 63	1.25 nm	$0.415 - 1.05\ \mu m$
AVIRIS	NASA JPL	Aerial	224	10 nm	$0.4 - 2.5\ \mu m$
HYDICE	Naval Research Lab	Aerial	210	7.6 nm	$0.4 - 2.5\ \mu m$
PROBE-1	Earth Search Science	Aerial	128	12 nm	$0.4 - 2.45\ \mu m$
CASI 550	ITRES Research Ltd	Aerial	288	1.9 nm	$0.4 - 1\ \mu m$
CASI 1500	ITRES Research Ltd	Aerial	288	2.5 nm	$0.4 - 1.05\ \mu m$
SASI 600	ITRES Research Ltd	Aerial	100	15 nm	$0.95 - 2.45\ \mu m$
TASI 600	ITRES Research Ltd	Aerial	64	250 nm	$8 - 11.5\ \mu m$
HyMap	Intergrated Spectronics	Aerial	125	17 nm	$0.4 - 2.5\ \mu m$
ROSIS-3	DLR	Aerial	115	4 nm	$0.43 - 0.85\ \mu m$
EPS-H	GER Corporation	Aerial	133	0.67 nm	$0.43 - 12.5\ \mu m$
EPS-A	GER Corporation	Aerial	31	23 nm	$0.43 - 12.5\ \mu m$
DAIS 7915	GER Corporation	Aerial	79	15 nm	$0.43 - 12.3\ \mu m$
AISA Eagle	Spectral Imaging	Aerial	244	2.3 nm	$0.4 - 0.97\ \mu m$
AISA Eaglet	Spectral Imaging	Aerial	200	-	$0.4 - 1.0\ \mu m$
AISA Hawk	Spectral Imaging	Aerial	320	8.5 nm	$0.97 - 2.45\ \mu m$
AISA Dual	Spectral Imaging	Aerial	500	2.9 nm	$0.4 - 2.45\ \mu m$
MIVIS	Daedalus	Aerial	102	20 nm	$0.43 - 12.7\ \mu m$
AVNIR	OKSI	Aerial	60	10 nm	$0.43 - 1.03\ \mu m$

The high capability of the HSI sensors enables the acquisition of images in which an individual pixel is a vector with very high spatial-spectral resolution [22, 23] as shown in Tables 2.1 and 2.2. This unprecedented high spectral-spatial resolution has opened the door to a series of civilian and military applications among which refer to; agriculture assessment, land use, environmental and ecological monitoring, mineral exploitation, man-made materials detection, and identification, change detection and observation, target detection, and recognition, target activities recognition, surveillance, ground cover classification, and

natural minerals identification [24, 25]. Underlying all these applications is the fact that all substances scatter electromagnetic energy at specific wavelengths in distinctive patterns related to their molecular composition [26].

2.2 Scope of Hyperspectral Imaging

The majority of image processing and analysis methods dealing with HSI can be classified as follows:

1. Detect known and unknown materials and objects in a given scene.
2. Classification and segmentation of the HSI's into the regions where the material or objects are predominant.
3. Estimate the materials or objects and the respective area fractions that they occupy within a pixel. This is so-called Hyperspectral Unmixing.

However, the HSI Dataset representation involves an array of spectral measurements on the natural scene where each of them corresponds to a pixel. This most elementary unit of the image provides a piece of extremely local information. Furthermore, besides the scale issue, the pixel-based representation also suffers from the lack of structure. As a result, HSI processing at the pixel level has to face major difficulties in terms of scale: the scale of representation is most of the time far too low with respect to the interpretation or decision scale.

As earlier explained, HSI sensors collect multivariate discrete images in a series of narrow and contiguous wavelength bands. The resulting HSI cube contains numerous bands in which each of them depicting the scene as viewed with a given wavelength λ. This whole set of images can be seen as a three-dimensional data cube where each pixel is characterized by a discrete spectrum related to the light absorption and/or scattering properties of the spatial region that it represents. Figure 2.2 shows an illustration of different HSI cubes.

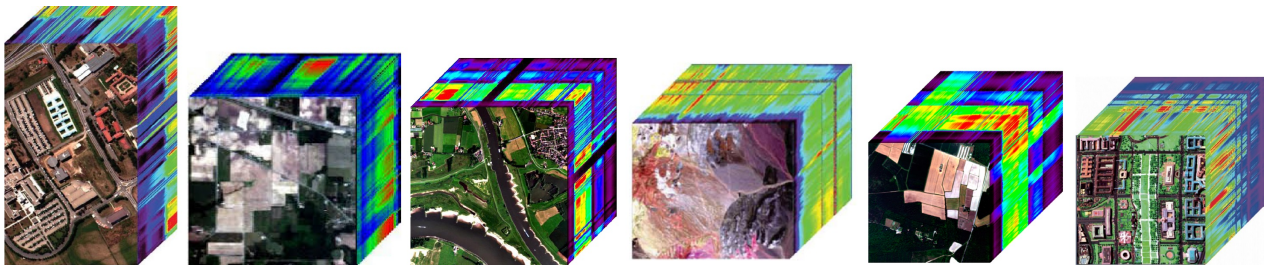


FIGURE 2.2: Hyperspectral Image Cubes.

2.3 Applications of Hyperspectral Imaging

The information provided by the HSI cube is a huge amount of data that cannot be fully exploited using traditional image analysis methods. Hence, given the wide range of real-life applications, for instance; civilian and military applications among which refer to; agriculture assessment, land use, environmental and ecological monitoring, mineral exploitation, man-made materials detection, and identification, change detection and observation, target detection, and recognition, target activities recognition, surveillance, ground cover classification, and natural minerals identification [24, 25], a great deal of research is devoted to the field of HIS data pre and/or post-processing [27, 28]. The number and variety of processing tasks in HSI are also enormous. However, the majority of algorithms can be organized according to the following specific tasks [29].

1. **Classification** consists of assigning a unique label (class representation) to each pixel of a hyperspectral cube [29].
2. **Dimensionality Reduction** consists of reducing the dimensionality of the input scene to facilitate subsequent processing tasks [30].
3. **Spectral Unmixing** consists of estimating the fraction of the pixel area covered by each material present in the scene [31].
4. **Target and Anomaly Detection** consist of searching the pixels of a hyperspectral cube for rare (either known or unknown) spectral signatures [25].
5. **Change Detection** consists of finding the significant (i.e., important to the user) changes between two hyperspectral scenes of the same geographic region acquired at different times [32].

This dissertation mainly focused on HSIC which is mainly focused on the integration of supervised and semi-supervised classification techniques for classification performance and generalization improvement using different classifiers i.e. generative, discriminative, ensemble, and parametric classifiers. In HSIC, we are generally, given a set of observations (i.e. possibly mixed pixel vectors). The goal of the HSIC algorithm is to assign a unique label to each sample so that it is well defined by a given class in a computationally efficient fashion [33].

2.4 Problem Formulation

A feature is a characteristic, value, or aspect of an object where several features of an object make up a feature vector. The set of available feature vectors spans the feature space. 2D or

3D subsets of features or the projection on feature space can be visualized as scatter plots. In Figure 2.3 one can easily observe an illustration of a feature vector, its feature space and a scatter plot in 2D.

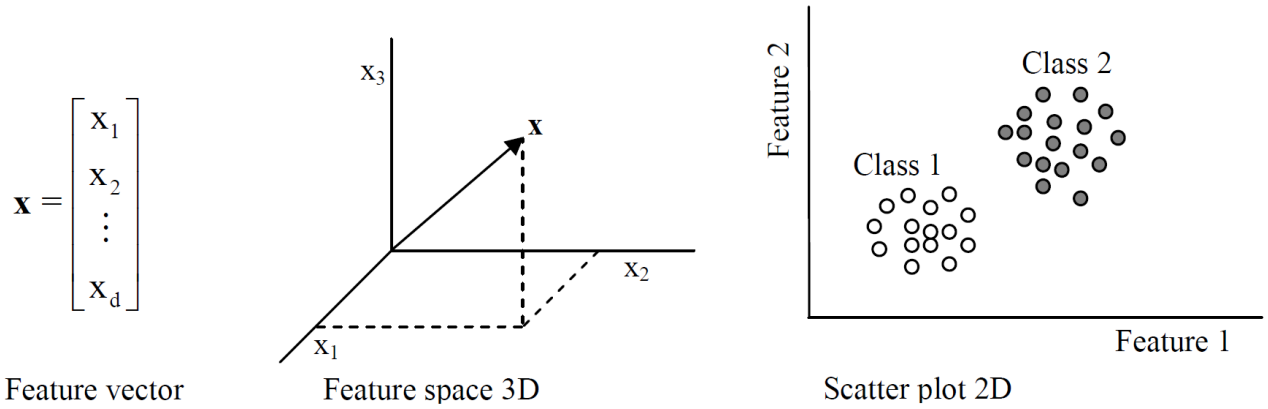


FIGURE 2.3: Relationship between feature vector and feature space [34].

An HSI cube can be modeled with Euclidean space where the number of bands is the dimension of the space and the pixels in the image are represented as points in that particular space. In supervised classification, each pixel of the cube is labeled as representing a particular ground cover or class taking the information provided by the training samples. The training can be established using maps, site visits, or aerial photography. The parameters of a particular classifier are calculated from these training samples.

However, in recent years, HSI data has become increasingly larger in both number of pixels per image (Spatial Resolution) and several bands (Spectral Resolution). As the number of spectral bands increases the separability of classes also increases but so does the number of statistical parameters defining the classes. Since there are only a fixed number of training samples available for deriving the statistical parameters, therefore, at some point the accuracy of the estimation must begin to decrease. An optimal value of dimensions and training samples is shown to exist in any given practical circumstance and depending upon the nature of the problem, more dimensions do not necessarily lead to better results in terms of accuracy.

Such high dimensionality of HSI datasets makes it difficult to analyze due to several reasons. Among high dimensionality, we can say that a lot of features increase the noise factor and hence the error factor that there are not enough observations to get good estimates or that most data is scattered and the bands in Hyperspectral data are highly correlated.

As an interesting note about the observation, consider a sphere of radius r inscribed inside a hypercube of dimension d with sides of length $2r$. The volume of the hypercube is $2r^d$ where d is the number of dimensions. It is possible to find that the volume of the sphere as;

$$\frac{(2r)^d \pi^{\frac{d}{2}}}{d\,\Gamma\{\frac{d}{2}\}} \tag{2.1}$$

Therefore, the proportion of the volume of the square that is inside the sphere is shown in equation (2.2) [35];

$$\lim_{d\to\infty} f(x)\frac{\pi^{\frac{d}{2}}}{d\Gamma^{\frac{d}{2}}} \to 0 \tag{2.2}$$

It looks like that in high dimensionality, the data accumulated in the corners. The work [36] referring to the computational complexity of searching the neighborhood of data points in high-dimensional settings was the first to put forward the term *curse of dimensionality* to describe the problem of data sparseness.

In addition to [36], the work [5] conducted a statistical analysis showing how the accuracy of a classifier depends on the number of labeled training samples and the number of bands which is known as the curse of dimensionality. Many works have dealt with the high dimensionality phenomenon for the last four decades. Recently, the work [37] proposed a general non-parametric method trying to avoid or reduce the Hughes effect. In contrast to the above, the work [38] presents a hybrid algorithm and as they claim, "no existing algorithm is entirely satisfactory isolation, but that a carefully designed combination can overcome the weaknesses of each".

The work [39] presents a methodology for band selection for the HSI cube, tailored to target detection applications that choose a subset of bands that maximized an objective function suitable for target detection. However, in [40], the authors propose two methods for dimensionality reduction of HSI data via spectral feature extraction and compared them to the traditional methods for finding relevant bands to determine optical regions. Moreover, instead of optimizing separability criteria, the overall classification accuracy of a validation dataset is used to decide which disjoint optical regions yield maximum accuracy.

In the last decade, a number of band selection, feature selection/extraction, and feature learning-based classification techniques have been proposed which proves the increasing interest in the more detailed analysis of HSI cube but in reduced dimensions. These investigations include but not limited to Linear Discriminant Analysis [40–44], Discriminant Neighbor Embedding [45], Stepwise Linear Discriminant Analysis [46], Local Discriminant Embedding [47], Marginal Fisher Analysis [48, 49], Exponential Local Discriminant Embedding [49], Double Adjacency Graph-based Discriminant Neighborhood Embedding [50],

Laplacian Linear Discriminant Analysis [51], Local and Global Structure Preservation Feature Selection [52], Similarity Preserving Feature Learning [53], Principal Component Analysis [54, 55], Locally Linear Embedding [56], ISOMAP [57], Laplacian Eigen Map [58], Unsupervised Discriminant Projection [59], Neighborhood Preserving Embedding [60], Locality Preserving Projections [61], Sammon Projection, Incremental Semi-Supervised Low-Rank Representation Graph [62], Sparse Probability Graph [63] and Graph-based Constrained Semi-Supervised Learning [64].

All the above-discussed methods select a subset of HSI data for classification purposes. However, there is a high probability that these methods lose important spatial information while reducing dimensionality. Such as geometrical representation of original HSI space i.e. spatial coordinates of original data. Another way around, these methods select a few numbers of bands from the entire HSI space that may discard some important bands which may contain more information about one particular material or object than other material or object of interest. Moreover, there is another issue related to the number of bands to be select for classification. In a nutshell, the following challenges that come across:

- **Complex Training Process**: Training of Deep Neural Network (DNN) and optimization by tuning parameters is an NP-complete problem where the convergence of the optimization process is not guaranteed [65]. Therefore, it is assumed that training of DNN is very difficult [66] especially in the case of HSI when a large number of parameters need to be adjusted/tuned.

- **Limited Availability of Training Data**: As discussed above, supervised DNN requires a considerably large amount of training data otherwise their tendency to overfit increases significantly [67] leads to the Hughes phenomena. The high dimensional characteristic of HSI coupled with a small amount of labeled training data makes the DNNs ineffective for HSIC as it demands a lot of adjustments during the training phase [68].

- **Model's Interpretability**: The training procedure of DNNs is difficult to interpret and understand. The black box kind of nature is considered as a potential weakness of DNNs and may affect the design decisions of the optimization process. Although, a lot of work has been done to interpret the model's internal dynamics.

- **High Computational Burden**: One of the main challenges of DNN is dealing with a big amount of data that involves increased memory bandwidth, high computational cost, and storage consumption [69]. However, advanced processing techniques like parallel and distributed architectures [70, 71] and high-performance computing (HPC) [72] make it possible for DNNs to process large amounts of data.

- **Training Accuracy Degradation**: It is assumed that deeper networks extract more rich features from data [73], however, this is not true for all systems to achieve higher accuracy by simply adding more layers. Because by increasing the network's depth, the problem of exploding or vanishing gradient becomes more prominent [74] and affects the convergence of the model [73].

Keeping in mind the aforementioned issues and conditions, this dissertation makes several contributions and investigates the behavior and performance, in terms of computational cost and classification accuracy, of the most common and widely used classification algorithms in the HSI domain under different experimental setups. Moreover, this thesis developed several integrated methodologies, for instance, 1): A fast and compact 3D CNN model to overcome the computational complexity of DL, 2): A Hybrid CNN (3D followed by 2D CNN) to further improve the generalization performance in a computationally efficient fashion. 3): Soft labeling technique (Regularization) for DL to avoid non-uniform distribution over labels which leads to improving the accuracy of Deep models, finally, 4): a fully automatic approach for the selection of most informative and heterogeneous samples for training using a novel Spectral Angle Mapper (SAM) based objective function for the computation of attribute profiles in a computationally efficient fashion.

Chapter 3

Literature Review – Traditional to Deep Models

This chapter enlists a systematic overview of DL for HSIC and compared state-of-the-art strategies of the said topic. Primarily, this chapter encapsulates the main challenges of traditional machine learning for HSIC and then acquaint the superiority of DL to address these problems. This chapter breakdown the state-of-the-art DL frameworks into spectral features, spatial features, and together spatial-spectral features to systematically analyze the achievements of these frameworks. Moreover, this chapter shall consider the fact that DL requires a large number of labeled training examples whereas acquiring such a number for HSIC is challenging in terms of time and cost. Therefore, this chapter discusses some strategies to improve the generalization performance of DL strategies which can provide some future guidelines.

3.1 Motivation

HSI has been utilized for several real-world applications including but not limited to the atmosphere, ecology, urban, agriculture, geology and mineral exploration, coastal zone, marine, forestry i.e. track forest health, water quality and surface contamination, inland waters and wetlands, snow and ice, biological and medical (A few applications have been shown in Figure 3.1). There are several military applications in camouflage, landmine detection, and littoral zone mapping. HSI has been used in space, air, and underwater vehicles to acquire detailed spectral information for a wide range of uses as well [75–77]. Therefore, it is quite important to detach the surface features where each feature has a different spectrum band. HSI can capture more than 200 spectral bands which helps practitioners to discriminate objects that were not possible before.

Infield collection and spectral library indexing of ground truth signatures for any of the said applications are critical for many reasons. For instance, the spectral information of vegetation is prejudiced by a wide range of environmental situations that make it challenging to satisfactorily represent variability without the collection of site-specific field spectra. Thus,

FIGURE 3.1: Various real-world applications of HSI.

considering the aforementioned limitations, HSI analysis is categorized into the following main streams: dimensionality reduction [2, 78, 79], spectral unmixing [80–85], change detection [86–88] classification [6, 89–91], feature learning for classification [92–94], restoration and denoising [95, 96], resolution enhancement [97, 98]. Figure 3.2 shows an exponentially growing trend in literature published per year for HSI analysis-related tasks and applications.

This chapter specifically focuses on HSIC, which has achieved a phenomenal interest of the research community due to its broad applications in the areas of land use and land cover [99–102], environment monitoring and natural hazards detection [103, 104], vegetation mapping [105, 106] and urban planning. HSIC methodologies exploit machine learning algorithms to perform the classification task [107, 108]. These methods are outlined in various comprehensive reviews published during/in the last decade [90, 109–116]. Nevertheless, continuous advancements in the field of Machine Learning (ML) provide improved methods from time to time. The development of DL models is one of such revolutionary advancements in ML, which has proven to provide improved HSIC results [89, 117–119].

This chapter aims to give an overview of the widely used DL-based techniques to perform HSIC. Specifically, this chapter first summarizes the main challenges of HSIC which cannot be effectively overcome by traditional ML, and later enlist the advantages of DL to handle the above-mentioned issues. At a later stage, this chapter builds a framework that divides the corresponding works into:

1. Spectral and spatial feature learning, individually, and

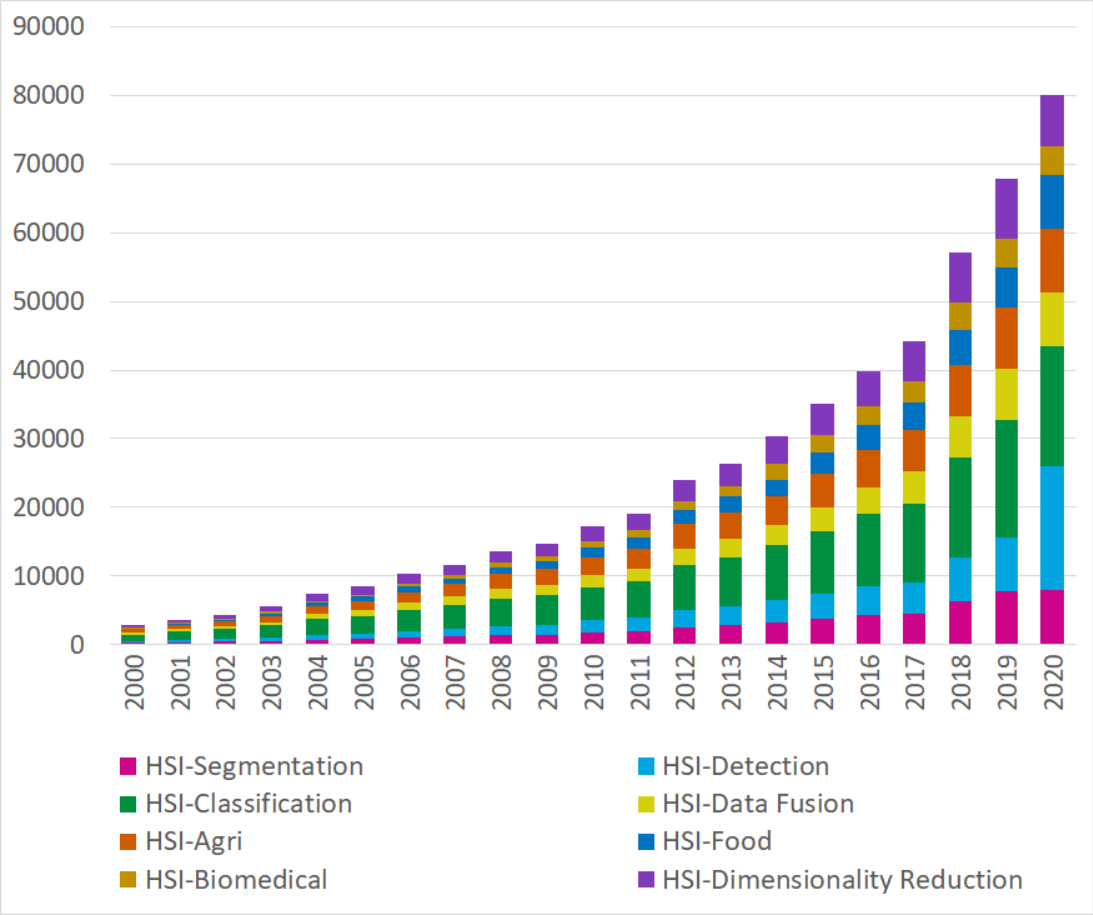


FIGURE 3.2: Various HSI related articles published per year, the results (including patents and citations) were sorted by relevance].

2. Spectral-spatial feature learning to systematically review the achievements in DL-based HSIC.

3. Future research stems to improve the generalization performance and robustness of DL models while considering the limited availability of reliable training samples.

3.2 Hyperspectral Image Classification (Background and Challenges)

3.2.1 Traditional to DL Models

The main task of HSIC is to assign a unique label to each pixel vector of HSI cube based on its spectral or spectral-spatial properties. Mathematically, an HSI cube can be represented as $\mathbf{X} = [x_1, x_2, x_3, \ldots, x_B]^T \in \mathcal{R}^{B\times(N\times M)}$, where B represent total number of spectral bands consisting of $(N \times M)$ samples per band belonging to $\mathbf{Y}$ classes where $x_i =$

$[x_{1,i},\ x_{2,i},\ x_{3,i}, \ldots, x_{B,i}]^T$ is the i^{th} sample in the HSI cube with class label $y_i \in \mathcal{R}^Y$. The classification problem can be considered as an optimization one, in which a mapping function $f_c(.)$ takes the input data **X** and after applying some transformations over it, obtains the corresponding label **Y**, to reduce the gap between obtained output and the actual one [68].

$$Y = f_c(X, \theta) \tag{3.1}$$

where θ is certain adjustable parameter that may be required to apply transformations on input data **X** such that $f_c : X \rightarrow Y$.

In literature, substantial work has been done on HSIC and there is a growing trend in the development of such techniques as shown in Figure 3.3. Most HSIC frameworks seemed to be influenced by the methodologies used in the computer vision domain [68]. Traditional machine learning-based HSIC approaches use hand-crafted features to train the classifier. These methods generally rely on utilizing engineering skills and domain expertise to design several human-engineered features, for instance, shape, texture, color, shape, spectral and spatial details. All these features are basic characteristics of an image and carry effective information for image classification. Commonly used hand-crafted feature extraction and classification methods include: texture descriptors such as Local Binary Patterns (LBPs) [120], Histogram of Oriented Gradients (HOG) [121], Global Image Scale-invariant Transform / Global Invariant Scalable Transform (GIST) [122], Pyramid Histogram of Oriented Gradients (PHOG), Scale-invariant Feature Transform (SIFT) [123], Random Forests [124], kernel-based Support Vector Machine (SVM) [125], K-nearest Neighbours (KNN), and Extreme Learning Machine (ELM).

Color histograms are simple and effective handcrafted features used for an image classification task. They are easy to compute and invariant to small changes in images i.e. translation and rotation. The major drawback of a color histogram is that it does not provide spatial contextual information, hence it becomes difficult to distinguish between objects of the same color but different distribution. Moreover, color histograms are sensitive to variance in illumination. HOG features represent the histogram of edge orientations of spatial sub-regions. It can effectively extract the edge and local shape details and has been utilized in various remote sensing related works [102, 126, 127].

Scale-invariant Feature Transform (SIFT) is a broadly used robust feature descriptor applied to image classification tasks [128–131]. The advantage of the SIFT descriptor is that it is invariant to the changes in image scale, rotation, illumination, and noise. SIFT is used to extract local features that describe a specific point in the image. The disadvantage of SIFT is that it is mathematically complex which increases its computational cost. GIST represents the global description of important aspects of an image that is the scales and orientations (gradient information) of various subregions of an image. GIST builds a spatial envelope in

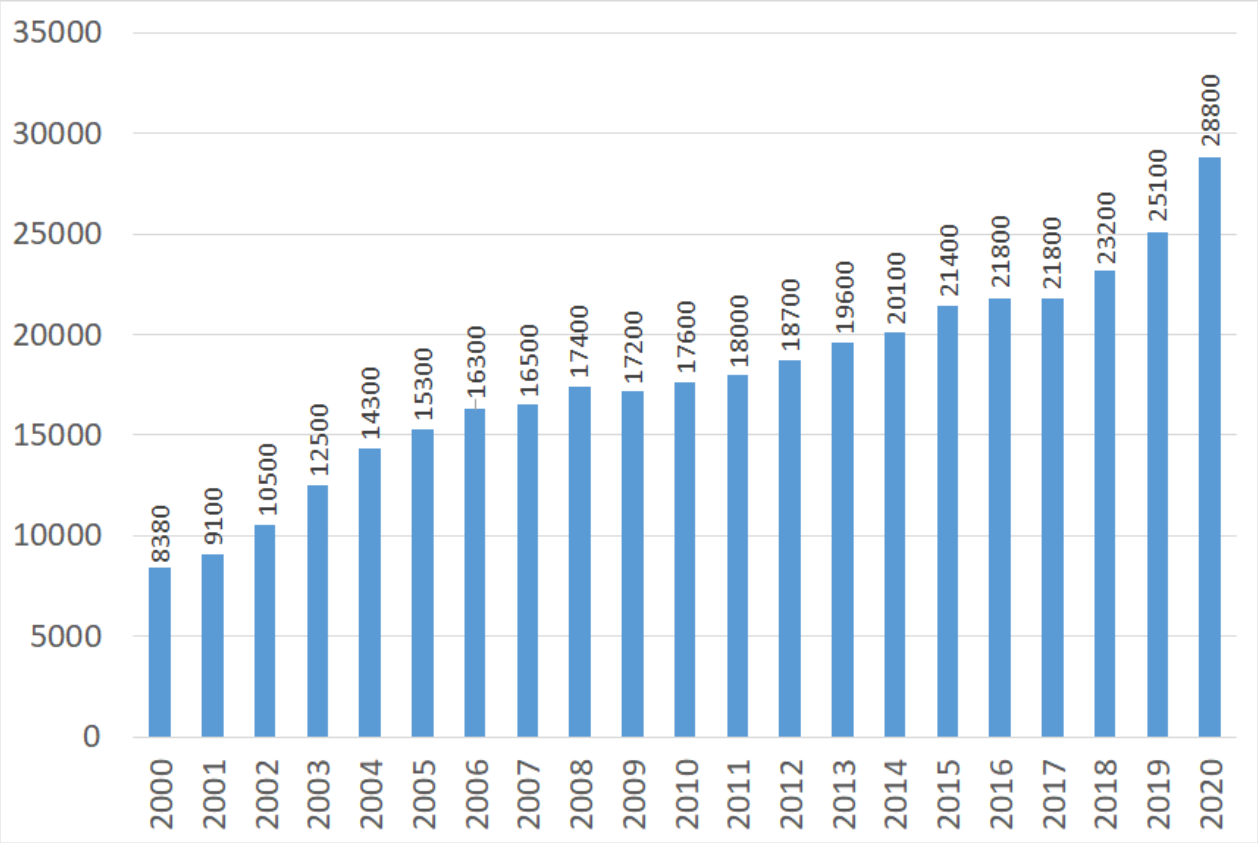


FIGURE 3.3: Remote sensing/Hyperspectral Image Classification related articles published per year. [Source: Google Scholar, the results (including patents and citations) were sorted by relevance].

terms of different statistical properties like roughness, openness, and ruggedness, etc [132]. Texture descriptors such as local binary patterns (LBPs) are used for remote sensing image analysis [120, 133]. LBPs are used to describe the texture around each pixel by choosing pixels from the square neighborhood and gray level values of all neighborhood pixels are thresholded with respect to the central pixel.

The color histograms, GIST, and texture descriptors are global features that represent certain statistical characteristics of an image like color, texture [134, 135], and spatial structure [122]. While HOG and SIFT are local features that describe geometrical information. Usually they are used to construct bag-of-visual-words (BoVW) models [100, 101, 104, 131, 136–141] and HOG feature-based models [102, 142]. Some popular feature encoding or pooling strategies to enhance the performance of BoVW are Fisher vector coding [120, 143, 144], Spatial Pyramid Matching (SPM) [145], and Probabilistic Topic Model (PTM) [141, 146–148]. A single feature is insufficient to represent the whole image information, hence a combination of these features is used for image classification [100, 136, 146, 148–154].

Hand-crafted features can effectively represent the various attributes of an image, hence work well with the data being analyzed. However, these features may be insubstantial in the case of real data, therefore it is difficult to fine-tune between robustness and discriminability

as the set of optimal features considerably vary between different data. Furthermore, human involvement in designing the features considerably affects the classification process, as it requires a high level of domain expertise to design hand-crafted features.

To mitigate the limitations of hand-crafted feature designing, a deep feature learning strategy was proposed by Hinton and Salakhutdinov in 2006 [155]. Deep learning (DL) based methods can automatically learn the features from data in a hierarchical manner, to construct a model with growing semantic layers until a suitable representation is achieved. Such models have shown great potential for feature representation in remote sensing image classification [156, 157].

DL architectures can learn the behavior of any data without any prior knowledge regarding the statistical distribution of the input data [66] and can extract both linear and non-linear features of input data without any pre-specified information. Such systems are capable of handling HSI data in both spectral and spatial domains individually, and also in a coupled fashion. DL systems possess a flexible architecture in terms of types of layers and their depth and are adaptive to various machine learning strategies like supervised, semi-supervised, and unsupervised techniques.

3.2.2 Hyperspectral Data Characteristics and DL Challenges

Despite the above-discussed DL potentials, there are still some challenges that need to be considered while applying DL to HSI data. Most of these challenges are related to the characteristics of HSI data i.e. hundreds of contiguous and narrow spectral channels with very high spectral resolution and low spatial resolution throughout the electromagnetic spectrum coupled with limited availability of training data. Although the pixels with rich spectral information are useful for classification purposes, however, the computation of such data takes a lot of time and resources.

Furthermore, processing such high-dimensional data is a somewhat complex task due to an increased number of parameters. This is known as the curse of dimensionality which considerably influences the classification performance especially in the case of supervised learning [158]. Since the size of training data is not adequate/insufficient and/or not reliable (i.e. the training samples may not provide any new information to the model or may have similar patterns/structures) to properly train the classifier which may lead the model to overfit. This is known as the Hughes phenomena [159] which occurs when labeled training data is significantly smaller than the number of spectral bands present in the data. Lack of labeled HSI data is a major issue in HSIC as labeling of HSI is a time-consuming and expensive task because it usually requires human experts or investigation of real-time scenarios.

In addition to high dimensionality, HSIC suffers from various other artifacts like high intra-class variability due to unconfined variations in reflectance values caused by several environmental interferers and degradation of data caused by instrumental noise while capturing the data [160]. Furthermore, the addition of redundant bands due to HSI instruments affects the computational complexity of the model. Spectral mixing is another challenge related to the spatial resolution of HSI. HSI pixels with low to average spatial resolution cover vast spatial regions on the surface of earth leading to mixed spectral signatures which result in high inter-class similarity in border regions. As a result, it becomes difficult to identify the materials based on their spectral reflectance values [72]. Following are some main challenges that come across when DL is applied to HSIC:

- **Complex Training Process**: Training of Deep Neural Network (DNN) and optimization by tuning parameters is an NP-complete problem where the convergence of the optimization process is not guaranteed [65]. Therefore, it is assumed that training of DNN is very difficult [66] especially in the case of HSI when a large number of parameters need to be adjusted/tuned.

- **Limited Availability of Training Data**: As discussed above, supervised DNN requires a considerably large amount of training data otherwise their tendency to overfit increases significantly [67] leads to the Hughes phenomena. The high dimensional characteristic of HSI coupled with a small amount of labeled training data makes the DNNs ineffective for HSIC as it demands a lot of adjustments during the training phase [68].

- **Model's Interpretability**: The training procedure of DNNs is difficult to interpret and understand. The black box kind of nature is considered as a potential weakness of DNNs and may affect the design decisions of the optimization process. Although, a lot of work has been done to interpret the model's internal dynamics.

- **High Computational Burden**: One of the main challenges of DNN is dealing with a big amount of data that involves increased memory bandwidth, high computational cost, and storage consumption [69]. However, advanced processing techniques like parallel and distributed architectures [70, 71] and high-performance computing (HPC) [72] make it possible for DNNs to process large amounts of data.

- **Training Accuracy Degradation**: It is assumed that deeper networks extract more rich features from data [73], however, this is not true for all systems to achieve higher accuracy by simply adding more layers. Because by increasing the network's depth, the problem of exploding or vanishing gradient becomes more prominent [74] and affects the convergence of the model [73].

3.3 Hyperspectral Data Representation

Hyperspectral data is represented in the form of a $3D$ hypercube, $X \in \mathcal{R}^{B\times(N\times M)}$, which contains $1D$ spectral and $2D$ spatial details of a sample where B represents the total number of spectral bands and N and M are spatial components i.e., width and height, respectively.

3.3.1 Spectral Representation

In such representations, each pixel vector is isolated from other pixels and processed based on spectral signatures only which means the pixel is represented only in spectral space $x_i \in \mathcal{R}^B$. Where B can either be the actual number of spectral channels or just relevant spectral bands extracted after some Dimensionality Reduction (DR) method. Usually, instead of using original spectral bands, a low dimensional representation of HSI is preferred for data processing in order to avoid redundancy and achieve better class separability, without considerable loss of useful information.

DR approaches for spectral HSI representation can either be supervised or unsupervised. Unsupervised techniques transform the high dimensional HSI into a low dimensional space without using the class label information, for example, Principal Component Analysis (PCA) and Locally Linear Embedding [161]. On the other hand, supervised DR methods utilize labeled samples to learn the data distribution i.e. to keep data points of the same classes near to each other and separate the data points of different classes. For instance, linear discriminant analysis (LDA), local Fisher discriminant analysis (LFDA) [162], local discriminant embedding (LDE) [163] and nonparametric weighted feature extraction (NWFE) [164]. LDA and LDFA provide better class separability by maximizing the inter-class distance of data points and minimizing the intra-class distance. However, due to the spectral mixing effect, in which the same material may appear with different spectra or different materials may have the same spectral signatures, it becomes difficult to differentiate among different classes based on the spectral reflectance values alone.

3.3.2 Spatial Representation

To deal with the limitations of spectral representation, another approach is to exploit the spatial information of the pixels, in which pixels in each band are represented in the form of a matrix, $x_i \in \mathcal{R}^{N\times M}$. Due to high spatial correlation, neighboring pixels have higher probabilities to belong to the same class. Therefore, in the case of spatial representation, neighboring pixels' information is also considered and the neighborhood of a pixel can be determined using kernel or pixel centric window [165]. Some common methods to extract spatial information from HSI cube are morphological profiles (MPs), texture features (like

Gabor filters, gray-level co-occurrence matrix (GLCM), and local binary pattern (LBP), etc.) and DNN based methods. Morphological profiles are capable of extracting geometrical characteristics. Few extensions of MPs include extended morphological profiles (EMPs) [166] and multiple-structure-element morphological profiles [167].

The texture of the image provides useful spatial contextual information of HSI. For instance, a Gabor filter, a texture analysis technique, can efficiently obtain textural information at various scales and orientations. Similarly, LBP can provide rotation-invariant spatial texture representation. The GLCM can effectively determine the spatial variability of HSI by exploiting the relative positions of neighborhood pixels. The DNNs can also extract spatial information of HSI by considering the pixel as an image patch instead of representing it as a spectral vector. The spatial information contained in HSI can also be extracted by combining various of the afore discussed methods. For instance, [168]combined Gabor filter and differential morphological profiles [169] to extract local spatial sequential features for a recurrent neural network (RNN) based HSIC framework.

3.3.3 Spectral-Spatial Representation

This representation jointly exploits both spectral and spatial information of data. In such approaches, a pixel vector is processed based on spectral features while considering spatial-contextual information. The strategies that simultaneously use both spectral and spatial representations of HSI, either concatenate the spatial details with spectral vector [114, 170] or process the $3D$ HSI cube to preserve the actual structure and contextual information [171].

In literature, all these HSI representations are widely exploited for HSIC. Most of the DNNs for pixel-wise classification utilized the spectral representation of HSIs [172, 173]. However, to mitigate the limitations of spectral representation, many efforts have been made to incorporate the spatial information [174, 175]. Recently, joint exploitation of both spectral and spatial features has gained much popularity and led to improved classification accuracy [2, 117, 176–179]. These HSI feature exploitation approaches, for HSIC, are further discussed in the following sections.

3.4 Learning Strategies

DL models can adopt various learning strategies that can be broadly categorized into the following:

3.4.1 Supervised Learning

In a supervised learning approach, the model is trained based on the labeled training data which means training data is comprised of a set of inputs and their corresponding outputs or class labels. During the training phase, the model iteratively updates its parameters in order to predict the desired outputs accurately. In the testing phase, the model is tested against the new input/test data in order to validate its ability to predict the correct labels. If trained sufficiently, the model can predict the labels of new input data. However, supervised learning of DNNs requires a lot of labeled training data to fine-tune the model parameter. Therefore, they are best suited to scenarios where plentiful labeled data is available. The details of various supervised learning techniques for DNNs will be explained in the respective sections.

3.4.2 Unsupervised Learning

In contrast to the supervised learning approach, unsupervised learning techniques learn from the input data with no explicit labels associated with it. These approaches try to identify the underlying statistical structure of input representations or patterns in the absence of corresponding labels. As there is no ground truth available for the training data so it might be difficult to measure the accuracy of the trained model. However, such learning strategies are useful in the cases where we want to learn the inherent structure of such datasets which have a scarcity of training data. The PCA is an unsupervised learning technique that can be used to learn a low-dimensional representation of the input. Similarly, k-means clustering is another unsupervised learning method that groups the input data into homogeneous clusters.

3.4.3 Semi-supervised Learning

The semi-supervised learning technique is halfway between unsupervised and supervised approaches. It learns from the partially labeled Datasets that are a small amount of labeled training data that can be utilized to label the rest of the unlabeled data. These techniques effectively utilize all available data instead of just labeled data, therefore, these techniques have gained much popularity among the research community and are being widely used for HSIC [180–183]. The details of these methods are briefly described in section 3.10.

3.5 Development of DNNs (Types of Layers)

In the following, we review recent developments of some widely used DNN frameworks for HSIC. We specifically surveyed the literature published from 2017 onward. DNNs exhibit a great variety of flexible and configurable models for HSIC that allow the incorporation of several types of layers. Few widely used types of layers are explained in the following subsection.

A layer is the key building block of DNN and the type of layer has a decisive impact in terms of feature processing. A layer takes the weighted input, processes it through linear or non-linear transformation, and outputs these values to the next layer. Generally, a layer is a uniform, as it has a single activation function. The first layer of the network is known as the input layer and the last layer as an output layer. All other layers in the network, in between the input and output layers, are known as hidden layers. These layers progressively find different features in the input data by performing various transformations. The choice of layer type depends on the task at hand, as some layers perform better for some tasks than others. The most commonly used layers for HSIC are explained below.

3.5.1 Fully Connected Layers

A fully connected (FC) layer connects every neuron in the lower layer to every neuron in the upper/next layer. Mostly, they are used as the last few layers of a model usually after convolution/pooling layers. FC takes the output of the previous layer and assigns weights to predict the probabilities for class labels. Due to a large number of connections, a large number of parameters need to be adjusted which significantly increases the computational overhead. Moreover, due to a large number of parameters, the model becomes more sensitive to overfitting [99]. However, to mitigate the effect of overfitting, a dropout method is introduced in [184].

3.5.2 Convolutional Layers

The convolutional (CONV) layer convolve the input data or feature maps from a lower layer with the filters (kernels). The filter contains weights whose dot product is calculated with the subset of input data by moving it across the width, height, and depth of the input region. The output of the filter is known as a feature map. CONV layer provides spatial invariance via a local connectivity approach in which the neuron in the feature map connects to a subset of input from the previous layer rather than connecting to every neuron. This reduces the number of parameters that need to train. To further reduce the number of parameters, the

CONV layer uses the mechanism of parameter sharing in which the same weights are used in a particular feature map.

3.5.3 Activation Layers

Activation layers are assumed to be a feature detector stage of DNNs [185]. FC and CONV layers provide linear representations of input data or it can be said that they work similarly to linear regressors and data transformed by these layers is considered to be at the feature extraction stage [68]. Therefore, to learn non-linear features of data, an activation layer must be used after FC and CONV layers. In the activation layer, feature maps from previous layers go through an activation function to form an activation map. Some commonly used activation functions are sigmoid, hyperbolic tangent (tanh), rectified linear unit (ReLU), and softmax. However, in HSI analysis, softmax and ReLU are widely employed activation functions [68]. Figure 3.4 presents a graphical representation of a few commonly utilized activation functions.

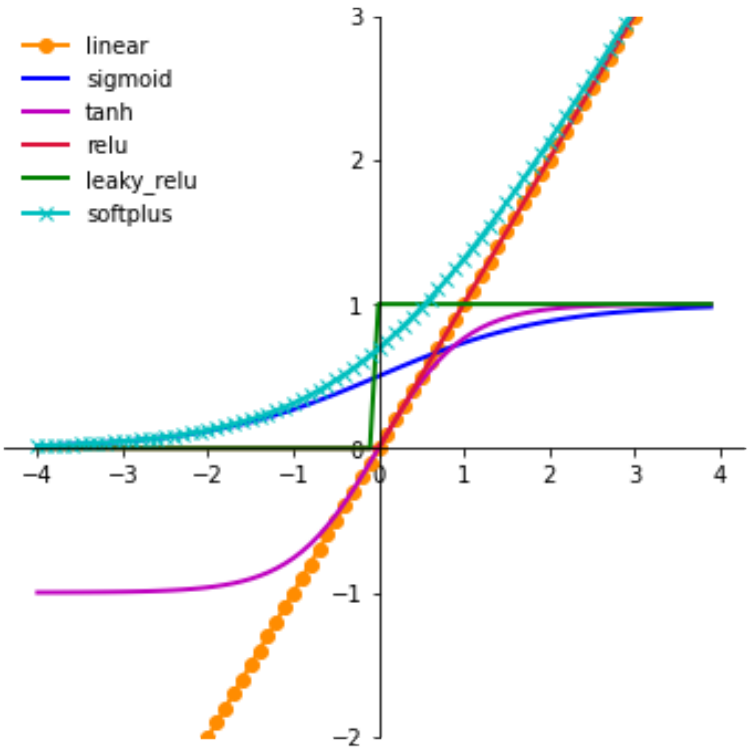


FIGURE 3.4: Graphical representation of various commonly used activation functions

3.5.4 Pooling or Sub-sampling layers

The pooling layer, also known as the sub-sampling or down-sampling layer, takes a certain input volume and reduces it to a single value as shown in Figure 3.5. This provides invariance to small distortions in the data. The pooling layer helps the model to control overfitting as the size of data and model parameters both are reduced which also leads to a decrease in the computational time. The commonly used down-sampling operations are max-pooling, average-pooling, and sum-pooling. Recently, a pooling technique, wavelet-pooling is introduced in [186] whose performance is commensurable to max-pooling and average-pooling. Alternatively, [187] proposed another trend in which the pooling layer is replaced by the CONV layer of increased filter stride.

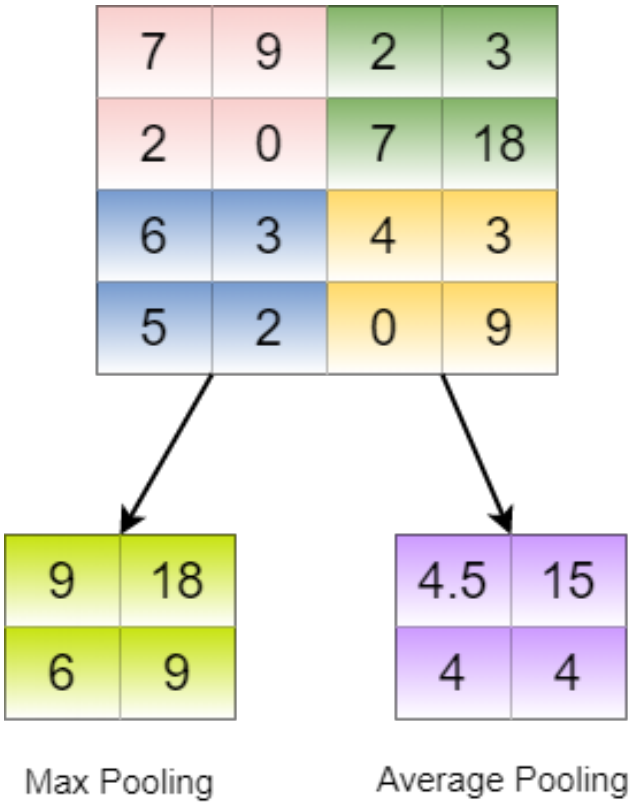


FIGURE 3.5: Max-pooling and average-pooling operations of down-sampling/pooling layer

3.6 Convolutional Neural Network (CNN)

The architecture of the CNN is inspired by the biological visual system presented in [188]. Following the natural visual recognition mechanism proposed by Hubel and Wiesel [188], Neocognitron [189] is regarded as the first hierarchical, position-invariant model for pattern recognition [190] which can be considered as the predecessor of CNN [191]. The architecture

of CNN can be divided into two main stages: one is Feature Extraction (FE) network and the other is a classification based on the feature maps extracted in the first stage.

