A survey on the classification of hyperspectral image data

Simon Vermeir, Ward Vercouter, Gerbrand De Laender, Dries Vansteelant

November 2020

Abstract

Hyperspectral imaging (HSI) is an imaging technique that captures more than just the visible spectrum; each pixel in a hyperspectral image comprises a spectrum that contains a wide and contiguous range of wavelengths with a fine spectral resolution. As opposed to classical RGB images, which only bear information about the visible spectrum (i.e. wavelengths we perceive as being red, green and blue), hyperspectral images are able to capture unique details in the scene under consideration, such as chemical decomposition and physical properties. Hyperspectral imaging is therefore a promising technique in the area of remote sensing, whose goal is to acquire information about objects without making physical contact with it. In this paper, the classification of different types of land cover and objects within hyperspectral images is considered. Classification of HSI images poses many challenges, among which the most prominent are the high dimensionality of the data and the scarcity of labeled training samples. This work provides a structured overview of HSI classification techniques along with the most common datasets and classification performance metrics. Typical challenges in HSI classification are also outlined, along with some state-of-the-art techniques that try to solve these. Ultimately, a comparison of the classification performance of some of these techniques is given, along with a personal take on future developments in the area of HSI classification.

1 Introduction

Since the 1970s, images in which pixels contain more information than just the visible spectrum, have been acquired and analysed by scientists [1]. The first Earth-observing satellite, launched in 1972 with the intent to study and monitor our planet's landmasses was NASA's Landsat 1 [2]. It was able to obtain information on agriculture, forestry, geology, hydrology, geography etc. by employing a multispectral scanner (MSS) that recorded data in four spectral bands: a green, a red and two infrared bands. Since then, capture and imaging modalities, processing hardware and processing techniques have made a huge leap forward and the resulting amount of information that can be extracted from these types of images has grown significantly.

Whereas multispectral imaging (or *multiband imaging*) measures a limited amount of discrete and spaced spectral bands, hyperspectral imaging measures continuous spectral bands with a fine wavelength resolution. Therefore, a pixel in a hyperspectral image is often referred to as a *hyperpixel* and contains information from twenty up to several hundreds of bands [3]. An example of a hyperspectral image (visualised as a 3D cube) is shown in Figure 1. The goal of hyperspectral image classification is to determine what type of ground cover or object is captured for each pixel of the image. This task may be more challenging than it seems at first glance: what if, for instance, one pixel covers more than one type of ground cover? What about distortions caused by the capturing equipment and atmosphere? The classification is further impeded by the limited amount of labelled training samples (*ground truth* images) and the scarcity of openly available hyperspectral image datasets in general, giving rise to the so called *Hughes phenomenon* which will be explained in section 3. To address these and several other challenges, numerous techniques have been developed, of which the most promising are based on deep learning.

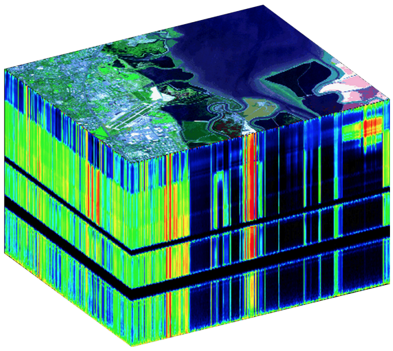

Figure 1: Example of a hyperspectral image [4]

This paper starts by giving an introduction to the most widely used hyperspectral image datasets, followed by an overview of the most prominent hyperspectral image classification techniques and with some commonly used metrics to evaluate their performance. In Section 3, the largest challenges in hyperspectral image classification are introduced and some state-of-the-art solutions are discussed. In Section 4 a comparison of the classification performance of some of these techniques is given, along with a personal take on future developments in the area of hyperspectral image classification. Section 5 concludes this paper.

2 Overview

This section aims at summarising the most prevalent approaches for hyperspectral image classification. First, an overview of the most prevalent datasets is given. Next, insight is given into the historical developments within this research domain using a timeline. Afterwards, classical learning approaches will be discussed, which form the baseline and inspiration for more advanced deep learning-based methods, which will be detailed after it. Finally, the most widely used metrics to evaluate the performance of these methods are presented.

2.1 Overview of HSI data

Hyperspectral images can be seen as three-dimensional data cubes. Two dimensions of this cube convey spatial information, the third dimension consists of (hundreds of) spectral bands that range from the visible light (∼400nm) to shortwave infrared (∼1000nm). The vast amount of information in a hyperspectral image makes it appealing for land class target detection (i.e. finding a specific type of soil) and classification applications. The spatial information can be used to extract properties such as size, shape and texture while the spectral information captures physical structure and chemical composition.

The most widely used datasets for HSI classification are named *Indian Pines*, *Salinas*, *University of Pavia*, *Kennedy Space Center (KSC)*, *Botswana* and *Data Fusion Contest (DFC) 2018*. The properties of the datasets are presented in Table 1 and include the total number of pixels in the image, the number of spectral bands, the range of these bands, the ground sample distance (GSD, i.e. the spatial resolution), the number of labels (i.e. labeled pixels), the number of ground cover classes and the capturing method. The datasets are ranked by prevalence within the citations of this paper: the *Indian Pines* and *University of Pavia* appear in 90.5% of them, followed by *Salinas* at 66.7% and *Kennedy Space Center* at 33.3%.

Figure 2 shows a false colour map of the *Salinas* dataset along with the corresponding ground truth map. The use of false colours is required, as hyperspectral images contain spectral information in regions that are invisible to the human eye. The 16 different classes, which are manually labelled, are represented using different colours. The figure also illustrates another challenge for classification: the imbalance between the classes. Minority classes (e.g. `Lettuce_romaine_7wk`) tend have lower classification accuracies when compared to larger classes (e.g. `Grapes_untrained`) if this is not taken into account during the training process. A stratified sampling strategy during training (i.e. balancing the number of training samples for every class) can counter this issue, but has the drawback that not all labelled samples of the majority classes are used.

2.2 Timeline

The historical evolution of popular HSI classification techniques is visualised in Figure 3 using a timeline that is based on Web of Science data. It shows the first occurrence of a given classification technique that gave rise to research in subsequent years. Hidden Markov random fields, for example, were already introduced around 2004, but only broke through about 10 years later [6]. The techniques shown will be discussed in the next two subsections and the timeline can therefore be used as a guidance. However, the broad trends will already be discussed in the next paragraphs.

The oldest widely used classification technique, *unmixing*, is purely based on linear algebra and was therefore already introduced mid 1990s. The technique was superseded by classical machine learning techniques such as *K-means clustering*, *support vector machines (SVM)*, *decision trees* and *Bayesian models* in the early 2000s. These techniques classify each pixel independently, thereby discarding any spatial information. Other spectral classifiers followed, such as *manifold learning* techniques, *random forests* and *multinomial logistic regression (MLR)*. It was only around 2013 when spatial-spectral classifiers, which also take into account the spatial information, became popular, among which *hidden Markov random fields (HMRF)* [7].

At the same time, deep learning approaches were being proposed to tackle the issue of the limited number of training samples. Because of their better classification performance, they became the de facto standard for hyperspectral image classification. Most modern techniques are mostly based on *convolutional neural networks (CNNs)*, which were introduced around 2015. They form the basis for the current state of the art, including techniques such as *residual networks (ResNet)*, *capsule networks (CapsNet)* and *generative adversarial networks (GAN)*, which are able to achieve near-perfect classification scores.

Table 1: Most widely used datasets in HSI classification, adapted from [5].

Dataset	Pixels	Bands	Range	GSD	Labels	Classes	Mode
Pavia (U & C)	991,040	103	0.43-0.85 μm	1.3 m	50,232	9	Aerial
Indian Pines	21,025	224	0.4-2.5 μm	20 m	10,249	16	Aerial
Salinas	111,104	227	0.4-2.5 μm	3.7 m	54,129	16	Aerial
KSC	314,368	176	0.4-2.5 μm	18 m	5,211	13	Aerial
Botswana	377,856	145	0.4-2.5 μm	30 m	3,248	14	Satellite
DFC 2018	5,014,744	48	0.38-1.05 μm	1 m	547,807	20	Aerial