The FE network consists of multiple hierarchically stacked CONV, activation, and pooling layers. The CONV layer extracts the features from input data by convolving a learned kernel with it. On each CONV layer, the kernel is spatially shared with whole input data which reduces the model's complexity and the network becomes easier to train as the number of parameters that need to be fine-tuned is reduced. Convolved results are then passed through an activation layer which adds nonlinearities in the network to extract non-linear features of the input. This is achieved by applying a non-linear function to the convolved results. Afterward, the resolution of the feature map is reduced by applying a pooling operation to achieve shift-invariance. Generally, the pooling layer is added with every CONV layer followed by the activation function.

The classification stage consisting of FC layers and a Softmax operator gives the probability of input pattern belonging to a specific class based on the feature maps extracted at the FE stage. FC layer connects every single neuron in the previous layer to every neuron in the current layer. In [192] and [193], the authors proposed that the FC layer can be disregarded by using a global average pooling layer. Softmax is commonly used for classification tasks [194–196] however, many works have also utilized SVM [197, 198] for this purpose.

In the following, we reviewed three types of CNN architectures for HSIC: i) Spectral CNN, ii) Spatial CNN and iii) Spectral-spatial CNN. Figure 3.6 illustrates the general architecture of these three frameworks.

3.6.1 Spectral CNN Frameworks for HSIC

Spectral CNN models only consider $1D$ spectral information ($x_i \in \mathcal{R}^B$) as input, where B could either be the original number of spectral bands or the appropriate number of bands extracted after some dimensionality reduction method. In [199], a CNN structure was proposed to mitigate the overfitting problem and achieved a better generalization capability by utilizes 1×1 convolutional kernels and enhanced dropout rates. Moreover, a global average pooling layer is used in place of a fully connected layer in order to reduce the network parameters. To reduce high correlation among HSI bands [193] proposed a CNN architecture for HSIC which fully utilized the spectral information by transforming the $1D$ spectral vector to a $2D$ feature matrix and by cascading composite layers consisting of 1×1 and 3×3 CONV layers, the architecture achieved the feature reuse capability. Similar to [193, 199] also utilized the global average pooling layer to lower the network's training parameters and to extract high dimensional features.

In [200] authors presented a hybrid model for HSIC in which the first few CONV layers are employed to extract position invariant middle-level features and then recurrent layers

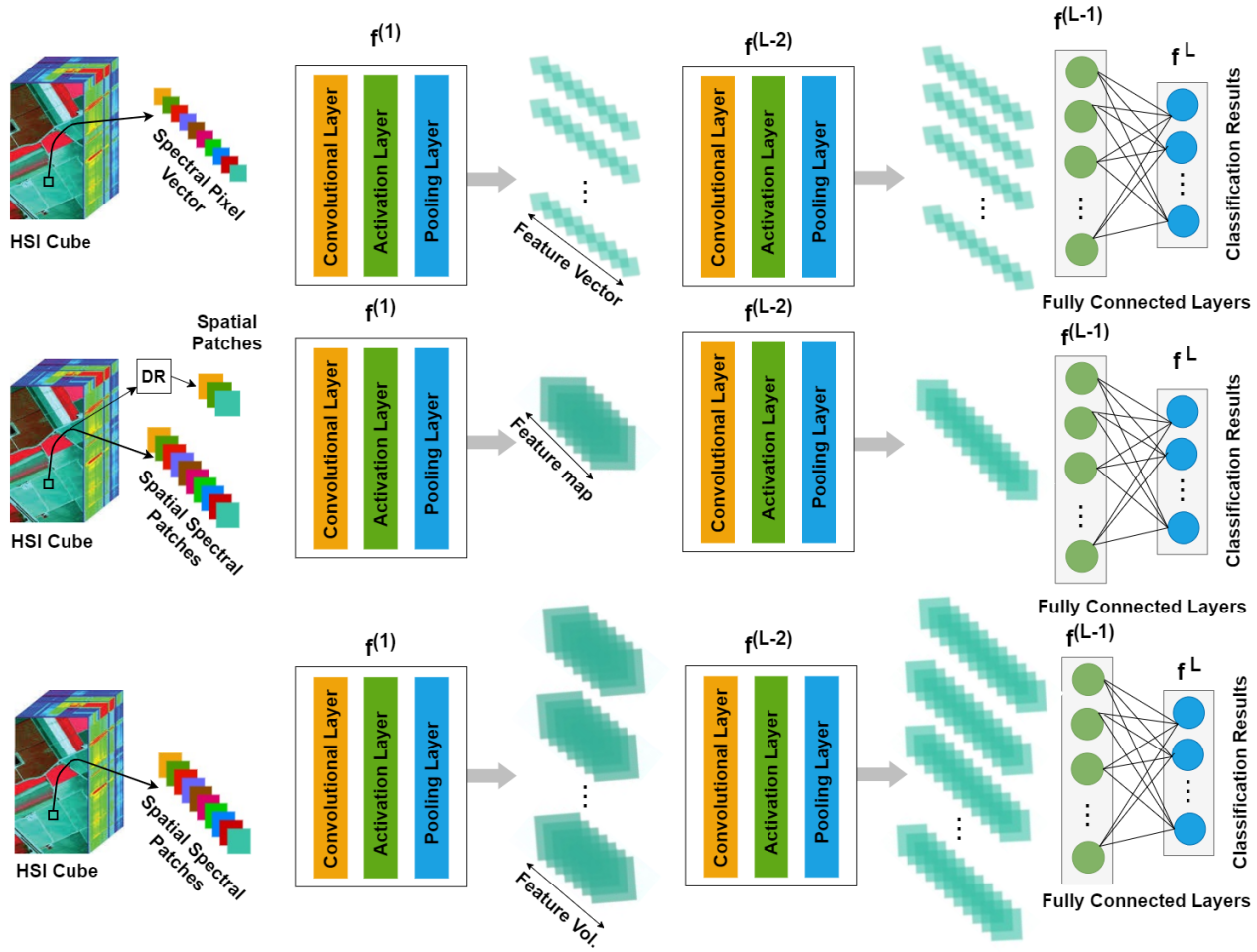


FIGURE 3.6: General architecture of Spectral CNN, Spatial CNN and Spectral-spatial CNN frameworks for HSIC.

are used to extract spectral-contextual details. Similarly, [172] used a hybrid architecture for classifying healthy and diseased Wheat heads. For the input layer, they transform spectral information into a $2D$ data structure. In [201] CNN proved to be more effective as compared to SVM and KNN for the spectral-based identification of rice seed's variety. A similar application of CNN was explored in [173] where various varieties of Chrysanthemum were identified using spectral data of the first five PCs of Principal component analysis (PCA). PCA is a dimensionality reduction method that is widely used in many DL applications to handle/preprocess high dimensional data. In [202] PCA was utilized to preprocess medical HSI and then the fusion of CNN kernels with Gabor kernels using dot product is used for classification.

The study [203] analyzed another dimensionality reduction technique Dynamic Mode Decomposition (DMD) which converted $3D$ HSI data to $2D$ and then this data is fed to vectorized CNN (VCNN) for classification. To overcome the noise effect in pixel-wise HSIC, a method of averaged spectra is used in [204] where an averaged spectra of a group of pixels

belonging to bacterial colonies is extracted for further analysis.

3.6.2 Spatial CNN frameworks for HSIC

Spatial CNN models only consider spatial information and to extract the spatial information from HSI data, dimensionality reduction (DR) methods are employed on spectral-domain to lower the dimensionality of original HSI data. For instance, [205] used PCA to extract the first PC with refined spatial information and fed it to a fully CNN framework for classification. Similarly, [206] trained a spatial-based $2D$-CNN with one PC. In [207], PCA whitened input data considering three PCs is fed to a random patches network as a $2D$-CNN classification framework.

The method proposed in [208] cropped the patches from $2D$ input images (i.e. images from the different spectral bands) to train a $2D$-CNN architecture that learns the data-adaptive kernels by itself. Furthermore, some authors also proposed the utilization of hand-crafted features along with spectral-domain reduction. For example, [209] combined the Gabor filtering technique with $2D$-CNN for HSIC to overcome the overfitting problem due to limited training samples. The Gabor filtering extracts the spatial details including edges and textures which effectively reduce the overfitting problem. The work [210] proposed a deformable HSIC network based on the concept of deformable sampling locations which can adaptively adjust their size and shape in accordance with HSI's spatial features. Such sampling locations are created by calculating $2D$ offsets for every pixel in the input image through regular convolutions by taking into account three PCs. These offsets are able to cover the locations of similar neighboring pixels possessing similar characteristics. Then structural information of neighboring pixels is fused to make deformable feature images. Regular convolution employed on these deformable feature images can extract more effective complex structures.

3.6.3 Spectral-Spatial CNN frameworks for HSIC

Spectral-spatial pixel-wise HSIC can be achieved by integrating spatial features into spectral information. For instance, [211] presented an improved pixel pair feature (PPF) [212] approach called spatial pixel pair feature which is different from traditional PPFs with respect to two main aspects: one is the selection of pixel pair that is only the pixel from the immediate neighborhood of central pixel can be used to make a pair, second is the label of pixel pair would be as of central pixel. To extract discriminative joint representation [213] introduced a supervised spectral-spatial residual Network (SSRN) that uses a series of 3D

convolutions in the respective spectral and spatial residual blocks. An efficient deep 3*D*-CNN framework was proposed in [214] that simultaneously exploits both spectral and spatial information for HSIC. Similarly, to reflect the variations of spatial contexture in various hyperspectral patches, [215] implemented an adaptive weight learning technique instead of assigning fixed weights to incorporate spatial details. Besides this, to make the convolutional kernel more flexible [179] explored a new architectural design that can adaptively find adjustable receptive filed and then an improved spectral-spatial residual network for joint feature extraction. The discriminative power of the extracted features can be further improved by combining both the max and min convolutional features before the ReLU nonlinearity reported in [216] for the classification task.

The deeper networks may suffer from the issues of overfitting and gradient vanishing problems due to the smaller number of available labeled training samples and to overcome this shortcoming the lightweight CNN's gain good attention in HSIC communities. The paper [217] introduced an end-to-end 3D lightweight convolutional neural network to tackle the limited numbers of training samples for HSI classification. To reduce the large gap between the massive trainable parameters and the limited labeled samples [218] proposed to extract the spatial-spectral Schroedinger eigenmaps (SSSE) joint spatial-spectral information, and then further reduced the dimensionality using compression technique. Approximately 90% of trainable weights of the total parameters are used immediately after the flatten operation i.e., in the fully connected layer, whereas the remaining only 10% weights are used on the previous convolutional layers of the whole network. To overcome the paper [219] introduced a lightweight bag-of-feature learning paradigm into an end-to-end spectral-spatial squeeze-and-excitation residual network for HSIC.

The morphological operations i.e., erosion and dilation are powerful nonlinear feature transformations that are widely used to preserve the essential characteristics of shape and structural information of an image. Inspired by these the paper [220] introduced a new end-to-end morphological convolutional neural network (MorphCNN) for HSIC which utilizes both the spectral and spatial features by concatenating the outputs from spectral and spatial morphological blocks extracted in a dual-path fashion.

The work [215] proposed a two-stage framework for joint spectral-spatial HSIC which can directly extract both spectral and spatial features instead of independently concatenating them. The first stage of the proposed network is comprised of a CNN and softmax normalization that adaptively learns the weights for input patches and extracts joint shallow features. These shallow features are then fed to a network of Stacked Autoencoder (SAE) to obtain deep hierarchical features and final classification is performed with a Multinomial Logistic Regression (MLR) layer. A 3*D*-CNN model was introduced in [221] to jointly exploit spectral-spatial features from HSI and to validate its performance comparison is performed

with spectral-based DBN, SAE, and 2D-spatial CNN for HSIC. The work [222] introduced a bilinear fusion mechanism over the two branches of squeeze operation based on the global and max-pooling whereas the excitation operation is performed with the fused output of squeeze operation.

The work [223] proposed a deep multiscale spectral-spatial feature extraction approach for HSIC which can learn effective discriminant features from the images with high spatial diversity. The framework utilizes the Fully Convolutional Network (FCN) to extract deep spatial information and then, these features are fused with spectral information by using a weighted fusion strategy. Finally, pixel-wise classification is performed on these fused features.

In [224] a dual-channel CNN framework was implemented for spectral-spatial HSIC. In the proposed approach, 1D-CNN is used to hierarchically extract spectral features and 2D-CNN to extract hierarchical spatial features. These features are then combined together for the final classification task. Furthermore, to overcome the deficiency of training data and to achieve higher classification accuracy, the proposed framework is supported by a data augmentation technique that can increase the training samples by a factor of 6. In [225], a multiscale 3D deep CNN is introduced for end-to-end HSIC which can jointly learn both 1D spectral and 2D multiscale spatial features without any pre-processing or post-processing techniques like PCA, etc. In order to reduce the band redundancy or noise in HSI, [226] explored a novel architecture for HSIC by embedding a band attention module in the traditional CNN framework. The study [227] proposed an HSIC architecture in which PCA transformed images are used to obtain multi-scale cubes for handcrafted feature extraction by utilizing multi-scale covariance maps which can simultaneously exploit spectral-spatial details of HSI. These maps are then used to train the traditional CNN model for classification.

The work [228] combined CNN with metric learning-based HSIC framework which first utilizes CNN to extract deep spatial information using the first three PCs extracted by PCA. Then, in a metric learning-based framework, spectral and spatial features are fused together for spectral-spatial feature learning by embedding a metric learning regularization factor for the classifier's training (SVM).

Similarly, [229] combines multi-scale convolution-based CNN (MS-CNN) with diversified deep metrics based on determinantal point process (DPP) [230] priors for (1-D spectral, 2-D spectral-spatial, and 3-D spectral-spatial) HSIC. Multiscale filters are used in CNN to obtain multi-scale features and DPP-based diversified metric transformation is performed to increase the inter-class variance and decrease intra-class variance, and better HSI representational ability. Final classification maps are obtained by using a softmax classifier.

In recent work, [231] an HSIC framework is proposed to extract multi-scale spatial features by constructing a three-channel virtual RGB image from HSI instead of extracting the first three PCs through PCA. The purpose of using a three-channel RGB image is to utilize existing networks trained on natural images to extract spatial features. For multi-scale feature extraction, these images are passed to a fully convolutional network. These multi-scale spatial features are fused and further joined with PCS extracted spectral features for final classification via SVM.

A two-branch (spectral and spatial) DNN for HSIC was introduced in [232]. The spatial branch consists of a band selection layer and a convolutional and de-convolutional framework with skip architecture to extract spatial information of HSI, and in the spectral branch, a contextual DNN is used to extract spectral features. The paper [233] introduced an adaptive band selection based semi-supervised $3D$-CNN to jointly exploit spectral-spatial features whereas [234] explored dual-attention based autoencoder-decoder network for unsupervised hyperspectral band selection and then joint feature extraction for land cover class prediction. Similarly, in [235] spectral-spatial features are simultaneously exploited in an unsupervised manner using a $3D$ convolution autoencoder. A hybrid $3D-2D$-CNN architecture was presented by [236] in which $3D$-CNN is first used to extract joint spectral-spatial features and then $2D$-CNN is further used to obtain more abstract spatial contextual features. The study [237] proposed a Bayesian HSIC architecture that combines CNN with Markov random field. The CNN first extracts joint spectral-spatial features and then a smooth MRF prior is placed on class labels to further refine the spatial details.

3.6.4 Future directions for CNN-based HSIC

In the preceding section, we have reviewed the recent developments of CNNs for HSIC. Although CNN's based HSIC frameworks have achieved great success with respect to classification performance, there are still many aspects that need further investigation. For instance, there is a need to further work on such models that can jointly employ spatial and spectral information for HSIC. Many of the above-surveyed frameworks use dimensionality reduction methods to achieve better spectral-spatial representation but such approaches discard useful spectral information of HSI. Hence the development of robust HSIC approaches that can preserve spectral information is required. However, processing of such approaches increases the computational burden, and the training process becomes slower, therefore, parallel processing of such networks using FPGAs and GPUs is desired in order to achieve the computationally fast models, that can even be suitable for mobile platforms, without the performance degradation. Moreover, as the CNNs are becoming deeper and deeper, more labeled training data is required for accurate classification, and as discussed before, there

is a lack of labeled training data in HSI. In order to overcome this issue, more research is required to integrate the CNN with unsupervised or semi-supervised approaches.

3.7 Autoencoders (AE)

Autoencoder (AE) is a popular symmetrical neural network for HSIC due to its unsupervised feature learning capability. AE itself does not perform a classification task instead it gives a compressed feature representation of high-dimensional HSI data. AE consists of an input layer, one hidden or encoding layer, one reconstruction or decoding layer, and an output layer as shown in Figure 3.7. AE is trained on input data in such a manner to encode it into a latent representation that is able to reconstruct the input. To learn a compressed feature representation of input data, AE tries to reduce the reconstruction error that is minimizing the difference between the input and the output.

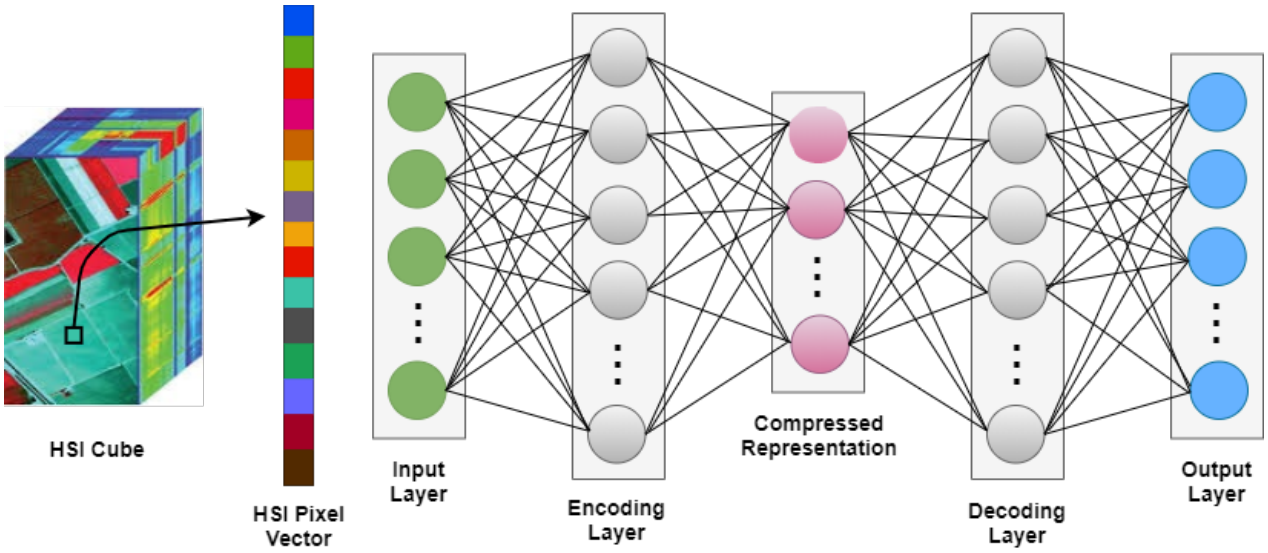


FIGURE 3.7: A general Autoencoder Architecture

Whereas, the Stacked Autoencoder (SAE) is built by stacking multiple layers of AEs in such a way that the output of one layer is served as an input of the subsequent layer. Denoising autoencoder (DAE) is a variant of AE that has a similar structure as AE except for the input data. In DAE, the input is corrupted by adding noise to it, however, the output is the original input signal without noise. Therefore, DAE, different from AE, has the capability to recover original input from a noisy input signal.

To learn high-level representation from data, the work [238] proposed a combination of multi-layer AEs with maximum noise fraction which reduces the spectral dimensionality of HSI, while a softmax logistic regression classifier is employed for HSIC. The study reported

in [239] combined multi-manifold learning framework proposed by [240] with Counteractive Autoencoder [241] for improved unsupervised HSIC. The work [242] jointly exploited spectral-spatial features of HSI through an unsupervised feature extracting framework composed of recursive autoencoders (RAE) network. It extracts the features from the neighborhood of the target pixel and weights are assigned based on the spectral similarity between target and neighboring pixels. A two-stream DNN with a class-specific fusion scheme was introduced in [243] which learns the fusion weights adaptively. One stream composed of stacked denoising auto-encoder is used to extract spectral features and the second stream is implemented to extract spatial information using Convolutional Neural Network (CNN), while final classification is performed by fusing the class prediction scores obtained from the classification results of both streams.

Another work proposed a hybrid architecture for multi-feature based spectral-spatial HSIC which utilizes Principle Component Analysis (PCA) for dimensionality reduction, guided filters [244] to obtain spatial information and sparse AE for high-level feature extraction. The framework proposed in [245] exploited both spectral and spatial information for HSIC by adopting batch-based training of AEs and features are generated by fusing spectral and spatial information via a mean pooling scheme. Another work [246] developed a spectral-spatial HSIC framework by extracting appropriate spatial resolution of HSI and utilization of stacked sparse AE for high-level feature extraction followed by Random Forest (RF) for the final classification task.

Similarly, [247] also used stacked sparse AE for various types of representation that is spectral-spatial and multi-fractal features along with other higher-order statistical representations. A combination of SAE and extreme learning machine was proposed in [248] for HSIC, which segments the features of the training set and transform them via SAE, after transformation, feature subsets are rearranged according to the original order of the training set and fed to extreme learning machine-based classifiers, while Q-statistics is used for final classification result. This processing of feature subsets helps to improve variance among base classifiers [248]. Similarly, in a recent work [249] implemented a computationally efficient multi-layer extreme learning machine-based AE which learns the features in three folds, as proposed in [250] for HSIC.

To overcome the issue of high intra-class variability and high inter-class similarity in HSI, [251] developed a stacked Autoencoder (SAE) based HSIC which can learn compact and discriminative features by imposing a local fisher discriminant regularization. Similarly, in the latest work [252] a k-sparse denoising AE is spliced with and spectral–restricted spatial features that overcome the high intra-class variability of spatial features for HSIC. The study [253] proposed an HSIC architecture that first makes the spectral segments of HSI based on mutual information measure to reduce the computation time during feature extraction

via SAE, while spatial information is incorporated by using extended morphological profiles (EMPs) and SVM/RF is used for final classification. Recently, [254] used SAE for the classification of an oil slick on the sea surface by jointly exploit spectral-spatial features of HSI.

3.7.1 Future Directions for AE-based HSIC

In the above section, we have surveyed the recent developments of AEs based techniques for HSIC. Although such frameworks provide powerful predictive performance and show good generalization capabilities, more sophisticated work is still desired. Many of the discussed approaches do not fully exploit abundant spatial information so further techniques need to be developed that can fully employ joint spatial and spectral information for HSIC. Moreover, the issue of high intra-class variability and high inter-class similarity in HSI also hinders the classification performance. Many of the above-reviewed works have addressed this issue but further research to overcome this aforesaid issue is required. One direction could be further exploring approaches like pre-training, co-training, and adaptive neural networks, etc for AE-based HSIC frameworks.

3.8 Deep Belief Network (DBN)

Deep Belief Network (DBN) [255] is a hierarchical deep DNN that learns the features from input in an unsupervised, layer-by-layer approach. The layers in DBN are built using Restricted Boltzmann Machine (RBM) comprised of a two-layer architecture in which visible units are connected to hidden units [256] as shown in Figure 3.8.

A detailed overview of RBM can be found at [256]. To extract more comprehensive features from input data, the hidden unit of one RBM can be feed to the visible units of other RBM. This type of layer-by-layer architecture builds a DBN, which is trained greedily and can capture deep features from HSI. The architecture of three-layer DBN is shown in Figure 3.9.

In literature, several works implemented DBN for HSIC. For instance, [257] used DBN for land cover classification by combining spectral-spatial information and made a comparison with some other classification approaches. The usual learning process of DBN involves two steps: one is unsupervised pre-training with unlabeled samples and the second is supervised fine-tuning with the help of labeled samples. However, this training process may result in two problems: first, multiple hidden units may tend to respond similarly [258] due to co-adaptation [259] and second is linked with the sparsity and selectivity of activations neurons that are some neurons may always be dead or always responding [260]. To mitigate

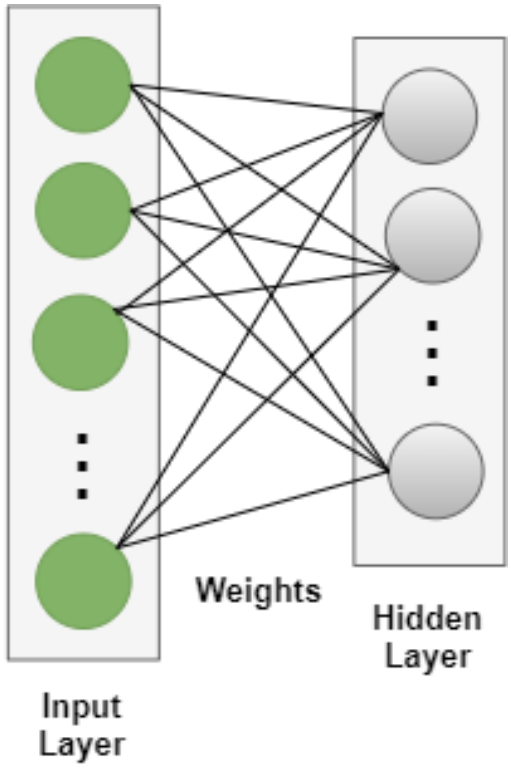


FIGURE 3.8: Basic architecture of RBM

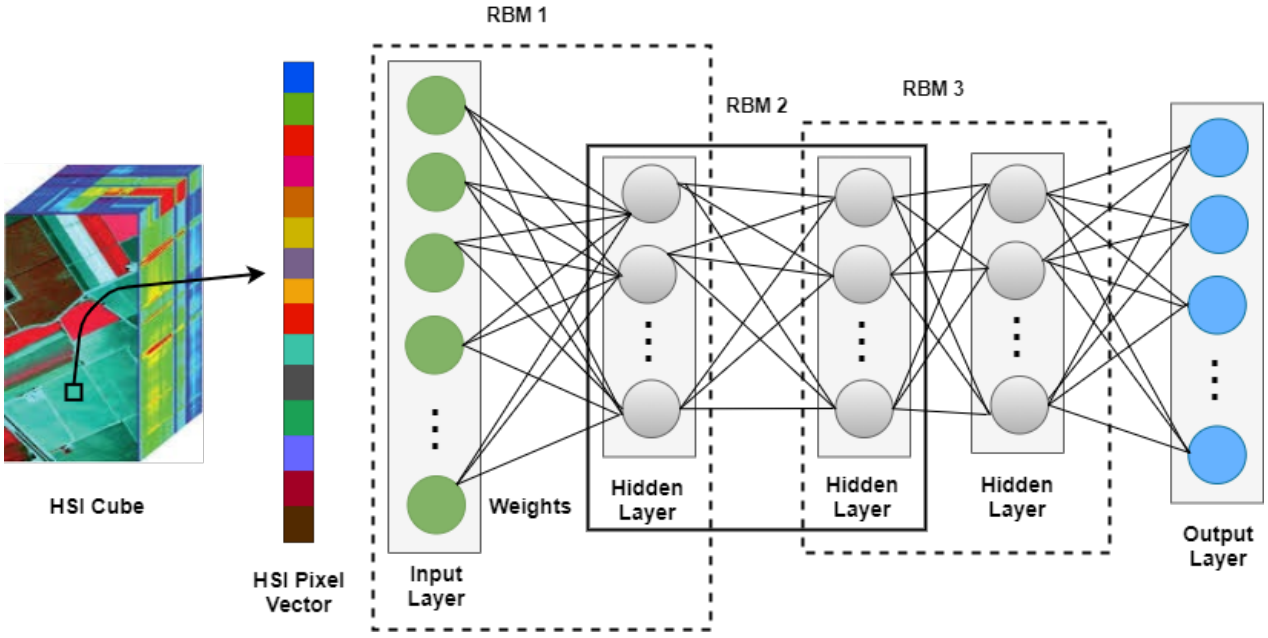


FIGURE 3.9: A three layer DBN architecture

these two problems, [261] introduced a diversified DBN model through regularizing the pre-training and fine-tuning process by imposing a diversity prior to enhancing the DBN's classification accuracy for HSI.

To extract efficient texture features for the HSIC, the work [262] proposed a DBN based texture feature enhancement framework that combines band grouping and sample band selection approach with a guided filter to enhance the texture features, which are then learned by a DBN model and final classification results are obtained by a softmax classifier. The work [263] implemented a parallel layers framework consisting of Gaussian-Bernoulli RBM which extracts high-level, local invariant, and nonlinear features from HSI and a logistic regression layer is used for classification.

To improve the classification accuracy, some works are considered to jointly exploit the spectral and spatial information contained in HSI. For instance, [264] introduced a DBN framework with the logistics regression layer and verified that the joint exploitation of spectral-spatial features leads to improved classification accuracy. Similarly, [265] proposed a spectral-spatial graph-based RBM method for HSIC which constructs the spectral-spatial graph through joint similarity measurement based on spectral and spatial details, then an RBM is trained to extract useful joint spectral-spatial features from HSI, and finally, these features are passed to a DBN and logistic regression layer for classification.

3.8.1 Future directions for DBN-based HSIC

In the preceding section, we have reviewed the latest developments of DBN-based HSIC frameworks. We have observed that relative to other DNNs, very few works have utilized the DBNs for HSIC. Therefore, there is a need to further explore the DBN-based robust techniques that can jointly employ spatial and spectral features for HSIC. In addition, another research direction can be the regularization of the pretraining and fine-tuning processes of DBN to efficiently overcome the issue of dead or potentially over-tolerant (always responding) neurons.

3.9 Recurrent Neural Network (RNN)

The architecture of the Recurrent Neural Network (RNN) (Shown in Figure 3.10) comprises loop connections, where the node activation of the next step depends on the previous step [266]. Therefore, RNNs are capable of learning temporal sequences. RNN models process the spectral information of HSI data as time sequence considering the spectral bands as time steps [267]. There are three basic models of RNN as follows;

1. Vanilla
2. Long-Short-Term Memory (LSTM)
3. Gated Recurrent Unit (GRU)

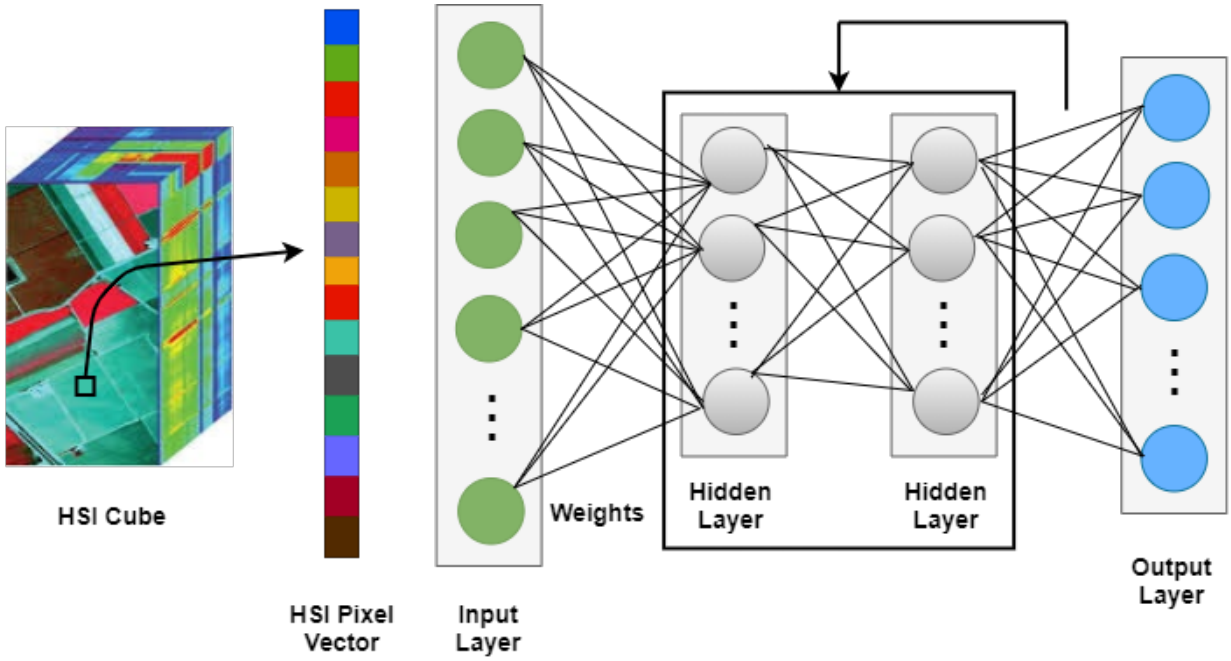


FIGURE 3.10: RNN architecture

Vanilla is the simplest RNN model and leads to information degradation while processing high-dimensional data. LSTM models composing of two states overcome this issue by controlling the information flow through three gates: input, forget, and output gates. It learns the relevant information over time by discarding the extraneous information. However, the gate controlling strategy makes the LSTM a considerably complex approach. GRU variant of LSTM enjoys the simplicity of the Vanilla model and provides high performance similar to LSTM. GRU is a simpler version of LSTM which modifies the input and forget gate as an update (z_t) and reset (r_t) gate and removes the output gate. A comparison of LSTM and GRU's internal architecture is presented in Figure 3.11.

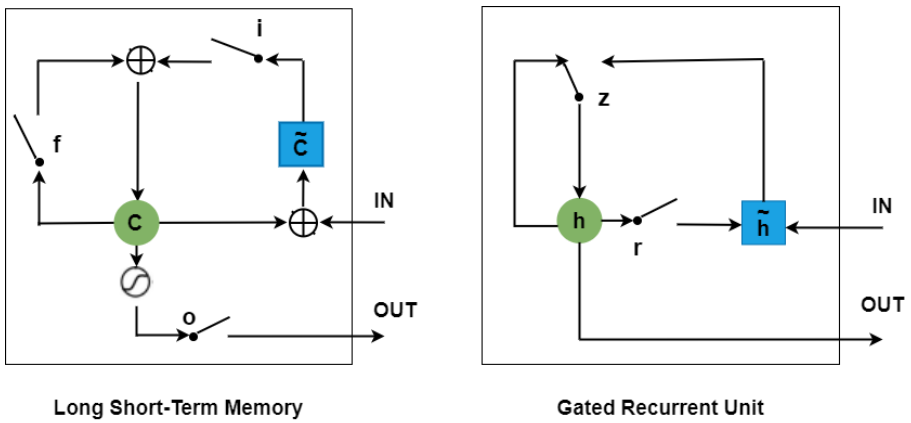


FIGURE 3.11: Internal architecture of LSTM and GRU

For the first time, [268] proposed an RNN based HSIC framework with a novel activation

function (parametric rectified tanh) and GRU, which utilizes the sequential property of HSI to determine the class labels. In [168] a local spatial sequential (LSS) method based RNN framework was introduced which first extracts low-level features from HSI by using Gabor filter and differential morphological profiles [169] and then fuse these features to obtain LSS features from the proposed method, these LSS features are further passed to an RNN model to extract high-level features, while a softmax layer is used for final classification.

Keeping in view the usefulness of spatial information to achieve improved classification accuracies, [269] proposed a spectral-spatial LSTM based network that learns spectral and spatial features of HSI by utilizing two separate LSTM followed softmax layer for classification, while a decision fusion strategy is implemented to get joint spectral-spatial classification results. Similarly, [270] proposed a patch-based RNN with LSTM cells that incorporate multi-temporal and multi-spectral information along with spatial characteristics for land cover classification.

In literature, several works proposed Convolutional Neural Network (CNN) based hybrid RNN architectures (CRNN) for HSIC. For instance, [200] implemented a convolutional RNN in which the first few CONV layers are employed to extract position invariant middle-level features, and then recurrent layers are used to extract spectral-contextual details for HSIC. Similarly, [271] utilized such a model for semi-supervised HSIC by using pseudo labels. The study [272] suggested an HSIC framework in which CNN is used to extract spatial features from HSI, then these features are passed to a GRU-based fusion network that performs feature level and decision level fusion.

Similarly, Luo, et.al., [273] exploited both spectral and spatial information contained in HSI by combining CNN with parallel GRU-based RNN which simplifies the training of GRU and improves performance. Bidirectional Convolutional LSTM (CLSTM) was proposed in [178] to jointly exploit spectral-spatial feature of HSI for classification. In, [274] combined multiscale local spectral-spatial features extracted by $3D$-CNN with a hierarchical RNN which learns the spatial dependencies of local spectral-spatial features at multiple scales. Recurrent $2D$-CNN and recurrent $3D$-CNN for HSIC were proposed in [275] and along with an interesting comparison of these frameworks with their corresponding $2D$ and $3D$-CNN models, which validates the superiority of recurrent CNN. The work [276] integrated CNN with CLSTM in which a $3D$-CNN model is used to capture low-level spectral-spatial features and CLSTM recurrently analyzes this low-level spectral-spatial information. Recently, [277], introduced a cascade RNN for HSIC which consist of two layers of GRU-based RNN, the first layer is used to reduce the redundant spectral bands and the second layer is used to learn the features from HSI, furthermore, few convolutional layers are employed to incorporate the rich spatial information contained in HSI.

3.9.1 Future directions for RNN-based HSIC

In the above section, we have surveyed the recent developments of AEs based techniques for HSIC. Although RNN-based HSIC frameworks have attracted considerable attention to the remote sensing community and achieved great success with respect to classification performance, there are still many aspects that need further investigation. For instance, the construction of sequential input data for RNN. Most of the surveyed methods considered HSI pixel as a sequential point that is the pixel from each spectral band that forms a data sequence. However, This increases the length of RNN's input sequence considerably large which can lead to an overfitting issue. Moreover, processing such large data sequences increases the computational time and the learning process becomes slower. Therefore, the use of parallel processing tools needs to be further investigated to achieve good generalization performance of RNN-based HSIC. In addition, approaches like a grouping of spectral bands to decrease the data sequence length and utilization of the entire spectral signature to better discriminate between various classes can further be explored to construct the sequential input of the RNN model. Another interesting future direction may involve the implementation of RNN-based HSIC frameworks in a real multi-temporal HSI context.

3.10 Strategies for Limited Labeled Samples

Although DNNs have been successfully exploited for the task of HSIC however, they require a considerably large amount of labeled training data. However, as discussed earlier, the collection of labeled HSI is very critical and expensive due to numerous factors that either demand human experts or exploration of real-time scenarios. The limited availability of labeled training data hinders classification performance. To overcome the aforesaid issue, many effective strategies have been proposed in the literature. In this section, we will briefly discuss some of these strategies while focusing on active learning algorithms.

3.10.1 Data Augmentation

To combat the issue of limited training samples, data augmentation is proven to be an effective tool for HSIC. It generates new samples from the original training samples without introducing additional labeling costs. Data augmentation approaches can be categorized into two main strategies as i) data wrapping; ii) oversampling [278]. Data wrapping usually encodes several invariances (translational, size, viewpoint, and/or illumination) by conducting geometric and color-based transformations while preserving the labels, and oversampling-based augmentation methods inflate the training data by generating synthetic

samples based on original data distributions. Oversampling techniques include mixture-based instance generation, feature space augmentations [278], and Generative Adversarial Networks (GANs) [279].

Referring to HSIC literature, several data augmentation-based frameworks have been employed to improve the classification performance by avoiding potential overfitting, which is generally caused by the limited availability of training data. For instance, [280] enhanced the training data by using three data augmentation operations (flip, rotate, and translation), and then this enhanced data is exploited to train CNN for HSIC. The paper [281] presented a comprehensive comparison of various extensively utilized HSI data augmentation techniques and proposed a pixel-block pair-based data augmentation that utilized both spectral and spatial information of HSI to synthesis new instances, to train a CNN model for HSIC. The work [194] compared the classification performance of their diverse region-based CNN framework with and without data augmentation techniques and demonstrated that the data augmentation leads to higher classification accuracies. Similarly, in another comparison [282], data augmentation based CNN exhibited a 10% increase in HSIC accuracy when compared to a PCA based CNN model.

The above-discussed methods utilize offline data augmentation techniques that increase the training data by creating new instances during/before the training process of a model. Recently, a novel data augmentation framework for HSI is proposed in [283] which, rather than inflating the training data, generates the samples at test time, and a DNN trained over original training data along with a voting scheme is used for the final class label. To improve the generalization capability of DNN models, [283] also proposed two fast data augmentation techniques for high-quality data syncretization. A similar PCA-based online data augmentation strategy is proposed in [284] which also synthesis new instances during the inference, instead of training.

3.10.2 Semi-supervised/Unsupervised Methods

Semi-supervised learning (SSL) approaches learn data distribution by jointly exploiting both labeled and unlabeled data. These techniques expand the training data by utilizing unlabeled samples along with labeled ones in order to construct a relationship between feature space and class labels. Several SSL-based HSIC frameworks have been proposed in the literature that can mainly be categorized as follows: i) Co-training, ii) Self-training, iii) Generative adversarial networks (GANs), iv) Graph-based SSL models and v) Semi-supervised SVM. A recent comprehensive survey on these SSL techniques can be found in [285]. Moreover, another in-depth survey of SSL approaches is also presented in [286].

The SSL-based HSIC techniques are briefly summarized in [287], where authors also made a detailed comparison of these methods. The method presented in [271] used pseudo

or cluster-labeled samples to pre-train a CRNN for HSIC and small-sized labeled data is used to fine-tune the network. Similarly, [181] proposed a semi-supervised HSIC framework that exploits PCA and extended morphological attribute profiles to extract pseudo-labeled samples which are fed to a CNN-based deep feature fusion network. The work [288] proposed a dual strategy co-training approach based on spectral and spatial features of HSI. Similarly, [289] separately pre-trained two SAEs, one using spectral and the other using spatial features of HSI, and fine-tuning is achieved via a co-training approach. [290] proposed a region information-based self-training approach to enhance the training data. A graph-based self-training framework was developed in [291] where initial sampling is achieved through subtractive clustering. Recently, [182] improved the HSIC performance by pseudo-labeling the unlabeled samples through a clustering-based self-training mechanism and regulate the self-training by employing spatial constraints.

3.10.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), proposed by [292], are comprised of two neural networks, one is known as a generator and the other is known as a discriminator (Figure 3.12). GANs can learn to replicate the samples by exploiting the data distribution details.

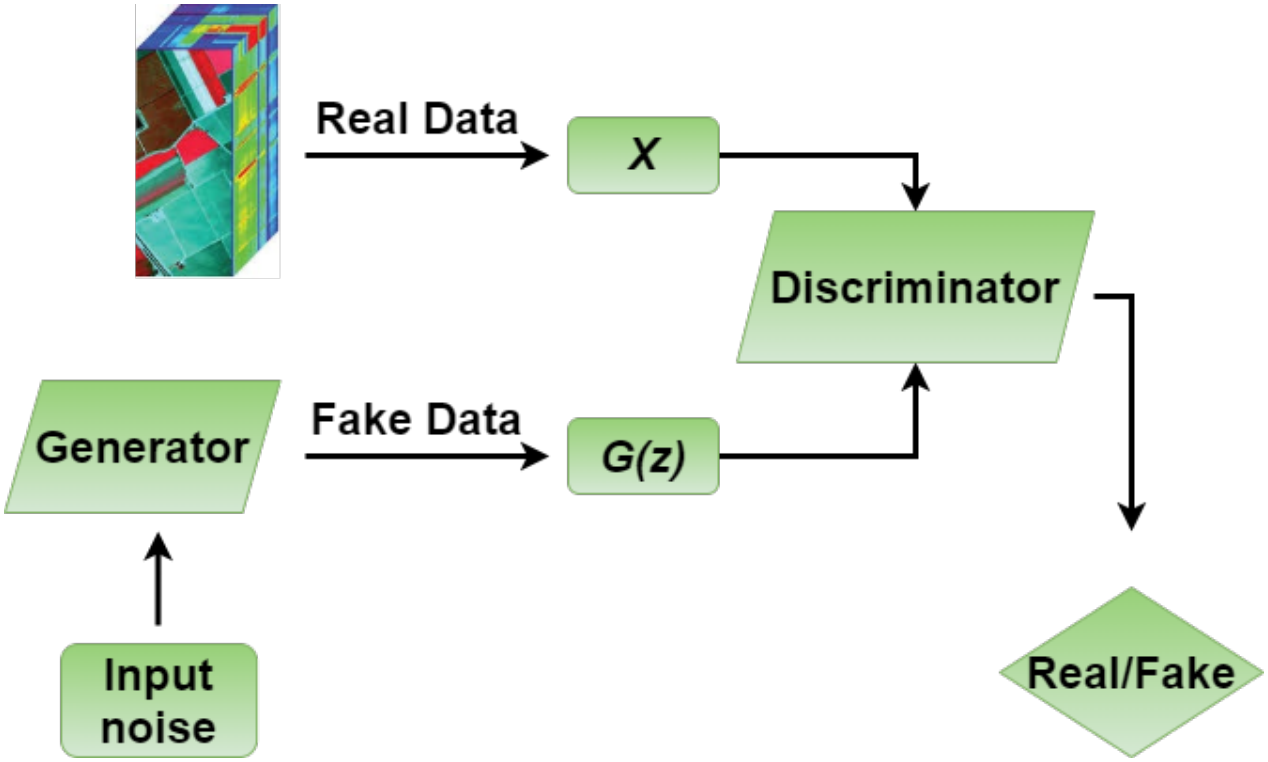


FIGURE 3.12: A general architecture of generative adversarial network (GAN)

The work [293] proposed a spectral feature-based GAN for SSL-based HSIC. Similarly, [294] proposed a GAN-based spectral-spatial HSIC framework. Similarly, [295] developed a CNN-based $1D$-GAN and $3D$-GAN architectures to enhance the classification performance. A 1D customized GAN is used to generate the spectral features [296], which is further used by CNN for feature extraction, and then majority voting is performed HSIC. Very recently, [297] introduced a spatial-spectral multi-class GAN (MSGAN) which utilizes two generators to produce spatial and spectral information with the help of multiple adversarial objectives.

To address the data imbalance problem for HSI classification [298] proposed a new semi-supervised model which combines GAN with conditional random fields (CRFs). Similarly, [299] investigated a Caps-TripleGAN model which effectively generates new samples using a 1D structure triple generative adversarial network (TripleGAN) and classifying the generated HSI samples using the capsule network (CapsNet). In [300] proposed to utilize a 3D CNN-based generator network and a 3D deep residual network-based discriminator network for HSIC.