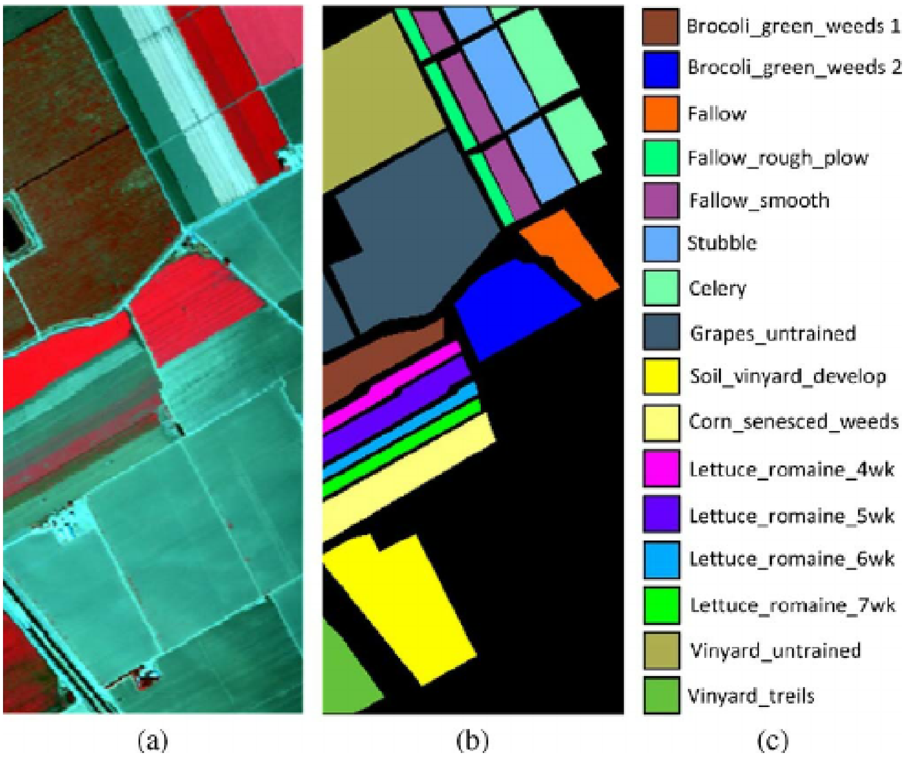

Figure 2: False colour map (a), ground truth map (b) and labels (c) from the Salinas dataset

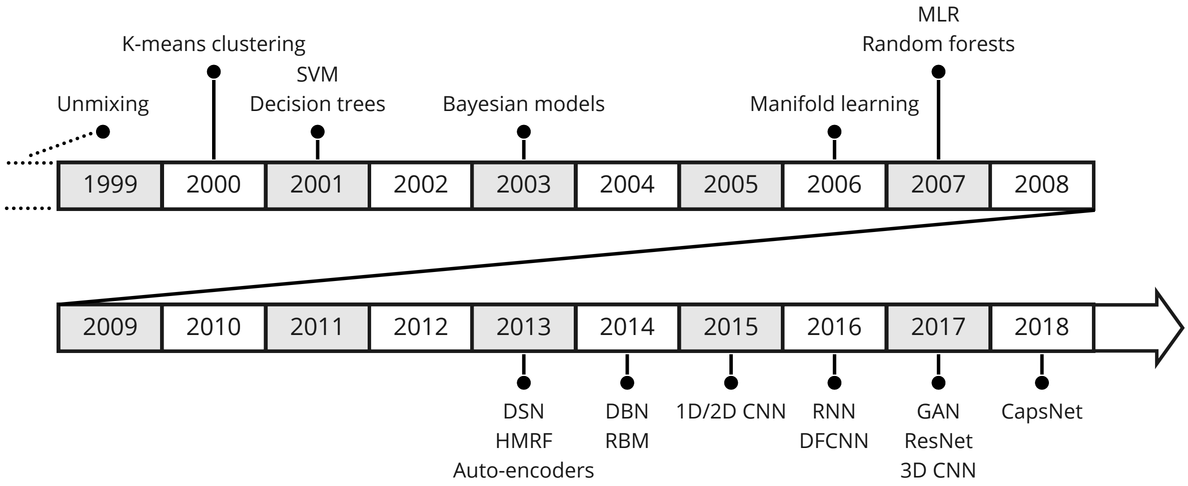

Figure 3: Timeline of various methods

2.3 Classical learning approaches

In the literature, different kinds of supervised and unsupervised approaches have been proposed to classify HSI data. A tree-based overview of classical learning approaches is given in Figure 4, and includes both spectral and spatial-spectral classifiers. The former classification method does not take into account the spatial relationship between hyperspectral pixels, making them less robust and less efficient because they cannot exploit the structural similarity between neighbouring pixels. In this paper, techniques that are not based on deep neural networks are considered as a "classical" approach.

2.3.1 Supervised methods

Supervised methods usually provide the most accurate results by learning the data relations from a given training set that contains ground-truth information. A wide range of traditional, supervised machine learning patterns have been successfully applied for hyperspectral image classification, including support vector machines (SVM) [8], multinomial logistic regression (MLR) [9], decision trees and random forests [10], Bayesian models [11] and manifold learning [12]. MLR and SVMs are both able to handle big datasets with a relatively low number of training samples and can be used for both spectral and spatial-spectral classification, for example by using specialised SVM kernels. Manifold learning tries to simplify the representation space by assuming that the original datasets lie on a common manifold in order to tackle the high dimensionality of the HSI data.

2.3.2 Unsupervised methods

Unsupervised methods, while usually being less accurate, do not require a supervised training phase, making them appealing for scenarios where poor prior knowledge of the scene is available. In such cases, they can provide insight into the complex HSI data by uncovering hidden data interactions and correlations that can be used for both segmentation and classification purposes. *Unmixing* [13] is a classical unsupervised technique that assumes that a hyperspectral pixel can be seen as a simple linear combination of the spectra of the pure materials that make up a scene. By feeding the model with pure material spectra, it is usually possible to apply linear algebra inversion techniques and numerical methods to find the most likely composition of the hyperspectral pixel [14]. Other unsupervised methods include clustering and segmentation methods such as *K-means clustering* [15] and *hidden Markov random fields (HMRF)* [16], and aim to improve the classification accuracy of spectral classifiers by incorporating spatial dependences. The latter is a statistical segmentation approach that generalises the idea of a hidden Markov model and assumes that if a hyperspectral pixel has a certain label, that neighbouring pixels also are very likely to have the same label.

2.4 Deep learning approaches

A drawback of these classical segmentation and classification approaches is that they require lots of feature engineering to improve their performance. Deep learning techniques, on the other hand, are focused on representational learning, i.e. is the automatic design of a feature space that is tailored to the objective task at hand. The joint optimisation of both representation and classification allows to achieve better performing models. An overview of the most prevalent deep learning classification methods is given in Figure 5.

2.4.1 Supervised methods

Supervised or *discriminative* deep networks work with labelled information with the goal to categorise new input data into these labels. Within the deep learning methods, they are often most efficient and most flexible to train and test. The supervised deep networks can be seen as a non-linear mapping from the feature space to the label space, which can allow higher levels of expressibility by using a hierarchy of layers that are connected through linear or non-linear activation functions.

Deep fully-connected neural networks (DFCNN) are a logical replacement of the standard shallow learning classifiers, because the principle remains the same. However, DFCNNs can model the classification task in a finer way and with a better discriminating power [17][18]. Another type of artificial neural networks are the *recurrent neural networks (RNN)*, in which connections between nodes form a directed graph along a temporal sequence. Using a memory of the past information, the model allows to exhibit temporal dynamic behaviour that can be used to process time series. By treating the spectral dimension of the HSI data as a time series, i.e. a sequence of reflectances, RNNs can be used to perform spectral classification [19][20]. It is also possible to assemble basic function modules or classifiers and stack them up to each other in order to learn complex tasks. This is the idea behind *deep stacking networks*, which have proven their validity for hyperspectral image classification [21].

The most widely used supervised deep learning technique are the *convolutional neural networks (CNN)*. They take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. On the scale of connectedness and complexity of deep neural networks, CNNs are on the lower extreme. Convolutional neural networks are composed by a set of blocks that transform input volume to an output volume of neuron activations, which will serve as the input to the next block. In contrast to multilayer perceptrons (MLP), the neurons of a block

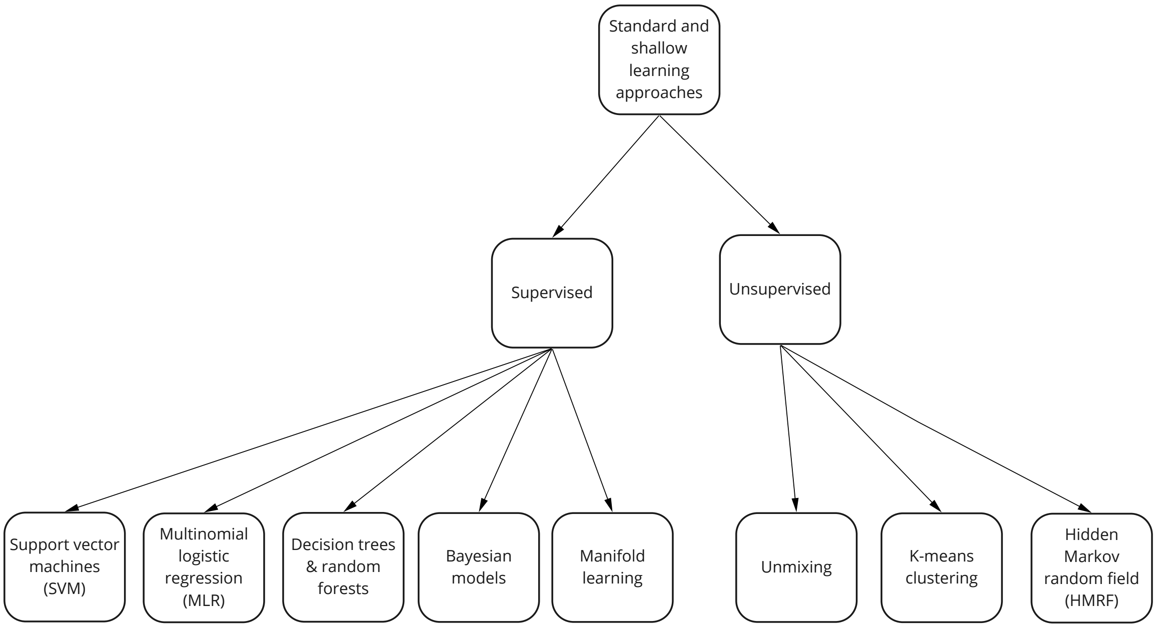
Standard and shallow learning approaches
Supervised
Unsupervised
Support vector machines (SVM)
Multinomial logistic regression (MLR)
Decision trees & random forests
Bayesian models
Manifold learning
Unmixing
K-means clustering
Hidden Markov random field (HMRF)