To learn high-level contextual features combination of both capsule network and convolutional long short-term memory (ConvLSTM) based discriminator model has been proposed in [301] for HSIC. [302] proposed to addresses the scarcity of training examples by utilizing a GAN model where the performance of the discriminator is further improved by an auxiliary classifier to produce more structurally coherent virtual training samples. Besides this, to enhance the model performance [303] proposed a generative adversarial minority oversampling-based technique for addressing the longstanding problem of class-wise data imbalanced imposed by HSIC.

3.10.4 Transfer Learning

Transfer learning enhances the performance of a model by using prior knowledge of a relevant primary task to perform a secondary task. In other words, information extracted from the relevant source domain is transferred to the target domain to learn unseen/unlabeled data. Therefore, transfer learning can be effectively employed in domains with insufficient or no training data. Based on the availability of labeled training instances, transfer learning frameworks can further be categorized as supervised or unsupervised transfer learning. Generally, both source and target domains are assumed to be related but not exactly similar. However, they may follow different distributions as in the case of HSIC where categories of interest are the same but data in two domains may vary due to different acquisition circumstances.

In DNN based HSIC, the model learns features in a hierarchical manner, where lower layers usually extract generic features, when trained on various images. Therefore, the features learned by these layers can be transferred to learn a new classifier for the target dataset. For

instance, [304] pertained to a two-branch spectral-spatial CNN model with an ample amount of training data from other HSIs and then applied the lower layers of the pre-trained model to the target network for the robust classification of target HSI. To learn the target-specific features, higher layers of the target network are randomly initialized and the whole network is fine-tuned by utilizing limited labeled instances of target HSI. Similarly, [305] proposed a suitable method to pre-train and fine-tune a CNN network to utilize it for the classification of new HSIs. The study [306] combined data augmentation and transfer learning approaches to combat the shortage of training data in order to improve HSIC performance.

As discussed before, data in source and target domain may vary in many aspects, for instance, in the case of HSIs, the dimensions of two HSIs may vary due to the acquisition from different sensors. Handling such cross-domain variations and transferring the knowledge between them is known as heterogeneous transfer learning (a detailed survey of such methods can be found in [307]). In HSIC literature, several works have been proposed to bridge the gap for transferring the knowledge between two HSIs, with varying dimensions and/or distributions.

For example, [308] proposed an effective heterogeneous transfer learning-based HSIC framework that works well with both homogeneous and heterogeneous HSIs, and [309] used an iterative re-weighting mechanism-based heterogeneous transfer learning for HSIC. Similarly, a recent work [310] proposed a band selection-based transfer learning approach to pre-train a CNN, which retains the same number of dimensions for various HSIs. Furthermore, [311] proposed an unsupervised transfer learning technique to classify completely unknown target HSI and [312] demonstrate that the networks trained on natural images can enhance the performance of transfer learning for remote sensing data classification as compared to the networks trained from scratch using smaller HSI data.

3.10.5 Active Learning

Active Learning (AL) iteratively enhances the predictive performance of a classifier by actively increasing the size of training data, for each training iteration, by utilizing an unlabeled pool of samples. In each iteration, AL enhances the training dataset by actively selecting the most valuable instances from the pool of unlabeled data and an oracle (Human or machine-based) assigns the true class labels to these instances. Finally, these useful instances are added to the existing training dataset and the classifier is retrained on this new training dataset. The process continues until a stopping criterion, that maybe the size of the training dataset, the number of iterations, or the desired accuracy score, is achieved. A general framework of AL is illustrated in Figure 3.13.

The selection of the most useful/effective samples is made in such a way that the samples should be informative and representative of the overall input distribution in order to

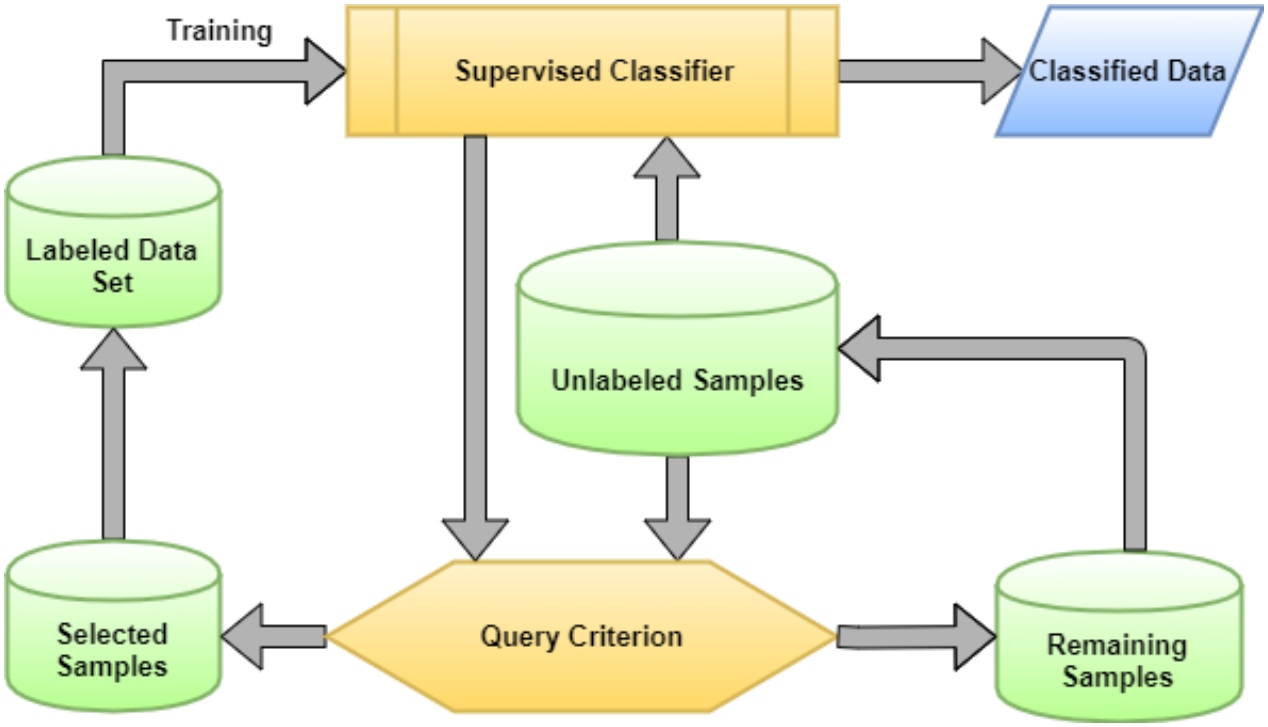


FIGURE 3.13: A general overview of Active Learning

improve accuracy. Based on the criteria of adding new instances to the training set, AL frameworks can be designated as either stream-based or pool-based. In stream-based selection, one instance at a time is drawn from an actual set of unlabeled samples and the model decides whether to label it or not based on its usefulness. While in pool-based strategy, samples are queried from a pool/subset of unlabeled data based on ranking scores computed from various measures to evaluate the sample's usefulness. The work [313] found that streamed-based selection gives poorer learning rates as compared to pool-based selection as the former tends to query extra instances. In pool-based selection, it is important to incorporate diversity in the pool of samples, in order to avoid redundancy within the pool of samples. Generally, the following three aspects are focused on while selecting/querying the most valuable samples: heterogeneity behavior, model's performance, and representativeness of samples. A brief introduction of these sampling approaches is given below:

Heterogeneity-based selection

These approaches select the samples that are more heterogeneous to the already seen instances with respect to model diversity, classification uncertainty, and contention between a committee of various classifiers. Uncertainty sampling, expected model change, and query-by-committee are examples of heterogeneity-based models.

- **Uncertainty Sampling**: In this approach, the classifier iteratively tries to query the label of those samples for which it is most uncertain while predicting the label. The selection of new instances is based on ranking scores against a specified threshold and the instances with scores closest to that threshold are queried for labels. One simple example of such a scheme could be implementing the probabilistic classifier on a sample in a scenario of binary classification and query its label if the predicted class probability is close to 0.5.

- **Query-by-Committee**: Such heterogeneity-based approaches perform the sampling process based on the dissimilarities in the predictions of various classifiers trained on the same set of labeled samples. A committee of various classifiers trained on the same set of training data is used to predict the class labels of unlabeled samples and the samples for which classifiers differ more are selected for querying labels. The committee of different classifiers can either be built by using ensemble learning algorithms like Bagging and Boosting [314] or by changing the model parameters [315]. Generally, a less number of diverse classifiers is adequate for constructing a committee [314, 316].

- **Expected Model Change**: Such a heterogeneity-based approach chooses the instances which result in a significant change from the current model in terms of the gradient of the objective function. Such techniques attempt to query the label for those instances that are considerably different from the current model. These sampling techniques only fit the models which follow gradient-based training procedures/optimization.

Performance-based Selection

Such methods consider the effect of adding queried samples to the model performance. They try to optimize the performance of the model by reducing variance and error. There are two types of performance-based sampling:

- **Expected Error Reduction**: This approach is interrelated to uncertainty sampling in such a way that uncertainty measures maximize the label uncertainty of the sample to be queried for the label while expected error reduction reduces the label uncertainty of the queried sample. Referring to the already discussed example of the binary classification problem, the expected error reduction approach would choose the samples with a probability far away from 0.5 in order to reduce the error rate. Such techniques are also known as the greatest certainty models [315].

- **Expected Variance Reduction**: Reducing the variance of the model is guaranteed to reduce future generalization error [317]. Therefore, expected variance reduction techniques attempt to indirectly reduce the generalization error by minimizing the model

variance. Such approaches query the instances that result in the lowest model variance. The Fisher information ratio is a well-known variance minimization framework.

Representativeness-based selection

Heterogeneity-based models are prone to include outlier and controversial samples but performance-based approaches implicitly avoid such samples by estimating future errors. Representative sampling tends to query such instances that are representative of the overall input distribution, hence, avoid outliers and unrepresentative samples. These approaches weigh the dense input region to a higher degree while the querying process. Density-weighted techniques like information density are examples of representativeness sampling approaches that consider the representativeness of samples along with heterogeneity behavior, and are also known as hybrid models [315].

Recently, AL has been intensively utilized in HSIC. [318] proposed a feature-driven AL framework to define a well-constructed feature space for HSIC. [319] proposed a Random Forest-based semi-supervised AL method that exploits spectral-spatial features to define a query function to select the most informative samples as target candidates for the training set.

Spatial information has been intensively exploited in many AL-based HSIC. For instance, [320] presented an AL framework that splice together the spectral and spatial features of superpixels. Similarly, [321] considered the neighborhood and superpixel information to enhance the uncertainty of queried samples. In recent work, [322] exploited the attribute profiles to incorporate spatial information in an AL-based HSIC framework.

Batch-mode AL frameworks have been widely employed to accelerate the learning process. Such approaches select a batch of samples, in each iteration, to be queried for a label. Therefore, the diversity of the samples is extremely critical in batch mode AL techniques in order to avoid redundancy. A multi-criteria batch-mode AL method proposed by [323] defines a novel query function based on diversity, uncertainty, and cluster assumption measures. These criteria are defined by exploiting the properties of KNN, SVM, and K-means clustering respectively, and finally, genetic algorithms are used to choose the batch of most effective samples. Similarly, [324] proposed a regularized multi-metric batch-mode AL framework for HSIC that exploits various features of HSI.

A multiview AL (MVAL) framework was proposed in [325] that analyzes the object from various views and measure the informativeness of the sample through multiview Intensity-based query criteria. Similarly, [326] also exploited the concept of multiview learning using the Fisher Discriminant Ratio to generate multiple views. In another work, [327] proposed a novel adaptive MVAL framework for HSIC which jointly exploits the spatial and spectral features in each view. Recently, [328] proposed an MVAL technique that utilizes pixel-level,

subpixel-level, and superpixel-level details to generate multiple views for the purpose of HSIC. Moreover, the proposed method exploits joint posterior probability estimation and dissimilarities among multiple views to query the representative samples.

In the HSIC literature, several works have combined the AL and DNN. For instance, [329] joined autoencoder with AL technique and [330] proposed a DBN-based AL framework for HSIC. Similarly, [331] coupled Bayesian CNN with AL paradigm for the purpose of spectral-spatial HSIC. Recently, [332] proposed a CNN-based AL framework to better exploit the unlabeled samples for HSIC.

Many works integrated AL with transfer learning for the purpose of HSIC. For example, [333] proposed an AL-based transfer learning framework that extracts the salient samples and exploits high-level features to correlate the source and target domain data. Another work, [334] proposed a stacked sparse AE-based active transfer learning technique that jointly utilizes both spectral and spatial features for HSIC. Another work [335] combined domain adaptation and AL methods based on multiple kernels for HSIC.

AL-based HSIC offers some sophisticated frameworks to enhance the generalization capabilities of models. For instance, [6] proposed a fuzziness-based AL method to improve the generalization performance of discriminative and generative classifiers. The method computes the fuzziness-based distance of each instance and estimated class boundary, and the instances having greater fuzziness values and smaller distances from class boundaries are selected to be the candidates for the training set. Recently, [336] proposed a non-randomized spectral-spatial AL framework for multiclass HSIC that combines the spatial prior Fuzziness approach with Multinomial Logistic Regression via a Splitting and Augmented Lagrangian classifier. The authors also made a comprehensive comparison of the proposed framework with state-of-the-art sample selection methods along with diverse classifiers.

3.11 Concluding Remarks

The rich information contained in HSI data is a captivating factor that constitutes the utilization of HSI technology in real-world applications. Moreover, advances in machine learning methods strengthen the deployment potentials of such technologies. This chapter surveyed recent developments of Hyperspectral Image Classification (HSIC) using state of the art Deep Neural Networks (for instance, Auto-encoder (AE), Deep Belief Network (DBN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Transfer Learning (TL), Few-shot Learning (FSL), Active/Self Learning (AL/SL), and Data Augmentation (DA)) in a variety of learning schemes (specifically, supervised, semi-supervised and unsupervised learning). In addition, this chapter also analyzed the strategies to overcome the

challenges of limited availability of training data like Data Augmentation, Few-shot Learning (FSL), Transfer Learning, and Active Learning, etc.

Although the current HSIC techniques reflect a rapid and remarkable sophistication of the task, further developments are still required to improve the generalization capabilities. The main issue of deep neural network-based HSIC is the lack of labeled data. HSI data is infamous due to the limited availability of labeled data and deep neural networks demand a sufficiently large amount of labeled training data. Section 3.10 discussed some widely used strategies to combat the aforesaid issue but significant improvements are still needed to efficiently utilize limited available training data. One direction to solve this problem could be to explore the integration of various learning strategies discussed in section 3.10 to cash in the joint benefits. One more way is to exploit a few-shot or K-shot learning approaches that can accurately predict the class labels with only a few labeled samples. Moreover, there is a need to focus on the joint exploitation of spectral-spatial features of HSI to complement classification accuracies achieved from the aforementioned HSIC frameworks. Another future potential of HSIC is computationally efficient architectures. Therefore, the issue of the high computational complexity of deep neural networks is of paramount importance and it is crucial to implement parallel HSIC architectures to speed up the processing of deep neural networks to meet the computational stipulation of time-critical HSI applications. In this direction, high-performance computing platforms and specialized hardware modules like graphical processing units (GPUs) and field-programmable gate arrays (FPGAs) can be used to implement the parallel HSIC frameworks. Hence, to assimilate aforesaid aspects in the development of a new HSIC framework is to appropriately utilize the limited training samples while considering joint spectral-spatial features of HSI and maintaining the low computational burden.

Chapter 4

A Fast and Compact 3D CNN

This chapter proposes a 3D CNN model that utilizes both spatial-spectral feature maps to improve the performance of HSIC. For this purpose, the HSI cube is first divided into small overlapping 3D patches, which are processed to generate 3D feature maps using a 3D kernel function over multiple contiguous bands of the spectral information in a computationally efficient way. In brief, an end-to-end trained model requires fewer parameters to significantly reduce the convergence time while providing better accuracy than existing models.

4.1 Motivation

The simplest way to improve the HSIC performance is to design a classifier that should incorporate both spectral and spatial information. Spatial information is considered as additional discriminatory information associated with the size, shape, and structure of the object which, if provided correctly, brings more competitive results. Spatial-spectral classifiers can generally be classified into two groups. The first category explores spatial and spectral information separately. The spatial information is extracted in advance using entropy [337], morphological operations [338, 339], low rank representation [340], attribute profiles [341] and fuzziness [342]. Later this information is combined with spectral information to perform pixel-level classification.

The second category fuses spatial-spectral information to get joint features [343], for instance, 3D wavelet, scattering wavelet and Gabor filter [344, 345] are generated at different frequencies and scales to extract the joint spatial-spectral features for classification. HSI is in $3D$ cubes thus the former category results in several 3D features, i.e., spatial-spectra feature cubes comprising key information, thus preserving joint spatial-spectral correlations that enable the extracted features to produce better results. However, the classical feature extraction methods are based on shallow learning and handcrafted features which largely depend on domain knowledge [346]. Accordingly, Deep models have been used to automatically learn low to high-level features from raw HSI data which have attained incredible success for HSIC.

The last few years witnessed an intensive improvement in CNN for HSIC where the spatial features are tailored by a 2*D* CNN model [347–349]. These spatial features are usually extracted separately that, to some extent, void the reason to jointly exploit the spatial-spectral information for HISC. A hybrid spectral CNN for HSIC has been proposed in [346], in which the authors proposed a 3*D* CNN followed by a spatial 2*D* CNN model. The 3*D* convolutional layers facilitate the spectral-spatial feature representation whereas 2*D* convolutional layers are used to learn abstract level information. The hybrid model produces better results as compared to the conventional 3D models but still lacks at extracting the abstract level spatial information. Recently, Paoletti et. al., [350, 351] proposed two deep pyramidal residual networks for HSI feature extraction and classification. The former work only considered spectral information for HSIC whereas the latter considered both the spectral-spatial capsule network for feature learning and classification. Chen et. al., proposed a 3*D* CNN model for feature extraction and classification [352]. Similarly, Zhong et. al., [353] proposed a spatial-spectral residual network for HSIC in which the residual blocks used identity mapping to connect 3D convolutional layers. Mou et. al., [354] proposed an unsupervised HSIC to further explore the residual CNNs. The review of the literature revealed several shortcomings, including but not limited to;

1. Though CNNs have become a promising method for HSIC, their memory requirement and high computational complexity make it challenging to accelerate their performance. This work investigates their application to HSIC targetting high accuracy but under controlled computational cost, in terms of the time, it takes for them to converge. To achieve this, our work progressively modifies a baseline model while preserving its accuracy and reducing its time complexity.

2. Preserving channel relationship information is a challenging problem. CNN models are usually trained on reshaped spectral bands or use single band (gray-scale) information (containing different properties), failing to extract the “fine structural/spatial information of HSI". Furthermore, the high inter-class similarity, intra-class variability, overlapping, and nested regions of HSI data make classification a challenging problem. To overcome the said issue, the proposed architecture first divides the HSI cube into small overlapping 3D patches. These patches are processed to generate 3D feature maps using 3*D* kernel function over multiple contiguous bands to preserve the joint spatial and spectral information for the feature learning process which exploits important discriminatory information for HSIC.

3. As a preprocessing step, incremental Principle Component Analysis (iPCA) is employed to reduce the redundancy among the bands to process the few important wavelengths out of the entire HSI cube. Finally, to increase the number of spatial-spectral

feature maps, four 3*D* convolutional layers are deployed to ensure that the model can discriminate the spatial information within different spectral bands without any loss.

In a nutshell, our end-to-end trained model requires fewer parameters, which significantly reduces the time it takes for the model to converge without compromising its accuracy, which is better than the existing models, as evident by our experimental results.

4.2 Proposed Methodology

Let us assume a HSI can be expressed as $X = [x_1, x_2, x_3, ..., x_L]^T \in R^{L\times(N\times M)}$ consisting of $N \times M$ samples associated with C classes per band with total L bands, in which each sample is represented as (x_i, y_j), where y_j is the class label of x_i sample. The HSI pixels exhibit high inter-class similarity, high intra-class variability, overlapping, and nested regions. To overcome the aforesaid issues, iPCA is applied to eliminate the redundant bands. iPCA reduces the number of bands (L to B, where $B \ll L$) while maintaining the spatial dimensions as shown in Figure 4.1.

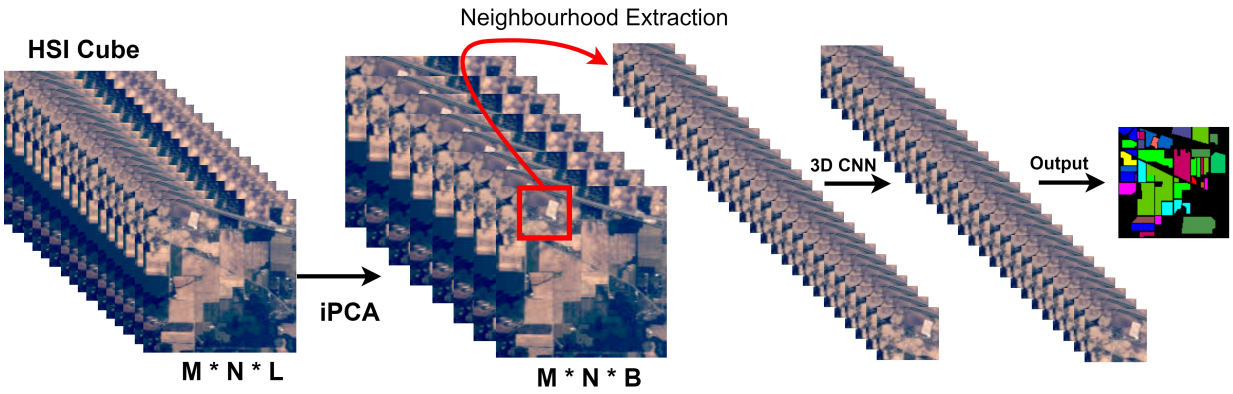


FIGURE 4.1: Proposed 3D CNN Model for HSIC. 3D CNN Model details, i.e., the number of 3D Convectional and fully connected layers, can be found in Table 4.1.

In order to pass the HSI cube to the model, it must be divided into a small overlapping 3*D* spatial patches on which the ground labels are formed based on the central pixel, as shown in Figure 4.2. The process creates neighboring patches $P \in R^{S\times S\times B}$ centered at the spatial location (a, b) covering $(S \times S)$ spatial windows [346]. The total of n patches given by $(M\tilde{}S + 1) \times (N\tilde{}S + 1)$. Thus, these patches cover the width from $\frac{a-(S-1)}{2}$ to $\frac{a+(S-1)}{2}$ and height from $\frac{b-(S-1)}{2}$ to $\frac{b+(S-1)}{2}$.

The input patches are first convolved with a 3D kernel function which computes the sum of the dot product between kernel function and input patch [346, 355]. Later these learned features are processed through an activation function that introduces the nonlinearity. The

TABLE 4.1: Layer based Summary of our Proposed 3D CNN Model architecture shown in Figure 4.2 with Window Size set as 11×11.

Layer	Output Shape	# of Parameters
Input Layer	(11, 11, 20, 1)	0
Conv3D_1 (Conv3D)	(9, 9, 14, 8)	512
Conv3D_2 (Conv3D)	(7, 7, 10, 16)	5776
Conv3D_3 (Conv3D)	(5, 5, 8, 32)	13856
Conv3D_4 (Conv3D)	(3, 3, 6, 64)	55360
Flatten_1 (Flatten)	(3456)	0
Dense_1 (Dense)	(256)	884992
Dropout_1 (Dropout)	(256)	0
Dense_2 (Dense)	(128)	32896
Dropout_2 (Dropout)	(128)	0
Dense_3 (Dense)	(# of Classes)	774
In total, **994,166** trainable parameters are required		

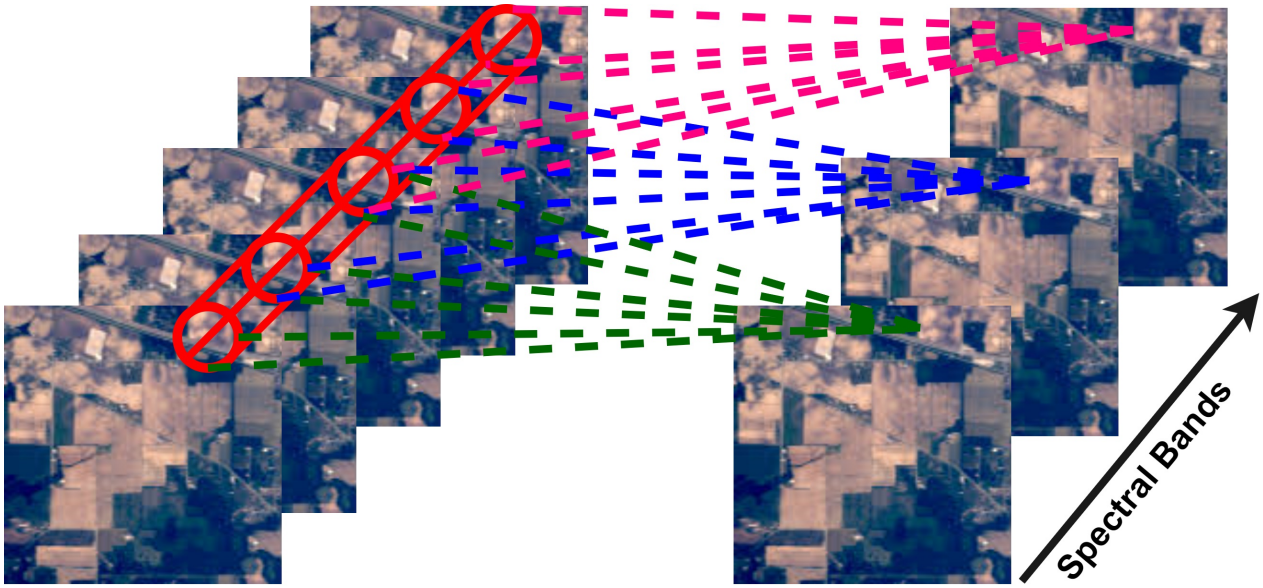


FIGURE 4.2: 3D Convolution Operation

activation values at spatial position (x, y, z) in the i^{th} layer and j^{th} feature map is denoted as $v_{i,j}^{x,y,z}$. Thus, the final model can be created as follows:

$$v_{i,j}^{x,y,z} = \mathcal{F}\left(\sum_{\tau=1}^{d_{i-1}} \sum_{\lambda=-\nu}^{\nu} \sum_{\rho=-\gamma}^{\gamma} \sum_{\phi=-\delta}^{\delta} w_{i,j,\tau}^{\lambda,\rho,\phi} \times v_{(i-1),\tau}^{(x+\lambda),(y+\rho),(z+\phi)} + b_{i,j}\right) \tag{4.1}$$

where $\mathcal{F}$ is an activation function, d_{i-1} be the number of 3D feature maps at $(i-1)^{th}$ layer and $w_{i,j}$ be the depth of the kernel, $b_{i,j}$ is the bias, $2\delta + 1$, $2\gamma + 1$ and $2\nu + 1$ be the height, width and depth of the kernel.

In short, the proposed 3D CNN convolutional kernels are as follows: $3D\_conv\_layer1 =$

$8 \times 3 \times 3 \times 7 \times 1$ where $K_1^1 = 3, K_2^1 = 3$ and $K_3^1 = 7$. $3D\_conv\_layer2 = 16 \times 3 \times 3 \times 5 \times 8$ where $K_1^2 = 3, K_2^2 = 3$ and $K_3^2 = 5$. $3D\_conv\_layer3 = 32 \times 3 \times 3 \times 3 \times 16$ where $K_1^3 = 3, K_2^3 = 3$ and $K_3^3 = 3$ and finally $3D\_conv\_layer4 = 64 \times 3 \times 3 \times 3 \times 16$ where $K_1^3 = 3, K_2^3 = 3$ and $K_3^3 = 3$. To increase the number of spatial-spectral feature maps, four 3D convolutional layers are deployed before the flatten layer to make sure the model is able to discriminate the spatial information within different spectral bands without any loss. Further details regarding the proposed model can be found in Table 4.1. The total number of parameters (i.e., tune-able weights) of our proposed 3D CNN model is 994, 166. The weights are initially randomized and optimized using Adam optimizer back-propagation with a soft-max loss function. The weights are updated using a mini-batch of size 256 with 50 epochs without batch normalization and augmentation.

Chapter 5

Regularized Hybrid CNN Feature Hierarchy

This chapter proposed an idea to enhance the generalization performance of CNN for HSIC using soft labels that are a weighted average of the hard labels and uniform distribution over ground labels. The proposed method helps to prevent CNN from becoming over-confident. We empirically show that in improving generalization performance, regularization also improves model calibration which significantly improves beam-search.

5.1 Motivation

CNN models can be categorized into two groups, i.e., single and two-stream, more information regarding single or two-stream methods can be found in [279]. This chapter explicitly investigates a single-stream method similar to the works proposed in [225, 352, 356–363]. Irrespective of the single or two-stream methods, all DL frameworks are sensitive to the loss which needs to be minimized [364]. Several classical works showed that the gradient descent to minimize cross-entropy performs better in terms of classification and has fast convergence, however, to some extent, leads to the overfitting [365]. Several regularization techniques such as dropout [366], L1, L2 [367], etc., have been used to overcome the overfitting issues together with several other exotic objectives performed exceptionally well than the standard cross-entropy [368]. Recently, a work [369] proposed a regularization technique that improves the accuracy significantly by computing cross-entropy with a weighted mixture of targets with uniform distribution instead of hard-coded targets.

Since then, regularization has been known to improve the classification performance [370]. However, the original idea was used to improve the classification performance of only the inception model on ImageNet data [369]. Despite this, various image classification models have used regularization [371, 372]. Though the regularization technique is a widely

used trick to improve the classification performance and to speed up the convergence process, however, it has not been much explored for HSIC, and above all, it has not been much explored regarding when and why regularization should work.

Considering the aforesaid issues, this chapter proposed a novel idea to enhance the generalization performance of CNN for HSIC using soft labels that are a weighted average of the hard labels and uniform distribution over target labels. The proposed method helps to prevent CNN from becoming over-confident.

5.2 Proposed Methodology

Let us assume that the Hyperspectral data can be represented as $R^{(M\times N)\times B^*} = [r_1, r_2, r_3, \ldots, r_S]^T$, where B^* be the total number of bands. $(M \times N)$ are the samples per band belonging to Y classes and $r_i = [r_{1,i},\ r_{2,i},\ r_{3,i},\ \ldots, r_{B^*,i}]^T$ is the i^{th} sample in the Hyperspectral Data. Suppose $(r_i, y_i) \in (\mathcal{R}^{M\times N\times B^*}, \mathcal{R}^Y)$, where y_i is the class label of the i^{th} sample. For HSI classification with Y candidate labels, for example, lets assume $(r_i, y_i) \in (\mathcal{R}^{M\times N\times B^*}, \mathcal{R}^Y)$, where y_i is the class label of the r_i sample belonging to the training set and the ground truth distribution p over labels $p(y|r_i)$ and $\sum_{y=1}^{Y} p(y|r_i) = 1$. One can have a model with parameters θ that predicts the predicted label distribution as $q_\theta(y|r_i)$ and of course $\sum_{y=1}^{Y} q_\theta(y|r_i) = 1$. Thus the cross entropy in this particular case would be:

$$H_i(p, q_\theta) = \sum_{y=1}^{Y} p(y|r_i)\ log\ q_\theta(y|r_i) \tag{5.1}$$

If one has $M \times N$ instance in the training set, then the loss function would be:

$$L = H_i(p, q_\theta) \tag{5.2}$$

$$L = -\sum_{i=1}^{M\times N}\sum_{j=1}^{Y} p(y|r_i)\ log\ 1_\theta(y|r_i) \tag{5.3}$$

However, in nature the $p(y|r_i)$ would be a one-hot-encoded vector [373, 374], which can be defined as:

$$p(y|r_i) = \begin{cases} 1 & if y = y_i \\ 0 & otherwise \end{cases} \tag{5.4}$$

Based on the above objective, one can reduce the loss function as:

$$L = \sum_{i=1}^{M \times N} H_i(p, q_\theta) \tag{5.5}$$

$$L = -\sum_{i=1}^{M \times N} \sum_{y=1}^{Y} p(y|r_i)\ log\ q_\theta(y|r_i) \tag{5.6}$$

$$L = -\sum_{i=1}^{M \times N} p(y_i|r_i)\ log\ q_\theta(y_i|r_i) \tag{5.7}$$

$$L = -\sum_{i=1}^{M \times N} log\ q_\theta(y_i|r_i) \tag{5.8}$$

Minimizing the above loss function is equivalent to do maximum likelihood estimation over the training set. However, during optimization, it is possible to minimize L to almost 0, if and only if, all the instances in the dataset do not have conflicting labels [1] This is due to $q_\theta(y_i|r_i)$ is computed from soft-max as:

$$_\theta(y_i|r_i) = \frac{exp(z_{y_i})}{\sum_{j=1}^{Y} exp(z_j)} \tag{5.9}$$

where z_i be the logit for candidate class i. The consequence of using one-hot-encoding is $exp(z_{y_i})$ will be extremely large and $exp(z_j)$ where $j \neq y_i$ will be extremely small. Given a non-conflicting dataset, the ultimate model will classify every training instance correctly with the confidence of almost 1. This is certainly a signature of overfitting, and the overfitted model does not generalize well. Thus, this work introduces a regularization technique $\mu(y|r_i)$ (noise distribution) irrespective to traditional techniques proposed in literature [375–377] for deep models [378]. Thus the new HSI ground truths (r_i, y_i) would be:

$$p'(y|r_i) = (1 - \varepsilon)p(y|r_i) + \varepsilon\mu(y|r_i) \tag{5.10}$$

$$f(x) = \begin{cases} 1 - \varepsilon + \varepsilon\mu(y|r_i) & if\ y = y_i \\ \varepsilon\mu(y|x_i) & otherwise \end{cases} \tag{5.11}$$

where $\varepsilon \in [0, 1]$ is a weight factor, and note that $\sum_{y=1}^{Y} p'(y|r_i) = 1$. These new ground truths has been used in loss function instead of one hot-encoding [379].

$$L' = -\sum_{i=1}^{M \times N} \sum_{y=1}^{Y} p'(y|r_i) \log q_\theta(y|r_i) \tag{5.12}$$

[1] Conflicting labels means, there are two examples with the extract same features from the dataset, but their ground truth labels are different.

$$L' = -\sum_{i=1}^{M\times N}\sum_{y=1}^{Y}\left[(1-\varepsilon)p(y|r_i) + \varepsilon\mu(y|r_i)\right]\log q_\theta(y|r_i) \tag{5.13}$$

$$\begin{aligned} L' = \sum_{i=1}^{M\times N}\Bigg\{(1-\varepsilon)\Big[-\sum_{y=1}^{Y} p(y|r_i)\log q_\theta(y|r_i)\Big] + \\ \varepsilon\Big[-\sum_{y=1}^{Y}\mu(y|x_i)\log q_\theta(y|r_i)\Big]\Bigg\} \end{aligned} \tag{5.14}$$

$$L' = \sum_{i=1}^{M\times N}\Big[(1-\varepsilon)H_i(p,q_\theta) + \varepsilon H_i(u,q_\theta)\Big] \tag{5.15}$$

where L' be the loss function and p' be the estimated probabilities. One can observe that each ground truth, the loss contribution is a mixture of entropy between predicted distribution ($H_i(p,q^\theta)$) and the one hot-encoding, and the entropy between the predicted distribution ($H_i(\mu,q^\theta)$) and the noise distribution. While training, $H_i(p,q^\theta) = 0$ if the model learns to predict the distribution confidently, however, $H_i(\mu,q^\theta)$ will increase dramatically. To overcome this phenomenon, we used a regularizer $H_i(\mu,q^\theta)$ to prevent the model from predicting too confidently. In practice, $\mu(y|r)$ is a uniform distribution that does not depend on Hyperspectral data. That is to say $\mu(y|r) = \frac{1}{Y}$.

In a nutshell, the details of 3D/2D convolutional layers and kernels are as follows: 3D conv layer 1 = $8 \times 5 \times 5 \times 7 \times 1$ i.e. $K_1^1 = 5, K_2^1 = 5$ and $K_3^1 = 7$. $3D\_conv\_layer\_2 = 16 \times 5 \times 5 \times 5 \times 8$ i.e. $K_1^2 = 5, K_2^2 = 5$. $K_3^2 = 5$. $3D\_conv\_layer\_3 = 32 \times 3 \times 3 \times 3 \times 16$ i.e. $K_1^3 = 3, K_2^3 = 3$ and $K_3^3 = 3$. $3D\_conv\_layer\_4 = 64 \times 3 \times 3 \times 3 \times 32$ i.e. $K_1^3 = 3, K_2^3 = 3$ and $K_3^3 = 3$. $2D\_conv\_layer\_5 = 128 \times 3 \times 3 \times 64$ i.e. $K_1^2 = 3$ and $K_2^2 = 3$. Three 3D convolutional layers are employed to increase the number of spectral-spatial feature maps and one 2D convolutional layer is used to discriminate the spatial features within different spectral bands while preserving the spectral information. Initially, the weights are randomized and then optimized using back-propagation with the Adam optimizer by using the loss function presented in equation 5.15. Further details regarding the Hybrid CNN architecture in terms of types of layers, dimensions of output feature maps and number of trainable parameters can be found in [380] also shown in Table 5.1.

TABLE 5.1: Regularized Hybrid CNN Model

Layers	Output Shape	Parameters
Input	$15 \times 15 \times 15 \times 1$	0
Conv-3D	$11 \times 11 \times 9 \times 8$	1408
Conv-3D	$7 \times 7 \times 5 \times 16$	16016
Conv-3D	$5 \times 5 \times 3 \times 32$	13856
Conv-3D	$3 \times 3 \times 1 \times 64$	55360
Reshape	$3 \times 3 \times 64$	0
Conv-2D	$1 \times 1 \times 128$	73856
Flatten	128	0
Dense	256	33024
Droupout (0.4%)	256	0
Dense	128	32896
Droupout (0.4%)	128	0
Output	# of classes	2064
Total # of trainable parameters = 228,480		

Chapter 6

Artifacts of Dimension Reduction on Hybrid CNN

3D CNNs are computationally expensive and 2D CNN alone cannot efficiently extract discriminating spectral-spatial features. Therefore, to overcome these challenges, this chapter presents a compact hybrid CNN model which overcomes the aforementioned challenges by distributing spatial-spectral feature extraction across 3D and 2D layers. An intensive preprocessing (several dimensional reduction methods) has been carried out to improve the classification results and to reduce the computational time.

6.1 Motivation

The classification performance of DL can be enhanced by considering two aspects: dimensionality reduction and utilization of spatial information. Dimensionality reduction is an important preprocessing step to reduce the spectral redundancy that subsequently results in less processing time and enhanced classification accuracy. Dimensionality reduction methods transform the high-dimensional data into a low-dimensional space whilst preserving the potential spectral information [381]. Whereas, the spatial information can improve the discriminative power of the classifier by considering the neighboring pixels' information.

Thus, processing spatial-spectral information together would be considered as a viable approach for HSIC. The spectral-spatial classification approaches can be categorized into two groups. The first group excavates for both spectral and spatial features individually. Spatial information is educed in advanced using various methods for instance, morphological operations [382], attribute profiles [383] and entropy [384, 385] etc., and then spliced together with spectral information for pixels-wise classification. The other group coalesces spectral and spatial information to acquire joint features like Gabor filter and wavelets [386, 387] are constructed at various scales to simultaneously extract spectral-spatial features for classification.

However, the handcrafted features and usually extract shallow features and rely on a high level of domain knowledge for feature designing [388]. To overcome these limitations, end-to-end models (i.e., feature extraction/learning and classification) such as Convolutional Neural Network (CNN) have been widely used to automatically learn the low and high-level representation of HSI in a hierarchical manner [363, 389]. CNN-based HSIC improves the generalization capabilities and predictive performance [264, 390]. CNN-based HSIC architectures have attracted prevalent attention due to substantial performance gain, in which 2D CNNs are used for spatial feature extraction and to extract both spectral and spatial features of HSI, many variants of 3D CNN have been proposed [214, 221, 225, 391, 392]. However, 3D CNN is computationally complex and 2D CNN alone cannot efficiently extract discriminating spectral features.

To overcome the aforesaid challenges, a hybrid CNN Feature Hierarchy is used that splices together 3D components with 2D components. The aim is to synergies the competencies of 3D and 2D CNN to obtain important discriminating spectral-spatial features of HSI for classification. Prior to the feature extraction, we incorporated dimensionality reduction as a preprocessing step and comprehensively investigated the impact of various state-of-the-art dimensionality reduction approaches on the performance of the hybrid CNN model. Also, we evaluated the impact of different input window sizes on the performance of the end model.

6.2 Proposed Methodology

Let us assume that the HSI cube can be represented as $X = [x_1, x_2, x_3, \ldots, x_S]^T \in \mathcal{R}^{S\times(C\times D)}$, where S denotes total number of spectral bands and $(C \times D)$ are the samples per band belonging to Y classes and $x_i = [x_{1,i},\ x_{2,i},\ x_{3,i},\ \ldots, x_{S,i}]^T$ is the i^{th} sample in the HSI cube. Suppose $(x_i, y_i) \in (\mathcal{R}^{S\times(C\times D)}, \mathcal{R}^Y)$, where y_i is the class label of the i^{th} sample [393].

Due to the spectral mixing effect which induces high intra-class variability and inter-class similarity in X, it becomes difficult to classify various materials based on their spectral signatures. To combat the aforesaid issue, we used dimensionality reduction as a preprocessing step to eliminate the spectral redundancy which reduces the number of spectral bands ($S \rightarrow B$, where $B \ll S$) while keeping the spatial dimensions unimpaired. Subsequently, this also results in a reduced computational overhead owing to a lower-dimensional feature subspace. In this work, we evaluated the effectiveness of the following dimensionality reduction methods for the hybrid CNN model.

6.2.1 Principle Component Analysis (PCA)

Principal Component Analysis (PCA) is based on the orthogonal transformation that computes linearly uncorrelated variables, known as principal components (PCs), from possibly correlated data. The first PC is the projection in the direction of the highest variance and it gradually decreases as we move towards the last PC. The transformation of the original image to PCs is the Eigen decomposition of the covariance matrix of mean-centered HSI data [389]. Eigen decomposition of covariance matrix i.e. finding the eigenvalues along with their corresponding eigenvectors is $E = ADA^T$ where $A = [a_1, a_2, a_3 \ldots, a_X]$ is a transformation matrix and $D = diag[\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_X]$ is a diagonal matrix of eigenvalues of covariance matrix. The linear HSI transformation is defined as $H = AX$.

6.2.2 Incremental PCA (iPCA)

Generally, PCA is performed in batch mode i.e., all the data is simultaneously available to compute the projection matrix. In order to find the updated PCs after incorporating the new data into the existing training set, PCA needs to be retrained with complete data. To combat this limitation, an incremental PCA (iPCA) approach is used that can be categorized as either covariance-based iPCA or covariance-free iPCA method. Covariance-based methods are further divided into two approaches. In the first approach, the covariance matrix is computed using existing training data and then the matrix is updated whenever new data samples are added. In the second approach, a reduced covariance matrix is computed using previous PCs and the new training data. Covariance-free iPCA methods update the PCs without computing the covariance methods, however, such methods usually face convergence problems in case of high dimensional data [117, 394].

6.2.3 Sparse PCA (SPCA)

The conventional PCA has a limitation that the PCs are the linear combinations of all input features/predictors or in other words, all the components are nonzero and direct interpretation becomes difficult. Therefore, to improve interpretability, it is desirable to use sparsity-promoting regularizers. In this regard, sparse PCA (SPCA) has emerged as an effective technique that finds the linear combinations of a few input features i.e. only a few active (nonzero) coefficients. SPCA works well in the scenarios where input features are redundant, that is, they do not contribute to identifying the underlying rational model structure [395].

6.2.4 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a mathematical technique that decomposes a matrix into three different matrices. It is knowns as truncated SVD when used for dimensionality reduction [396]. This matrix decomposition is represented as $X = PSQ^T$ where P and Q are orthogonal matrices of left and right singular vectors and S is a diagonal matrix having singular values as its diagonal entries. An SVD reduced X is obtained by taking into account the contribution of only the first k eigenimages, computed as $X_{SVD} = \sum_{i=1}^{k} P_i S_i Q_i^T$.

6.2.5 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is one of the popular approaches among other dimensionality reduction methods, that extracts statistically independent components (ICs) through a linear or non-linear transformation that minimizes the mutual information between ICs or maximizes the likelihood or non-Gaussianity of ICs. It transforms the HSI into a lower-dimensional feature space by comparing the average absolute weight coefficients for each spectral band of HSI and retain only those independent bands which contain maximum information [397]. Given n-dimensional data X, the main task of ICA is to find the linear transformation W such that $H = WX$ where H has statistically independent components.

6.2.6 Hybrid CNN

In order to pass the reduced HSI data cube to the Hybrid model, it is divided into multiple small overlapping 3D patches, and the class labels of these patches are decided based on the label of central pixel. The 3D neighboring patches $P \in \mathcal{R}^{(W \times W) \times B}$ are formed that are centered at spatial position (a, b), covering the $W \times W$ windows. The total number of 3D spatial patches created from X is given by $(M - W + 1) \times (N - W + 1)$. These 3D patches centered at location (a, b) represented by $P_{(a,b)}$ covers the width from $\frac{a-(W-1)}{2}$ to $\frac{a+(W-1)}{2}$ and height from $\frac{b\tilde{}(W-1)}{2}$ to $\frac{b+(W-1)}{2}$ and all B spectral bands obtained after dimensionality reduction method.

In 2D CNN, the input data is convolved with the 2D kernel function that computes the sum of the dot product between the input and the 2D kernel function. The kernel is stridden over the input in order to cover the whole spatial dimension. Then these convolved features are processed through an activation function that helps to learn non-linear features of data by introducing non-linearity in the model. In case of 2D convolution, the activation value of j^{th} feature map at spatial location (x, y) in the i^{th} layer, denoted by $v_{i,j}^{x,y}$, can be formulated as follows:

$$v_{i,j}^{x,y} = \mathcal{F}(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho} \times v_{i-1,\tau}^{x+\sigma,y+\rho}) \tag{6.1}$$

where $\mathcal{F}$ is the activation function, d_{l-1} is the number of feature map at $(l-1)^{th}$ layer and the depth of kernel $w_{i,j}$ for j^{th} feature map at i^{th} layer, $b_{i,j}$ denotes the bias parameter for j^{th} feature map at i^{th} layer, $2\gamma + 1$ and $2\sigma + 1$ be the width and height of the kernel.