Figure 4: Classical learning based techniques HSI classification techniques

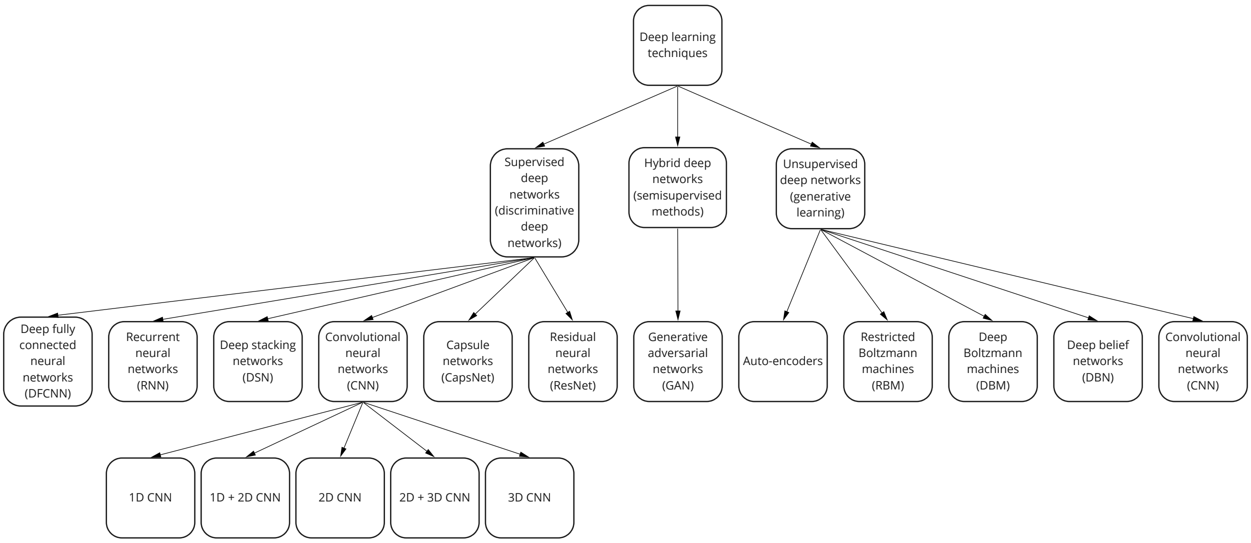
Deep learning techniques
Supervised deep networks (discriminative deep networks)
Hybrid deep networks (semisupervised methods)
Unsupervised deep networks (generative learning)
Deep fully connected neural networks (DFCNN)
Recurrent neural networks (RNN)
Deep stacking networks (DSN)
Convolutional neural networks (CNN)
Capsule networks (CapsNet)
Residual neural networks (ResNet)
Generative adversarial networks (GAN)
Auto-encoders
Restricted Boltzmann machines (RBM)
Deep Boltzmann machines (DBM)
Deep belief networks (DBN)
Convolutional neural networks (CNN)
1D CNN
1D + 2D CNN
2D CNN
2D + 3D CNN
3D CNN

Figure 5: Deep learing based HSI techniques classification

are not fully connected to all neurons of the previous layer. CNNs can therefore be seen as regularised versions of multilayer perceptrons that are much less prone to overfitting [22].