Whereas, the 3D convolutional process first computes the sum of the dot product between input patches and 3D kernel function i.e. the 3D input patches are convolved with 3D kernel function [363, 389]. Later these feature maps are passed through an activation function to induce non-linearity. The Hybrid model generates the features maps of the 3D convolutional layer by using the 3D kernel function over B spectral bands, extracted after dimensionality reduction, in the input layer. In 3D convolutional process, the activation value at spatial location (x, y, z) at the i^{th} layer and j^{th} feature map can be formulated as [389, 398]:

$$v_{i,j}^{x,y} = \mathcal{F}(b_{i,j} + \sum_{\tau=1}^{d_{l-1}} \sum_{\lambda=-v}^{v} \sum_{\rho=-\gamma}^{\gamma} \sum_{\sigma=-\delta}^{\delta} w_{i,j,\tau}^{\sigma,\rho,\lambda} \times v_{i-1,\tau}^{x+\sigma,y+\rho,z+\lambda}) \tag{6.2}$$

where all the parameters are the same as defined in Equation 6.1 except $2v + 1$ which is the depth of 3D kernel along a spectral dimension.

The details of 3D convolutional kernels are as follows: $3D\_conv\_layer\_1 = 8 \times 3 \times 3 \times 7 \times 1$ i.e. $K_1^1 = 3, K_2^1 = 3$ and $K_3^1 = 7$. $3D\_conv\_layer\_2 = 16 \times 3 \times 3 \times 5 \times 8$ i.e. $K_1^2 = 3, K_2^2 = 3$ and $K_3^2 = 5$. $3D\_conv\_layer\_3 = 32 \times 3 \times 3 \times 3 \times 16$ i.e. $K_1^3 = 3, K_2^3 = 3$ and $K_3^3 = 3$. The details of 2D convolutional kernel are: $2D\_conv\_layer\_1 = 64 \times 3 \times 3 \times 96$ i.e. $K_1^4 = 3$ and $K_2^4 = 3$. Three 3D convolutional layers are employed to increase the number of spectral-spatial feature maps and one 2D convolutional layer is used to discriminate the spatial features within different spectral bands while preserving the spectral information.

Further details regarding the Hybrid CNN architecture in terms of types of layers, dimensions of output feature maps, number of trainable parameters, and layer-wise hierarchy are shown in Figure 6.1. The base model was proposed by Roy et.al [389] and explored in this work with several dimensional reduction methods to further validate which dimensionality reduction method works better with different settings. Initially, the weights are randomized and then optimized using back-propagation with the Adam optimizer by using the Softmax loss function. The network is trained for 50 epochs using a mini-batch size of 256 and without any batch normalization and data augmentation.

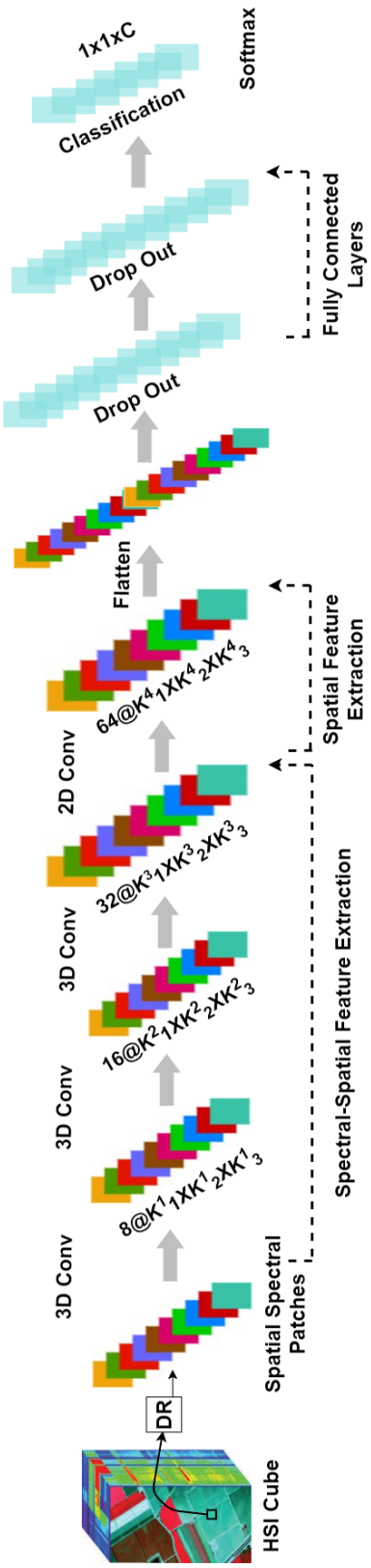


FIGURE 6.1: Hybrid CNN Feature Hierarchy framework for HSIC.

Chapter 7

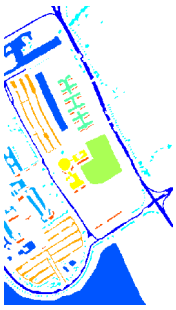
Spectral Angle Mapper for Spatial-Spectral Classification

Acquisition of labeled data for supervised HSIC is expensive in terms of both time and costs. Moreover, manual selection and labeling are often subjective and tend to induce redundancy into the classifier. Active learning (AL) can be a suitable approach for HSIC as it integrates data acquisition to the classifier design by ranking the unlabeled data to provide advice for the next query that has the highest training utility. However, multiclass AL techniques tend to include redundant samples into the classifier to some extent. This chapter addresses such a problem by introducing an AL pipeline that preserves the most representative and spatially heterogeneous samples. The adopted strategy for sample selection utilizes fuzziness to assess the mapping between actual output and the approximated a-posteriori probabilities, computed by a marginal probability distribution based on discriminative random fields. The samples selected in each iteration are then provided to the spectral angle mapper-based objective function to reduce the inter-class redundancy.

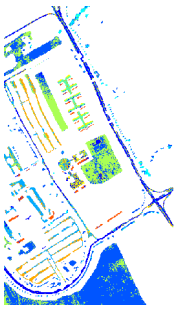
7.1 Motivation

Supervised classification methods are widely adopted in the analysis of HSI datasets. These methods include, for example, multinomial logistic regression [9], random forests [10], ensemble learning [11], deep learning [12], support vector machine (SVM) [13], and k-nearest neighbors (KNN) [6]. However, supervised classifiers often underperform due to the Hughes phenomenon [5], also known as the issue of dimensionality, which occurs whenever the number of available labeled training samples is considerably lower than the number of spectral bands required by the classifier [6]. Figure 7.1 (Loss of accuracy in terms of ground maps) and Table 7.1 (Loss of accuracy in terms of overall and kappa (κ) for a different number of labeled training samples i.e., 1% and 10% respectively) illustrates the loss in the predictive performance of such classification methods for a particular ground image (Pavia University) when using two different sample size.

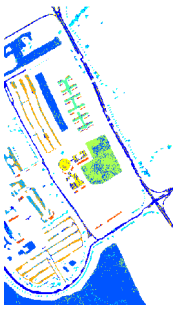
(A) Pavia Image



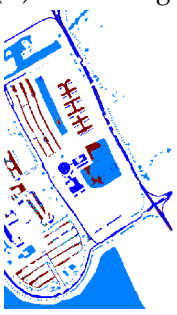
(B) True Ground Truths



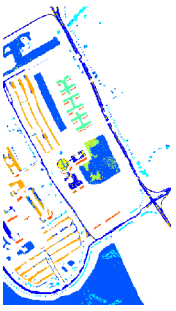
(C) SVM (1% Training)



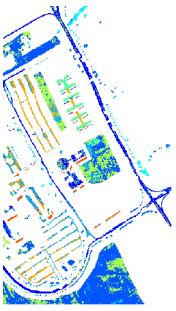
(D) SVM (10% Training)



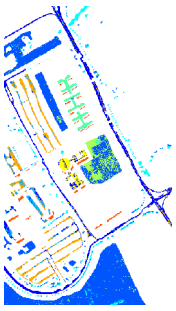
(E) KNN (1% Training)



(F) KNN (10% Training)



(G) LB (1% Training)



(H) LB (10% Training)

FIGURE 7.1: (**a**) Pavia University image; (**b**) True ground truths differentiate nine different classes; (**c**) SVM trained with 1% randomly selected training samples; (**d**) SVM trained with 10% randomly selected training samples; (**e**) KNN trained with 1% randomly selected training samples; (**f**) KNN trained with 10% randomly selected training samples; (**g**) Logistic boost (LB) trained with 1% randomly selected training samples; (**h**) LB trained with 10% randomly selected training samples.

TABLE 7.1: Classification accuracies in-terms of overall and kappa (κ) obtained by three different classifiers with two different number of randomly selected labeled training samples, i.e., 1% and 10% respectively. All these classifiers are trained and tested using 5-fold cross validation, and from results one can conclude that SVM produce better results when trained with 10% randomly selected training samples. However, the obtained results are not good enough to identify the ground materials accurately, therefore, further investigations are required.

Classifiers	**1% Training Samples**		**10% Training Samples**	
	kappa (κ)	**Overall**	**kappa (κ)**	**Overall**
SVM	0.5853%	0.6818%	0.7804%	0.8365%
KNN	0.4897%	0.6488%	0.6791%	0.7691%
LB	0.5216%	0.6488%	0.7531%	0.8365%

The limited availability of labeled training data in the HSI-domain is one of the motivations for the utilization of semi-supervised learning [399]. Examples of such methods include kernel techniques [400] such as SVM, Tri-training [401] algorithms which generate three classifiers from the original labeled samples, then these classifiers are refined using unlabeled samples in the tri-training process, and Graph-based learning [2, 402]. A major limitation of such approaches, however, is the low predictive performance when utilizing a small number of training samples within high dimensionality, as commonly observed in

HSI classification [29, 92] as shown in Figure 7.1 and Table 7.1.

Active learning (AL) is a class of semi-supervised learning methods that tackles the limitations as mentioned earlier [403, 404]. The main component of an AL method is the iterative utilization of the training model to acquire new training samples to be entered into the training set for the next iteration [405]. AL methods can be pool-based or stream-based depending on how they enter new data to the training set, and employ measures like uncertainty, representativeness, inconsistency, variance, and error to rank and select new samples [406]. Despite the gained success, there are still particular characteristics that can cause AL to present inflated false discovery rate and low statistical power [6]. These characteristics include (i) sample selection bias; (ii) high correlation among the bands; and (iii) non-stationary behavior of unlabeled samples.

Alternatives of sample selection method utilized in AL and corresponding references to the literature are shown in Table 7.2. Table 7.2 classifies the references in the literature according to the information utilized by the sample selection methods, being either spectral (consider only the wavelength of the pixel) or spectral-spatial (pixel location in addition to the wavelength). The latter class is particularly relevant in the HSI domain as the acquisition of training samples depends on a large degree on the spatial distribution of the queried samples. However, only a few studies have integrated spatial constraints into AL methods [407, 408]. Table 7.2 provides a unified summary of existing sample selection methods and the information they use along with references to their respective papers.

TABLE 7.2: Different sample selection methods used in Active Learning frameworks for hyperspectral image classification in the recent years.

Sample Selection Methods	**References**	
	Spectral	**Spectral–Spatial**
Random selection	[6, 403, 409]	[410]
Mutual information	[411, 412]	[413]
Breaking ties	[414]	[415, 416]
Modified breaking ties	[415]	[415, 416]
Uncertain sampling	[417, 418]	[419–421]
Fisher information ratio	[422]	[423]
Fuzziness information	[6]	——-
Query by committee	[316]	——-

Tuia et al. [409] presented a detailed survey on AL methods addressing HSI analysis and contrasted non-probabilistic methods, which assume that all query classes are known before the initialization, to probabilistic approaches that allow the discovery of new classes. The latter class was also pointed as more suitable for cases when the prior assumption is no longer fulfilled [424]. In addition to probabilistic and non-probabilistic AL methods,

large margin heuristics have been utilized as the base learner to combine the benefits of HSI analysis and AL [409, 425]. A particular approach for selecting samples that have achieved remarkable results for several applications is query by committee (QBC) [426]. Contrarily from previous methods, QBC selects samples based on the maximum disagreement of an ensemble of classifiers. Overall, these sampling methods suffer from high computational complexity due to the iterative training of the classifier for each sample [409].

Pool-based AL, also known as batch-mode AL, addresses the high computational complexity observed in the aforementioned methods by concomitantly considering the uncertainty (spectral information) and diversity (spatial information) of the selected samples [417]. Seminal work was presented by Munoz-Mari et al. [427], which highlighted the benefits of integrating spatial-contextual information to AL even when the distribution of queried samples in the spatial space is ignored. This method was later expanded to include the position of selected samples in the feature space [413]. One of the outcomes from such a transformation is the point-wise dispersed distribution in the spatial domain, which incurs the risk of revisiting the same geographical location several times, especially in the HSI-domain [413].

A considerable amount of research has been conducted on AL in recent years, often analyzing only spectral properties, whilst ignoring spatial information that plays a vital role in HSIC as shown in [413, 428]. Spatial and spectral HSIC can achieve higher performance than its pixel-wise counterpart as it utilizes not only information of spectral signature but also from spatial domain [428]. Thus, the combination of spatial and spectral information for AL represents a novel and promising contribution yet to be explored in the HSI domain.

Thus, this chapter introduces a customized AL pipeline for HSI to reduce sample selection bias whilst maintaining the data stability in the spatial domain. The presented pipeline distinguishes from standard AL methods in three relevant aspects. First, instead of simply using the uncertainty of samples to select new samples, it utilizes the fuzziness measure associated with the confidence of the training model in classifying those samples correctly. Second, it couples samples' fuzziness with their diversity to select new training samples which simultaneously minimizes the error among the training samples while maximizing the spectral angle between the selected sample and the existing training samples. In this chapter, instead of measuring angle-based distances among all new samples and all existing training samples, a reference sample is selected from within the training set against which the diversity of the new samples is measured. This achieves the same goal while reducing the computational overhead as the size of the training set is always much smaller than the validation set which is the source of new samples. Thirdly, the proposed Fuzziness and Spectral Angle Mapper (FSAM) method keep the pool of new samples balanced, giving equal representation to all classes, which is achieved via softening the thresholds at run time.

7.2 Proposed Methodology

This chapter addresses the small sample problem when classifying high dimensional HSI data by defining an AL scheme selecting a pool of diverse samples by taking into account two main criteria. The first step is to compute the fuzziness of samples, which is associated with the confidence of the trained model in properly classifying the unseen samples. The second is the diversity of the samples, thus reducing the redundancy among the selected samples. The combination of two criteria results in the selection of a pool of potentially most informative and diverse samples in each iteration.

Although there have been lots of different sampling methods (few mentioned in Table 7.2), uncertainty remains one of the most popular methods that can be used to select the informative samples [6, 429]. Usually, the most uncertain samples have similar posterior probabilities for the two most possible classes [429]. Thus, a probabilistic model could be directly used to evaluate the uncertainty of unlabeled sample [429].

However, assessing the uncertainty of a sample is not as straightforward when one is using non-probabilistic (NP) classifiers because their output does not exist in the form of posteriori probabilities [406, 429]. The output of such classifiers can be manipulated to obtain an approximation of posteriori probability functions for the classes being trained [406].

Suppose $X = [x_1, x_2, x_3, \ldots, x_L]^T \in \mathcal{R}^{L\times(M\times N)}$ is an HSI cube which is composed of L spectral bands and $(M \times N)$ samples per band belonging to C classes where $x_i = [x_{1,i},\ x_{2,i}\ ,\ x_{3,i},\ \ldots, x_{L,i}]^T$ is the i^{th} sample in the cube. Let us assume $(x_i, y_i) \in \mathcal{R}^{(M\times N)} \times \mathcal{R}^C$, where y_i is the class label of the i^{th} sample. Let us further assume that n finite (limited) number of labeled training samples are selected from X to create the training set $D_T = \{(x_i, y_i)\}_{i=1}^{n}$. The rest of the samples form the validation set $D_V = \{(x_i, y_i)\}_{i=1}^{m}$. Please note that $n \ll m$, and $(D_T \cap D_V) = \emptyset$.

An NP classifier trained on D_T when tested on D_V would produce an output matrix μ of $m \times C$ dimensions containing NP outputs of the classifier. Let μ_{ij} be the j^{th} output (membership for the j^{th} class) for i^{th} sample. There are several methods proposed in the literature to transform such NP outputs into the posteriori probabilities [406, 429]. Such methods are computationally complex in two folds. First these methods need to compute the Bayesian decision for each samples x_i choosing the category y_j having the largest discriminant function $f_j(x_i) = p(\mu_j|x_i)$. Secondly, these methods assume that the training outputs are restricted as $\{0, 1\}$. However, these methods also consider manipulating each Bayes rule using Jacobin's derivation over the limit theorem on infinite number samples to approximate the posterior probabilities in the least-squares sense, i.e., $f(x_i, \mu_j) = p(\mu_j, x_i)$ [429].

In order to overcome the above-mentioned difficulties, in this chapter, we used marginal probability distribution [413] which is obtained from the D_V information in the HSI data,

serves as an engine in which our AL pipeline can exploit both the spatial and spectral information in the data. The posteriori class probabilities are modeled with the discriminative random field [413, 430] in which the association potential is linked with a discriminative, generative, ensemble, and signal hidden layer feed-forward neural network-based classifiers. Thus, the posteriori probabilities are computed as similar to the work [413]. From these posteriori probabilities we obtained the membership matrix which should satisfy the following properties [6]:

$$\sum_{j=1}^{C} \mu_{ij} = 1 \ \ and \ \ 0 < \sum_{i=1}^{N} \mu_{ij} < 1 \tag{7.1}$$

In Equation (7.1), $\mu_{ij} \in [0, 1]$ and $\mu_{ij} = \mu_j(x_i)$ is a function that represents the membership of i^{th} sample x_i to the j^{th} class [6]. For the true class, the posteriori probability would be approximated as close to 1, whereas, if the output is small (wrong class), the probability would be approximated as close to 0. However, AL methods do not require accurate probabilities, but only need a ranking of the samples according to their posteriori probabilities which would help to estimate the fuzziness [429] and the output of the sample.

The fuzziness $(E(\mu))$ upon $(M \times N)$ samples for C classes from the membership matrix (μ_{ij}) can then be defined as expressed in Equation (7.2) which must satisfy the properties defined in [431, 432].

$$E(\mu) = \frac{-1}{N \times C} \sum_{i=1}^{N} \sum_{j=1}^{C} \Big[\mu_{ij} log(\mu_{ij}) + (1 - \mu_{ij}) log(1 - \mu_{ij}) \Big] \tag{7.2}$$

Then, we first associate $E(\mu)$, predicted class labels, and actual class labels with D_V and then sort the D_V in descending order based on the fuzziness values. We then heuristically select the $\hat{m}$ number of misclassified samples which have higher fuzziness, where $\hat{m} \ll m$. The proposed strategy keeps the pool of $\hat{m}$ new samples balanced, giving equal representation to all classes, which is achieved via softening the thresholds at run time.

Next, the spectral angular mapper (SAM) (More information about spectral angle mapper (SAM) function can be found in the following papers [433–435]) function is used to discriminate the samples within the same class to minimize the redundancy among the pool of $\hat{m}$ selected samples. SAM is an automated method for directly comparing sample spectra to known spectra. It treats both spectra as vectors and calculates the spectral angle between them. It is insensitive to illumination since it uses only the vector direction and not the vector length [436]. The output of SAM is an image showing the best match of each pixel at each spatial location [437]. This approach is typically used as a first cut for determining the mineralogy and works well in the area of homogeneous regions.

In this chapter, SAM takes the arc-cosine-based dot product between the test spectrum which have higher fuzziness $D_V^H = \{(x_{ij}^l), y_j)\}_{i=1}^{\hat{m}}$ to a reference (training samples) spectrum $D_T = \{(x_{pj}^l, y_j)\}_{p=1}^{n}$, where $j \in \{1, 2, 3, \ldots, C\}$ and $l \in \{1, 2, 3, \ldots, L\}$ where L is the total number of bands in HSI dataset, with the following objective functions:

$$\angle(\alpha_j) = \min_{p \in n} \left(cos^{-1} \frac{\sum_{l=1}^{L} x_{pj}^l \cdot x_{pj}^l}{\sqrt{\sum_{l=1}^{L} (x_{pj}^l)^2} \sqrt{\sum_{l=1}^{L} (x_{pj}^l)^2}} \right) \tag{7.3}$$

Equation (7.3) aims to compute the spectral difference among all the training samples for C classes, respectively. We then select one reference spectrum from each class which minimizes the angular distance among others within the same class, i.e., the sample which is more similar to others in the given class. This process will return the number of reference spectrum up to the number of classes in HSI. We then pick one reference spectrum from $\angle(\alpha_j)$ to compare with all the selected test spectrum for the same class and account the angular distance among them in $\angle(\beta_{ij})$ as shown in Equation (7.4).

$$\angle(\beta_{ij}) = \sum_{j=1}^{C} \sum_{i=1}^{\hat{m}} \left(cos^{-1} \frac{\sum_{l=1}^{L} \left(\angle(\alpha_j)\right) \cdot x_{ij}^l}{\sqrt{\sum_{l=1}^{L} (\angle(\alpha_j))^2} \sqrt{\sum_{l=1}^{L} (x_{ij}^l)^2}} \right) \tag{7.4}$$

$$Ind_{(D_V^H)} = \underset{i \in (D_V^H)|\mathcal{X}}{\operatorname{argmax}} \left(\phi\big(\angle(\beta_{ij})\big) \right) \tag{7.5}$$

where $Ind_{(D_V^H)}$ denotes the induces of samples which have higher fuzziness, $D_V^H|\mathcal{X}$ represents the index of samples of D_V^H that are not contained in $\mathcal{X}$, ϕ provides the trade-off between diversity, and $\mathcal{X}$ denotes the index of the unlabeled sample that will be included in the pool. Please note that here we used a soft threshing scheme to balance the number of classes in both training and selected samples. The proposed pipeline systematically selects the $(h/(number\ of\ classes))$ higher fuzziness samples from D_V^H for each class, if one or more classes are missed in the pool of selected samples. This process is repeated until the cardinality of $\mathcal{X}$ is equal to h, i.e., $|\mathcal{X}| = h$, where h is the size of pool. This technique guarantees that the selected samples in $\mathcal{X}$ are diverse regarding their angles to all the others in $(\angle(\beta_{ij}))$. Since the initial size of $\mathcal{X}$ is zero, thus, the first sample included in $\mathcal{X}$ is always the higher fuzziness sample from $E(\mu)$.

There are several advantages of using fuzziness information carried out through SAM as query function: (i) easy to implement; (ii) robust in mapping the spectral similarity for reference to higher fuzziness test spectrum only; (iii) powerful because it represents the influence of shading effects to accentuate the selected test reflectance characteristics [435].

On the other hand, the main drawback of SAM is spectral mixture problems, i.e., SAM assumes that the reference spectrum chosen to classify the HSI represents the pure spectrum

which is not the case in our problem. Such problems occur when the HSI is in low or medium spatial resolution. Furthermore, as we know, the surface of the earth is heterogeneous and complex in many ways, thus, containing many mixed samples. The spectral confusion in samples can lead to overestimation or underestimation errors for spectral signatures.

This is not the case of the proposed solution, since we iteratively select the reference spectrum from each class using Equation (7.3) as a pure spectrum and comparing this with the selected test spectrum respectively using Equation (7.4) with the help of whiting parameter to minimize the redundancy among the selected samples. The complete workflow of our proposed pipeline is described in Algorithm 1 and Figure 7.2.

Algorithm 1: Pseudo-code of our Proposed FSAM Algorithm.

Data: D_T, D_V training and test set, respectively.
Initialization: $\mathcal{X}$; $\hat{m}$ = number of samples to select; ϕ; h;
while $|D_T| \leq Threshold$ **do**
  Train and Test ELM, SVM, kNN and EL;
  $\mu_{ij} \leftarrow$ Compute the membership matrix;
  $E(\mu) \leftarrow$ Compute the fuzziness;
  $D_V^H \leftarrow$ Associate the fuzziness, actual and predicted class and spatial information with D_V and sort in descending order;
  Select $\hat{m}$ misclassified samples from D_V^H ;
  $\angle(\alpha_j) \leftarrow$ Compute the spectral angle between training samples;
  while $|\mathcal{X}| \leq h$ **do**
    $\angle(\beta_{ij}) \leftarrow$ Compute the spectral angle between reference spectrum and $\hat{m}$ selected samples;
    $Ind_{D_V^H} \leftarrow$ Pick the index of most diverse samples.;
    $\mathcal{X} \leftarrow$ index of selected samples;
  end
  Pick $\mathcal{X}$ samples from D_V, add them to D_T, and remove from D_V;
end

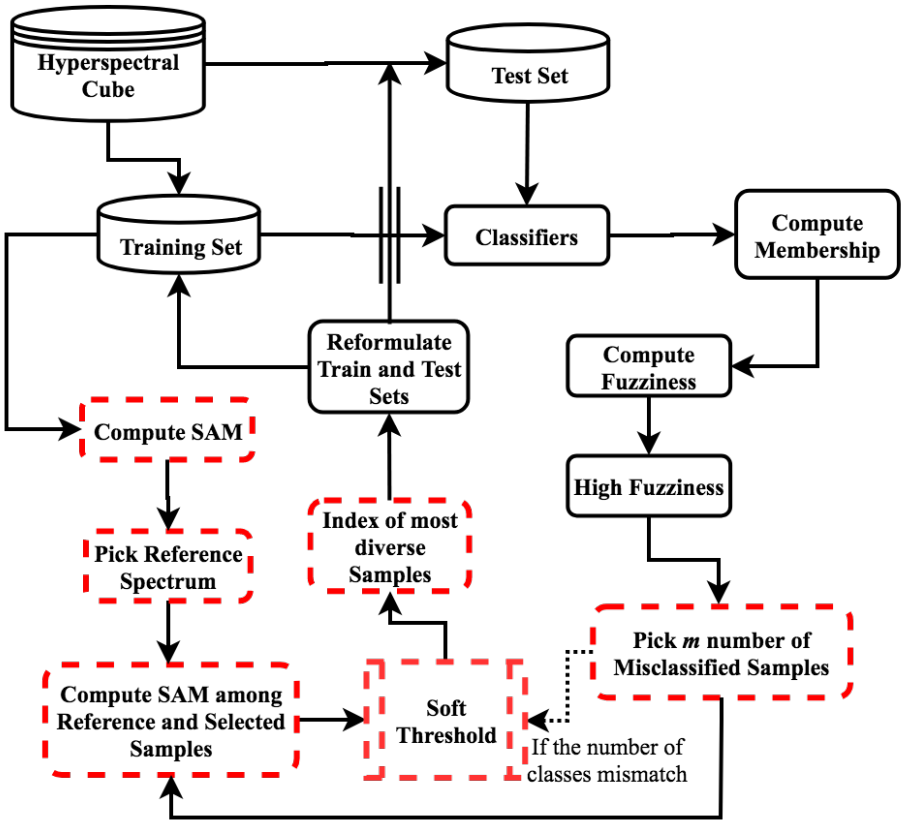


FIGURE 7.2: Proposed FSAM-AL Pipeline in which red labeled boxes represent where we contribute.

Chapter 8

Experimental Evaluation

8.1 Experimental Datasets

This section discusses several benchmark real HSI datasets used in this dissertation. All these datasets are being acquired by different airborne and satellite sensors. These sensors include Airborne Visible Infrared Imaging Spectrometer (AVIRIS), Reflective Optics System Imaging Spectrometer (ROSIS), and National Aeronautics and Space Administration (NASA) EO-1 Satellite Hyperion sensor. These datasets include Salinas-A, Salinas, Kennedy Space Center, and Indian Pines datasets acquired by AVIRIS sensor, Pavia University and Pavia Center datasets acquired by ROSIS sensor, Botswana dataset acquired by Satellite Hyperion sensor. Evaluating ROSIS and Hyperion sensors datasets are a more challenging classification problem dominated by complex urban classes and nested regions than AVIRIS sensor datasets. Table 8.1 provides a summary description of each dataset used in this thesis.

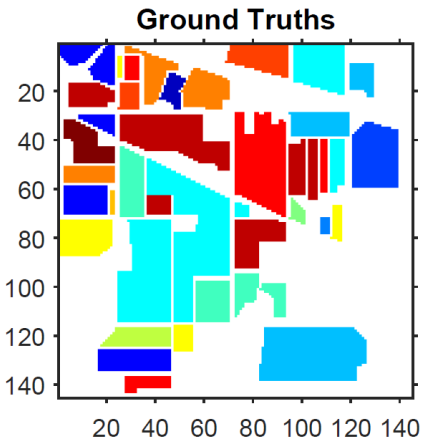
TABLE 8.1: Summary of the HSI datasets.

Dataset	Year	Source	Spatial dimensions	Spectral	Wavelength	Samples	Classes	Sensor	Resolution
Botswana	2001-2004	NASA EO-1	1496×256	242 bands	400-2500	3248	14	Satellite	30
Indian Pines	1992	NASA AVIRIS	145×145	220 bands	400 - 2500	10249	16	Aerial	20
Salinas	1998	NASA AVIRIS	512×217	224 bands	360 - 2500	54129	16	Aerial	3.7
Pavia University	2001	ROSIS-03 sensor	610×610	115 bands	430 - 860	42776	9	Aerial	1.3
Pavia Center	2001	ROSIS-03 sensor	1096×1096	102 bands	430 - 860	7256	9	Aerial	1.3

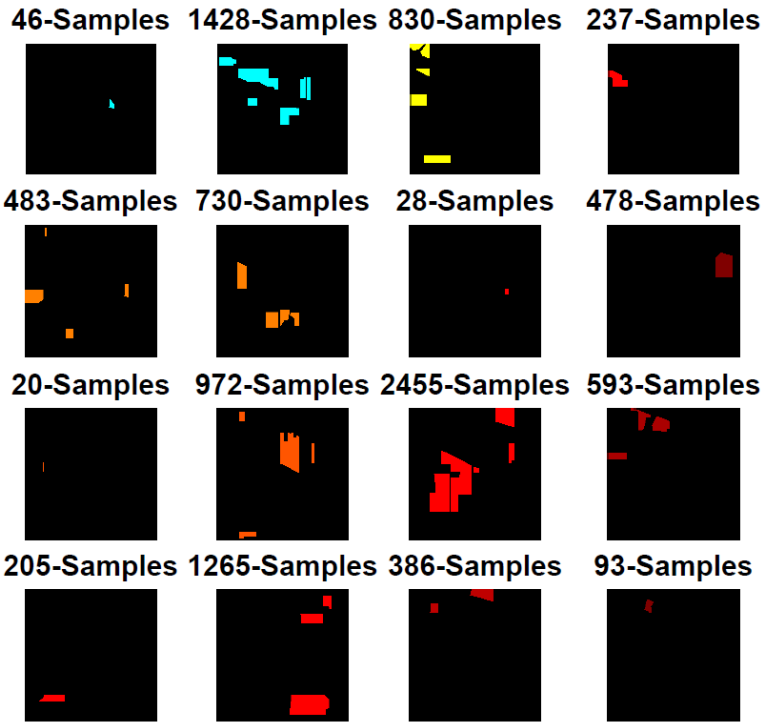
8.1.1 Indian Pines Dataset

Indian Pines (IP) data was gathered by airborne visible infrared spectroscopy (AVIRIS) sensor over the Indian Pines test site in North Western Indiana. Indian Pines dataset consists of (145×145) pixels and 224 spectral reflectance bands (channels) in the wavelength range $(0.4 - 2.5)10^{-6}$ meters. Indian Pines scene is a subset of a larger one. The Indian Pines scene contains $2/3$ agriculture and $1/3$ forest or other natural perennial vegetation. There are two major dual-lane highways, a rail line, as well as some low-density housing, other built structures, and smaller roads. Since the Indian Pines scene was taken in June some of the crops

present, corn, soybeans, are in the early stages of growth with less than 5% coverage. The ground truth available is designated into sixteen classes and is not all mutually exclusive. We have also reduced the number of bands to 200 by removing bands covering the region of water absorption. The removed bands are [104-108], [150-163], and 220. Per class, ground truths are presented in Figure 8.1b and full ground truths are presented in Figure 8.1a. Class description with the number of samples and class names are described in Table 8.2.



(A) Ground Truth for Indian Pines



(B) Per Class Ground Truth for Indian Pines

FIGURE 8.1: Ground Truth for IP.

8.1.2 Salinas-A Dataset

The Salinas-A (SLA) scene was collected by the AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution with 3.7 meter per pixel with 224 bands. The area is covered by 86 $\times$ 83 samples. As with the Indian Pines dataset, we discarded the 20 water absorption bands which are [108-112], [154-167], and 224. This image was available only as at-sensor radiance data. It includes vegetables, bare soils, and vineyard fields. Salinas-A ground truths contain 6 classes. Per class, ground truths are presented in Figure 8.2b and full ground truths are presented in Figure 8.2a. Class description with the number of samples and class names are described in Table 8.3.

TABLE 8.2: Class Description for IP Dataset

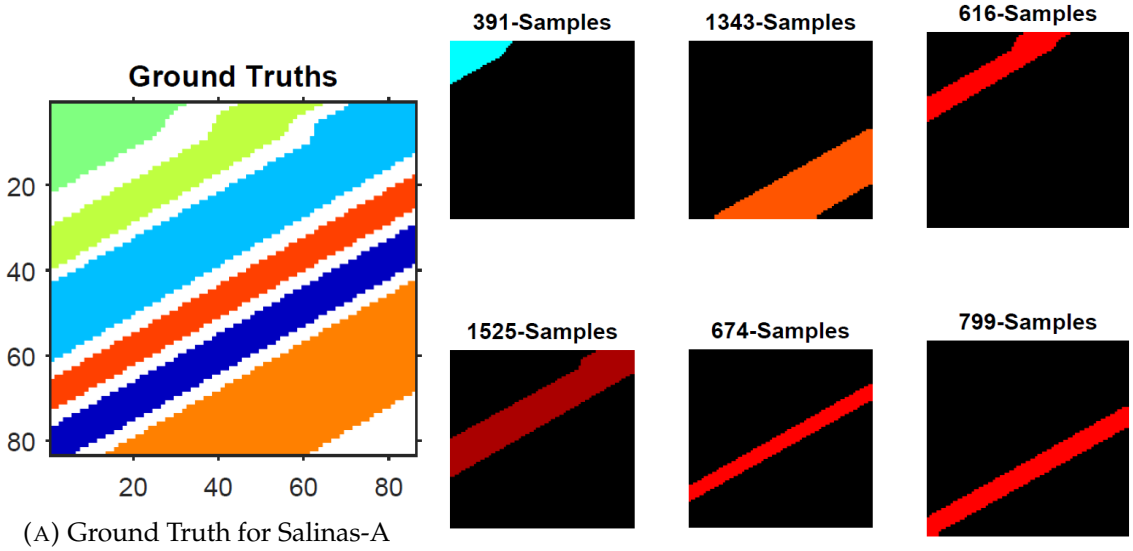
Class	Class Name	Samples in Class
1	Alfalfa	46
2	Corn notill	1428
3	Corn mintill	830
4	Corn	237
5	Grass pasture	483
6	Grass trees	730
7	Grass pasture mowed	28
8	Hay windrowed	478
9	Oats	20
10	Soybean notill	972
11	Soybean mintill	2455
12	Soybean clean	593
13	Wheat	205
14	Woods	1265
15	Buildings Grass Trees Drives	386
16	Stone Steel Towers	93

TABLE 8.3: Class Description for SLA Dataset

Class	Class Name	Samples in Class
1	Brocoli green weeds 1	391
2	Corn senesced green weeds	1343
3	Lettuce romaine 4wk	616
4	Lettuce romaine 5wk	1525
5	Lettuce romaine 6wk	674
6	Lettuce romaine 7wk	799

8.1.3 Salinas Dataset

The Salinas (SA) scene was collected by the AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution with 3.7 meter per pixel with 224 bands. The area is covered by 512×217 samples. As with the Indian Pines dataset, we discarded the 20 water absorption bands which are [108-112], [154-167], and 224. This image was available only as at-sensor radiance data. It includes vegetables, bare soils, and vineyard fields. Salinas-A ground truths contain 16 classes. Full ground truths are presented in Figure 8.3a and per class, ground truths are presented in Figure 8.3b. Class description with the number of training samples and class names are described in Table 8.4.



(A) Ground Truth for Salinas-A

(B) Per Class Ground Truth for Salinas-A

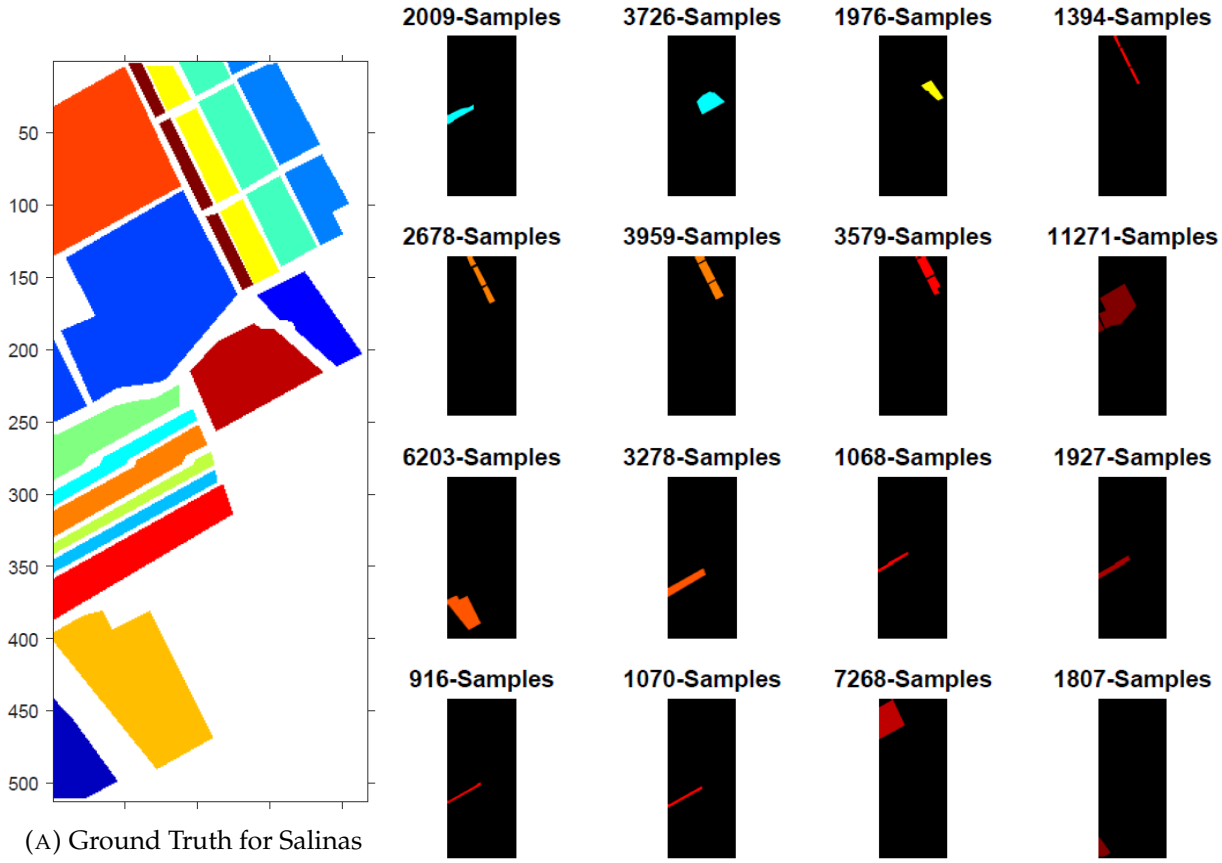
FIGURE 8.2: Ground Truth for SLA.

TABLE 8.4: Class Description for SA Dataset

Class	Class Name	Samples in Class
1	Brocoli green weeds 1	2009
2	Brocoli green weeds 2	3726
3	Fallow	1976
4	Fallow rough plow	1394
5	Fallow smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes untrained	11271
9	Soil vinyard develop	6203
10	Corn senesced green weeds	3278
11	Lettuce romaine 4wk	1068
12	Lettuce romaine 5wk	1927
13	Lettuce romaine 6wk	916
14	Lettuce romaine 7wk	1070
15	Vinyard untrained	7268
16	Vinyard vertical trellis	1807

8.1.4 Pavia Center Dataset

Pavia Center (PC) scene is acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The number of spectral bands is 102 for Pavia Center. Pavia Center is a 1096×1096 pixels image, but some of the samples in both images contain no information and have to be discarded before the analysis.



(A) Ground Truth for Salinas

(B) Per Class Ground Truth for Salinas

FIGURE 8.3: Ground Truth for SA.

The geometric resolution is 1.3 meters. Pavia Center image ground truths differentiate 9 classes. It can be seen the discarded samples in the figures as abroad black strips. Per class, ground truths are presented in Figure 8.4b and full ground truths are presented in Figure 8.4a. Class description with the number of training samples and class names are described in Table 8.5.

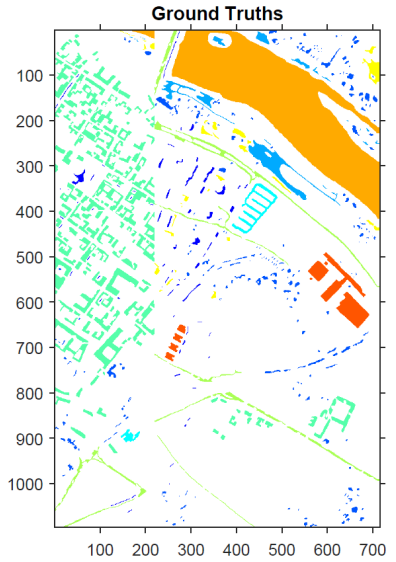
8.1.5 Pavia University Dataset

Pavia University (PU) scene is acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy. The number of spectral bands is 103 for Pavia University. Pavia University is a 610×610 pixels image, but some of the samples in both images contain no information and have to be discarded before the analysis.

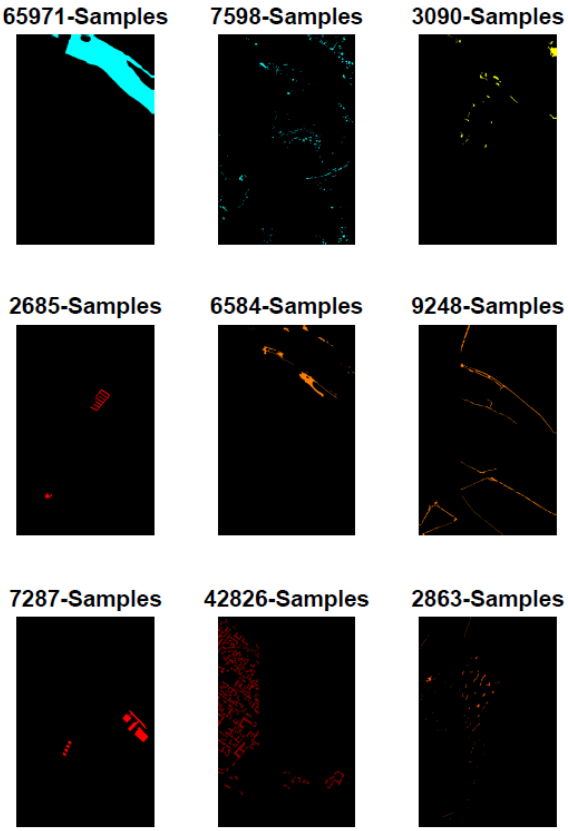
The geometric resolution is 1.3 meters. Pavia University image ground truths differentiate 9 classes. It can be seen the discarded samples in the figures as abroad black strips. Per class, ground truths are presented in Figure 8.5b and full ground truths are presented

TABLE 8.5: Class Description for PC Dataset

Class	Class Name	Samples in Class
1	Water	824
2	Trees	820
3	Asphalt	816
4	Self-Blocking Bricks	808
5	Bitumen	808
6	Tiles	1260
7	Shadows	476
8	Meadows	824
9	Bare Soil	820



(A) Ground Truth for Pavia Center



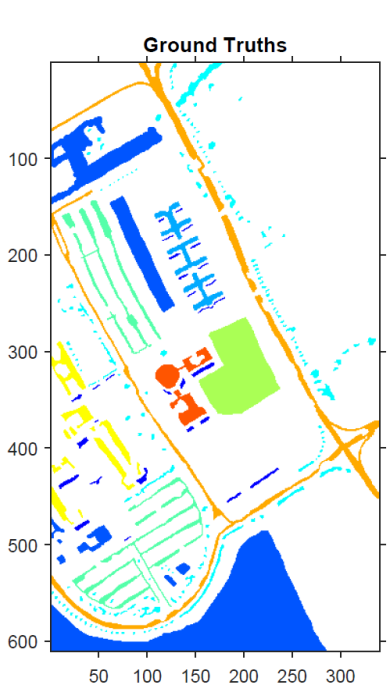
(B) Per Class Ground Truth for Pavia Center

FIGURE 8.4: Ground Truth for PC.

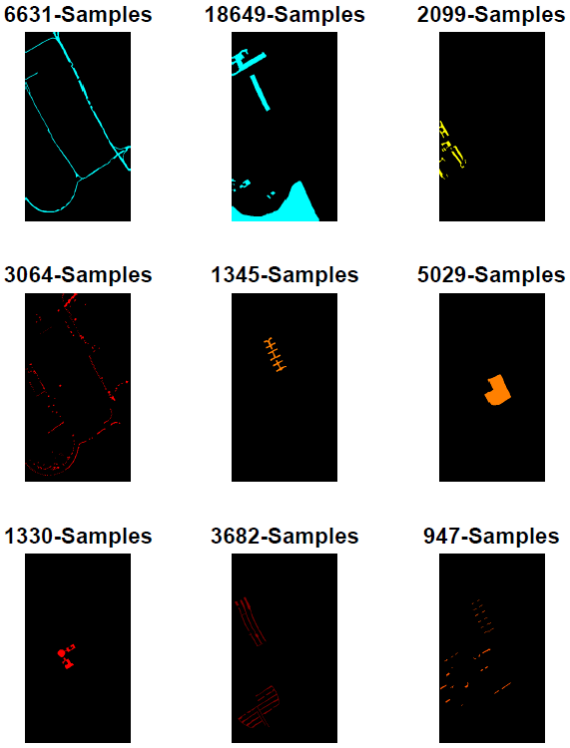
in Figure 8.5a. Class description with the number of training samples and class names are described in Table 8.6.

TABLE 8.6: Class Description for PU Dataset

Class	Class Name	Samples in Class
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947



(A) Ground Truth for Pavia University



(B) Per Class Ground Truth for Pavia University

FIGURE 8.5: Ground Truth for PU.

8.1.6 Botswana Dataset

The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana (BS) in 2001-2004. The Hyperion sensor on EO-1 acquires data at 30 m pixel resolution over a 7.7 km strip in 242 bands covering the 400-2500 nm portion of the spectrum in 10 nm windows.

Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220].

The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta. Per class, ground truths are presented in Figure 8.6b and full ground truths are presented in Figure 8.6a. Class description with the number of training samples and class names are described in Table 8.7.

TABLE 8.7: Class Description for BS Dataset

Class	Class Name	Samples in Class
1	Water	270
2	Hippo grass	101
3	Floodplain grasses1	251
4	Floodplain grasses2	215
5	Reeds1	269
6	Riparian	269
7	Firescar2	259
8	Island interior	203
9	Acacia woodlands	314
10	Acacia shrublands	248
11	Acacia grasslands	305
12	Short mopane	181
13	Mixed mopane	268
14	Exposed soils	95

8.2 Performance Evaluation Metrics

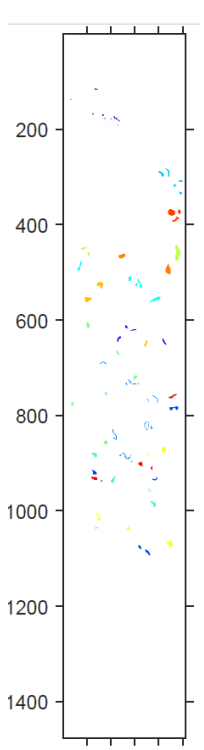
The following accuracy metrics have been used to validate the claims made in this dissertation. The accuracy metrics include Kappa (κ)[1], Average Accuracy (AA)[2], and Overall Accuracy (OA)[3]. All these metrics are computed using the following equations.

$$Kappa\ (\kappa) = \frac{P_o - P_e}{1 - P_e} \tag{8.1}$$

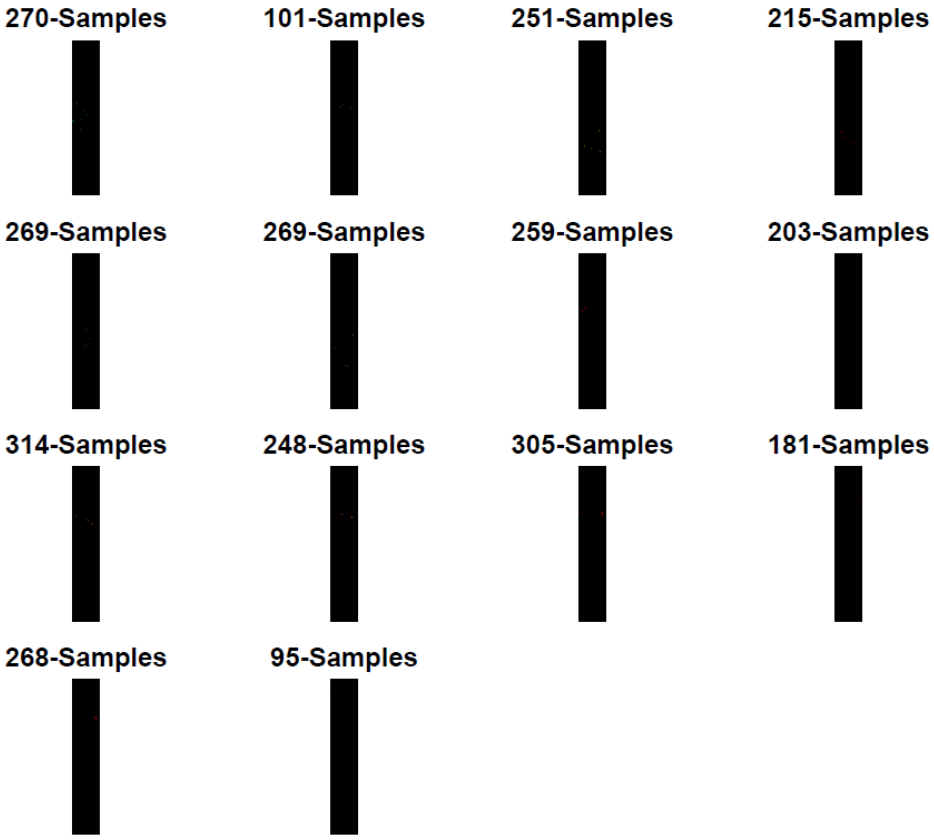
[1] κ is known as a statistical metric that considered the mutual information regarding a strong agreement among classification and ground-truth maps

[2] AA represents the average class-wise classification performance

[3] OA is computed as the number of correctly classified examples out of the total test examples



(A) Ground Truth for Botswana



(B) Per Class Ground Truth for Botswana

FIGURE 8.6: Ground Truth for BS.

where

$$P_e = P^+ + P^-$$

$$P^+ = \frac{TP + FN}{TP + FN + FP + TN} \times \frac{TP + FN}{TP + FN + FP + TN}$$

$$P^- = \frac{FN + TN}{TP + FN + FP + TN} \times \frac{FP + TN}{TP + FN + FP + TN}$$

$$P_o = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Overall\ (OA) = \frac{1}{K} \sum_{i=1}^{K} TP_i \tag{8.2}$$

Moreover, to validate the experimental results, several statistical tests such as Precision (P_r), Recall (R_c), and F1 Score are also considered. The said metrics are computed using the following mathematical expressions:

$$P_r = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i} \tag{8.3}$$

$$R_c = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \tag{8.4}$$

$$F1 = \frac{2 \times (R_c \times P_r)}{(R_c + P_r)} \tag{8.5}$$

where K be the total number of classes present in HSI dataset, TP and FP are true and false positive, TN and FN are true and false negative, respectively.

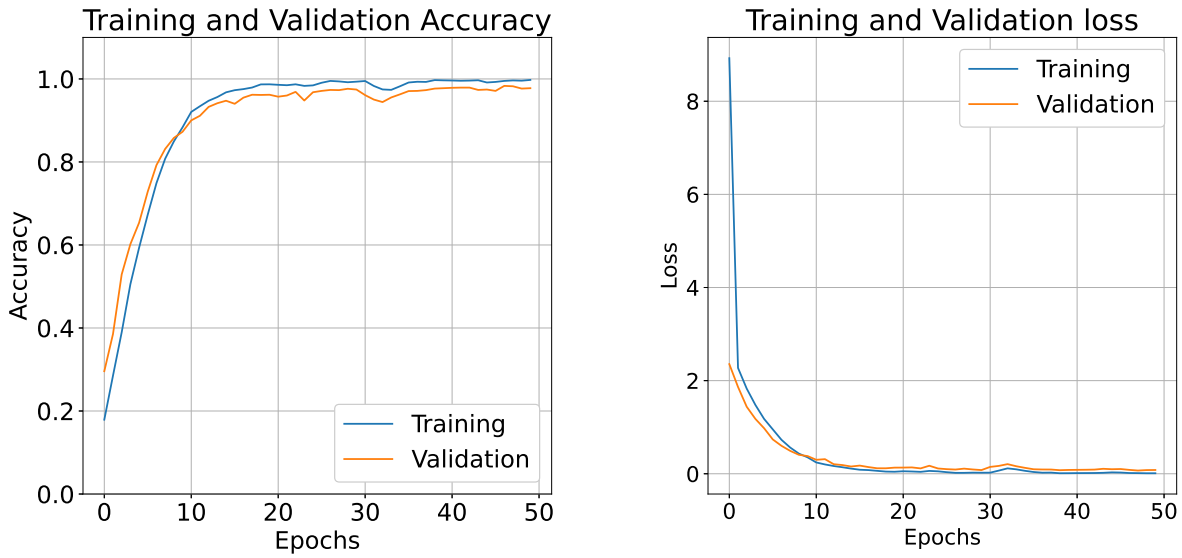
8.3 Experimental Results for A Fast and Compact 3D CNN

All the experiments were performed on an online platform known as Google Colab [438]. Google Colab is an online platform that requires a good speed of the internet to run any environment. Google Colab provides an option to execute the codes on python 3 notebook with Graphical Processing Unit (GPU), 25 GB of Random Access Memory (RAM) and 358.27 GB of could storage for data computation. In all the experiments, the initial Test/Train set is divided into a 30/70% ratio on which Training samples (70% of the entire population) are further divided into 50/50% for the Training and Validation set.

For fair comparisons, the learning rate for all the experiments is set to 0.001, $relu$ as an activation function is used for all layers except last on which $softmax$ is used, the patch sizes are set as $11 \times 11 \times 20$, $13 \times 13 \times 20$, $15 \times 15 \times 20$, $17 \times 17 \times 20$, $19 \times 19 \times 20$, $21 \times 21 \times 20$ and $25 \times 25 \times 20$, respectively with 20 most informative bands selected by iPCA method. For evaluation purposes, the Average Accuracy (AA), Overall Accuracy (OA), and Kappa (κ) coefficient have been computed from the confusion matrices. AA represents the average class-wise classification performance, OA is computed as the number of correctly classified examples out of the total test examples, and finally, κ is known as a statistical metric that considered the mutual information regarding a strong agreement among classification and ground-truth maps. Along with OA, AA, and κ metrics, several statistical tests are also being considered such as F1-Score, Precision, and Recall.

The convergence loss and accuracy of our proposed 3D CNN model for a 50 number of epochs are shown in Figure 8.7. From these figures, one can conclude that the proposed model converged almost around 20 epochs.

Whereas, the computational time of our proposed model is shown in Table 8.8 which reveals a fast convergence and computational efficiency of our proposed model. The computational time highly depends on the speed of the internet and available RAM.



(A) Accuracy (B) Loss

FIGURE 8.7: Accuracy and Loss for Training and Validation sets on IP Dataset with 11×11 window patch correspond to the 50 number of Epochs.

TABLE 8.8: Computational time in minutes for all the experimental datasets with several window sizes.

Dataset	Proposed with Several Window Sizes							2D CNN	3D-CNN	MS-3D-CNN
	11×11	13×13	15×15	17×17	19×19	21×21	25×25			
SL-A	0.22	0.23	0.56	0.28	0.98	0.37	0.45	—	—	—
SL	1.34	1.41	1.60	2.00	3.17	2.63	3.52	2.2	74.0	25.5
IP	0.33	0.33	0.61	0.78	0.62	0.58	0.76	1.9	15.2	14.1
PU	2.16	5.26	1.35	2.00	2.46	2.14	2.83	1.8	58.0	20.3

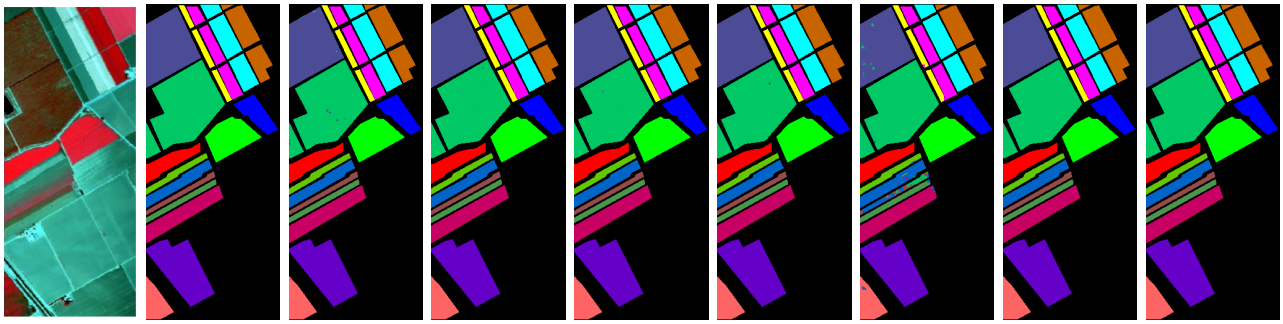
TABLE 8.9: Impact of window size on our proposed model

Window	PU			IP			SA			SL-A		
	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
11×11	99.94	99.89	99.92	88.65	83.52	87.11	99.80	99.91	99.78	100	100	100
13×13	99.81	99.65	99.75	95.38	94.14	94.72	99.93	99.94	99.93	100	100	100
15×15	99.85	99.62	99.80	93.69	93.09	92.79	99.99	99.99	99.99	100	100	100
17×17	99.05	98.49	98.75	91.80	91.74	90.62	99.95	99.97	99.95	99.93	99.93	99.92
19×19	99.93	99.78	99.91	93.13	93.42	92.15	98.04	94.02	97.81	100	100	100
21×21	99.78	99.43	99.72	94.34	91.31	93.52	99.99	99.99	99.99	100	100	100
25×25	98.79	97.67	98.39	97.75	96.17	97.44	99.96	99.93	99.95	100	100	100

The accuracy analysis i.e., OA, AA, and κ based on the impact of spatial dimensions processed by the proposed model is presented in Table 8.9. While looking into the Table 8.9, one can conclude that the window size of 11×11 is enough for Pavia University, Salinas, and Salinas-A dataset whereas the window size of 13×13 and 25×25 both works almost

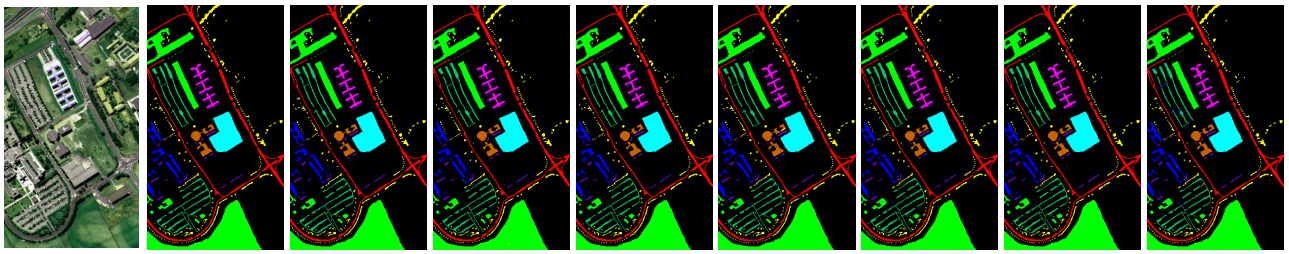
(A) IP (B) GT (C) 11×11 (D) 13×13 (E) 15×15 (F) 17×17 (G) 19×19 (H) 21×21 (I) 25×25

FIGURE 8.8: **IP Dataset** Ground Truths for different spatial dimensions processed through our proposed model.



(A) SA (B) GT (C) 11×11 (D) 13×13 (E) 15×15 (F) 17×17 (G) 19×19 (H) 21×21 (I) 25×25

FIGURE 8.9: **SA Dataset** Ground Truths for different spatial dimensions processed through our proposed model.



(A) PU (B) GT (C) 11×11 (D) 13×13 (E) 15×15 (F) 17×17 (G) 19×19 (H) 21×21 (I) 25×25

FIGURE 8.10: **PU Dataset** Ground Truths for different spatial dimensions processed through our proposed model.

(A) SLA (B) GT (C) 11×11 (D) 13×13 (E) 15×15 (F) 17×17 (G) 19×19 (H) 21×21 (I) 25×25

FIGURE 8.11: **SLA Dataset** Ground Truths for different spatial dimensions processed through our proposed model.

the same. Furthermore, the classification maps (geographical locations for each class) according to the different number of window sizes (spatial dimensions) are shown in Figures 8.8-8.11. In regards to comparison, the proposed model is compared with several state-of-the-art methods published in recent years. From experimental results listed in Table 8.10 one can conclude that the proposed model has competitive results and to some extent better

in regards to the other methods. The comparative methods includes Multi-scale-3D-CNN [439], 3D/2D-CNN [355, 359]. All the comparative methods are being trained according to the settings mentioned in their respective papers. Experiments listed in Table 8.10 shows the proposed method improves the results significantly than the state-of-the-art methods with even fewer training samples, number of convolutional layers, number of filters, number of epochs, and above all, in less computational time.

TABLE 8.10: Comparative evaluations with State-of-the-art methods while considering 11×11 Spatial dimensions with even less number of training samples (i.e., 60/40% (train/test) and 70/30% (train/validation)).

Dataset	MS-3D-CNN			3D-CNN			2D-CNN			**Proposed**		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
PU	95.95	97.52	93.40	96.34	97.03	94.90	96.63	94.84	95.53	**98.40**	**97.89**	**97.89**
IP	81.39	75.22	81.20	82.62	76.51	79.25	80.27	68.32	75.26	**97.75**	**94.54**	**97.44**
SA	94.20	96.66	93.61	85.00	89.63	83.20	96.34	94.36	95.93	**98.06**	**98.80**	**97.85**

8.4 Concluding Remarks for A Fast and Compact 3D CNN

The proposed model provided state-of-the-art experimental results in a computationally efficient fashion on four HSI benchmark datasets which resolved the problem of inter-class similarity and high intra-class variability using 3D convolution-based spatial-spectral information. To summarize, the proposed end-to-end trained 3D CNN has fewer parameters, better recognition accuracy, and fast convergence time than existing 2D/3D CNN models. The experimental results reveal that the proposed method outperformed state-of-the-art methods on various public benchmarks while being less complex than the conventional 3D CNN models.

8.5 Experimental Results for Regularized Hybrid CNN Feature Hierarchy

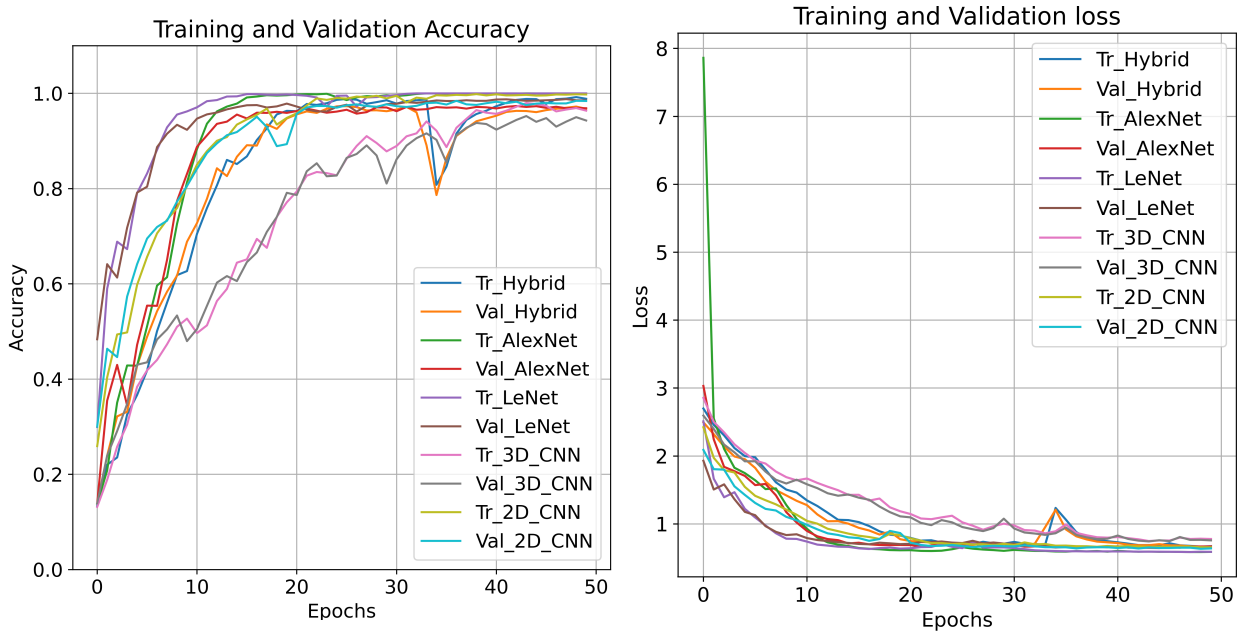
The experiments have been conducted on three real HSI datasets, namely, IP, SA, and PU. These datasets are acquired by two different sensors i.e, Reflective Optics System Imaging Spectrometer (ROSIS) and Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [380]. The experimental results explained in this work have been obtained through Google Colab [438] which is an online platform to execute any python environment while having a good internet speed to execute the code. Google Colab provides the option to execute many versions of python, Graphical Process Unit (GPU), up to 358+ GB of cloud storage, and 25 GB of Random Access Memory (RAM).

In all the experiments, the initial size of the train/validation/test sets is set to 25%/25%/50% to validate the proposed model as well as several other state-of-art-deep models. The baseline models include AlexNet (5 Convolutional layers (96, 256, 384, 384, 256 filters while each layer has filter sizes as (7,7), (5,5), (3,3), (3,3), and (3,3)), 1 pooling layer (after first Convolutional layer), flatten layer, dense layers with 4096 units and after each dense layer a dropout layer has been used with 0.5% and finally an output layer has been used with the total number of classes to predict.) [440], LeNet (2 Convolutional layers (32 and 64 filters while each layer has filter sizes as (5,5) and (3,3)), 1 pooling layer (after first Convolutional layer), flatten layer, dense layer with 100 units and after dense layer, an output layer has been used with the total number of classes to predict.) [441], 2D CNN (4 Convolutional layers (8, 16, 32, and 64 filters while each layer has filter size (3,3)), flatten layer, 2 dense layers with 256 and 100 units and after each dense layer, a dropout layer has been used with 0.4% and finally, an output layer has been used with the total number of classes to predict.) [442], 3D CNN (4 Convolutional layers (8, 16, 32, and 64 filters while each layer has filter sizes as (3,3,7), (3,3,5), (3,3,3), and (3,3,3)), flatten layer, 2 dense layers with 256 and 128 units and after each dense layer, a dropout layer has been used with 0.4% and finally, an output layer has been used with the total number of classes to predict) [356], and 3D/2D hybrid model [380].

For fair comparison purposes, the learning rate for all these models including hybrid models is set to 0.001, Relu as the activation function for all layers except the output layer on which Softmax is used, patch size is set of 15, and for all the experiments, 15 most informative bands have been selected using PCA to reduce the computational load. The convergence, accuracy, and loss of our proposed regularization technique with several CNN models for 50 epochs are presented in Figure 8.12. From loss and accuracy curves, one can conclude that the regularization has faster convergence.

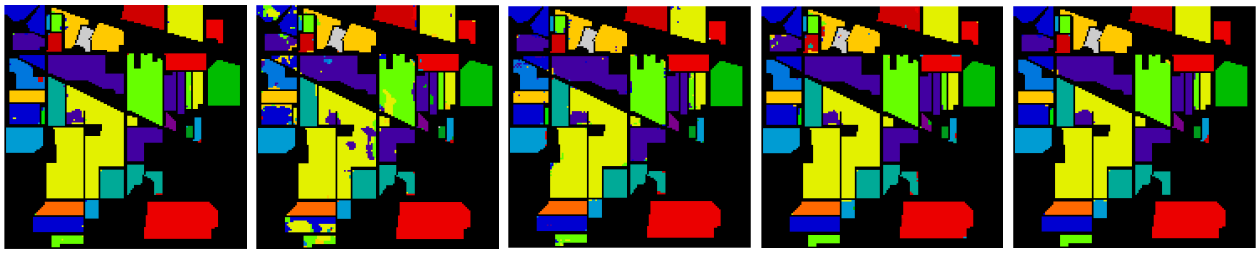
Table 8.11 and Figure 8.13 presents in-depth comparative accuracy analysis on the IP dataset. Table 8.12 and Figure 8.14 presents in-depth comparative accuracy analysis on the PU dataset. Table 8.13 and Figure 8.15 presents in-depth comparative accuracy analysis on the SA dataset.

In all experimental results, the training, validation, and test sets are selected using a 5-fold cross-validation process with 25, 25, and 50% samples for training, validation, and test sets, respectively. The hybrid and all other competing models are trained using 15×15 patch size because, the classification performance strongly depends on the patch size, in which if the patch size is too big, then the model may take pixels from various classes whereas if the patch size is too small, the model may decrease the inter-class diversity in samples. Hence in both cases, the ultimate result will be in terms of a higher misclassification rate, lead to low generalization performance. Therefore, an appropriate patch size must need to opt before the final experimental setup. The patch size selected in these experiments is selected based



(A) Accuracy (B) Loss

FIGURE 8.12: Accuracy and Loss for Training and Validation sets on IP for 50 epochs.



(A) 2D (B) 3D (C) A.Net (D) LeNet (E) Hybrid

FIGURE 8.13: **IP:** Classification accuracy: Fig. 8.13a: 2D-CNN = 98.94%, Fig. 8.13b: 3D CNN = 91.57%, Fig. 8.13c: AlexNet = 97.65%, Fig. 8.13d: LeNet = 98.14%, and Fig. 8.13e Hybrid = **99.29%**.

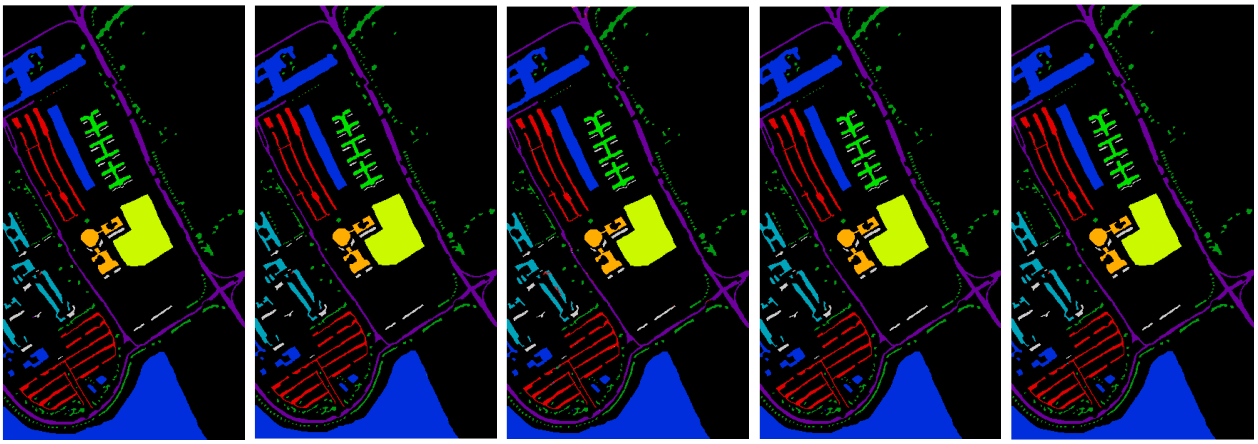
on the hit and trial method (i.e., provided the best accuracy).

The experimental results on benchmark HSI datasets are presented in Table 8.14. From these results, one can conclude that the proposed label smoothing process significantly improves the performance, in terms of accuracy, speed of convergence, and computational time. For comparison purposes, the framework, i.e., label smoothing for the Hybrid CNN model is compared with various state-of-the-art works published in recent years. From the experimental results presented in Table 8.14, one can conclude that label smoothing with Hybrid CNN has obtained better results as compared to the state-of-the-art frameworks and to some extent outperformed with respect to the other models.

The comparative models include Support Vector Machine (SVM) with and without any grid optimization. Multi-layer Perceptron (MLP) having four fully connected layers with dropout. 2-D CNN model proposed by Sharma et.al. [357]. Semi-supervised CNN model

TABLE 8.11: **IP:** Performance analysis of different state-of-the-art models trained using the label smoothing technique.

Class	Train/Val/Test	2D	3D	AlexNet	LeNet	Hybrid
Alfalfa	11/12/23	100	91.3043	95.6521	82.6086	100
Corn-notill	357/357/714	98.3193	93.5574	97.3389	97.0588	98.8795
Corn-mintill	207/208/415	99.5180	66.7469	98.3132	99.5180	99.5180
Corn	59/59/118	94.0677	90.6779	93.2203	99.1525	100
Grass-pasture	121/121/242	98.3471	97.1074	96.2809	94.2148	96.2809
Grass-trees	182/183/365	98.9041	97.5342	98.3561	98.9041	99.7260
Grass-mowed	7/7/14	92.8571	92.8571	100	100	100
Hay-windrowed	119/120/239	100	100	100	100	100
Oats	5/5/10	70	0	100	70	100
Soybean-notill	243/243/486	98.5596	82.3045	93.6213	99.3827	97.9423
Soybean-mintill	614/614/1228	99.6742	92.4267	97.8013	99.9185	99.8371
Soybean-clean	148/149/297	97.6430	98.6531	96.9696	95.2861	99.6632
Wheat	51/51/102	99.0196	98.0392	100	99.0196	99.0196
Woods	316/317/633	99.8420	99.3680	99.5260	98.8941	99.8420
Buildings	96/97/193	99.4818	90.6735	100	91.1917	99.4818
Stone-Steel	23/23/46	100	97.8260	100	93.4782	100
Training Time		55.6695	250.1662	919.5566	61.8763	248.5993
Test Time		1.4897	4.0402	5.6891	1.2752	3.9997
Overall Accuracy		98.9463	91.5707	97.6585	98.14634	**99.2975**
Average Accuracy		98.7980	86.8173	97.9425	94.9142	**99.3869**
Kappa (κ)		96.6396	90.3561	97.3312	97.8853	**99.1990**



(A) 2D (B) 3D (C) A.Net (D) LeNet (E) Hybrid

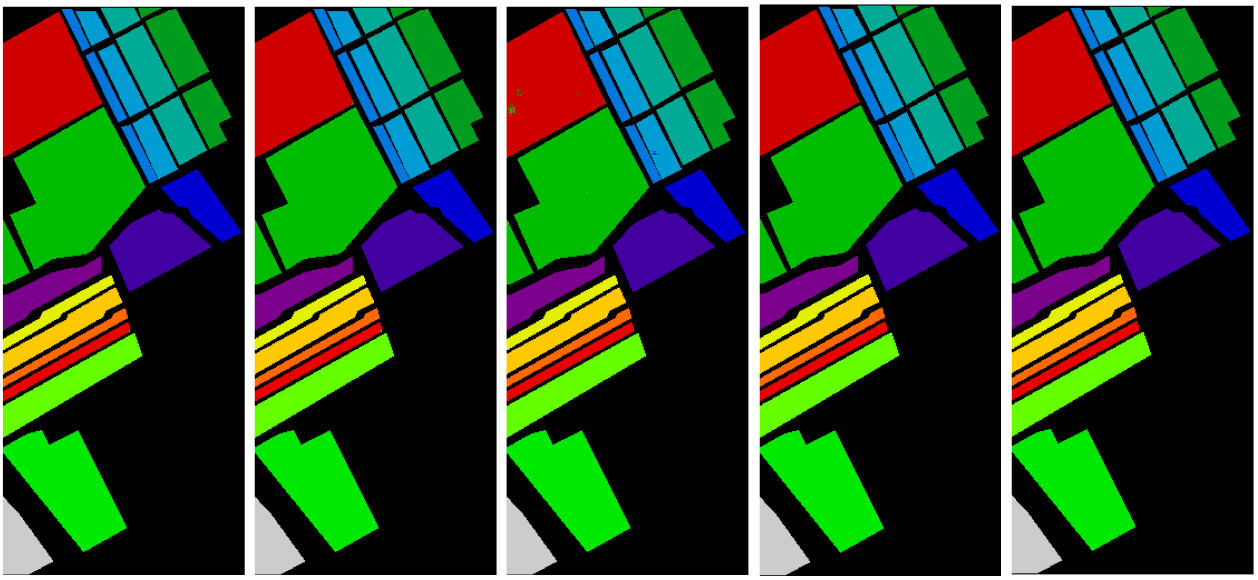
FIGURE 8.14: **PU:** Classification accuracy: Fig. 8.14a: 2D-CNN = 99.9070%, Fig. 8.14b: 3D CNN = 99.9256%, Fig. 8.14c: AlexNet = 99.0768%, Fig. 8.14d: LeNet = 99.9318%, and Fig. 8.14e: Hybrid = **99.9628%**.

TABLE 8.12: **PU:** Performance analysis of different state-of-the-art models trained using the label smoothing technique.

Class	Train/Val/Test	2D	3D	AlexNet	LeNet	Hybrid
Asphalt	1658/1658/3316	100	100	98.9143	100	100
Meadows	4662/4662/9324	100	99.9892	100	100	100
Gravel	524/525/1049	99.5233	99.3326	95.5195	99.4280	99.6186
Trees	766/766/1532	99.6736	100	98.8250	99.7389	100
Painted	336/337/673	100	100	100	100	100
Soil	1257/1257/2514	100	100	99.9602	100	100
Bitumen	332/333/665	100	100	99.6992	100	100
Bricks	920/921/1841	99.8913	99.8913	97.9359	100	99.8913
Shadows	237/237/74	99.3670	99.5780	98.5232	99.7890	100
Training Time		296.0174	1145.9233	4716.4900	308.6389	1143.4996
Test Time		5.7400	14.8634	24.1701	4.5054	15.5904
Overall Accuracy		99.9298	99.9438	99.3033	99.9485	**99.9719**
Average Accuracy		99.8283	99.8657	98.8197	99.8839	**99.9455**
Kappa (κ)		99.9070	99.9256	99.0768	99.9318	**99.9628**

TABLE 8.13: **SA:** Performance analysis of different state-of-the-art models trained using the label smoothing technique.

Class	Train/Val/Test	2D	3D	AlexNet	LeNet	Hybrid
Weeds 1	502/502/1005	100	100	100	100	100
Weeds 2	931/931/1863	100	100	100	100	100
Fallow	494/494/988	100	100	100	100	100
Fallow rough plow	348/348/698	100	100	100	100	100
Fallow smooth	669/669/1340	100	100	99.7012	100	100
Stubble	990/990/1980	100	100	100	100	100
Celery	894/894/1790	99.9441	100	100	100	100
Grapes untrained	2817/2818/5636	99.9822	100	99.9822	100	100
Soil vinyard develop	1550/1551/3102	100	100	100	100	100
Corn Weeds	819/820/1639	100	100	100	100	100
Lettuce 4wk	267/267/534	100	100	100	100	100
Lettuce 5wk	481/482/963	100	100	100	100	100
Lettuce 6wk	229/229/458	100	100	100	100	100
Lettuce 7wk	267/268/535	100	99.6261	100	100	100
Vinyard untrained	1817/1817/3634	99.8130	99.9174	99.1744	100	100
Vinyard trellis	451/452/904	100	100	100	100	100
Training Time	—	257.9992	1256.1199	4667.9047	288.3353	1267.9766
Test Time	—	7.3995	16.8058	27.8388	6.3860	19.2670
Overall Accuracy	—	99.9889	99.9815	99.8706	100.0	**100.0**
Average Accuracy	—	99.9837	99.9714	99.9286	100.0	**100.0**
Kappa (κ)	—	99.9876	99.9794	99.8559	100.0	**100.0**



(A) 2D (B) 3D (C) A.Net (D) LeNet (E) Hybrid

FIGURE 8.15: **SA:** Classification accuracy: Fig. 8.15a: 2D-CNN = 99.9876%, Fig. 8.15b: 3D CNN = 99.9794%, Fig. 8.15c: AlexNet = 99.8559%, Fig. 8.15d: LeNet = 100.0%, and Fig. 8.15e Hybrid = **100.0%**.

proposed by Liu et.al. [358]. A 3-D CNN model proposed by Hamida et. al., [359]. A hybrid CNN model proposed by Lee et.al. [360] consists of two 3D and eight 2D convolutional layers. A simple and compact 3D CNN model proposed by Chen et.al., [352] consists of three 3D convolutional layers. A lightweight 3D CNN model proposed by Li et.al. [361] consists of two 3D convolutional layers and a fully connected layer. Li's work is different from traditional 3D CNN models as it uses fixed spatial-sized 3D convolutional layers with slight changes in spectral depth. Multi-scale-3D-CNN [225], a fast and compact 3D-CNN (FC-3D-CNN) [356] and three different versions of Hybrid Depth-Separable Residual Network [362].

All the comparative models are being trained as per the settings mentioned in their respective papers except for the number of dimensions and patch size (i.e., 15 dimensions selected using PCA, and 15×15 path size). The experimental results listed in Table 8.14 show that the proposed framework has significantly improved the classification results as compared to the other methods with fewer training samples.

8.6 Concluding Remarks for Regularized Hybrid CNN Feature Hierarchy

This section proposed the use of an entropy-based regularization process to improve the generalization performance using soft labels. These soft labels are the weighted average of

TABLE 8.14: Experimental Comparison with State-of-the-art models.

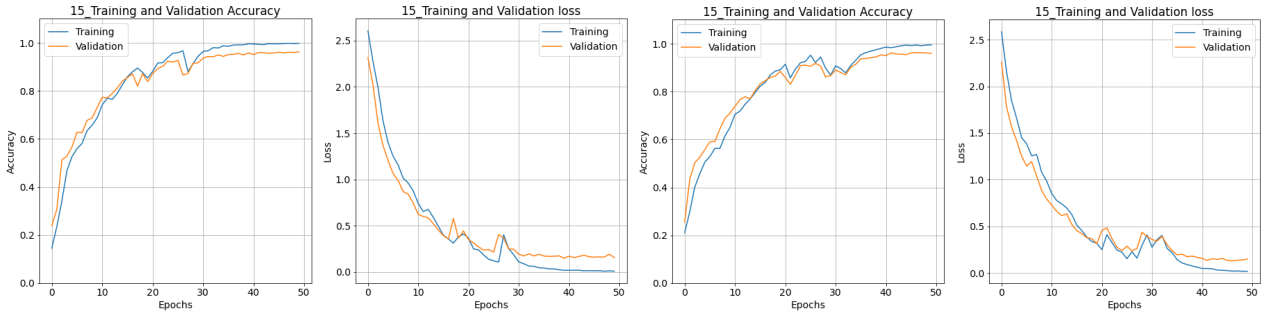
Methods	Salinas Full Scene			IP		
	OA	AA	Kappa	OA	AA	Kappa
MLP	79.79	67.37	77.40	87.57	89.07	85.80
SVM-Grid	67.39	45.89	62.80	87.93	88.02	86.20
SVM	92.95	94.60	92.11	85.30	79.03	83.10
FC-3D-CNN [356]	98.06	98.80	97.85	98.20	96.46	97.95
Xie et al. [357]	93.35	91.88	92.60	95.64	96.01	95.10
Liu et al. [358]	84.27	79.10	82.50	89.56	89.32	88.10
3D-CNN [359]	85.00	89.63	83.20	82.62	76.51	79.25
Lee et al. [360]	84.14	73.27	82.30	87.87	83.42	86.10
Chen et al. [352]	86.83	92.08	85.50	93.20	95.51	92.30
Li [361]	88.62	86.84	87.40	94.22	96.71	93.40
MS-3D-CNN [225]	94.69	94.03	94.10	91.87	92.21	90.80
Zhao et al. [362]	98.89	98.88	98.85	95.86	96.08	95.09
SyCNN-S [363]	97.44	98.46	97.20	95.90	97.84	95.30
SyCNN-D [363]	97.76	98.95	97.50	96.13	98.08	95.60
SyCNN-ATT [363]	98.92	99.35	98.80	97.31	98.43	96.90
Regularized AlexNet	99.87	99.92	99.85	97.65	97.94	97.33
Regularized LeNet	**100.0**	**100.0**	**100.0**	98.14	94.91	97.88
Regularized 2D	99.98	99.98	99.98	98.94	98.79	96.63
Regularized 3D	99.98	99.97	99.97	91.57	86.81	90.35
Regularized Hybrid	**100.0**	**100.0**	**100.0**	**99.29**	**99.38**	**99.19**

the hard labels and uniform distribution over entire ground truths. The entropy-based regularization process prevents CNN from becoming over-confident while learning and predicting thus improves the model calibration and beam-search. Extensive experiments have confirmed that the proposed pipeline outperformed several state-of-the-art methods.

8.7 Experimental Results for Artifacts of Dimension Reduction on Hybrid CNN

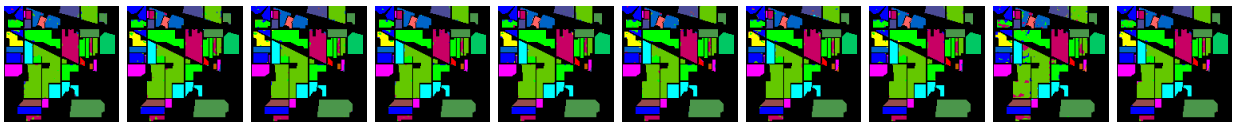
In all the experiments, each dataset is initially divided into a 50/50% ratio for the training and test set and then the training set is further split into a 50/50% ratio for training and validation samples. In all the experiments, learning rate is set to 0.001 and relu activation function used for all the layers except the last layer where softmax is applied. Spatial dimensions of 3D input patches for all datasets are set as as $9 \times 9 \times 15$, $11 \times 11 \times 15$, $9 \times 9 \times 18$, $11 \times 11 \times 18$, $9 \times 9 \times 21$, $11 \times 11 \times 21$, $9 \times 9 \times 24$, $11 \times 11 \times 24$, and $9 \times 9 \times 27$, $11 \times 11 \times 27$, where 15, 18, 21, 24 and 27 are the number of most informative bands extracted by PCA, iPCA, SPCA, ICA, and SVD, respectively.

The convergence loss and accuracy of Hybrid CNN for 50 epochs with two different patch sizes are illustrated in Figure 8.16. From these accuracy and loss curves, one can deduce that the model is converged almost around the 35 epoch for both 9×9 and 11×11 window sizes. A detailed experimental results on IP dataset is shown in Table 8.15 and Figure 8.17. Moreover, per class statistical significance is shown in Table 8.16. A detailed experimental result on SA is shown in Table 8.17 and Figure 8.18. Moreover, the statistical significance is shown in Table 8.18. A detailed experimental result on the PU is shown in Table 8.19 and Figure 8.19. Moreover, the statistical significance is shown in Table 8.20.



(A) Accuracy, $9 \times 9 \times 15$ (B) Loss, $9 \times 9 \times 15$ (C) Accuracy, $11 \times 11 \times 15$ (D) Loss, $11 \times 11 \times 15$

FIGURE 8.16: Accuracy and Loss for Training and Validation sets on IP for 50 number of epochs with two different spatial dimensions (9×9 and 11×11) and 15 number of dimensions.



(A) 15, 9 (B) 15, 11 (C) 18, 9 (D) 18, 11 (E) 21, 9 (F) 21, 11 (G) 24, 9 (H) 24, 11 (I) 27, 9 (J) 27, 11

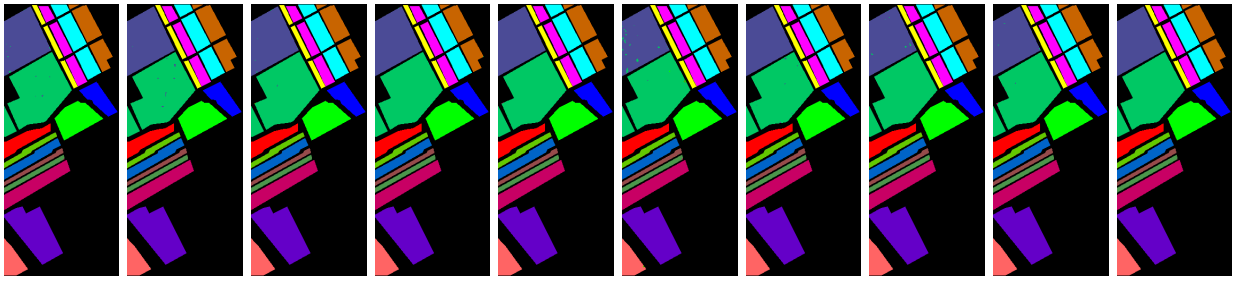
FIGURE 8.17: Classification results of **IP** for different number of dimensions (15, 18, 21, 24, 27 dimensions selected using PCA) with 9×9 and 11×11 patch sizes respectively.

TABLE 8.15: Kappa, Overall and Average accuracy for **IP** with different number of dimensions (15, 18, 21, 24, 27) and different number of patch sizes (i.e., 9×9 and 11×11).

Method	**15 Bands**		**18 Bands**		**21 Bands**		**24 Bands**		**27 Bands**	
	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$
PCA	96.75	96.84	96.75	97.00	97.55	96.95	97.53	97.24	91.54	97.00
	97.15	97.23	97.15	97.37	97.85	97.33	97.83	97.58	92.57	97.37
	97.54	95.57	97.51	93.39	97.37	96.98	97.49	95.96	91.53	96.95
iPCA	62.30	35.54	18.81	66.48	83.29	73.33	38.91	86.50	0.00	84.72
	66.87	47.47	36.27	70.71	85.40	76.82	49.89	88.18	23.96	86.63
	56.35	25.41	12.49	46.91	67.58	50.32	31.16	81.44	6.25	75.68
SPCA	73.20	75.66	76.12	81.86	68.95	11.82	73.96	79.81	0.00	75.88
	76.72	78.87	79.08	84.16	72.88	28.45	77.23	82.44	23.96	78.79
	65.43	56.02	64.46	74.97	59.91	12.27	65.39	60.81	6.25	53.97
ICA	68.52	71.24	65.13	71.36	79.84	84.42	91.83	90.23	79.94	90.55
	72.72	75.10	69.76	75.36	82.42	86.40	92.84	91.41	82.48	91.75
	60.79	59.92	57.74	61.50	70.80	81.28	86.59	85.93	74.69	85.49
SVD	14.77	0.00	0.00	0.00	0.00	0.00	0.00	29.94	0.00	0.00
	30.79	23.96	23.96	23.96	23.96	23.96	23.96	43.28	23.96	23.96
	12.34	6.25	6.25	6.25	6.25	6.25	6.25	18.70	6.25	6.25

TABLE 8.16: Class Names, Total Samples, Train, Validation and Test Sample numbers (Tr, Val, Te) along with the Statistical Test (Precision (P_r), Recall (R_c) & F1-Score (F1)) results for **IP dataset** with PCA as dimensional reduction method over two window sizes i.e., $W_1 = 9 \times 9$ and $W_2 = 11 \times 11$.