CNNs can be further decomposed in the context of HSI classification, depending on whether they use spectral, spatial or spatial-spectral features. Considering only the spectral data gives rise to a 1D CNN architecture: the inputs to the network are $N \times 1$ input vectors, with N the number of spectral bands. Each pixel is therefore assigned a single land cover (material) type [23]. In contrast, models can also only take into account the spatial information in the hyperspectral image, resulting in a 2D CNN architecture: the inputs to the network are $K \times K$ patches of neighbouring pixels. To add spectral information to these models, a pre-processing step can for example flatten the spectral dimension of the hypercube [24] or force it into a three-channel image such that traditional 2D CNNs can be borrowed from traditional computer vision applications [25]. It is also possible to combine spatial and spectral classifiers, resulting in a 1D + 2D CNN architecture. In this case, 1D convolution kernels are iteratively applied to the spectral dimension, while 2D kernels are being used for the spatial dimensions. Eventually, a third classifier or a stack of fully connected layers can perform the final classification step [26]–[28]. Taking full advantage of the spatial-spectral characteristics of the 3D remote sensing data can be done by using a three-dimensional kernel that operates on blocks of size $K \times K \times N$, which results in a 3D CNN architecture. By processing the HSI hypercube as a whole, higher classification accuracies can be achieved than using lower-dimensional convolutional neural networks [29]. However, too deep or too shallow networks may lead to poor classification performance. Ge et al. [30] recently proposed a multibranch neural network that combines 2D and 3D CNNs to extract image features. It combines a 2D CNN that under-utilises the inter-band correlation of HSIs with a 3D CNN that depends on a more complex model, achieving even better classification scores.

A drawback of higher dimensional CNNs is that they require a lot of training data to capture the full relationship between the input and the labels. More recent research has been been focusing towards reducing this need of training data, which is very scarce in the context of HSI.

Capsule networks (CapsNet) is such a technique that requires less training data. The network is able to capture (parts of) objects in the scene, along with parameters that describe it (the associated *instantiation parameters*). Also, it is able to capture spatial hierarchies within the image [31]. The functioning of capsule networks is detailed in Section 3.1.3.

Another technique that is being adopted is the use of *residual networks (ResNet)*. Very deep neural networks are difficult to train because of problems related to vanishing and exploding gradients: the accuracy of a network first saturates and then degrades rapidly as more layers are added. This is because of the nature of gradient-based learning methods: in a network of N hidden layers, N derivatives of the loss function will be multiplied together. If the derivatives are large, the gradient will increase exponentially while propagating down the model until it eventually "explodes". Alternatively, the gradient tends to "vansish" when the derivatives are small. ResNets try to tackle this issue by introducing skip-connections to jump over some layers. The activation from one layer is hence fed to another layer, typically two or three layers deeper in the neural network. Such a residual unit can be replicated many times, which allows to train very deep networks (often more than 100 layers) that can outperform standard deep CNNs in HSI analysis and classification [32]. ResNets have also been used in combination with *fully convolutional networks (FCN)*, resulting in a lot less parameters to tune and a speedup in inference time as FCNs are able to predict all hyperspectral pixels in an input patch, rather than just the central pixel [33].

2.4.2 Unsupervised methods

Unsupervised deep networks (*generative networks*) look for patterns between pixels through capturing high-order correlation of data [22]. A possible approach to unsupervised spatial-spectral classification is to use one of the following unsupervised techniques to one-dimensional standard descriptors. The descriptor vector typically consists of the full radiometric spectrum of a pixel, to which the most prominent spatial features (e.g. using PCA), calculated on a local neighbourhood, are concatenated [5].

auto-encoders are neural networks in which the desired output is equal to the input and passes through a single layer of neurons (linear) or multiple layers (non-linear). They effectively learn efficient data codings and turn out to be more efficient than standard PCA. auto-encoders can for example be stacked to denoise images and play a crucial role in reducing the dimensionality of hyperspectral data [34]. Other unsupervised techniques include *restricted Boltzmann machines (RBM)* and *deep Boltzmann machines (DBM)*, which are variants of regular Boltzmann machines that can learn a probability distribution over its set of inputs [35]. In RBMs, neurons must form a bipartite graph, whereas in DBMs, there are multiple layers of hidden random variables. Furthermore, simple unsupervised networks such as auto-encoders and RBMs can be used to compose a *deep belief network (DBN)*, where each hidden layer of a sub-network serves as the visible layer for the next. The core of DBNs is a greedy learning algorithm that optimizes the network weights layer by layer. Li et al. [36], for example, apply this technique on airborne HSI data for feature extraction and classification. Finally, it should be noted that convolutional neural networks can also be used for unsupervised learning. Romero et al. [37] for example use a single-layer CNN in an efficient algorithm for unsupervised learning of sparse features.

2.4.3 Semi-supervised methods

A third and more recent deep learning approach is making use of both labelled and unlabelled data for training. The most promising HSI classification technique in this category is the use of *generative adversarial networks (GAN).* A GAN uses labelled training data to generate new samples that have the same properties (and hence look at least superficially authentic to human observers) as the original data. Internally, they consist of 2 networks, a discriminator and a generator network. The generator tries to generate new samples while the discriminator tries to distinguish real samples from the generated samples. Several methods have been proposed in the literature that outperform state-of-the-art HSI classification techniques by combining GANs and CapsNets [38]–[40].