Class	Tr, Val, Te	15 Bands			18 Bands			21 Bands			24 Bands			27 Bands		
		P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1
		W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2		
C1 [4]	11, 12, 23	1.00/1.00	0.96/0.91	0.98/0.95	1.00/1.00	0.96/0.91	0.98/0.95	1.00/1.00	0.96/0.87	0.98/0.93	1.00/1.00	1.00/0.96	1.00/0.98	1.00/0.96	1.00/0.96	1.00/0.96
C2 [5]	357, 357, 714	0.96/0.96	0.96/0.98	0.96/0.97	0.95/0.95	0.95/0.96	0.95/0.96	0.98/0.96	0.96/0.97	0.97/0.96	0.99/0.98	0.96/0.97	0.97/0.97	0.94/0.98	0.85/0.94	0.89/0.96
C3 [6]	208, 207, 415	0.96/0.97	0.98/0.98	0.97/0.97	0.94/0.95	0.99/0.98	0.96/0.96	0.95/0.95	0.99/0.97	0.97/0.96	0.96/0.95	0.95/0.98	0.97/0.96	0.86/0.94	0.97/0.98	0.91/0.96
C4 [7]	59, 60, 118	0.99/1.00	0.95/0.89	0.97/0.94	0.97/1.00	0.89/0.86	0.93/0.92	0.94/0.96	0.93/0.92	0.94/0.94	1.00/0.94	0.92/0.96	0.96/0.95	0.96/1.00	0.84/0.93	0.90/0.96
C5 [8]	121, 120, 242	1.00/0.99	0.98/0.95	0.99/0.97	0.99/0.99	0.96/0.99	0.97/0.99	1.00/0.99	0.97/0.98	0.98/0.99	0.97/0.99	0.97/0.95	0.97/0.97	1.00/0.98	0.93/0.95	0.97/0.97
C6 [9]	183, 182, 365	1.00/0.99	0.99/1.00	0.99/1.00	1.00/0.99	1.00/1.00	1.00/1.00	1.00/0.99	1.00/1.00	1.00/1.00	0.98/0.99	1.00/1.00	0.99/1.00	0.98/1.00	0.99/1.00	0.99/1.00
C7 [10]	7, 7, 14	1.00/0.92	1.00/0.86	1.00/0.89	1.00/1.00	1.00/1.00	1.00/1.00	0.93/0.93	1.00/1.00	0.97/0.97	0.64/0.93	1.00/1.00	0.78/0.97	0.82/0.74	1.00/1.00	0.90/0.85
C8 [11]	120, 119, 239	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
C9 [12]	5, 5, 10	0.91/0.90	1.00/0.90	0.95/0.90	0.83/1.00	1.00/0.40	0.91/0.57	1.00/1.00	0.90/1.00	0.95/1.00	1.00/1.00	0.90/0.70	0.95/0.82	0.83/0.90	0.50/0.90	0.62/0.90
C10 [13]	243, 243, 486	0.96/0.95	0.93/0.94	0.94/0.95	0.99/0.98	0.93/0.93	0.96/0.95	0.99/0.95	0.95/0.94	0.97/0.95	0.96/0.99	0.96/0.93	0.96/0.96	0.78/0.97	0.90/0.94	0.83/0.95
C11 [14]	614, 613, 1228	0.97/0.98	0.98/0.97	0.97/0.98	0.98/0.98	0.97/0.98	0.98/0.98	0.98/0.98	0.99/0.98	0.98/0.98	0.98/0.98	0.99/0.99	0.98/0.98	0.94/0.97	0.91/0.99	0.93/0.98
C12 [15]	148, 148, 297	0.94/0.92	0.94/0.96	0.94/0.94	0.93/0.94	0.98/0.98	0.95/0.96	0.96/0.95	0.98/0.94	0.97/0.94	0.96/0.95	0.99/0.97	0.97/0.96	0.92/0.95	0.86/0.97	0.89/0.96
C13 [16]	51, 52, 102	1.00/1.00	0.98/1.00	0.99/1.00	0.97/1.00	1.00/0.98	0.99/0.99	1.00/0.96	0.99/0.99	1.00/0.98	1.00/1.00	0.98/0.99	0.99/1.00	1.00/1.00	0.98/0.98	0.99/0.99
C14 [17]	316, 316, 633	0.99/0.99	0.99/0.99	0.99/0.99	0.99/0.99	0.99/1.00	0.99/0.99	0.99/1.00	0.99/1.00	0.99/1.00	1.00/0.99	1.00/1.00	1.00/1.00	0.97/0.99	1.00/1.00	0.98/1.00
C15 [18]	96, 97, 193	0.93/0.97	0.97/0.96	0.95/0.97	0.93/0.97	0.98/0.98	0.95/0.97	0.93/0.99	0.98/0.95	0.95/0.97	1.00/0.92	0.99/0.98	0.99/0.95	0.96/0.95	0.91/0.98	0.94/0.96
C16 [19]	23, 24, 46	0.92/0.92	1.00/1.00	0.96/0.96	0.92/0.94	1.00/1.00	0.96/0.97	0.90/0.92	1.00/1.00	0.95/0.96	0.98/0.92	1.00/1.00	0.99/0.96	0.78/0.96	1.00/1.00	0.88/0.98



(A) 15, 9 (B) 15, 11 (C) 18, 9 (D) 18, 11 (E) 21, 9 (F) 21, 11 (G) 24, 9 (H) 24, 11 (I) 27, 9 (J) 27, 11

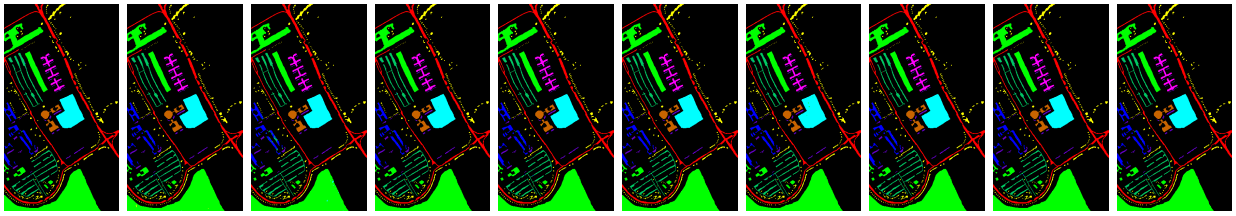
FIGURE 8.18: Classification results for **SA** for different number of dimensions (15, 18, 21, 24, 27 dimensions selected using PCA) with different number of patch sizes (9×9 and 11×11).

TABLE 8.17: Kappa, Overall and Average accuracy for **SA dataset** with different number of dimensions (15, 18, 21, 24, 27) and different number of patch sizes (9×9 and 11×11).

Method	**15 Bands**		**18 Bands**		**21 Bands**		**24 Bands**		**27 Bands**	
	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$	$\mathbf{9 \times 9}$	$\mathbf{11 \times 11}$
PCA	99.86	99.89	99.93	99.96	99.98	99.58	99.91	99.86	99.93	99.99
	99.87	99.90	99.93	99.97	99.99	99.62	99.92	99.88	99.93	99.99
	99.90	99.97	99.97	99.97	99.98	99.68	99.95	99.90	99.97	99.98
iPCA	97.42	21.48	91.66	42.17	97.70	1.57	93.46	84.19	84.19	91.70
	97.69	33.73	92.53	49.45	97.93	21.91	94.13	85.82	85.84	92.54
	93.35	18.66	90.33	33.55	97.61	7.37	91.46	72.51	70.23	85.30
SPCA	95.97	33.53	96.57	88.31	99.31	83.12	96.65	69.92	97.04	81.52
	96.38	39.80	96.92	89.51	99.38	85.00	97.00	73.45	97.34	83.77
	92.13	30.96	97.25	87.04	99.32	74.04	92.26	52.74	97.46	72.27
ICA	99.08	99.56	99.17	99.93	99.07	99.77	99.72	99.75	99.83	99.93
	99.17	99.60	99.25	99.93	99.17	99.80	99.75	99.78	99.85	99.94
	99.46	99.81	99.58	99.94	99.42	99.90	99.87	99.87	99.89	99.91
SVD	98.02	0.00	97.14	0.00	82.59	94.16	96.22	16.91	71.79	0.00
	98.22	20.82	97.43	13.43	84.44	94.76	96.61	32.28	74.82	20.82
	98.40	6.25	98.54	6.25	73.60	91.05	91.93	12.50	68.44	6.25

TABLE 8.18: Class Names, Total Samples, Train, Validation and Test Sample numbers (Tr, Val, Te) along with the Statistical Test (Precision (P_r), Recall (R_c) & F1-Score (F1)) results for **SA** with PCA as dimensional reduction method over two window sizes ($W_1 = 9 \times 9$ and $W_2 = 11 \times 11$).

Class	Tr, Val, Te	15 Bands			18 Bands			21 Bands			24 Bands			27 Bands		
		P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1
		W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2		
C1 [20]	502, 503, 1004	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1 .00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C2 [21]	932, 931, 1863	1.00/1.00	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	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C3 [22]	494 , 494 , 988	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C4 [23]	348, 349, 697	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C5 [24]	669, 670, 1339	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C6 [25]	990, 989, 1980	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C7 [26]	895, 894, 1790	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C8 [27]	2818, 2817, 5636	1.00/1.00	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	1.00/1.00	1.00/1.00	1.00/1.00
C9 [28]	1550, 1551, 3102	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C10 [29]	820, 819, 1639	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C11 [30]	267, 267, 534	1.00/1.00	1.00/1.00	1.00/1.00	0.99/1.00	1.00/1.00	1.00/1.00	1.00/0.99	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
C12 [31]	482, 482, 963	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C13 [32]	229, 229, 458	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/0.98	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
C14 [33]	267, 268, 535	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	1.00/0.98	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
C15 [34]	1817, 1817, 3634	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	1.00/0.98	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
C16 [35]	452, 452, 903	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00



(A) 15, 9 (B) 15, 11 (C) 18, 9 (D) 18, 11 (E) 21, 9 (F) 21, 11 (G) 24, 9 (H) 24, 11 (I) 27, 9 (J) 27, 11

FIGURE 8.19: Classification results of **PU** for different number of dimensions (15, 18, 21, 24, 27 dimensions selected using PCA) with different number of patch sizes (9×9, and 11×11).

TABLE 8.19: Kappa, Overall and Average accuracy for **PU** with different number of bands (15, 18, 21, 24, 27) and different number of patch sizes (9×9 and 11×11).

Method	**15 Bands**		**18 Bands**		**21 Bands**		**24 Bands**		**27 Bands**	
	9×9	**11×11**	**9×9**	**11×11**	**9×9**	**11×11**	**9×9**	**11×11**	**9×9**	**11×11**
PCA	56.97	99.61	58.56	99.69	56.77	99.76	59.40	99.77	64.80	99.68
	61.93	99.71	63.39	99.77	61.74	99.82	63.62	99.83	68.73	99.76
	44.55	99.54	47.32	99.55	46.56	99.69	55.41	99.64	57.55	99.59
iPCA	39.35	0.00	67.85	99.11	48.34	99.40	55.18	98.77	60.45	98.93
	46.43	43.59	71.34	99.33	54.30	99.55	60.02	99.07	64.47	99.20
	31.76	11.11	65.84	98.72	37.98	99.11	43.91	98.59	51.85	98.98
SPCA	51.37	19.39	46.13	99.18	62.65	99.70	56.53	99.45	60.21	99.44
	56.45	41.50	52.53	99.38	66.50	99.78	61.28	99.59	64.74	99.57
	42.69	15.70	34.69	98.95	55.65	99.64	46.94	99.25	47.46	99.18
ICA	0.00	98.78	0.00	98.96	0.00	99.18	0.00	99.57	0.00	99.30
	17.81	99.08	17.81	99.21	17.81	99.38	17.81	99.67	17.81	99.47
	7.69	98.28	7.69	98.85	7.69	99.11	7.69	99.53	7.69	99.25
SVD	45.29	0.00	50.72	97.13	56.33	99.08	49.49	0.00	54.06	98.57
	52.57	43.59	57.06	97.83	61.17	99.30	56.10	43.59	59.59	98.92
	33.28	11.11	40.52	97.48	50.77	98.73	37.56	11.11	45.42	98.51

TABLE 8.20: Class Names, Total Samples, Train, Validation and Test Sample numbers (Tr, Val, Te) along with the Statistical Test (Precision (P_r), Recall (R_c) & F1-Score (F1)) results for **PU** with PCA as dimensional reduction method over two window sizes ($W_1 = 9 \times 9$ and $W_2 = 11 \times 11$).

Class	Tr, Val, Te	15 Bands			18 Bands			21 Bands			24 Bands			27 Bands		
		P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1	P_r	R_c	F1
		W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2			W_1/W_2		
C1 [36]	1657, 1658, 3316	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C2 [37]	4662, 4663, 9324	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C3 [38]	525, 525, 1049	0.99/0.98	0.99/0.98	0.99/0.98	1.00/0.99	0.88/0.99	0.93/0.99	0.99/1.00	0.96/0.98	0.98/0.99	0.98/1.00	0.99/0.99	0.99/0.99	0.99/0.99	0.99/0.98	0.99/0.99
C4 [39]	766, 766, 1532	1.00/1.00	0.99/1.00	1.00/1.00	1.00/1.00	0.99/0.99	0.99/1.00	1.00/1.00	1.00/0.99	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
C5 [40]	336, 336, 673	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C6 [41]	1258, 1257, 2514	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	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00
C7 [42]	333, 332, 665	1.00/0.99	1.00/1.00	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	0.99/1.00	1.00/1.00	1.00/1.00	0.99/0.99	1.00/1.00	0.99/1.00	0.99/0.99	1.00/1.00	1.00/1.00
C8 [43]	921, 920, 1841	0.99/0.99	0.99/0.99	0.99/0.99	0.93/0.99	1.00/0.99	0.96/0.99	0.98/0.99	0.99/1.00	0.99/0.99	1.00/0.99	0.98/1.00	0.99/0.99	0.99/0.99	0.99/0.99	0.99/0.99
C9 [44]	236, 237, 474	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00	0.99/0.99	1.00/1.00	1.00/1.00	1.00/1.00	1.00/1.00

In all the experiments, we evaluated the performance of our proposed model for a set of experiments that are initially analyzed using several dimensionality reductions approaches for hybrid CNN and assessed the performance against a different number of spectral bands (15, 18, 21, 24, and 27) extracted through PCA, iPCA, SPCA, SVD, and ICA methods. Later we examined the effect of input window size on the classification performance of the proposed model by choosing two different patch sizes (9×9 and 11×11).

The experimental results on benchmark HSI datasets are presented in Tables 8.17, 8.15 and 8.19 and Figures 8.17-8.19. From these results, one can conclude that for all the datasets, the proposed model performed significantly better with PCA as compared to the other well-known dimensionality reduction methods. However, from the experimental results, one can observe that the κ, OA, AA values remain almost the same with an increasing number of spectral bands extracted through dimensionality reduction techniques.

The classification performance of CNN-based HSIC models also relies on the input window size. If the patch size is too small, it decreases the inter-class diversity in samples and if the patch size is set larger then it may take in the pixels from various classes, hence, both cases result in misclassification. We evaluated the hybrid model against two window sizes i.e. $W_1 = 9 \times 9$ and $W_2 = 11 \times 11$. From the experimental results, it can be observed that there is a slight improvement in the classification results with increased window size for both Indian Pines (IP) and Salinas Full Scene (SFS) datasets. However, in the case of the Pavia University dataset, one can notice a considerable enhancement in the classification accuracy with 11×11 window patch as compared to 9×9.

For comparison purposes, the Hybrid model is compared with various state-of-the-art frameworks published in recent years. From the experimental results presented in Table **??** one can interpret that Hybrid CNN has obtained results comparable to the state-of-the-art frameworks and to some extent outperformed with respect to the other models. The comparative frameworks used in this section include Support Vector Machine (SVM) without any optimization. SVM is one of the most widely used classifiers for HSIC.

Moreover, the SVM-Grid approach was also tested which was optimized by the stochastic gradient descent algorithm. Multi-layer Perceptron (MLP) model with 4 fully connected layers with dropout is used for comparative purposes. The number of layers is set heuristically. MLP is considered as a baseline for many deep learning models. The work proposed by Sharma [357] consists of a 2D CNN model that is built upon 2D convolutional operations with band selection as preprocessing. The idea was initially proposed for HS imaging-based face detection and recognition. Liu et.al. [358] proposed a semi-supervised CNN model for HSIC. Liu's model consists of convolutional operation, clean and corrupted encoder, and decoder. The work Hamida et. al. [359] proposed a 3-D CNN model for HSIC that consists of four 3-D convolutional layers. The work [360] proposed a fully convolutional network

that doesn't consider any subsampling layer with arbitrary sizes. The model proposed by Lee et. al. consists of two 3-D and eight 2-D convolutional layers. The work proposed by Chen et.al. [352] deployed a simple and compact 3-D CNN model consists of three 3-D convolutional layers. The work [361] proposed a lightweight 3-D CNN model having two 3-D convolutional layers and a fully connected layer for HSIC. Li's work is different from other state-of-the-art 3-D CNN models. In this work, Li et.al. proposed the use of fixed spatial size 3-D convolutional layers with slight changes in spectral depth. Multi-scale-3D-CNN [225], a fast and compact 3D-CNN (FC-3D-CNN) [117] and Hybrid Depth-Separable Residual Network [362].

All the comparative models are being trained as per the settings mentioned in their respective papers except for the number of dimensions and patch size (i.e., 15 dimensions selected using PCA, and 11×11 path size is used for experimental purposes). The experimental results listed in Table 8.21 show that Hybrid CNN has significantly improved the classification results as compared to the other methods with fewer convolutional layers, number of filters, number of epochs, and even a small number of training samples. Moreover, the loss and accuracy trend between 3D and hybrid CNN is shown in Figure 8.20.

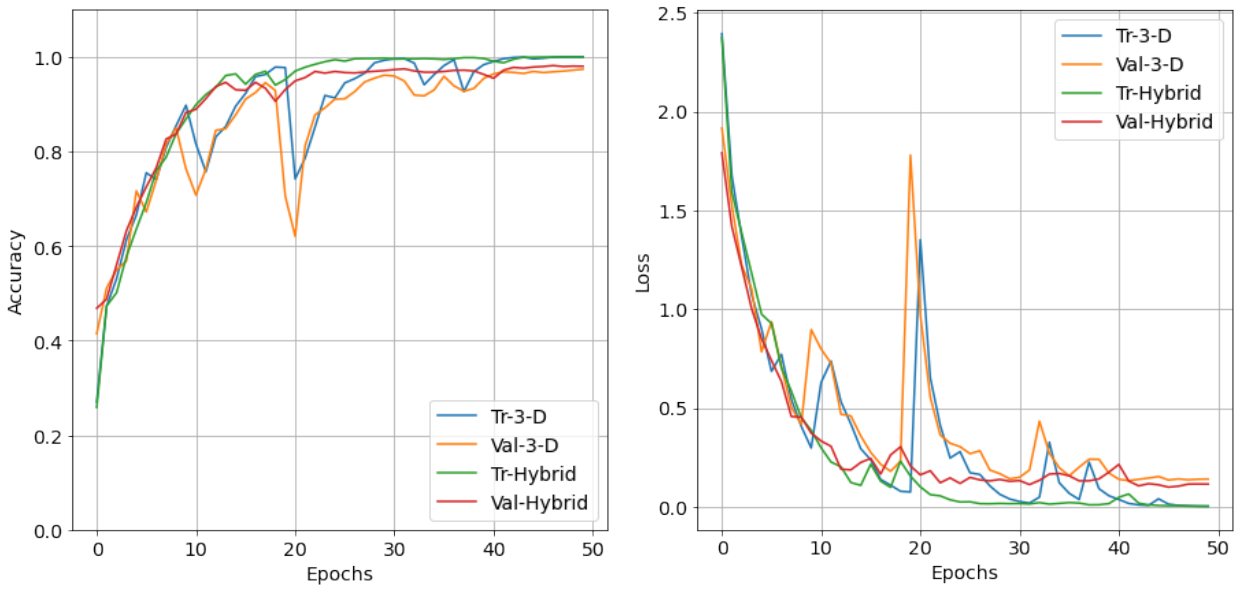


FIGURE 8.20: Accuracy and Loss for training and validation sets for FC-3D-CNN and Hybrid CNN model. The significant improvement in convergence can be observed.

TABLE 8.21: Comparative evaluations with State-of-the-art methods while considering 11×11 Spatial dimensions with even less number of training samples. Where SVM stands for Support Vector Machine, MLP stands for Multi-Layer Perceptron, MS-3D-CNN stands for Multi-Scale 3D CNN, FC-3D-CNN stands for fast and compact 3D CNN, SyCNN stands for Synergistic CNN model, and rest abbreviations are as follows: Simple Synergistic CNN (SyCNN-S), Deep SyCNN with Data Interaction Module (SyCNN-D), and Deep SyCNN-attention Network (SyCNN-ATT).

Methods	Indian Pines			Salinas Full Scene		
	OA	AA	Kappa	OA	AA	Kappa
SVM	85.30	79.03	83.10	92.95	94.60	92.11
SVM-Grid	87.93	88.02	86.20	67.39	45.89	62.80
MLP	87.57	89.07	85.80	79.79	67.37	77.40
3D-CNN	82.62	76.51	79.25	85.00	89.63	83.20
SyCNN-S [363]	95.90	97.84	95.30	97.44	98.46	97.20
SyCNN-D [363]	96.13	98.08	95.60	97.76	98.95	97.50
SyCNN-ATT [363]	97.31	98.43	96.90	98.92	99.35	98.80
2D-CNN [221]	80.27	68.32	75.26	96.34	94.36	95.93
MS-3D-CNN [225]	91.87	92.21	90.80	94.69	94.03	94.10
FC-3D-CNN [117]	98.20	96.46	97.95	98.06	98.80	97.85
Xie et.al. [357]	95.64	96.01	95.10	93.35	91.88	92.60
Liu et.al. [358]	89.56	89.32	88.10	84.27	79.10	82.50
Hamida [359]	86.99	90.16	85.20	76.22	62.82	73.10
Lee et.al. [360]	87.87	83.42	86.10	84.14	73.27	82.30
Chen et.al. [352]	93.20	95.51	92.30	86.83	92.08	85.50
Li [361]	94.22	96.71	93.40	88.62	86.84	87.40
Zhao et.al. [362]	95.86	96.08	95.09	98.89	98.88	98.85
Proposed	**98.26**	**96.94**	**98.02**	**99.89**	**99.90**	**99.97**

8.8 Concluding Remarks for Artifacts of Dimension Reduction on Hybrid CNN

HSIC is a challenging task due to the spectral mixing effect which induces high intra-class variability and inter-class similarity. 2D CNNs are utilized for spatial feature extraction and classification, whereas several variants of 3D CNN are used for joint spectral-spatial feature extraction and classification. However, 3D CNNs are computationally complex and 2D CNN alone cannot efficiently extract discriminating spectral features. Therefore, to overcome these challenges, this chapter proposed a Hybrid CNN feature hierarchy that provided outstanding classification results on benchmark HSI datasets.

8.9 Experimental Results for Spectral Angle Mapper for Spatial-Spectral Classification

The performance of the FSAM-AL pipeline is validated on five benchmark HSI datasets acquired by two different sensors, e.g., Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Reflective Optics System Imaging Spectrometer (ROSIS). These datasets include SLA, SA, KSC, PU, and PC.

We evaluated the FSAM pipeline against four different classifiers: extreme learning machine (ELM) [443], Support Vector Machine (SVM), k-nearest neighbor (kNN), and Ensemble Learning (EL). These classifiers are chosen because they have been extensively studied in the literature for HSIC and rigorously utilized for comparison purposes. Furthermore, our goal is to show that the proposed method can work well with a diverse set of classifiers.

To further validate the real-time applicability of FSAM, we compared it against four benchmark sample selection methods, namely: Random Sampling (RS), Mutual Information (MI), Breaking Ties (BT), and Modified Breaking Ties (MBT).

1. **Random Sampling (RS)** [6, 409] method relies on the random selection of the samples without considering any specific conditions.

2. **Mutual Information (MI)** [413] of two samples is a measure of the mutual dependence between the two samples.

3. **Breaking Ties (BT)** [414] relies on the smallest difference of the posterior probabilities for each sample. In multiclass settings, BT can be applied by calculating the difference between the two highest probabilities. As a result, BT finds the samples minimizing the distance between the first two most probable classes. The BT method generally focuses on the boundaries comprising many samples, possibly disregarding boundaries with fewer samples.

4. **Modified Breaking Ties (MBT)** [415, 416] includes more diversity in the sampling process as compared to BT. The samples are selected by maximizing the probability of the largest class for each class. MBT takes into account all the class boundaries by cyclically conducting the sampling, making sure that the MBT does not get trapped in any class whereas BT could be trapped in a single boundary.

In all experiments, the initial training size is set as 100 samples from an entire HSI data. In each iteration, the size of the training set increases with h = 1% actively selected samples by the FSAM pipeline. The best part of FSAM is that there are no hyper-parameters that need to be tuned except classification methods. In ELM, the hidden neurons are systematically

selected from the range of [1 - 500]. Similarly, in kNN, the nearest neighbors are set to k = [2 - 20], SVM is tested with polynomial kernel function, and ensemble learning classifiers are trained using a tree-based model with [1 - 100] a number of trees. All such parameters are carefully tuned and optimized during the experimental setup. All these experiments are carried out using MATLAB (2014b) on an Intel Core i5 3.20 GHz CPU with 12 GB of RAM.

Here we performed a set of experiments to evaluate the FSAM pipeline using both ROSIS and AVIRIS sensors datasets. Evaluating ROSIS sensor datasets is a more challenging classification problem dominated by complex urban classes and nested regions than AVIRIS. Here we evaluate the influence of the number of labeled samples on the classification performance achieved by several classifiers. Figures 8.21 and 8.22 shows the overall and kappa (κ) accuracy as a function of the number of labeled samples obtained by FSAM, i.e., fuzziness and SAM diversity-based active selection of most informative and diverse samples in each iteration. These labeled samples were selected by machine–machine interaction which significantly reduces the cost in terms of labeled collection through human supervisors which is the key aspect of automatic AL methods. The plots are shown in Figures 8.21 and 8.22 and generated based on only selected samples in contrast to the entire population which reveals clear advantages of using fewer labeled samples for the FSAM pipeline.

From Figures 8.21 and 8.22, it can be observed that FSAM greatly improved the accuracy. The results also reveal that SVM and LB outperformed other classifiers in most cases, whereas, as expected, KNN provides lower classification accuracy than SVM and LB, since the candidates are more relevant when the samples are acquired from the class boundaries. Furthermore, it can also be observed that SVM always performed better than KNN, ELM, and ensemble learning classifiers. ELM could perform better with more hidden neurons on more powerful machines. For instance, when the 2% of labeled samples were used, the performance has been significantly increased in contrast to the 1% of actively selected samples. These observations confirm that FSAM can greatly improve the results obtained by different classifiers based on a small portion of the entire population, i.e., the classifiers trained using a limited number of selected labeled samples can produce better generalization performance rather than selecting the bulk amount of label training samples.

It is perceived from Figures 8.21 and 8.22 that by including the samples back to the training set, the classification results are significantly improved for all the classifiers. Moreover, it can be seen that SVM and ELM classifiers are more robust than ensemble and KNN classifiers. For example, with 1% actively selected samples in the ELM classifier case, only 2% difference in classification with a different number of samples can be observed, however, for the KNN and SVM classifiers, the difference is quite high. Similar observations can be made for ensemble models.

In order to present the classification results in geographical fashioned for both ROSIS

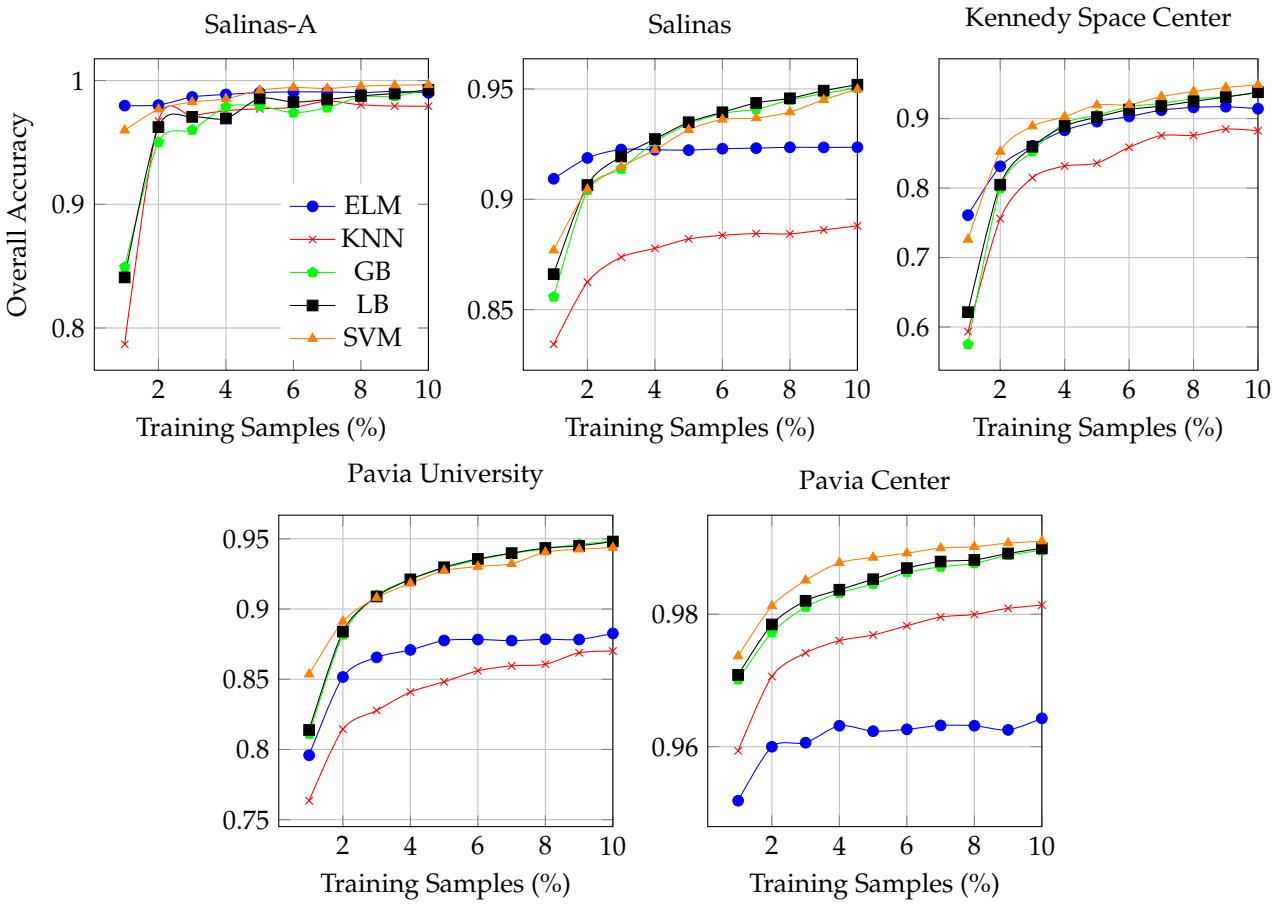


FIGURE 8.21: Overall accuracy with a different number of training samples (%) selected in each iteration from different datasets. It is perceived from the above figure that by including the samples back to the training set, the classification results are significantly improved for all the classifiers. Moreover, it can be seen that SVM and ELM classifiers are more robust. For example, with 2% actively selected samples in the ELM classifier case, only 2% difference in the classification with a different number of samples can be observed, however, for the KNN and SVM classifiers, the difference is quite high.

and AVIRIS sensors datasets, Figures 8.23–8.27 shows ground truths segmentation of all experimental datasets used in this work. These ground truths are generated using 2% of actively selected samples by the FSAM pipeline. In all the experiments, we provide the number of labeled training samples and the test samples which indicate the number of true versus estimated labels used in the experiments. It can be observed from the listed results, that our proposed fuzziness and diversity-based active labeled sample selection pipeline is quite robust as it achieved higher classification results which are way better or at least comparable with several state-of-the-art AL methods.

To better analyze the performance of FSAM on ROSIS and AVIRIS datasets, Table 8.22 shows the statistical significance in terms of recall, precision, and F1-score tests. The experiments shown in Table 8.22 are performed with 2% of actively selected labeled samples

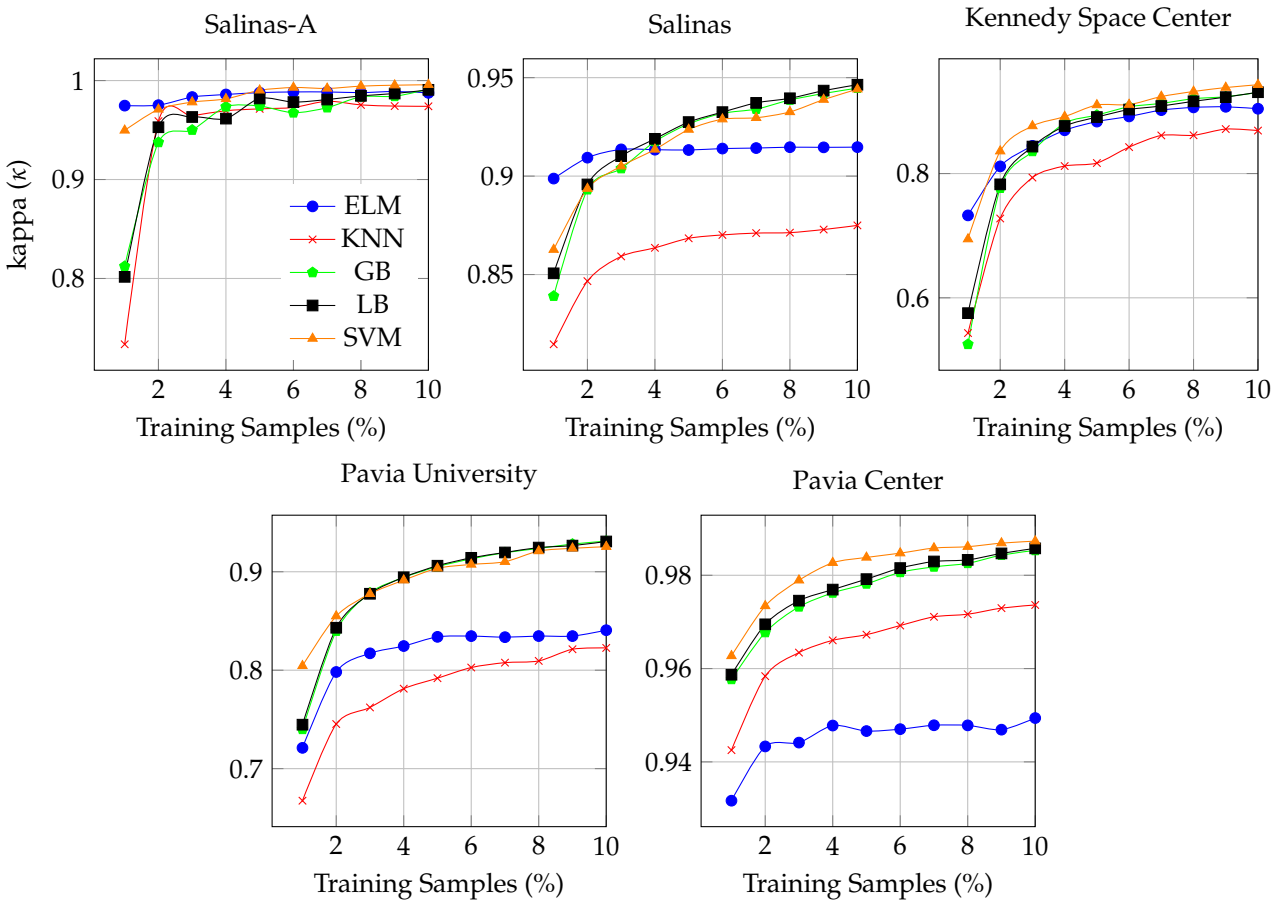


FIGURE 8.22: Kappa (κ) accuracy with different number of training samples (%) selected in each iteration from Salinas-A, Salinas, Kennedy Space Center, Pavia University, and Pavia Center datasets respectively. It is perceived from the above figure that by including the samples back to the training set, the classification results in terms of kappa κ are significantly improved for all the classifiers. Moreover, it can be seen that SVM and ELM classifiers are more robust than ensemble and KNN classifiers. For example, with 2% actively selected samples in the ELM classifier case, only 2% difference in the classification with a different number of samples can be observed, however, for the KNN and SVM classifiers, the difference is quite high. Similar observations can be made for ensemble learning models.

(A) Band (B) GT (C) Train (D) Test (E) SVM (F) KNN (G) GB (H) LB (I) ELM

FIGURE 8.23: **SLA**: (**a**): Ground Band, (**b**): True Ground Truths, (**c**): Training Ground Truths, (**d**): Test Ground Truths, and ground truths predicted by (**e**): SVM, (**f**): KNN, (**g**): GB, (**h**): LB, and (**i**): ELM classifier with 2% of selected training samples.

from each class for all experimental datasets. Table 8.22 is produced to support the results shown in Figures 8.21–8.27 for both AVIRIS and ROSIS sensor datasets. The global recall, precision, and F1-score for each classifier of these results are obtained using 5 Monte Carlo

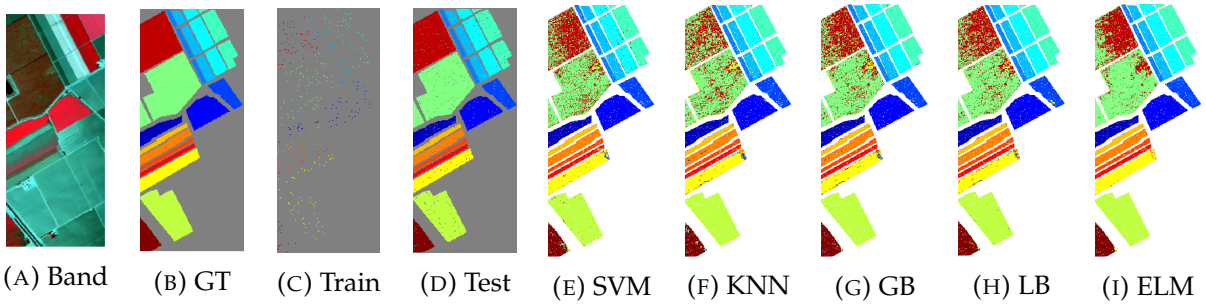


FIGURE 8.24: **SA**: (**a**) Ground Band, (**b**): True Ground Truths, (**c**): Training Ground Truths, (**d**): Test Ground Truths, and ground truths predicted by (**e**): SVM, (**f**): KNN, (**g**): GB, (**h**): LB, and (**i**): ELM classifier with 2% of selected training samples.

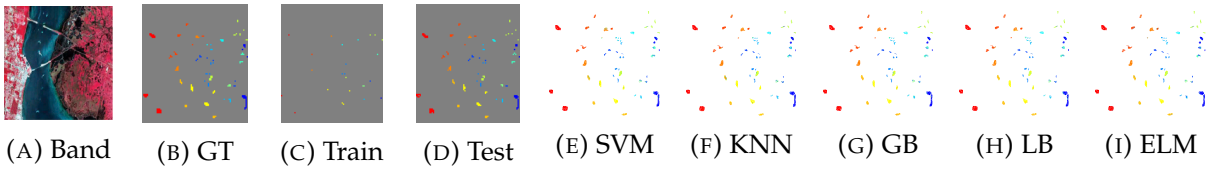


FIGURE 8.25: **KSC**: (**a**) Ground Band, (**b**): True Ground Truths, (**c**): Training Ground Truths, (**d**): Test Ground Truths, and ground truths predicted by (**e**): SVM, (**f**): KNN, (**g**): GB, (**h**): LB, and (**i**): ELM classifiers with 2% of selected training samples.

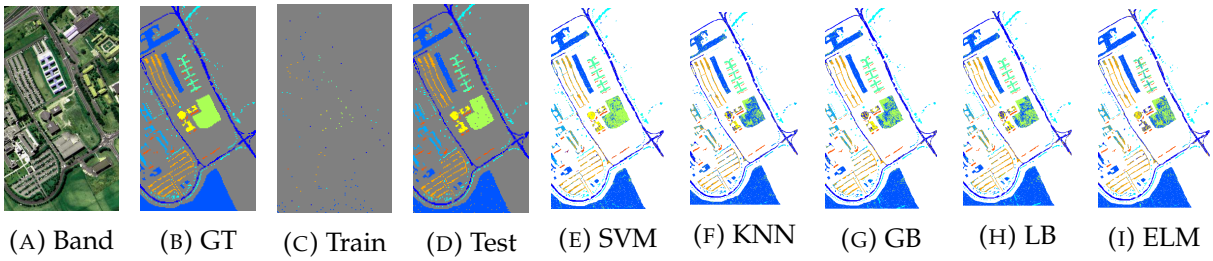


FIGURE 8.26: PU : (**a**) Ground Band, (**b**): True Ground Truths, (**c**): Training Ground Truths, (**d**): Test Ground Truths, and ground truths predicted by (**e**): SVM, (**f**): KNN, (**g**): GB, (**h**): LB, and (**i**): ELM classifier with 2% of selected training samples.

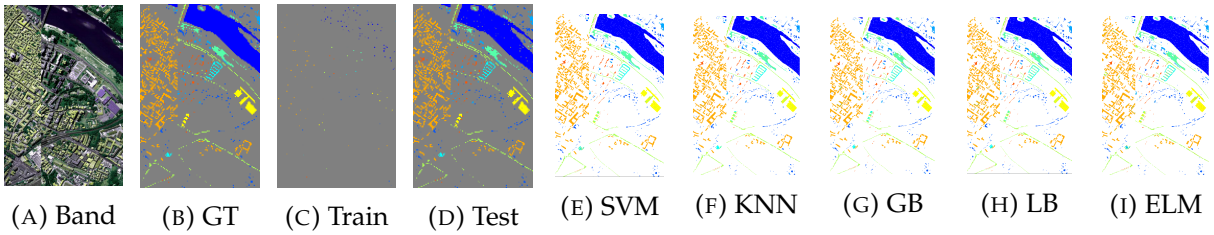


FIGURE 8.27: **PC**: (**a**) Ground Band, (**b**): True Ground Truths, (**c**): Training Ground Truths, (**d**): Test Ground Truths, and ground truths predicted by (**e**): SVM, (**f**): KNN, (**g**): GB, (**h**): LB, and (**i**): ELM classifier with 2% of selected training samples.

runs. Furthermore, these Tables show the statistical significance of FSAM in terms of recall, precision, and F1-score with the 99% confidence interval. The obtained values indicate the ability of FSAM to correctly identify the unseen samples in which each classifier was trained

on a very small amount of labeled training samples. For any good model, precision, recall, and F1-score values should be greater than 80% on average, and in our case, these values are almost above 80% for all experimental datasets and all classifiers, demonstrating that the proposed FSAM-AL pipeline is not classifier sensitive.

TABLE 8.22: Statistical applicability of our proposed FSAM samples selection method. Each classifier is trained with 2% of actively selected training samples.

Tests	ELM	KNN	GB	LB	SVM
			Salinas-A Dataset.		
Recall	0.9852 ± 0.0037	0.9489 ± 0.0528	0.9644 ± 0.0281	0.9650 ± 0.0315	0.9855 ± 0.0092
Precision	0.9903 ± 0.0029	0.9567 ± 0.0352	0.9649 ± 0.0275	0.9672 ± 0.0287	0.9885 ± 0.0071
F1 Score	0.9875 ± 0.0031	0.9459 ± 0.0609	0.9639 ± 0.0270	0.9654 ± 0.0300	0.9867 ± 0.0076
			Salinas Dataset.		
Recall	0.9544 ± 0.0032	0.9247 ± 0.0162	0.9551 ± 0.0153	0.9580 ± 0.0135	0.9583 ± 0.0115
Precision	0.9584 ± 0.0030	0.9189 ± 0.0117	0.9596 ± 0.0127	0.9603 ± 0.0117	0.9642 ± 0.0091
F1 Score	0.9552 ± 0.0026	0.9225 ± 0.0129	0.9570 ± 0.0138	0.9588 ± 0.0124	0.9611 ± 0.0101
			Kennedy Space Center Dataset.		
Recall	0.8220 ± 0.0518	0.7513 ± 0.0739	0.8158 ± 0.0839	0.8159 ± 0.0790	0.8546 ± 0.0579
Precision	0.8445 ± 0.0423	0.8375 ± 0.0638	0.8315 ± 0.0751	0.8344 ± 0.0683	0.8596 ± 0.0509
F1 Score	0.8260 ± 0.0469	0.7491 ± 0.0760	0.8183 ± 0.0815	0.8195 ± 0.0775	0.8542 ± 0.0533
			Pavia University Dataset.		
Recall	0.7782 ± 0.0283	0.7954 ± 0.0346	0.8838 ± 0.0391	0.8856 ± 0.0376	0.8858 ± 0.0287
Precision	0.8708 ± 0.0229	0.9508 ± 0.0090	0.9165 ± 0.0247	0.9146 ± 0.0252	0.8929 ± 0.0203
F1 Score	0.8107 ± 0.0255	0.8122 ± 0.0316	0.8973 ± 0.0336	0.8976 ± 0.0327	0.8886 ± 0.0241
			Pavia Center Dataset.		
Recall	0.8913 ± 0.0073	0.9281 ± 0.0151	0.9493 ± 0.0124	0.9509 ± 0.0120	0.9568 ± 0.0115
Precision	0.9187 ± 0.0066	0.9125 ± 0.0136	0.9469 ± 0.0112	0.9489 ± 0.0107	0.9572 ± 0.0119
F1 Score	0.8999 ± 0.0064	0.9248 ± 0.0129	0.9480 ± 0.0116	0.9498 ± 0.0112	0.9568 ± 0.0117

The most advanced developments in AL are single-pass context and hybrid AL. These techniques combine the concepts of incremental and adaptive learning from the field of online and traditional machine learning. These advancements have resulted in a substantial number of AL methods. The most classical and well studied AL methods include, for example, the works [444, 445] focused on online learning. These works are specifically designed for an online single-pass setting in which the data stream samples arrive continuously, thus, do not allow classifier re-training. Furthermore, these works focused on close concepts of conflict and ignorance. Conflict models how close a query point is to the actual class boundary and ignorance represents the distance between already seen training samples and a new sample.

Similar works proposed in [446, 447] focused only on early AL strategies such as early-stage experimental design problems. The TED method was proposed to select the samples using the robust AL method incorporated with structured sparsity-inducing norms to relax

the NP-hard objective of the convex formulation. Thus, these works only focused on selecting an optimal set of initial samples to kick-start the AL. However, the superiority of our proposed FSAM pipeline is that it shows state-of-the-art performance independent of how the initial labeled training samples are selected. Such methods can easily be integrated into the works which utilize the decision boundary-based sample selection methods.