2.5 Performance metrics

In image classification applications, three metrics are widely used: overall accuracy (A_O), average accuracy (A_A) and Cohen's kappa coefficient.
The overall accuracy is equal to the percentage of correctly classified pixels [30] as shown in the following equation.

$$A_O = \frac{N_c}{N_t},$$

with N_c equal to the amount of correctly classified samples and N_t equal to the total number of samples.
Average accuracy denotes the mean value of the overall accuracies measured over each class [30]. It is calculated as follows:

$$A_A = \frac{\sum_{c \in C} A_c}{|C|},$$

where A_c denotes the overall accuracy calculated from the samples in a specific class c and C is the set of all classes.
Lastly, the Kappa coefficient states the degree of agreement between the true values and the predicted values and is calculated as follows [41]:

$$\kappa = \frac{A_O - p_e}{1 - p_e},$$

where p_e denotes the chance agreement and is defined as

$$p_e = \frac{\sum_{i=1}^{|C|} N_{.i} N_{i.}}{(N_t)^2},$$

with $N_{.i}$ and $N_{i.}$ the sums of elements in the i-th column and the i-th row of the confusion matrix, respectively. N_t is equal to the total number of samples as mentioned earlier.

3 Challenges

Two major challenges in hyperspectral image classification, the imbalance between the high dimensionality of the data and the limited number of training samples available, are embodied in what is called the *Hughes phenomenon* [42]. It states that classification accuracy increases gradually with a growing number of spectral bands or dimensions, but decreases dramatically when the band number reaches some value. This is because the separability of classes increases with increasing dimensionality, but so does the number of statistical parameters defining the classes. Since there are only a fixed number of training samples for deriving the statistical parameters, at some point the accuracy of the estimation must begin to decrease [43]. This section presents and discusses several solutions that tackle this issue.

3.1 Limited training data

Supervised classifiers are often preferred over unsupervised ones because of their capacity to provide high accuracies, but these methods may be affected by the limited availability of training samples [42]. The lack of training data usually to overfitting if the model is complex enough. Typical approaches used to reduce overfitting for HSI classification include:

- Using a smaller kernel size in CNN methods [44].
- Introduce pooling layers in CNN methods to reduce the number of parameters in the network [30].
- The use of *dithering* can also mitigate the issue of overfitting in CNNs [45]. In dithering, additive Gaussian noise is added to the training samples to suppress inherent nonlinear distortion and aliasing introduced by the neural network, hence regularizing the CNN.
- For neural networks, using of a dropout mechanism that sets output of some randomly selected hidden neurons to 0 such that they are not used in the backpropagation. This is especially beneficial for fully connected layers. [42].
- Apply L2 regularisation, although existing methods do not seem to suffice [42].

Unfortunately, a recurring problem within HSI classification is the lack of good labelled data. The number of training samples in the HSI field is rather limited compared to the number of available spectral bands [32]. The reason for this is that labelling HSI data is a labour intensive process [40].

This fact typically results in an under-complete training process which is prone to overfitting, i.e., the *Hughes phenomenon.* Additionally, spectral redundancy and noise are often present in HSI, since contiguous bands tend to be highly correlated, and the physical limitations of the acquisition technology always introduces some sort of signal perturbations [32]. The robustness of a model can thus be improved by either augmenting the training set or using a model that requires less data.

3.1.1 General approaches

Yushi Chen Et Al. [46] uses radiation-based and mixture-based sampling as a form of image augmentation.
Radiation-based sampling changes the radiation within an image (light intensity), simulating the capturing procedure, since often times light conditions are different resulting in different images.
Mixture-based augmentation mixes two samples from the same class. Using ratios it is able to create a new sample. Mixture-based augmentation is often used in remote sensing because of the long distance between object and sensor.

Using their 3D CNN, they achieve state-of-the-art results. Combining both 3D CNN and the aforementioned techniques they show improvement over using 3D CNN alone. These augmentations are straightforward and effective, giving them a strong advantage.

Mercedes E. Paoletti [42] addresses limited availability of training samples by using semi-supervised and active learning methods. Such combinations have been done with morphological component analysis (MCA), which decomposes images into texture and cartoon parts or into a smoothness and texture component in case of multiple morphological component analysis (MMCA).