A novel tri-training semi-supervised HSIC method based on regularized local discriminant embedding feature extraction (RLDE) was proposed in [448]. In this work, the RLDE process is used for an optimal number of feature extraction to overcome the limitation of singular values and over-fitting of local Fisher discriminant analysis and local discriminant embedding. At a later stage, the AL method is used to select the informative samples from the candidate set. This work solves the singularity issues of LDA, however, this may include the redundant samples back to the training set which does not provide any new information to the classifier.

Spatial-spectral multiview 3D Gabor inspired AL for HSIC method was proposed in [449]. Trivial multiview AL methods can make a comprehensive analysis of both sample selection and object characterization in active learning by using several features of multiple views. However, multiview cannot effectively exploit spatial-spectral information by respecting the 3D nature of HSI, therefore, the sample selection method in multiview is only based on the disagreement of multiple views. To overcome such problems, J. Hu, et al. [449] proposed a two-step 3D Gabor-inspired multiview method for HSIC. The first step consists of the view generation step, in which a 3D Gabor filter was used to generate multiple cubes with limited bands and utilize the features assessment strategies to select cubes for constructing views. In a second stage, an AL method was presented which used both external and internal uncertainty estimation of views. More specifically, posterior probability distribution was used to learn the internal uncertainty of each independent view, and external uncertainty was computed using inconsistency between the views.

Of course, the frameworks proposed in the above papers can be easily integrated with our proposed FSAM sample selection method instead of selecting the samples based on uncertainty or tri-training methods. We initialize our active learning method from 100 number of randomly selected labeled training samples and we experimentally demonstrate that randomly increasing the size of the training set slightly increases the accuracy nevertheless the classifiers become computationally complex. Therefore, at the first step, we decided to separate the set of misclassified samples that have higher fuzziness values (samples fuzziness magnitude between 0.7–1.0). We then select a specific percentage of misclassified samples that have higher fuzziness to compute the spectral angle among the reference training samples. We then fused a specific percentage of selected samples with the original training set to retrain the classifier from scratch for better generalization and classification performance

on those samples which were initially misclassified by the same classifier.

More specifically, FSAM has been rigorously investigated through comparison against some significant works recently published in the HSI classification area, adopting different sample selection methods such as Random Sampling (RS), Mutual Information (MI), Breaking Ties (BT), Modified Breaking Ties (MBT), uncertainty, and fuzziness. This comparison is based on the Botswana HSI acquired by the NASA EO-1 Satellite Hyperion sensor.

The experiments are based on five Monte Carlo runs with 100 initial training samples selected from this dataset. In each iteration, the training set size has been increased of 50 samples selected by a specific method among the ones to be compared. The results thus obtained are presented in the Tables 8.23-8.27. Based on such results, we can argue that the FSAM pipeline outperforms the other solutions taken into account in these experiments. This is due to the dual soft thresholding method for the selection of the most informative as well as spatially heterogeneous labeled training samples. Furthermore, another benefit of the FSAM solution is that it systematically selects the most informative but least redundant labeled training samples by machine-machine interaction without involving any supervisor, automatically, while the other AL frameworks need that a supervisor selects the samples at each iteration, manually.

TABLE 8.23: Kappa (κ) accuracy obtained by **SVM Classifier** with different number of training samples selected in each iteration from BS dataset with different sample selection methods from literature.

Sample Selection Method	**Number of Training Samples**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
	Kappa Accuracy									
Random Sampling [409]	0.8156	0.8483	0.8738	0.8886	0.9005	0.9101	0.9170	0.9151	0.9163	0.9221
Mutual Information [411–413]	0.8149	0.8437	0.8602	0.8798	0.8863	0.9002	0.9108	0.9217	0.9195	0.9302
Breaking Ties [414]	0.8163	0.8316	0.8401	0.8561	0.8778	0.8919	0.9008	0.9014	0.9087	0.9128
Modified Breaking Ties [415, 416]	0.8156	0.8522	0.8563	0.8893	0.9007	0.9040	0.9068	0.9136	0.9138	0.9103
Fuzziness [6]	0.8174	0.8129	0.8422	0.8648	0.8755	0.8934	0.8989	0.8986	0.9156	0.9119
FSAM	0.8167	**0.8749**	**0.9027**	**0.9091**	**0.9493**	**0.9556**	**0.9668**	**0.9788**	**0.9928**	**0.9984**

By the Botswana dataset, we experimentally demonstrated that FSAM outperforms all other sample selection methods, i.e., RS, MI, BT, MBT, and Fuzziness in terms of accuracy, starting from the same classifiers and the same number of labeled training samples as shown in Tables 8.23-8.27. Furthermore, all these sample selection methods are more often subjective and tend to bring redundancy into the classifiers. Moreover, it reducing the generalization performance of the classifiers. More specifically, the number of samples required to learn a model in FSAM can be much lower than the number of selected samples. In such scenarios, there is a risk, however, that the learning model may get overwhelmed because of the uninformative or spatially miscellaneous samples selected by the query function.

TABLE 8.24: Kappa (κ) accuracy obtained by **ELM Classifier** with different number of training samples selected in each iteration from BS dataset with different sample selection methods from literature.

Sample Selection Method	**Number of Training Samples**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
	Kappa Accuracy									
Random Sampling [409]	0.8094	0.8253	0.8364	0.8564	0.8648	0.8730	0.8919	0.8958	0.9063	0.9140
Mutual Information [411–413]	0.8051	0.8246	0.8430	0.8538	0.8699	0.8772	0.8881	0.8962	0.8983	0.9070
Breaking Ties [414]	0.8051	0.8174	0.8392	0.8607	0.8680	0.8744	0.8819	0.8963	0.8927	0.9022
Modified Breaking Ties [415, 416]	0.7961	0.8216	0.8563	0.8654	0.8718	0.8769	0.8896	0.9012	0.9081	0.9149
Fuzziness [6]	0.7958	0.8224	0.8463	0.8513	0.8606	0.8733	0.8780	0.8841	0.9008	0.9083
FSAM	0.8021	**0.8385**	**0.8544**	**0.8846**	**0.8923**	**0.8968**	**0.9213**	**0.9355**	**0.9492**	**0.9551**

TABLE 8.25: Kappa (κ) accuracy obtained by **KNN Classifier** with different number of training samples selected in each iteration from BS dataset with different sample selection methods from literature.

Sample Selection Method	**Number of Training Samples**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
	Kappa Accuracy									
Random Sampling [409]	0.7854	0.8145	0.8158	0.8428	0.8547	0.8556	0.8603	0.8640	0.8695	0.8757
Mutual Information [411–413]	0.7854	0.8029	0.8154	0.8342	0.8485	0.8519	0.8592	0.8653	0.8727	0.8814
Breaking Ties [414]	0.7854	0.8205	0.8330	0.8466	0.8469	0.8541	0.8626	0.8731	0.8760	0.8757
Modified Breaking Ties [415, 416]	0.7854	0.8396	0.8463	0.8474	0.8584	0.8635	0.8685	0.8749	0.8813	0.8841
Fuzziness [6]	0.7854	0.8248	0.8298	0.8445	0.8529	0.8584	0.8628	0.8678	0.8702	0.8771
FSAM	0.7854	**0.8369**	**0.8512**	**0.8842**	**0.8919**	**0.9189**	**0.9236**	**0.9469**	**0.9584**	**0.9625**

TABLE 8.26: Kappa (κ) accuracy obtained by **GB Classifier** with different number of training samples selected in each iteration from BS dataset with different sample selection methods from literature.

Sample Selection Method	**Number of Training Samples**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
	Kappa Accuracy									
Random Sampling [409]	0.8140	0.8272	0.8342	0.8358	0.8479	0.8597	0.8649	0.8703	0.8675	0.8717
Mutual Information [411–413]	0.8139	0.7941	0.8118	0.8406	0.8620	0.8608	0.8687	0.8750	0.8787	0.8712
Breaking Ties [414]	0.8139	0.7875	0.8318	0.8355	0.8625	0.8691	0.8795	0.8895	0.8935	0.8928
Modified Breaking Ties [415, 416]	0.8139	0.8067	0.8404	0.852	0.8570	0.8612	0.8711	0.8772	0.8750	0.8795
Fuzziness [6]	0.8139	0.8470	0.8488	0.8524	0.8559	0.8658	0.8692	0.8755	0.8782	0.8853
FSAM	0.8140	0.8054	**0.8605**	**0.8852**	**0.9060**	**0.9268**	**0.9247**	**0.9280**	**0.9455**	**0.9592**

8.10 Concluding Remarks for Spectral Angle Mapper for Spatial-Spectral Classification

The classification of multiclass spatial-spectral HSI with a small labeled training sample size is a challenging task. To overcome this problem, this chapter introduces a customized AL

TABLE 8.27: Kappa (κ) accuracy obtained by **LB Classifier** with different number of training samples selected in each iteration from BS dataset with different sample selection methods from literature.

Sample Selection Method	**Number of Training Samples**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
	Kappa Accuracy									
Random Sampling [409]	0.8074	0.8105	0.8275	0.8359	0.8419	0.8429	0.8482	0.8620	0.8699	0.8768
Mutual Information [411–413]	0.8073	0.8212	0.83151	0.8366	0.8409	0.8442	0.8564	0.854	0.8571	0.8565
Breaking Ties [414]	0.8074	0.8155	0.8300	0.8293	0.8337	0.8467	0.8513	0.8490	0.8582	0.8679
Modified Breaking Ties [415, 416]	0.8074	0.8303	0.8402	0.8422	0.8484	0.8649	0.8704	0.8813	0.8816	0.8794
Fuzziness [6]	0.8073	0.8126	0.8236	0.8411	0.8447	0.8586	0.8626	0.8684	0.8669	0.8709
FSAM	0.8073	0.8118	0.8226	**0.8779**	**0.8993**	**0.9094**	**0.9216**	**0.9402**	**0.9475**	**0.9588**

pipeline for HSI to reduce the sample selection bias while maintaining the data stability in the spatial domain. The proposed FSAM pipeline differs from traditional AL methods in three relevant aspects. First, instead of simply using the uncertainty of samples to select new samples, it utilizes the fuzziness measure associated with the confidence of the training model in classifying those samples correctly.

Second, it couples the samples' fuzziness with their diversity to select new training samples which simultaneously minimizes the error among the training samples while maximizing the spectral angle between the selected sample and the existing training samples. For FSAM, instead of measuring angle-based distances among all new samples and all existing training samples, a reference sample is selected from within the training set against which the diversity of the new samples is measured. This achieves the same goal while reducing the computational overhead as the size of the training set is always much smaller than the validation set which is the source of new samples.

Thirdly, the FSAM keeps the pool of new samples balanced, giving equal representation to all classes, which is achieved via softening the thresholds at run time. Experimental results on five benchmark datasets demonstrate that the FSAM leads to an increased predictive power regarding kappa (κ) and overall accuracy, precision, recall, and F1-Score parameters.

A comparison of FSAM with state-of-the-art sample selection method is performed, confirming that the FSAM is effective in terms of overall accuracy and κ, also with few training samples. However, the main drawback of SAM is spectral mixture problems, i.e., SAM assumes that the reference spectra chosen to classify the HSI represent the pure spectra. Such a problem occurs when the HSI is in low or medium spatial resolution. Furthermore, as we know, the surface of the earth is widely heterogeneous and complex, thus containing many mixed samples. The spectral confusion in samples can lead to overestimation or underestimation errors for spectral signatures.

Future research direction aims to address such limitations to classify low or mid-spatial resolution HSI's in a computationally efficient way. Further work will be directed toward

testing the FSAM pipeline in different analysis scenarios dominated by the limited availability of training samples a priori.

Chapter 9

Conclusion

The rich information contained in HSI data is a captivating factor that constitutes the utilization of HSI technology in real-world applications. HIS Classification (HSIC) is a challenging task due to high inter-class similarity, high intra-class variability, overlapping, and nested regions. Thus, advances in machine learning methods strengthen the deployment potentials of such technologies. Though 2D Convolutional Neural Networks (CNNs) have emerged as a viable approach for HSIC, 3D CNNs are a better alternative because accurate HSIC depends on both Spectral-Spatial information. However, 3D CNN can be highly computational complex due to its volume and spectral dimensions. Therefore, this Thesis proposed several Deep Learning as well as Active Learning (AI) methods for HSIC to overcome the aforesaid issues. Moreover, the comparative results are intensively compared with state of the art Deep Neural Networks (for instance, Auto-encoder (AE), Deep Belief Network (DBN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Transfer Learning (TL), Few-shot Learning (FSL), Active/Self Learning (AL/SL), and Data Augmentation (DA)) in a variety of learning schemes (specifically, supervised, semi-supervised and unsupervised learning). In addition, this thesis also analyzed the strategies to overcome the challenges of limited availability of training data like AI, Data Augmentation, Few-shot Learning (FSL), Transfer Learning, and Active Learning, etc.

Although the current HSIC techniques reflect a rapid and remarkable sophistication of the task, further developments are still required to improve the generalization capabilities. The main issue of deep neural network-based HSIC is the lack of labeled data. HSI data is infamous due to the limited availability of labeled data and deep neural networks demand a sufficiently large amount of labeled training data. Thus this dissertation discussed some and proposed some strategies to combat the aforesaid issues, however, significant improvements are still needed to efficiently utilize limited available training data. One direction to solve this problem could be to explore the integration of various learning strategies to cash in the joint benefits. One more way is to exploit a few-shot or K-shot learning approaches that can accurately predict the class labels with only a few labeled samples.

Moreover, there is a need to focus on the joint exploitation of spectral-spatial features of

HSI to complement classification accuracies achieved from the aforementioned HSIC frameworks as proposed in this dissertation. Another potential issue of HSIC is computationally efficient architectures which have been overcome in this dissertation.

The issue of the high computational complexity of deep neural networks is of paramount importance and it is crucial to implement parallel HSIC architectures to speed up the processing of deep neural networks to meet the computational stipulation of time-critical HSI applications. In this direction, high-performance computing platforms and specialized hardware modules like graphical processing units (GPUs) and field-programmable gate arrays (FPGAs) can be used to implement the parallel HSIC frameworks. Hence, to assimilate aforesaid aspects in the development of a new HSIC framework is to appropriately utilize the limited training samples while considering joint spectral-spatial features of HSI and maintaining the low computational burden.

Bibliography

1. Schneider, A. & Feussner, H. *Chapter 5 - Diagnostic Procedures* 87–220. https://doi.org/10.1016/B978-0-12-803230-5.00005-1 (Institute of Minimally Invasive Interdisciplinary Therapeutic Interventions (MITI), Technische Universität München (TUM), Biomedical Engineering in Gastrointestinal Surgery, London, 2017).
2. Ahmad, M., Khan, A. M. & Hussain, R. Graph-based spatial-spectral feature learning for hyperspectral image classification. *IET Image Processing* **11,** 1310–1316 (Dec. 2017.).
3. Qu, Y., Qi, H. & Kwan, C. Unsupervised Sparse Dirichlet-Net for Hyperspectral Image Super-Resolution. *CVPR'18, CoRR* **abs/1804.05042.** arXiv: 1804.05042. http://openaccess.thecvf.com/content_cvpr_2018/papers/Qu_Unsupervised_Sparse_Dirichlet-Net_CVPR_2018_paper.pdf (2018).
4. Ahmad, M, Bashir, A. K. & Khan, A. M. Metric similarity regularizer to enhance pixel similarity performance for hyperspectral unmixing. *Optik-International Journal for Light and Electron Optics* **140C,** 86–95. https://doi.org/10.1016/j.ijleo.2017.03.051 (2017).
5. Hughes, G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* **14,** 55–63. ISSN: 0018-9448 (1968).
6. Ahmad, M. *et al.* Fuzziness-based active learning framework to enhance hyperspectral image classification performance for discriminative and generative classifiers. *PLoS One* **13** (2018.).
7. Liu, C., He, L., Li, Z. & Li, J. Feature-Driven Active Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 341–354. ISSN: 0196-2892 (2018).
8. Mountrakis, G., Im, J. & Ogole, C. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensingn* **66,** 247–259 (2011).
9. Li, J., Bioucas-Dias, J. M. & Plaza, A. Spectral-Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields. *IEEE Transactions on Geoscience and Remote Sensing* **50,** 809–823 (2012.).

10. Xia, J., Peijun, D., Xiyan, H. & Chanussot, J. Hyperspectral Remote Sensing Image Classification Based on Rotation Forest. *IEEE Geoscience and Remote Sensing Letters* **11,** 239–243 (2014).

11. Pan, B., Shi, Z. & Xu, X. Hierarchical Guidance Filtering-Based Ensemble Classification for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 4177–4189 (2017).

12. Pan, B., Shi, Z. & Xia, X. R-VCANet: A New Deep-Learning-Based Hyperspectral Image Classification Method. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 1975–1986 (2017).

13. Tan, K. & Du, P. Hyperspectral Remote Sensing Image Classification Based on Support Vector Machine. *Journal of Infrared and Millimeter Waves* **27,** 123–128 (2013.).

14. Macdonald, J. S., Ustin, S. L., Schaepman & Michael, E. The contributions of Dr. Alexander F.H. Goetz to imaging spectrometry. *Remote Sensing of Environment* **113,** S2–S4. `https://doi.org/10.5167/uzh-23319` (2009).

15. Goetz, A. F. H., Vane, G, Solomon, J. E. & Rock, B. N. Imaging Spectrometry for Earth Remote Sensing. *Science* **228,** 1147–1153. ISSN: 0036-8075. eprint: `http://science.sciencemag.org/content/228/4704/1147.full.pdf`. `http://science.sciencemag.org/content/228/4704/1147` (1985).

16. Knaeps, E., Sterckx, S., Bollen, M., Trouw, K. & Houthuys, R. Operational Remote sensing Mapping of Estuarine suspended Sediment concentrations (ORMES) (Jan. 2006).

17. Vane, G *et al.* The airborne visible infrared imaging spectrometer (AVIRIS). *Remote Sensing of the Environment* **44,** 127–143. `https://doi.org/10.1016/0034-4257(93)90012-M` (1993).

18. Green, R. O. *et al.* Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of the Environment* **65,** 227–248. `https://doi.org/10.1016/S0034-4257(98)00064-9` (1998).

19. Chevrier, M, Bannari, A, Deguise, J. C., McNairn, H & Staenz, K. *Hyperspectal narrowwavebands for discriminating crop residue from bare soil* in *IEEE International Geoscience and Remote Sensing Symposium* **4** (2002), 2202–2204.

20. Levesque, J & Staenz, K. *A method for monitoring mine tailings revegetation using hyperspectral remote sensing* in *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium* **1** (2004), 578.

21. Cocks, T, Jenssen, R, Stewart, A, Wilson, I & Shields, T. *The HYMAP airborne hyperspectral sensor, The system, calibration and performance* in *EARSEL 1998. 1st EARSEL Workshop on Imaging Spectroscopy (M. Schaepman, D. SchlÃ¤pfer, and K.I. Itten, Eds.)* (1998), 37–42.

22. Lillesand, T., Kiefer, R. W. & Chipman, J. *Remote Sensing and Image Interpretation* 7th. ISBN: 978-1-118-91947-7 (John Wiley and sons, 2015).

23. Oppenheimer, C, Rothery, D. A., Pieri, D. C., Abrams, M. J. & Carrere, V. Analysis of Airborne Visible/Infrared Imaging Spectrometer (AVTRIS) data of volcanic hot spots. *International Journal of Remote Sensing* **14,** 2919–2934. eprint: https://doi.org/10.1080/01431169308904411. https://doi.org/10.1080/01431169308904411 (1993).

24. Shaw, G & Burke, H. Spectral imaging for remote sensing. *Lincoln Laboratory Journal* **14,** 3–28. https://www.ll.mit.edu/publications/journal/pdf/vol14_no1/14_1remotesensing.pdf (2003).

25. Shaw, G & Manolakis, D. Signal processing for hyperspectral image exploitation. *IEEE Signal Processing Magazine* **19,** 12–16. ISSN: 1053-5888 (2002).

26. Hapke, B. *Reflectance spectroscopy. In Theory of Reflectance and Emittance Spectroscopy* 2nd, 369–411 (Cambridge University Press, 2012).

27. Plaza, A *et al. Advanced processing of hyperspectral images* in *2006 IEEE International Symposium on Geoscience and Remote Sensing* (2006), 1974–1978.

28. Valero, S, Salembier, P & Chanussot, J. Hyperspectral Image Representation and Processing With Binary Partition Trees. *IEEE Transactions on Image Processing* **22,** 1430–1443. ISSN: 1057-7149 (2013).

29. Zhang, L. & DU, B. Recent advances in hyperspectral image processing. *Geo-spatial Information Science* **15,** 143–156 (2012).

30. Richards, A. J. & Xiuping, J. *Remote Sensing Digital Image Analysis: An Introduction* 3rd (eds Ricken, D. E. & Gessner, W) ISBN: 3540648607 (Springer-Verlag, Berlin, Heidelberg, 1999).

31. Bioucas-Dias, J. M. *et al.* Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **5,** 354–379. ISSN: 1939-1404 (2012).

32. Bruzzone, L & Prieto, D. F. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing* **38,** 1171–1182. ISSN: 0196-2892 (2000).

33. Fauvel, M, Tarabalka, Y, Benediktsson, J. A., Chanussot, J & Tilton, J. C. Advances in Spectral-Spatial Classification of Hyperspectral Images. *Proceedings of the IEEE* **101,** 652–675. ISSN: 0018-9219 (2013).

34. Alonso, C. M., Malpica, J. A. & de Agirre, A. M. *Consequences of the hughes phenomenon on some classification techniques* in *2011 American Society for Photogrammetry and Remote Sensing Annual Conference* (2011), 1–9.

35. Stibor, T, Timmis, J & Eckert, C. *On the Use of Hyperspheres in Artificial Immune Systems as Antibody Recognition Regions* in *Artificial Immune Systems* (eds Bersini, H. & Carneiro, J.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006), 215–228. ISBN: 978-3-540-37751-1.

36. Wilcox, R. H. *Adaptive control processes:A guided tour* `https://doi.org/10.1002/nav.3800080314` (Princeton University Press, 1961).

37. Lavergnea, P & Patilea, V. Breaking the curse of dimensionality in nonparametric testing. *Journal of Econometrics* **143,** 103–122. `https://doi.org/10.1016/j.jeconom.2007.08.014` (2008).

38. Gheyas, I. A. & Smith, L. S. Feature subset selection in large dimensionality domains. *Pattern recognition* **43,** 5–13. `https://doi.org/10.1016/j.patcog.2009.06.009` (2010).

39. Diani, M, Acito, N, Greco, M & Corsini, G. *A New Band Selection Strategy for Target Detection in Hyperspectral Images* in *Artificial Immune Systems* (eds Bersini, H. & Carneiro, J.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008), 424–431.

40. Mojaradi, B, Abrishami-Moghaddam, H, Zoej, M. J. V. & Duin, R. P. W. Dimensionality Reduction of Hyperspectral Data via Spectral Feature Extraction. *IEEE Transactions on Geoscience and Remote Sensing* **47,** 2091–2105. ISSN: 0196-2892 (2009).

41. Yu, H & Yang, J. A direct lda algorithm for high-dimensional data with application to face recognition. *Pattern Recognition* **34,** 2067–2070 (2001).

42. Martinez, A. M. & Kak, A. C. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23,** 228–233. ISSN: 0162-8828 (2001).

43. Fukunaga, K. *Introduction to Statistical Pattern Recognition (2nd Ed.)* ISBN: 0-12-269851-7 (Academic Press Professional, Inc., San Diego, CA, USA, 1990).

44. Ye, J., Janardan, R, Park, C. H. & Park, H. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26,** 982–994. ISSN: 0162-8828 (2004).

45. Wei, Z., Xiangyang, X., Hong, L. & Yue-Fei, G. Discriminant Neighborhood Embedding for Classification. *Pattern Recogn.* **39,** 2240–2243. ISSN: 0031-3203. `http://dx.doi.org/10.1016/j.patcog.2006.05.011` (Nov. 2006).

46. Ahmad, M, Khan, A. M., Brown, J. A., Protasov, S & Khattak, A. M. *Gait fingerprinting-based user identification on smartphones* in *2016 International Joint Conference on Neural Networks (IJCNN)* (2016), 3060–3067.

47. Chen, H.-T., Chang, H.-W. & Liu, T.-L. *Local discriminant embedding and its variants* in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* **2** (2005), 846–853 vol. 2.

48. Yan, S *et al.* Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29,** 40–51. ISSN: 0162-8828 (2007).

49. Yan, S., Xu, D., Zhang, B. & Zhang, H.-J. *Graph embedding: a general framework for dimensionality reduction* in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* **2** (2005), 830–837 vol. 2.

50. Ding, C. T. & Zhang, L. Double adjacency graph-based discriminant neighborhood embedding. *Pattern Recognition* **48,** 1734–1742 (2015).

51. Chen, J, Ye, J & Li, Q. *Integrating Global and Local Structures: A Least Squares Framework for Dimensionality Reduction* in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), 1–8.

52. Liu, X., Wang, L., Zhang, J., Yin, J. & Liu, H. Global and Local Structure Preservation for Feature Selection. *IEEE Transactions on Neural Networks and Learning Systems* **25,** 1083–1095. ISSN: 2162-237X (2014).

53. Zhao, Z, Wang, L, Liu, H & Ye, J. On Similarity Preserving Feature Selection. *IEEE Transactions on Knowledge and Data Engineering* **25,** 619–632. ISSN: 1041-4347 (2013).

54. T, J. I. *Principal Component Analysis* (Springer Verlag, 1986).

55. Turk, M. A. & Pentland, A. P. *Face recognition using eigenfaces* in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1991), 586–591.

56. T, R. S. & K, S. L. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290,** 2323–2326 (2000).

57. B., T. J., de Silva Vin & C, L. J. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290,** 2319 (2000).

58. Mikhail, B. & Partha, N. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering* in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (MIT Press, Vancouver, British Columbia, Canada, 2001), 585–591. http://dl.acm.org/citation.cfm?id=2980539.2980616.

59. Yang, J, Zhang, D, y Yang, J & Niu, B. Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29,** 650–664. ISSN: 0162-8828 (2007).

60. He, X., Cai, D., Yan, S. & Zhang, H.-J. *Neighborhood preserving embedding* in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* **2** (2005), 1208–1213 Vol. 2.

61. He, X. & Niyogi, P. *Locality Preserving Projections* in *In Advances in Neural Information Processing Systems 16* (MIT Press, 2003).

62. Yang, S, Wang, X, Wang, M, Han, Y & Jiao, L. Semi-supervised low-rank representation graph for pattern recognition. *IET Image Processing* **7,** 131–136. ISSN: 1751-9659 (2013).

63. He, R, Zheng, W. S., Hu, B. G. & Kong, X. W. *Nonnegative sparse coding for discriminative semi-supervised learning* in *CVPR 2011* (2011), 2849–2856.

64. Zhang, Z, Zhao, M & Chow, T. W. S. Graph Based Constrained Semi-Supervised Learning Framework via Label Propagation over Adaptive Neighborhood. *IEEE Transactions on Knowledge and Data Engineering* **27,** 2362–2376. ISSN: 1041-4347 (2015).

65. Nguyen, Q. & Hein, M. Optimization landscape and expressivity of deep cnns. *arXiv preprint arXiv:1710.10928* (2017).

66. Chen, S. & Wang, Y. Convolutional neural network and convex optimization. *Dept. of Elect. and Comput. Eng., Univ. of California at San Diego, San Diego, CA, USA, Tech. Rep* (2014).

67. Erhan, D. *et al.* Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* **11,** 625–660 (2010).

68. Paoletti, M., Haut, J., Plaza, J & Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* **158,** 279–317 (2019).

69. Alom, M. Z. *et al.* A state-of-the-art survey on deep learning theory and architectures. *Electronics* **8,** 292 (2019).

70. Plaza, A., Valencia, D. & Plaza, J. An experimental comparison of parallel algorithms for hyperspectral analysis using heterogeneous and homogeneous networks of workstations. *Parallel Computing* **34,** 92–114 (2008).

71. Plaza, A., Plaza, J., Paz, A. & Sanchez, S. Parallel hyperspectral image and signal processing [applications corner]. *IEEE Signal Processing Magazine* **28,** 119–126 (2011).

72. Bioucas-Dias, J. M. *et al.* Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote sensing magazine* **1,** 6–36 (2013).

73. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

74. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **5,** 157–166 (1994).

75. *10 Important Applications of Hyperspectral Image* `https://grindgis.com/remote-sensing/10-important-applications-of-hyperspectral-image`. Accessed: 2020-03-10.

76. Xing, F. *et al.* Recent developments and applications of hyperspectral imaging for rapid detection of mycotoxins and mycotoxigenic fungi in food products. *Critical Reviews in Food Science and Nutrition* **59.** PMID: 28846441, 173–180. eprint: `https://doi.org/10.1080/10408398.2017.1363709`. `https://doi.org/10.1080/10408398.2017.1363709` (2019).

77. *Applications of Hyperspectral Image* `https://resonon.com/applications`. Accessed: 2020-03-10.

78. Haut, J. M., Paoletti, M. E., Plaza, J. & Plaza, A. Fast dimensionality reduction and classification of hyperspectral images with extreme learning machines. *Journal of Real-Time Image Processing* **15,** 439–462 (2018).

79. Ahmad, M., Lee, S., Ulhaq, D. & Mushtaq, Q. Hyperspectral Remote Sensing: Dimensional Reduction and End member Extraction. *International Journal of Soft Computing and Engineering (IJSCE)* **2,** 2231–2307 (May 2012).

80. Bioucas-Dias, J. M. *et al.* Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing* **5,** 354–379 (2012).

81. Ahmad, M., Khan, A. & Bashir, A. K. Metric similarity regularizer to enhance pixel similarity performance for hyperspectral unmixing. *Optik - International Journal for Light and Electron Optics* **140,** 86–95. ISSN: 0030-4026 (2017).

82. Zhong, Y. *et al.* Blind spectral unmixing based on sparse component analysis for hyperspectral remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **119,** 49–63 (2016).

83. Ahmad, M., Ihsan, D. & Ulhaq, D. *Linear Unmixing and Target Detection of Hyperspectral Imagery Using OSP* in *International Proceedings of Computer Science and Information Technology* (Jan. 2011), 179–183.

84. Ahmad, M., Ulhaq, D., Mushtaq, Q. & Sohaib, M. A New Statistical Approach for Band Clustering and Band Selection Using K-Means Clustering. *International Journal of Engineering and Technology* **3,** 606–614 (Dec. 2011).

85. Ahmad, M., Ulhaq, D. & Mushtaq, Q. *AIK Method for Band Clustering Using Statistics of Correlation and Dispersion Matrix* in *International Conference on Information Communication and Management* (Jan. 2011), 114–118.

86. Stein, D. W. *et al.* Anomaly detection from hyperspectral imagery. *IEEE signal processing magazine* **19,** 58–69 (2002).

87. Xu, Y., Wu, Z., Li, J., Plaza, A. & Wei, Z. Anomaly detection in hyperspectral images based on low-rank and sparse representation. *IEEE Transactions on Geoscience and Remote Sensing* **54,** 1990–2000 (2015).

88. Li, S., Zhang, K., Hao, Q., Duan, P. & Kang, X. Hyperspectral anomaly detection with multiscale attribute and edge-preserving filters. *IEEE Geoscience and Remote Sensing Letters* **15,** 1605–1609 (2018).

89. Ghamisi, P., Plaza, J., Chen, Y., Li, J. & Plaza, A. J. Advanced spectral classifiers for hyperspectral images: A review. *IEEE Geoscience and Remote Sensing Magazine* **5,** 8–32 (2017).

90. Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J. & Tilton, J. C. Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE* **101,** 652–675 (2012).

91. Ahmad, M., Shabbir, S., Oliva, D., Mazzara, M. & Distefano, S. Spatial-prior Generalized Fuzziness Extreme Learning Machine Autoencoder-based Active Learning for Hyperspectral Image Classification. *Optik-International Journal for Light and Electron Optics.* `https://www.sciencedirect.com/science/article/abs/pii/S0030402619316109` (2020).

92. Ahmad, M. *et al. Unsupervised geometrical feature learning from hyperspectral data* in *2016 IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2016)* (2016), 1–6.

93. Ahmad, M., Protasov, S. & Khan, A. M. Hyperspectral Band Selection Using Unsupervised Non-Linear Deep Auto Encoder to Train External Classifiers. *CoRR* **abs/1705.06920.** arXiv: 1705.06920. http://arxiv.org/abs/1705.06920 (2017).

94. Ahmad, M. *et al.* Segmented and Non-Segmented Stacked Denoising Autoencoder for Hyperspectral Band Reduction. *Optik - International Journal for Light and Electron Optics* **180,** 370–378. ISSN: 0030-4026 (2018).

95. Wei, W., Zhang, L., Tian, C., Plaza, A. & Zhang, Y. Structured sparse coding-based hyperspectral imagery denoising with intracluster filtering. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 6860–6876 (2017).

96. Zhang, H., He, W., Zhang, L., Shen, H. & Yuan, Q. Hyperspectral image restoration using low-rank matrix recovery. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 4729–4743 (2013).

97. Yi, C., Zhao, Y.-Q., Yang, J., Chan, J. C.-W. & Kong, S. G. Joint hyperspectral superresolution and unmixing with interactive feedback. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 3823–3834 (2017).

98. Yi, C., Zhao, Y.-Q. & Chan, J. C.-W. Hyperspectral image super-resolution based on spatial and spectral correlation fusion. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 4165–4177 (2018).

99. Cheng, G., Han, J. & Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* **105,** 1865–1883 (2017).

100. Zhu, Q., Zhong, Y., Zhao, B., Xia, G.-S. & Zhang, L. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters* **13,** 747–751 (2016).

101. Wu, H., Liu, B., Su, W., Zhang, W. & Sun, J. Hierarchical coding vectors for scene level land-use classification. *Remote Sensing* **8,** 436 (2016).

102. Cheng, G. *et al.* Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **53,** 4238–4249 (2015).

103. Martha, T. R., Kerle, N., van Westen, C. J., Jetten, V. & Kumar, K. V. Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis. *IEEE Transactions on Geoscience and Remote Sensing* **49,** 4928–4943 (2011).

104. Cheng, G. *et al.* Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *International Journal of Remote Sensing* **34,** 45–59 (2013).

105. Mishra, N. B. & Crews, K. A. Mapping vegetation morphology types in a dry savanna ecosystem: integrating hierarchical object-based image analysis with Random Forest. *International Journal of Remote Sensing* **35,** 1175–1198 (2014).

106. Li, X. & Shao, G. Object-based urban vegetation mapping with high-resolution aerial photography as a single data source. *International journal of remote sensing* **34,** 771–789 (2013).

107. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review* **26,** 159–190 (2006).

108. Kotsiantis, S. B., Zaharakis, I & Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* **160,** 3–24 (2007).

109. Plaza, A. *et al.* Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment* **113,** S110–S122 (2009).

110. Zhang, L. & Du, B. Recent advances in hyperspectral image processing. *Geo-spatial Information Science* **15,** 143–156 (2012).

111. Ablin, R & Sulochana, C. H. A survey of hyperspectral image classification in remote sensing. *International Journal of Advanced Research in Computer and Communication Engineering* **2,** 2986–3000 (2013).

112. Camps-Valls, G., Tuia, D., Bruzzone, L. & Benediktsson, J. A. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE signal processing magazine* **31,** 45–54 (2013).

113. Chutia, D., Bhattacharyya, D., Sarma, K. K., Kalita, R & Sudhakar, S. Hyperspectral remote sensing classifications: a perspective survey. *Transactions in GIS* **20,** 463–490 (2016).

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

115. Ghamisi, P. *et al.* Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* **5,** 37–78 (2017).

116. Li, C. *et al.* Deep belief network for spectral–spatial classification of hyperspectral remote sensor data. *Sensors* **19,** 204 (2019).

117. Ahmad, M. *et al.* A Fast and Compact 3-D CNN for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters,* 1–5. ISSN: 1558-0571. `http://dx.doi.org/10.1109/LGRS.2020.3043710` (2020).

118. Petersson, H., Gustafsson, D. & Bergstrom, D. *Hyperspectral image analysis using deep learning—A review* in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (2016), 1–6.

119. Zhu, X. X. *et al.* Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* **5,** 8–36 (2017).

120. Huang, L., Chen, C., Li, W. & Du, Q. Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors. *Remote Sensing* **8,** 483 (2016).

121. Dalal, N. & Triggs, B. *Histograms of oriented gradients for human detection* in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* **1** (2005), 886–893.

122. Oliva, A. & Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* **42,** 145–175 (2001).

123. Lowe, D. G. *Object recognition from local scale-invariant features* in *Proceedings of the seventh IEEE international conference on computer vision* **2** (1999), 1150–1157.

124. Ham, J., Chen, Y., Crawford, M. M. & Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing* **43,** 492–501 (2005).

125. Camps-Valls, G. & Bruzzone, L. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **43,** 1351–1362 (2005).

126. Cheng, G. *et al. Object detection in VHR optical remote sensing images via learning rotation-invariant HOG feature* in *2016 4th International Workshop on Earth Observation and Remote Sensing Applications (EORSA)* (2016), 433–436.

127. Cheng, G., Zhou, P., Han, J., Guo, L. & Han, J. Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images. *IET Computer Vision* **9,** 639–647 (2015).

128. Azhar, R., Tuwohingide, D., Kamudi, D., Suciati, N., *et al.* Batik image classification using SIFT feature extraction, bag of features and support vector machine. *Procedia Computer Science* **72,** 24–30 (2015).

129. Zeglazi, O., Amine, A. & Rziza, M. *Sift descriptors modeling and application in texture image classification* in *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)* (2016), 265–268.

130. Xu, Y., Hu, K., Tian, Y. & Peng, F. *Classification of hyperspectral imagery using SIFT for spectral matching* in *2008 Congress on Image and Signal Processing* **2** (2008), 704–708.

131. Yang, Y. & Newsam, S. *Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery* in *2008 15th IEEE international conference on image processing* (2008), 1852–1855.

132. Nhat, H. T. M. & Hoang, V. T. *Feature fusion by using LBP, HOG, GIST descriptors and Canonical Correlation Analysis for face recognition* in *2019 26th International Conference on Telecommunications (ICT)* (2019), 371–375.

133. Roy, S. K., Chanda, B., Chaudhuri, B. B., Ghosh, D. K. & Dubey, S. R. Local morphological pattern: A scale space shape descriptor for texture classification. *Digital Signal Processing* **82,** 152–165 (2018).

134. Haralick, R. M., Shanmugam, K. & Dinstein, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 610–621 (1973).

135. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* **24,** 971–987 (2002).

136. Zhao, L., Tang, P. & Huo, L. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *Journal of Applied Remote Sensing* **10,** 035004 (2016).

137. Zhang, Y., Sun, X., Wang, H. & Fu, K. High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model. *IEEE Geoscience and Remote Sensing Letters* **10,** 1055–1059 (2013).

138. Yang, Y. & Newsam, S. *Bag-of-visual-words and spatial extensions for land-use classification* in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (2010), 270–279.

139. Xu, S., Fang, T., Li, D. & Wang, S. Object classification of aerial images with bag-of-visual words. *IEEE Geoscience and Remote Sensing Letters* **7,** 366–370 (2009).

140. Zhang, J., Li, T., Lu, X. & Cheng, Z. Semantic classification of high-resolution remote-sensing images based on mid-level features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **9,** 2343–2353 (2016).

141. Bahmanyar, R., Cui, S. & Datcu, M. A comparative study of bag-of-words and bag-of-topics models of EO image patches. *IEEE Geoscience and Remote Sensing Letters* **12,** 1357–1361 (2015).

142. Cheng, G., Han, J., Zhou, P. & Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing* **98,** 119–132 (2014).

143. Zhao, B., Zhong, Y., Zhang, L. & Huang, B. The Fisher kernel coding framework for high spatial resolution scene classification. *Remote Sensing* **8,** 157 (2016).

144. Hu, J., Xia, G.-S., Hu, F. & Zhang, L. A comparative study of sampling analysis in the scene classification of optical high-spatial resolution remote sensing imagery. *Remote Sensing* **7,** 14988–15013 (2015).

145. Lazebnik, S., Schmid, C. & Ponce, J. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories* in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* **2** (2006), 2169–2178.

146. Zhao, B., Zhong, Y., Xia, G.-S. & Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **54,** 2108–2123 (2015).

147. Kusumaningrum, R., Wei, H., Manurung, R. & Murni, A. Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image. *Journal of Applied Remote Sensing* **8,** 083690 (2014).

148. Zhong, Y., Zhu, Q. & Zhang, L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **53,** 6207–6222 (2015).

149. Yu, H., Yang, W., Xia, G.-S. & Liu, G. A color-texture-structure descriptor for high-resolution satellite image classification. *Remote Sensing* **8,** 259 (2016).

150. Risojević, V. & Babić, Z. Fusion of global and local descriptors for remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters* **10,** 836–840 (2012).

151. Mekhalfi, M. L., Melgani, F., Bazi, Y. & Alajlan, N. Land-use classification with compressive sensing multifeature fusion. *IEEE Geoscience and Remote Sensing Letters* **12,** 2155–2159 (2015).

152. Sheng, G., Yang, W., Xu, T. & Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *International journal of remote sensing* **33,** 2395–2412 (2012).

153. Xie, L., Wang, J., Zhang, B. & Tian, Q. Incorporating visual adjectives for image classification. *Neurocomputing* **182,** 48–55 (2016).

154. Zou, J., Li, W., Chen, C. & Du, Q. Scene classification using local and global features with collaborative representation fusion. *Information Sciences* **348,** 209–226 (2016).

155. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **313,** 504–507 (2006).

156. Zou, Q., Ni, L., Zhang, T. & Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters* **12,** 2321–2325 (2015).

157. Hu, F., Xia, G.-S., Hu, J. & Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing* **7,** 14680–14707 (2015).

158. Bellman, R. E. *Adaptive control processes: a guided tour* (Princeton university press, 2015).

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

160. Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566,** 195–204 (2019).

161. Fang, Y. *et al.* Dimensionality reduction of hyperspectral images based on robust spatial information using locally linear embedding. *IEEE Geoscience and Remote Sensing Letters* **11,** 1712–1716 (2014).

162. Sugiyama, M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of machine learning research* **8,** 1027–1061 (2007).

163. Chen, H.-T., Chang, H.-W. & Liu, T.-L. *Local discriminant embedding and its variants* in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* **2** (2005), 846–853.

164. Kuo, B.-C. & Landgrebe, D. A. Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing* **42,** 1096–1105 (2004).

165. Kumar, B., Dikshit, O., Gupta, A. & Singh, M. K. Feature extraction for hyperspectral image classification: a review. *International Journal of Remote Sensing* **41,** 6248–6287 (2020).

166. Benediktsson, J. A., Palmason, J. A. & Sveinsson, J. R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing* **43,** 480–491 (2005).

167. Gu, Y., Liu, T., Jia, X., Benediktsson, J. A. & Chanussot, J. Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **54,** 3235–3247 (2016).

168. Zhang, X. *et al.* Spatial sequential recurrent neural network for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11,** 4141–4155 (2018).

169. Pesaresi, M. & Benediktsson, J. A. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE transactions on Geoscience and Remote Sensing* **39,** 309–320 (2001).

170. Chen, Y., Zhao, X. & Jia, X. Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **8,** 2381–2392 (2015).

171. Paoletti, M. E., Haut, J. M., Plaza, J. & Plaza, A. Deep&dense convolutional neural network for hyperspectral image classification. *Remote Sensing* **10,** 1454 (2018).

172. Jin, X., Jie, L., Wang, S., Qi, H. J. & Li, S. W. Classifying wheat hyperspectral pixels of healthy heads and Fusarium head blight disease using a deep neural network in the wild field. *Remote Sensing* **10,** 395 (2018).

173. Wu, N., Zhang, C., Bai, X., Du, X. & He, Y. Discrimination of Chrysanthemum varieties using hyperspectral imaging combined with a deep convolutional neural network. *Molecules* **23,** 2831 (2018).

174. Li, Y., Xie, W. & Li, H. Hyperspectral image reconstruction by deep convolutional neural network for classification. *Pattern Recognition* **63,** 371–383 (2017).

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

176. Paoletti, M. E. *et al.* Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 740–754 (2018).

177. Acquarelli, J., Marchiori, E., Buydens, L., Tran, T. & Van Laarhoven, T. Spectral-spatial classification of hyperspectral images: Three tricks and a new learning setting. *Remote Sensing* **10,** 1156 (2018).

178. Liu, Q., Zhou, F., Hang, R. & Yuan, X. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. *Remote Sensing* **9,** 1330 (2017).

179. Roy, S. K., Manna, S., Song, T. & Bruzzone, L. Attention-Based Adaptive Spectral-Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* (2020).

180. Liu, B. *et al.* A semi-supervised convolutional neural network for hyperspectral image classification. *Remote Sensing Letters* **8,** 839–848 (2017).

181. Kang, X., Zhuo, B. & Duan, P. Semi-supervised deep learning for hyperspectral image classification. *Remote Sensing Letters* **10,** 353–362 (2019).

182. Wu, Y. *et al.* Semi-Supervised Hyperspectral Image Classification via Spatial-Regulated Self-Training. *Remote Sensing* **12,** 159 (2020).

183. Zhang, Z. Semi-supervised Hyperspectral Image Classification Algorithm based on Graph Embedding and Discriminative Spatial Information. *Microprocessors and Microsystems,* 103070 (2020).

184. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *Imagenet classification with deep convolutional neural networks* in *Advances in neural information processing systems* (2012), 1097–1105.

185. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).

186. Williams, T. & Li, R. *Wavelet Pooling for Convolutional Neural Networks* in *International Conference on Learning Representations* (2018). `https://openreview.net/forum?id=rkhlb8lCZ`.

187. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

188. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* **160,** 106–154 (1962).

189. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* **36,** 193–202 (1980).

190. Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* **2018** (2018).

191. Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern Recognition* **77,** 354–377 (2018).

192. Lin, M., Chen, Q. & Yan, S. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

193. Gao, H., Yang, Y., Li, C., Zhou, H. & Qu, X. Joint alternate small convolution and feature reuse for hyperspectral image classification. *ISPRS International Journal of Geo-Information* **7,** 349 (2018).

194. Zhang, M., Li, W. & Du, Q. Diverse region-based CNN for hyperspectral image classification. *IEEE Transactions on Image Processing* **27,** 2623–2634 (2018).

195. Zhao, W. *et al.* Superpixel-based multiple local CNN for panchromatic and multispectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 4141–4156 (2017).

196. Alhichri, H., Alajlan, N., Bazi, Y. & Rabczuk, T. *Multi-Scale Convolutional Neural Network for Remote Sensing Scene Classification* in *2018 IEEE International Conference on Electro/Information Technology (EIT)* (2018), 1–5.

197. Noor, M., Salwa, S., Ren, J., Marshall, S. & Michael, K. Hyperspectral image enhancement and mixture deep-learning classification of corneal epithelium injuries. *Sensors* **17,** 2644 (2017).

198. Leng, J., Li, T., Bai, G., Dong, Q. & Dong, H. *Cube-CNN-SVM: a novel hyperspectral image classification method* in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)* (2016), 1027–1034.

199. Yu, S., Jia, S. & Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **219,** 88–98 (2017).

200. Wu, H. & Prasad, S. Convolutional recurrent neural networks forhyperspectral data classification. *Remote Sensing* **9,** 298 (2017).

201. Qiu, Z. *et al.* Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences* **8,** 212 (2018).

202. Huang, Q., Li, W. & Xie, X. *Convolutional neural network for medical hyperspectral image classification with kernel fusion* in *BIBE 2018; International Conference on Biological Information and Biomedical Engineering* (2018), 1–4.

203. Charmisha, K., Sowmya, V & Soman, K. *Dimensionally reduced features for hyperspectral image classification using deep learning* in *International Conference on Communications and Cyber Physical Engineering 2018* (2018), 171–179.

204. Turra, G., Arrigoni, S. & Signoroni, A. *CNN-based identification of hyperspectral bacterial signatures for digital microbiology* in *International Conference on Image Analysis and Processing* (2017), 500–510.

205. Li, J. *et al.* Classification of hyperspectral imagery using a new fully convolutional neural network. *IEEE Geoscience and Remote Sensing Letters* **15,** 292–296 (2018).

206. Haut, J. M., Paoletti, M. E., Plaza, J., Plaza, A. & Li, J. Hyperspectral image classification using random occlusion data augmentation. *IEEE Geoscience and Remote Sensing Letters* **16,** 1751–1755 (2019).

207. Xu, Y., Du, B., Zhang, F. & Zhang, L. Hyperspectral image classification via a random patches network. *ISPRS journal of photogrammetry and remote sensing* **142,** 344–357 (2018).

208. Ding, C. *et al.* Convolutional neural networks based hyperspectral image classification method with adaptive kernels. *Remote Sensing* **9,** 618 (2017).

209. Chen, Y. *et al.* Hyperspectral images classification with Gabor filtering and convolutional neural network. *IEEE Geoscience and Remote Sensing Letters* **14,** 2355–2359 (2017).

210. Zhu, J., Fang, L. & Ghamisi, P. Deformable convolutional neural networks for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **15,** 1254–1258 (2018).

211. Ran, L., Zhang, Y., Wei, W. & Zhang, Q. A hyperspectral image classification framework with spatial pixel pair features. *Sensors* **17,** 2421 (2017).

212. Li, W., Wu, G., Zhang, F. & Du, Q. Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 844–853 (2016).

213. Zhong, Z., Li, J., Luo, Z. & Chapman, M. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 847–858 (2018).

214. Paoletti, M., Haut, J., Plaza, J & Plaza, A. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS journal of photogrammetry and remote sensing* **145,** 120–147 (2018).

215. Li, S., Zhu, X., Liu, Y. & Bao, J. Adaptive spatial-spectral feature learning for hyperspectral image classification. *IEEE Access* **7,** 61534–61547 (2019).

216. Roy, S. K., Paoletti, M. E., Haut, J. M., Hendrix, E. M. T. & Plaza, A. *A New Max-Min Convolutional Network for Hyperspectral Image Classification* in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (2021), 1–5.

217. Zhang, H. *et al.* Hyperspectral classification based on lightweight 3-D-CNN with transfer learning. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 5813–5828 (2019).

218. Jia, S. *et al.* A lightweight convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **59,** 4150–4163 (2020).

219. Roy, S. K., Chatterjee, S., Bhattacharyya, S., Chaudhuri, B. B. & Platoš, J. Lightweight spectral–spatial squeeze-and-excitation residual bag-of-features learning for hyperspectral classification. *IEEE Transactions on Geoscience and Remote Sensing* **58,** 5277–5290 (2020).

220. Roy, S. K., Mondal, R., Paoletti, M. E., Haut, J. M. & Plaza, A. J. Morphological Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2021).

221. Li, Y., Zhang, H. & Shen, Q. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing* **9,** 67 (2017).

222. Roy, S. K., Dubey, S. R., Chatterjee, S. & Chaudhuri, B. B. FuSENet: fused squeeze-and-excitation network for spectral-spatial hyperspectral image classification. *IET Image Processing* **14,** 1653–1661 (2020).

223. Jiao, L. *et al.* Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 5585–5599 (2017).

224. Zhang, H., Li, Y., Zhang, Y. & Shen, Q. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote sensing letters* **8,** 438–447 (2017).

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

226. Dong, H., Zhang, L. & Zou, B. Band Attention Convolutional Networks For Hyperspectral Image Classification. *arXiv preprint arXiv:1906.04379* (2019).

227. He, N. *et al.* Feature extraction with multiscale covariance maps for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 755–769 (2018).

228. Cheng, G., Li, Z., Han, J., Yao, X. & Guo, L. Exploring hierarchical convolutional features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 6712–6722 (2018).

229. Gong, Z., Zhong, P., Yu, Y., Hu, W. & Li, S. A CNN with multiscale convolution and diversified metric for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 3599–3618 (2019).

230. Zhong, P., Peng, N. & Wang, R. Learning to diversify patch-based priors for remote sensing image restoration. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **8,** 5225–5245 (2015).

231. Liu, L. *et al.* Multiscale Deep Spatial Feature Extraction Using Virtual RGB Image for Hyperspectral Imagery Classification. *Remote Sensing* **12,** 280 (2020).

232. Ma, X., Fu, A., Wang, J., Wang, H. & Yin, B. Hyperspectral image classification based on deep deconvolution network with skip architecture. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 4781–4791 (2018).

233. Sellami, A., Farah, M., Farah, I. R. & Solaiman, B. Hyperspectral imagery classification based on semi-supervised 3-D deep neural network and adaptive band selection. *Expert Systems with Applications* **129,** 246–259 (2019).

234. Roy, S. K., Das, S., Song, T. & Chanda, B. DARecNet-BS: Unsupervised Dual-Attention Reconstruction Network for Hyperspectral Band Selection. *IEEE Geoscience and Remote Sensing Letters* (2020).

235. Mei, S. *et al.* Unsupervised Spatial–Spectral Feature Learning by 3D Convolutional Autoencoder for Hyperspectral Classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 6808–6820 (2019).

236. Roy, S. K., Krishna, G., Dubey, S. R. & Chaudhuri, B. B. HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **17,** 277–281 (2020).

237. Cao, X. *et al.* Hyperspectral image classification with Markov random fields and a convolutional neural network. *IEEE Transactions on Image Processing* **27,** 2354–2367 (2018).

238. Zhu, J., Wu, L., Hao, H., Song, X. & Lu, Y. *Auto-encoder based for high spectral dimensional data classification and visualization* in *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)* (2017), 350–354.

239. Hassanzadeh, A., Kaarna, A. & Kauranne, T. *Unsupervised multi-manifold classification of hyperspectral remote sensing images with contractive Autoencoder* in *Scandinavian Conference on Image Analysis* (2017), 169–180.

240. Wang, Y., Jiang, Y., Wu, Y. & Zhou, Z.-H. *Multi-manifold clustering* in *Pacific Rim International Conference on Artificial Intelligence* (2010), 280–291.

241. Rifai, S., Vincent, P., Muller, X., Glorot, X. & Bengio, Y. *Contractive Auto-Encoders: Explicit Invariance during Feature Extraction* in *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Omnipress, Bellevue, Washington, USA, 2011), 833–840. ISBN: 9781450306195.

242. Zhang, X. *et al.* Recursive autoencoders-based unsupervised feature learning for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **14,** 1928–1932 (2017).

243. Hao, S., Wang, W., Ye, Y., Nie, T. & Bruzzone, L. Two-stream deep architecture for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 2349–2361 (2017).

244. He, K., Sun, J. & Tang, X. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence* **35,** 1397–1409 (2012).

245. Sun, X. *et al.* Encoding spectral and spatial context information for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **14,** 2250–2254 (2017).

246. Zhao, C. *et al.* Spectral-spatial classification of hyperspectral imagery based on stacked sparse autoencoder and random forest. *European journal of remote sensing* **50,** 47–63 (2017).

247. Wan, X., Zhao, C., Wang, Y. & Liu, W. Stacked sparse autoencoder in hyperspectral data classification using spectral-spatial, higher order statistics and multifractal spectrum features. *Infrared Physics & Technology* **86,** 77–89 (2017).

248. Lv, F., Han, M. & Qiu, T. Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder. *IEEE Access* **5,** 9021–9031 (2017).

249. Ahmad, M., Khan, A. M., Mazzara, M. & Distefano, S. *Multi-layer extreme learning machine-based autoencoder for hyperspectral image classification* in *Proceedings of the 14th International Conference on Computer Vision Theory and Applications (VISAPP'19), Prague, Czech Republic* (2019), 25–27.

250. Ahmad, M. *et al.* Segmented and non-segmented stacked denoising autoencoder for hyperspectral band reduction. *Optik* **180,** 370–378 (2019).

251. Zhou, P., Han, J., Cheng, G. & Zhang, B. Learning compact and discriminative stacked autoencoder for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 4823–4833 (2019).

252. Lan, R., Li, Z., Liu, Z., Gu, T. & Luo, X. Hyperspectral image classification using k-sparse denoising autoencoder and spectral–restricted spatial characteristics. *Applied Soft Computing* **74,** 693–708 (2019).

253. Paul, S. & Kumar, D. N. Spectral-spatial classification of hyperspectral data with mutual information based segmented stacked autoencoder approach. *ISPRS journal of photogrammetry and remote sensing* **138,** 265–280 (2018).

254. Liu, B., Zhang, Q., Ying, L., Chang, W. & Zhou, M. Spatial–Spectral Jointed Stacked Auto-Encoder-Based Deep Learning for Oil Slick Extraction from Hyperspectral Images. *Journal of the Indian Society of Remote Sensing* **47,** 1989–1997 (2019).

255. Hinton, G. E., Osindero, S. & Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation* **18,** 1527–1554 (2006).

256. Zhang, N., Ding, S., Zhang, J. & Xue, Y. An overview on restricted Boltzmann machines. *Neurocomputing* **275,** 1186–1199 (2018).

257. Ayhan, B. & Kwan, C. *Application of Deep Belief Network to Land Cover Classification Using Hyperspectral Images* in *Advances in Neural Networks - ISNN 2017* (eds Cong, F., Leung, A. & Wei, Q.) (Springer International Publishing, Cham, 2017), 269–276.

258. Shaham, U. *et al. A deep learning approach to unsupervised ensemble learning* in *International conference on machine learning* (2016), 30–39.

259. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

260. Xiong, H., Rodríguez-Sánchez, A. J., Szedmak, S. & Piater, J. Diversity priors for learning early visual features. *Frontiers in computational neuroscience* **9,** 104 (2015).

261. Zhong, P., Gong, Z., Li, S. & Sch¨onlieb, C.-B. Learning to diversify deep belief networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 3516–3530 (2017).

262. Li, J., Xi, B., Li, Y., Du, Q. & Wang, K. Hyperspectral classification based on texture feature enhancement and deep belief networks. *Remote Sensing* **10,** 396 (2018).

263. Tan, K., Wu, F., Du, Q., Du, P. & Chen, Y. A parallel gaussian–bernoulli restricted boltzmann machine for mining area classification with hyperspectral imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12,** 627–636 (2019).

264. Li, S. *et al.* Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 6690–6709 (2019).

265. Sellami, A & Farah, I. *Spectra-spatial Graph-based Deep Restricted Boltzmann Networks for Hyperspectral Image Classification* in *2019 PhotonIcs & Electromagnetics Research Symposium-Spring (PIERS-Spring)* (2019), 1055–1062.

266. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* **1,** 270–280 (1989).

267. Paoletti, M. E., Haut, J. M., Plaza, J. & Plaza, A. Scalable recurrent neural network for hyperspectral image classification. *The Journal of Supercomputing,* 1–17 (2020).

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

269. Zhou, F., Hang, R., Liu, Q. & Yuan, X. Hyperspectral image classification using spectral-spatial LSTMs. *Neurocomputing* **328,** 39–47 (2019).

270. Sharma, A., Liu, X. & Yang, X. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. *Neural Networks* **105,** 346 –355. ISSN: 0893-6080. `http://www.sciencedirect.com/science/article/pii/S0893608018301813` (2018).

271. Wu, H. & Prasad, S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing* **27,** 1259–1270 (2017).

272. Zhou, F., Hang, R., Liu, Q. & Yuan, X. *Integrating convolutional neural network and gated recurrent unit for hyperspectral image spectral-spatial classification* in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (2018), 409–420.

273. Luo, H. Shorten Spatial-spectral RNN with Parallel-GRU for Hyperspectral Image Classification. *arXiv preprint arXiv:1810.12563* (2018).

274. Shi, C. & Pun, C.-M. Multi-scale hierarchical recurrent neural networks for hyperspectral image classification. *Neurocomputing* **294,** 82–93 (2018).

275. Yang, X. *et al.* Hyperspectral image classification with deep learning models. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 5408–5423 (2018).

276. Seydgar, M., Alizadeh Naeini, A., Zhang, M., Li, W. & Satari, M. 3-D convolution-recurrent networks for spectral-spatial classification of hyperspectral images. *Remote Sensing* **11,** 883 (2019).

277. Hang, R., Liu, Q., Hong, D. & Ghamisi, P. Cascaded recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 5384–5394 (2019).

278. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6,** 60 (2019).

279. Jia, S. *et al.* A survey: Deep learning for hyperspectral image classification with few labeled samples. *Neurocomputing* **448,** 179–204. ISSN: 0925-2312. `https://www.sciencedirect.com/science/article/pii/S0925231221004033` (2021).

280. Yu, X., Wu, X., Luo, C. & Ren, P. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing* **54,** 741–758 (2017).

281. Li, W., Chen, C., Zhang, M., Li, H. & Du, Q. Data augmentation for hyperspectral image classification with deep cnn. *IEEE Geoscience and Remote Sensing Letters* **16,** 593–597 (2018).

282. Rochac, J. F. R., Zhang, N., Thompson, L. & Oladunni, T. *A Data Augmentation-Assisted Deep Learning Model for High Dimensional and Highly Imbalanced Hyperspectral Imaging Data* in *2019 9th International Conference on Information Science and Technology (ICIST)* (2019), 362–367.

283. Nalepa, J., Myller, M. & Kawulok, M. Training-and test-time data augmentation for hyperspectral image segmentation. *IEEE Geoscience and Remote Sensing Letters* (2019).

284. Nalepa, J., Myller, M. & Kawulok, M. Hyperspectral data augmentation. *arXiv preprint arXiv:1903.05580* (2019).

285. Van Engelen, J. E. & Hoos, H. H. A survey on semi-supervised learning. *Machine Learning* **109,** 373–440 (2020).

286. Pise, N. N. & Kulkarni, P. *A survey of semi-supervised learning methods* in *2008 International Conference on Computational Intelligence and Security* **2** (2008), 30–34.

287. Sawant, S. S. & Prabukumar, M. *Semi-supervised techniques based hyper-spectral image classification: a survey* in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)* (2017), 1–8.

288. Fang, B., Li, Y., Zhang, H. & Chan, J. C.-W. Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection. *Remote Sensing* **10,** 574 (2018).

289. Zhou, S., Xue, Z. & Du, P. Semisupervised stacked autoencoder with cotraining for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 3813–3826 (2019).

290. Li, F., Clausi, D. A., Xu, L. & Wong, A. ST-IRGS: A region-based self-training algorithm applied to hyperspectral image classification and segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 3–16 (2017).

291. Aydemir, M. S. & Bilgin, G. Semisupervised hyperspectral image classification using small sample sizes. *IEEE Geoscience and Remote Sensing Letters* **14,** 621–625 (2017).

292. Goodfellow, I. *et al. Generative adversarial nets* in *Advances in neural information processing systems* (2014), 2672–2680.

293. Zhan, Y., Hu, D., Wang, Y. & Yu, X. Semisupervised hyperspectral image classification based on generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters* **15,** 212–216 (2017).

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

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

296. Zhan, Y. *et al. Semi-supervised classification of hyperspectral data based on generative adversarial networks and neighborhood majority voting* in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (2018), 5756–5759.

297. Feng, J. *et al.* Classification of hyperspectral images based on multiclass spatial–spectral generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 5329–5343 (2019).

298. Zhong, Z., Li, J., Clausi, D. A. & Wong, A. Generative adversarial networks and conditional random fields for hyperspectral image classification. *IEEE transactions on cybernetics* (2019).

299. Wang, X., Tan, K., Du, Q., Chen, Y. & Du, P. Caps-TripleGAN: GAN-assisted CapsNet for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 7232–7245 (2019).

300. Xue, Z. Semi-supervised convolutional generative adversarial network for hyperspectral image classification. *IET Image Processing* **14,** 709–719 (2019).

301. Wang, W.-Y. *et al.* Generative Adversarial Capsule Network With ConvLSTM for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters* (2020).

302. Alipour-Fard, T. & Arefi, H. Structure Aware Generative Adversarial Networks for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13,** 5424–5438 (2020).

303. Roy, S. K., Haut, J. M., Paoletti, M. E., Dubey, S. R. & Plaza, A. Generative adversarial minority oversampling for spectral-spatial hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* (2021).

304. Yang, J., Zhao, Y.-Q. & Chan, J. C.-W. Learning and transferring deep joint spectral–spatial features for hyperspectral classification. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 4729–4742 (2017).

305. Windrim, L., Melkumyan, A., Murphy, R. J., Chlingaryan, A. & Ramakrishnan, R. Pretraining for hyperspectral convolutional neural network classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 2798–2810 (2018).

306. Liu, X., Sun, Q., Meng, Y., Fu, M. & Bourennane, S. Hyperspectral image classification based on parameter-optimized 3D-CNNs combined with transfer learning and virtual samples. *Remote Sensing* **10,** 1425 (2018).

307. Day, O. & Khoshgoftaar, T. M. A survey on heterogeneous transfer learning. *Journal of Big Data* **4,** 29 (2017).

308. Lin, J., Ward, R. & Wang, Z. J. *Deep transfer learning for hyperspectral image classification* in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)* (2018), 1–5.

309. Li, X., Zhang, L., Du, B., Zhang, L. & Shi, Q. Iterative reweighting heterogeneous transfer learning framework for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 2022–2035 (2017).

310. Liu, Y. & Xiao, C. *Transfer learning for hyperspectral image classification using convolutional neural network* in *MIPPR 2019: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications* **11432** (2020), 114320E.

311. Lin, J., He, C., Wang, Z. J. & Li, S. Structure preserving transfer learning for unsupervised hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **14,** 1656–1660 (2017).

312. Pires de Lima, R. & Marfurt, K. Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis. *Remote Sensing* **12,** 86 (2020).

313. Ganti, R. & Gray, A. *Upal: Unbiased pool based active learning* in *Artificial Intelligence and Statistics* (2012), 422–431.

314. Melville, P. & Mooney, R. J. *Diverse ensembles for active learning* in *Proceedings of the twenty-first international conference on Machine learning* (2004), 74.

315. Aggarwal, C. C., Kong, X., Gu, Q., Han, J. & Philip, S. Y. in *Data Classification: Algorithms and Applications* 571–605 (CRC Press, 2014).

316. Seung, H. S., Opper, M. & Sompolinsky, H. *Query by committee* in *Proceedings of the fifth annual workshop on Computational learning theory* (1992), 287–294.

317. Settles, B. *Active learning literature survey* tech. rep. (University of Wisconsin-Madison Department of Computer Sciences, 2009).

318. Liu, C., He, L., Li, Z. & Li, J. Feature-driven active learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 341–354 (2017).

319. Zhang, Y., Cao, G., Li, X., Wang, B. & Fu, P. Active Semi-Supervised Random Forest for Hyperspectral Image Classification. *Remote Sensing* **11,** 2974 (2019).

320. Guo, J., Zhou, X., Li, J., Plaza, A. & Prasad, S. Superpixel-based active learning and online feature importance learning for hyperspectral image analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 347–359 (2016).

321. Xue, Z., Zhou, S. & Zhao, P. Active learning improved by neighborhoods and superpixels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters* **15,** 469–473 (2018).

322. Bhardwaj, K., Das, A. & Patra, S. *Spectral–Spatial Active Learning with Attribute Profile for Hyperspectral Image Classification* in *International Conference on Intelligent Computing and Smart Communication 2019* (2020), 1219–1229.

323. Patra, S., Bhardwaj, K. & Bruzzone, L. A spectral-spatial multicriteria active learning technique for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 5213–5227 (2017).

324. Zhang, Z. & Crawford, M. M. A batch-mode regularized multimetric active learning framework for classification of hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing* **55,** 6594–6609 (2017).

325. Xu, X., Li, J. & Li, S. Multiview intensity-based active learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 669–680 (2017).

326. Pradhan, M. K., Minz, S. & Shrivastava, V. K. *Fisher discriminant ratio based multiview active learning for the classification of remote sensing images* in *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (2018), 1–6.

327. Zhang, Z., Pasolli, E. & Crawford, M. M. An Adaptive Multiview Active Learning Approach for Spectral–Spatial Classification of Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing* **58,** 2557–2570 (2019).

328. Li, Y., Lu, T. & Li, S. Subpixel-Pixel-Superpixel-Based Multiview Active Learning for Hyperspectral Images Classification. *IEEE Transactions on Geoscience and Remote Sensing* (2020).

329. Sun, Y., Li, J., Wang, W., Plaza, A. & Chen, Z. *Active learning based autoencoder for hyperspectral imagery classification* in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2016), 469–472.

330. Liu, P., Zhang, H. & Eom, K. B. Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 712–724 (2016).

331. Haut, J. M., Paoletti, M. E., Plaza, J., Li, J. & Plaza, A. Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 6440–6461 (2018).

332. Cao, X., Yao, J., Xu, Z. & Meng, D. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing* (2020).

333. Lin, J., Zhao, L., Li, S., Ward, R. & Wang, Z. J. Active-learning-incorporated deep transfer learning for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11,** 4048–4062 (2018).

334. Deng, C., Xue, Y., Liu, X., Li, C. & Tao, D. Active transfer learning network: A unified deep joint spectral–spatial feature learning model for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 1741–1754 (2018).

335. Deng, C., Liu, X., Li, C. & Tao, D. Active multi-kernel domain adaptation for hyperspectral image classification. *Pattern Recognition* **77,** 306–315 (2018).

336. Ahmad, M., Raza, R. A. & Mazzara, M. Multiclass Non-Randomized Spectral–Spatial Active Learning for Hyperspectral Image Classification. *Applied Sciences* **10,** 4739 (July 2020).

337. Tuia, D., Volpi, M., Mura, M. D., Rakotomamonjy, A. & Flamary, R. Automatic Feature Learning for Spatio-Spectral Image Classification With Sparse SVM. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 6062–6074 (2014).

338. Ghamisi, P., Dalla Mura, M. & Benediktsson, J. A. A Survey on Spectral–Spatial Classification Techniques Based on Attribute Profiles. *IEEE Transactions on Geoscience and Remote Sensing* **53,** 2335–2353 (2015).

339. Benediktsson, J. A., Palmason, J. A. & Sveinsson, J. R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing* **43,** 480–491 (2005).

340. Jia, S., Zhang, X. & Li, Q. Spectral–Spatial Hyperspectral Image Classification Using $\ell_{1/2}$ Regularized Low-Rank Representation and Sparse Representation-Based Graph Cuts. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **8,** 2473–2484 (2015).

341. Dalla Mura, M., Villa, A., Benediktsson, J. A., Chanussot, J. & Bruzzone, L. Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis. *IEEE Geoscience and Remote Sensing Letters* **8,** 542–546 (2011).

342. Ahmad, M. *et al.* Multiclass Non-Randomized Spectral–Spatial Active Learning for Hyperspectral Image Classification. *Applied Sciences* **10.** ISSN: 2076-3417. `https://www.mdpi.com/2076-3417/10/14/4739` (July. 2020).

343. Zhong, Y., Ma, A. & Zhang, L. An Adaptive Memetic Fuzzy Clustering Algorithm With Spatial Information for Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7,** 1235–1248 (2014).

344. Shen, L. & Jia, S. Three-Dimensional Gabor Wavelets for Pixel-Based Hyperspectral Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing* **49,** 5039–5046 (2011).

345. Qian, Y., Ye, M. & Zhou, J. Hyperspectral Image Classification Based on Structured Sparse Logistic Regression and Three-Dimensional Wavelet Texture Features. *IEEE Transactions on Geoscience and Remote Sensing* **51,** 2276–2291 (2013).

346. Roy, S., Krishna, G., Dubey, S. R. & Chaudhuri, B. HybridSN: Exploring 3-D-2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters* **17,** 277–281 (June 2019).

347. Li, Y. & He, L. *An improved hybrid CNN for hyperspectral image classification* in *Eleventh International Conference on Graphics and Image Processing (ICGIP 2019)* (eds Pan, Z. & Wang, X.) **11373** (SPIE, 2020), 485 –490. https://doi.org/10.1117/12.2557384.

348. Fang, B., Bai, Y. & Li, Y. Combining Spectral Unmixing and 3D/2D Dense Networks with Early-Exiting Strategy for Hyperspectral Image Classification. *Remote Sensing* **12,** 779 (Feb. 2020).

349. Huang, L. & Chen, Y. Dual-Path Siamese CNN for Hyperspectral Image Classification With Limited Training Samples. *IEEE Geoscience and Remote Sensing Letters,* 1–5 (2020).

350. Paoletti, M. E. *et al.* Deep Pyramidal Residual Networks for Spectral–Spatial Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 740–754 (2019).

351. Paoletti, M. E. *et al.* Capsule Networks for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **57,** 2145–2160 (2019).

352. Chen, Y., Jiang, H., Li, C., Jia, X. & Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing* **54,** 6232–6251 (2016).

353. Zhong, Z., Li, J., Luo, Z. & Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 847–858 (2018).

354. Mou, L., Ghamisi, P. & Zhu, X. X. Unsupervised Spectral–Spatial Feature Learning via Deep Residual Conv–Deconv Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 391–406 (2018).

355. Li, Y., Zhang, H. & Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sensing* **9,** 67 (Jan. 2017).

356. Ahmad, M. *et al.* A Fast and Compact 3-D CNN for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters,* 1–5 (2020).

357. Xie, Z. *et al.* Hyperspectral face recognition based on sparse spectral attention deep neural networks. *Opt. Express* **28,** 36286–36303. http://www.opticsexpress.org/abstract.cfm?URI=oe-28-24-36286 (2020).

358. Liu, B. *et al.* A semi-supervised convolutional neural network for hyperspectral image classification. *Remote Sensing Letters* **8,** 839–848 (Sept. 2017).

359. Ben Hamida, A., Benoit, A., Lambert, P. & Ben Amar, C. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 4420–4434 (2018).

360. Lee, H. & Kwon, H. *Contextual deep CNN based hyperspectral classification* in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2016), 3322–3325.

361. Li, Y., Zhang, H. & Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sensing* **9,** 67 (Jan. 2017).

362. Zhao, C., Zhao, H., Wang, G. & Chen, H. Hybrid Depth-Separable Residual Networks for Hyperspectral Image Classification. *Complexity* **2020,** 1–17 (Aug. 2020).

363. Yang, X. *et al.* Synergistic 2D/3D Convolutional Neural Network for Hyperspectral Image Classification. *Remote Sensing* **12,** 2033 (June 2020).

364. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. in *Neurocomputing: Foundations of Research* 696–699 (MIT Press, Cambridge, MA, USA, 1988). ISBN: 0262010976.

365. Sha, H., Al Hasan, M. & Mohler, G. Learning Network Event Sequences Using Long Short-Term Memory and Second-Order Statistic Loss. *Stat. Anal. Data Min.* **14,** 61–73. ISSN: 1932-1864. https://doi.org/10.1002/sam.11489 (2021).

366. Li, H.-C. *et al.* Recurrent Feedback Convolutional Neural Network for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters,* 1–5 (2021).

367. Lei, R. *et al.* A non-local capsule neural network for hyperspectral remote sensing image classification. *Remote Sensing Letters* **12,** 40–49. eprint: https://doi.org/10.1080/2150704X.2020.1864052. https://doi.org/10.1080/2150704X.2020.1864052 (2021).

368. Bi, H., Santos-Rodriguez, R. & Flach, P. *Polsar Image Classification via Robust Low-Rank Feature Extraction and Markov Random Field* in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium* (2020), 708–711.

369. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. *Rethinking the Inception Architecture for Computer Vision* in (June 2016).

370. Zhang, C. & Han, M. Multi-feature hyperspectral image classification with L2,1 norm constrained joint sparse representation. *International Journal of Remote Sensing* **42,** 4789–4808. eprint: https://doi.org/10.1080/01431161.2021.1890854. https://doi.org/10.1080/01431161.2021.1890854 (2021).

371. Real, E., Aggarwal, A., Huang, Y. & Le, Q. V. Regularized Evolution for Image Classifier Architecture Search. *Proceedings of the AAAI Conference on Artificial Intelligence* **33,** 4780–4789. https://ojs.aaai.org/index.php/AAAI/article/view/4405 (2019).

372. Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Learning Transferable Architectures for Scalable Image Recognition. *CoRR* **abs/1707.07012.** arXiv: 1707.07012. http://arxiv.org/abs/1707.07012 (2017).

373. Wang, J. *et al.* NAS-Guided Lightweight Multiscale Attention Fusion Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 1–14 (2021).

374. Yin, B. & Cui, B. *Multi-feature extraction method based on Gaussian pyramid and weighted voting for hyperspectral image classification* in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)* (2021), 645–648.

375. Zhou, H. *et al. Rethinking Soft Labels for Knowledge Distillation: A Bias-Variance Tradeoff Perspective* 2021. arXiv: 2102.00650 [cs.LG].

376. Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T. & Nguyen, H. Q. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing* **437,** 186–194. ISSN: 0925-2312. https://www.sciencedirect.com/science/article/pii/S0925231221000953 (2021).

377. Song, M., Zhao, Y., Wang, S. & Han, M. *Word Similarity Based Label Smoothing in Rnnlm Training for ASR* in *2021 IEEE Spoken Language Technology Workshop (SLT)* (2021), 280–285.

378. Xie, F., Gao, Q., Jin, C. & Zhao, F. Hyperspectral Image Classification Based on Superpixel Pooling Convolutional Neural Network with Transfer Learning. *Remote Sensing* **13.** ISSN: 2072-4292. https://www.mdpi.com/2072-4292/13/5/930 (2021).

379. Yang, X., Song, Z., King, I. & Xu, Z. *A Survey on Deep Semi-supervised Learning* 2021. arXiv: 2103.00550 [cs.LG].

380. Ahmad, M. *et al.* Hyperspectral Image Classification: Artifacts of Dimension Reduction on Hybrid CNN. *arXiv preprint arXiv:2101.10532* (2021).

381. Ahmad, M. *et al.* Segmented and non-segmented stacked denoising autoencoder for hyperspectral band reduction. *Optik* **180,** 370 –378. ISSN: 0030-4026. http://www.sciencedirect.com/science/article/pii/S0030402618316644 (2019).

382. Benediktsson, J. A., Palmason, J. A. & Sveinsson, J. R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing* **43,** 480–491 (2005).

383. Dalla Mura, M., Villa, A., Benediktsson, J. A., Chanussot, J. & Bruzzone, L. Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. *IEEE Geoscience and Remote Sensing Letters* **8,** 542–546 (2010).

384. Tuia, D., Volpi, M., Dalla Mura, M., Rakotomamonjy, A. & Flamary, R. Automatic feature learning for spatio-spectral image classification with sparse SVM. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 6062–6074 (2014).

385. Ahmad, M. Ground truth labeling and samples selection for Hyperspectral Image Classification. *Optik* **230,** 166267. ISSN: 0030-4026. http://www.sciencedirect.com/science/article/pii/S0030402621000103 (2021).

386. Shen, L. & Jia, S. Three-dimensional Gabor wavelets for pixel-based hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing* **49,** 5039–5046 (2011).

387. Qian, Y., Ye, M. & Zhou, J. Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features. *IEEE Transactions on Geoscience and Remote Sensing* **51,** 2276–2291 (2012).

388. Wang, J., Song, X., Sun, L., Huang, W. & Wang, J. A Novel Cubic Convolutional Neural Network for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13,** 4133–4148 (2020).

389. Roy, S. K., Krishna, G., Dubey, S. R. & Chaudhuri, B. B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters* **17,** 277–281 (2020).

390. Shabbir, S. & Ahmad, M. Hyperspectral Image Classification–Traditional to Deep Models: A Survey for Future Prospects. *arXiv preprint arXiv:2101.06116* (2021).

391. Mohan, A. & Venkatesan, M. HybridCNN based hyperspectral image classification using multiscale spatiospectral features. *Infrared Physics & Technology* **108,** 103326. ISSN: 1350-4495. https://www.sciencedirect.com/science/article/pii/S1350449519310485 (2020).

392. Wang, C., Ma, N., Ming, Y., Wang, Q. & Xia, J. Classification of hyperspectral imagery with a 3D convolutional neural network and JM distance. *Advances in Space Research* **64,** 886–899 (2019).

393. Ahmad, M. *et al.* Multiclass Non-Randomized Spectral–Spatial Active Learning for Hyperspectral Image Classification. *Applied Sciences* **10,** 4739 (2020).

394. Weng, J., Zhang, Y. & Hwang, W.-S. Candid covariance-free incremental principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25,** 1034–1040 (2003).

395. Wang, L., Xie, X., Li, W., Du, Q. & Li, G. *Sparse feature extraction for hyperspectral image classification* in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)* (2015), 1067–1070.

396. Sarker, Y. *et al. Regularized Singular Value Decomposition Based Multidimensional Convolutional Neural Network for Hyperspectral Image Classification* in *2020 IEEE Region 10 Symposium (TENSYMP)* (2020), 1502–1505.

397. Du, H., Qi, H., Wang, X., Ramanath, R. & Snyder, W. E. *Band selection using independent component analysis for hyperspectral image processing* in *32nd Applied Imagery Pattern Recognition Workshop, 2003. Proceedings.* (2003), 93–98.

398. Ahmad, M., Mazzara, M. & Distefano, S. Regularized CNN Feature Hierarchy for Hyperspectral Image Classification. *Remote Sensing* **13.** ISSN: 2072-4292. `https://www.mdpi.com/2072-4292/13/12/2275` (2021).

399. Persello, C. & Bruzzone, L. Active and Semisupervised Learning for the Classification of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 6937–6956 (2014.).

400. Yang, L., Yang, S., Jin, P. & Zhang, R. Semi-Supervised Hyperspectral Image Classification Using Spatio-Spectral Laplacian Support Vector Machine. *IEEE Geoscience and Remote Sensing Letters* **11,** 651–655 (2014.).

401. Z.H-Zhou, & Li, M. Tri-Training: Exploiting Unlabeled Data Using Three Classifiers. *IEEE Transactions on Knowledge and Data Engineering* **17,** 1529–1541 (2005.).

402. Ly, N. H., Qian, D. & Fowler, J. E. Sparse Graph-Based Discriminant Analysis for Hyperspectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 3872–3884 (2014.).

403. Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F. & Emery, W. J. Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **47,** 2218–2232 (2009.).

404. Dopido, I. *et al.* Semisupervised Self-Learning for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* **51,** 4032–4044 (2013.).

405. Yang, L., MacEachren, A. M., Mitra, P. & Onorati, T. Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS International Journal of Geo-Information* **7,** 65 (2018.).

406. Yu, H., Yang, X., Zheng, S. & Sun, C. Active Learning From Imbalanced Data: A Solution of Online Weighted Extreme Learning Machine. *IEEE Transactions on Neural Networks and Learning Systems,* 1–16. ISSN: 2162-237X (2018).

407. Pasolli, E., Melgani, F., Tuia, D., Pacifici, F. & Emery, W. J. *Improving active learning methods using spatial information* in *2011 IEEE International Geoscience and Remote Sensing Symposium* (2011), 3923–3926.

408. Liu, A., Jun, G. & Ghosh, J. *Active learning of hyperspectral data with spatially dependent label acquisition costs* in *2009 IEEE International Geoscience and Remote Sensing Symposium* **5** (2009), V–256–V–259.

409. Tuia, D., Volpi, M., Copa, L., Kanevski, M. & Munoz-Mari, J. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing* **5,** 606–617. ISSN: 1932-4553 (2011).

410. Pasolli, E., Melgani, F., Tuia, D., Pacifici, F. & Emery, W. J. SVM Active Learning Approach for Image Classification Using Spatial Information. *IEEE Transactions on Geoscience and Remote Sensing* **52,** 2217–2233. ISSN: 0196-2892 (2014).

411. MacKay, D. J. C. Information-Based Objective Functions for Active Data Selection. *Neural Computation* **4,** 590–604. ISSN: 0899-7667 (1992).

412. Krishnapuram, B. *et al.* in *Advances in Neural Information Processing Systems 17* (eds Saul, L. K., Weiss, Y. & Bottou, L.) 721–728 (MIT Press, 2005). `http://papers.nips.cc/paper/2719-on-semi-supervised-classification.pdf`.

413. Li, J., Bioucas-Dias, J. M. & Plaza, A. Spectral–Spatial Classification of Hyperspectral Data Using Loopy Belief Propagation and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing* **51,** 844–856. ISSN: 0196-2892 (2013).

414. Luo, T. *et al. Active learning to recognize multiple types of plankton* in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* **3** (2004), 478–481 Vol.3.

415. Li, J., Bioucas-Dias, J. M. & Plaza, A. Hyperspectral Image Segmentation Using a New Bayesian Approach With Active Learning. *IEEE Transactions on Geoscience and Remote Sensing* **49,** 3947–3960. ISSN: 0196-2892 (2011).

416. Shi, Q., Du, B. & Zhang, L. Spatial Coherence-Based Batch-Mode Active Learning for Remote Sensing Image Classification. *IEEE Transactions on Image Processing* **24,** 2037–2050. ISSN: 1057-7149 (2015).

417. Demir, B., Persello, C. & Bruzzone, L. Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **49,** 1014–1031. ISSN: 0196-2892 (2011).

418. Lewis, David, D. & A., G. W. *A Sequential Algorithm for Training Text Classifiers* in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Springer-Verlag New York, Inc., Dublin, Ireland, 1994), 3–12. ISBN: 0-387-19889-X. `http://dl.acm.org/citation.cfm?id=188490.188495`.

419. Di, W. & Crawford, M. M. Active Learning via Multi-View and Local Proximity Co-Regularization for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Signal Processing* **5,** 618–628. ISSN: 1932-4553 (2011).

420. Patra, S., Bhardwaj, K. & Bruzzone, L. A Spectral-Spatial Multicriteria Active Learning Technique for Hyperspectral Image Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10,** 5213–5227. ISSN: 1939-1404 (2017).

421. Li, J. *Active learning for hyperspectral image classification with a stacked autoencoders based neural network* in *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (2015), 1–4.

422. David, A. C., Ghahramani, Z. & I, I. J. M. Active Learning with Statistical Models. *journal of Artificial Intelligence Research* **4,** 129–145 (1996).

423. Rajan, S., Ghosh, J. & Crawford, M. M. An Active Learning Approach to Hyperspectral Data Classification. *IEEE Transactions on Geoscience and Remote Sensing* **46,** 1231–1242. ISSN: 0196-2892 (2008).

424. Haines, T. & Xiang, T. *Active Learning using Dirichlet Processes for Rare Class Discovery and Classification* in *Proceedings of the British Machine Vision Conference* http://dx.doi.org/10.5244/C (BMVA Press, 2011), 9.1–9.11. ISBN: 1-901725-43-X.

425. Michel, J., Malik, J. & Inglada, J. *Lazy yet efficient land-cover map generation for HR optical images* in *2010 IEEE International Geoscience and Remote Sensing Symposium* (2010), 1863–1866.

426. Borisov, A., Tuv, E. & Runger, G. *Active Batch Learning with Stochastic Query-by-Forest (SQBF)* in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010* **16** (PMLR, 2011), 59–69. http://proceedings.mlr.press/v16/borisov11a.html.

427. Munoz-Mari, J., Tuia, D. & Camps-Valls, G. Semisupervised Classification of Remote Sensing Images With Active Queries. *IEEE Transactions on Geoscience and Remote Sensing* **50,** 3751–3763. ISSN: 0196-2892 (2012).

428. He, L., Li, J., Liu, C. & Li, S. Recent Advances on Spectral–Spatial Hyperspectral Image Classification: An Overview and New Guidelines. *IEEE Transactions on Geoscience and Remote Sensing* **56,** 1579–1597. ISSN: 0196-2892 (2018).

429. Yu, H., Sun, C., Yang, W., Yang, X. & Zuo, X. AL-ELM: One uncertainty-based active learning algorithm using extreme learning machine. *Neurocomput.* **166,** 140–150. ISSN: 0925-2312. http://dx.doi.org/10.1016/j.neucom.2015.04.019 (Oct. 2015).

430. Kumar, S. & Hebert, M. Discriminative Random Fields. *International Journal of Computer Vision* **68,** 179–201. ISSN: 1573-1405. https://doi.org/10.1007/s11263-006-7007-9 (2006).

431. Luca, A. D. & Termini, S. A Definition of a Non-Probabilistic Entropy in the Setting of Fuzzy Sets Theory. *Journal of Information and Control* **20,** 301–312. ISSN: 1932-4553 (1972).

432. Yeung, D. S. & Trillas, E. *Measures of Fuzziness under Different Uses of Fuzzy Sets* in *Advances in Computational Intelligence* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012), 25–34. ISBN: 978-3-642-31715-6.

433. Yuhas, R. H., Goetz, A. F. H. & Boardman, J. W. *Discrimination Among Semi-Arid Landscape Endmembers Using the Spectral Angle Mapper (SAM) Algorithm* in *Summaries of the 4th JPL Airborne Earth Science Workshop, JPL Publication* (JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop, NASA; United States, 1992), 147–149. ISBN: SEE N94-16666 03-42. https://ntrs.nasa.gov/search.jsp?R=19940012238.

434. van der Meer, F. D., Vazquez-Torres, M. & van Dijk, P. M. Spectral characterization of ophiolite lithologies in the Troodos ophiolite complex of Cyprus and its potential in prospecting for massive sulphide deposits. *International journal of remote sensing* **18,** 1245–1257. ISSN: 0143-1161 (1997).

435. Carvalho, O. A. D. & Meneses, P. R. *Spectral Correlation Mapper (SCM); An Improvement on the Spectral Angle Mapper (SAM)* in *Summaries of the 9th JPL Airborne Earth Science Workshop, JPL Publication 00-18* (Summaries of the 9th JPL Airborne Earth Science Workshop, JPL Publication, NASA; United States, 2000), 9–p. https://pdfs.semanticscholar.org/a59e/7e00b3c007c9ec3370370e0f3e966c277724.pdf.

436. Remondino, F. *et al.* Review of Geometric and Radiometric Analyses of Paintings. *The Photogrammetric Record* **26,** 439–461. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1477-9730.2011.00664.x (2011).

437. D. Singh, R. S. Evaluation of EO-1 Hyperion Data for Crop Studies in Part of Indo-Gangatic Plains: A Case Study of Meerut District. *Advances in Remote Sensing* **4,** 263–269. eprint: http://file.scirp.org/Html/1-2630132_61577.htm (2015).

438. Carneiro, T. *et al.* Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access* **6,** 61677–61685 (2018).

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

440. Mei, S., Chen, X., Zhang, Y., Li, J. & Plaza, A. Accelerating Convolutional Neural Network-Based Hyperspectral Image Classification by Step Activation Quantization. *IEEE Transactions on Geoscience and Remote Sensing*, 1–12 (2021).

441. Yuan, Y., Wang, C. & Jiang, Z. Proxy-Based Deep Learning Framework for Spectral-Spatial Hyperspectral Image Classification: Efficient and Robust. *IEEE Transactions on Geoscience and Remote Sensing*, 1–15 (2021).

442. Li, H.-C. *et al.* Recurrent Feedback Convolutional Neural Network for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters*, 1–5 (2021).

443. Yang, J., Hualong, Y., Xibei, Y. & Xin, Z. *Imbalanced Extreme Learning Machine Based on Probability Density Estimation* in *Multi-disciplinary Trends in Artificial Intelligence* (eds Bikakis, A. & Zheng, X.) (Springer International Publishing, Cham, 2015), 160–167. ISBN: 978-3-319-26181-2.

444. Woodward, M. & Finn, C. Active One-shot Learning. *CoRR* **abs/1702.06559.** arXiv: 1702.06559. http://arxiv.org/abs/1702.06559 (2017).

445. Lughofer, E. Single-pass active learning with conflict and ignorance. *Evolving Systems* **3,** 251–271. ISSN: 1868-6486. https://doi.org/10.1007/s12530-012-9060-7 (2012).

446. Liu, W., Chang, X., Chen, L. & Yang, Y. *Early Active Learning with Pairwise Constraint for Person Re-identification* in *Machine Learning and Knowledge Discovery in Databases* (eds Ceci, M., Hollmén, J., Todorovski, L., Vens, C. & Džeroski, S.) (Springer International Publishing, Cham, 2017), 103–118. ISBN: 978-3-319-71249-9.

447. Nie, F., Wang, H., Huang, H. & Ding, C. *Early Active Learning via Robust Representation and Structured Sparsity* in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (AAAI Press, Beijing, China, 2013), 1572–1578. ISBN: 978-1-57735-633-2. http://dl.acm.org/citation.cfm?id=2540128.2540354.

448. Ou, D. *et al.* A Novel Tri-Training Technique for the Semi-Supervised Classification of Hyperspectral Images Based on Regularized Local Discriminant Embedding Feature Extraction. *Remote Sensing* **11.** ISSN: 2072-4292. http://www.mdpi.com/2072-4292/11/6/654 (2019).

449. Hu, J., He, Z., Li, J., He, L. & Wang, Y. 3D-Gabor Inspired Multiview Active Learning for Spectral-Spatial Hyperspectral Image Classification. *Remote Sensing* **10.** ISSN: 2072-4292. http://www.mdpi.com/2072-4292/10/7/1070 (2018).