More recent methods use generative models to simulate realistic hyperspectral scenes, such as Gaussian mixture models (GMM) and generative adversarial networks (GANs) [47]. They are able to synthesise hyperspectral pixels or patches of pixels from scratch. While the benefit of adding new training data generated by such a model is not really substantial at this point, these models will likely be used to estimate specific domain specific image transformations such as atmospheric corrections and transfer function estimation between sensors and image denoising.

Finally a new type of network called a capsule network (CapsNet) [48] promises an improvement over traditional CNN architectures even while dealing with limited training samples and providing better accuracies. The next two sections will cover GANs and CapsNets in more detail.

3.1.2 Generative adversarial networks

Ian Goodfellow et al. [47] introduced the Generative adversarial network or GAN for short. A GAN uses training data to generate new samples that have the same properties as the original data.

A GAN conists of 2 networks, a discriminator and a generator. The generator tries to generate new samples while the discriminator tries to distinguish real samples from generated samples. In this way the two networks contest with each other, thereby improving each other during training.

Odena et al. [49] used a method where the discriminator could also classify samples given labelled data. In this way a GAN offers two functions, generating samples and classifying samples.

Chongxuan Li et al. [29] extended GAN to triple GAN, by adding a third network called a classifier. Given labels, the classifier will classify either generated or true samples and classify them. This alleviates some of the pressure on the discriminator, allowing better results.

GAN and its variations have recently become popular for image augmentation. Making it an ideal candidate for HSI classification were a lack of data poses a problem.

Lin Zhu. et al. [38] introduced a GAN for the first time to do HSI classification. Both a 1D-GAN and 3D-GAN was introduced, the former being a spectral classifier and the latter a spectral-spacial classifier. The 1D-GAN uses one single spectral pixel to feed the network, leaving any spatial information behind, where the 3D-GAN uses a region of pixels to feed to the network. Both use a similar design based on the work of Ian Goodfellows [47] and Odena [49], as shown in Figure 6. Internally the discriminator and generator are both a type of CNN's. Hyperspectral images have high dimensionality since they consist of several spectral bands (ranging from twenty up to several hundreds of bands [3]). Therefore it is difficult to train the generator. Adjustments where made to deal with the high amount of redundancy. Notably principal component analyses (PCA) was done. For the 1D-GAN it has been found that extracting 10 principle components was the best. For 3D-GAN, 3 principle components proved sufficient because of the additional presence of spatial information.

Two networks need to be trained, therefore one drawback of using a GAN is training time. Nevertheless Lin Zhu et al. [38] were able to get competitive results using a GAN under the presence of limited training samples, highlighting the promise of using a GAN for HSI classification.

Finally, Xue Wang et al. [50] applied a Triple GAN in conjunction with a capsule network.

The input to the network is a concatenated 1D spectral vector and a 1D spatial vector. The 1D spectral vector are the spectral bands of 1 pixel. The spatial vector is obtained by applying PCA to the image, retaining three components. Afterwards using a patch of pixels as the 1D spatial vector. This approach proved to be unsuitable. The authors noted that an end-to-end approach is better, i.e. avoiding handcrafted features and using spectral-spatial features simultaneously.

Their approach showed that the combination of both

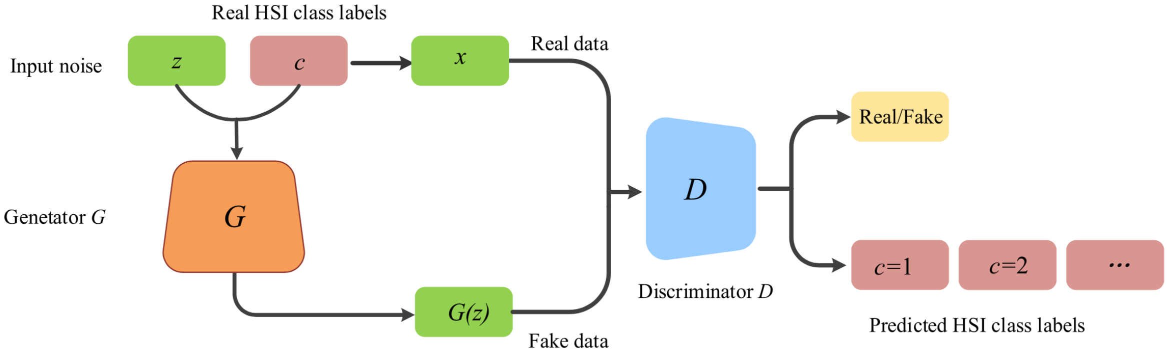

Figure 6: Architecture of a GAN for HSI classification, proposed by Lin Zhu. et al. [38]

networks resulted in better results when only 10% labelled samples were present compared to using only a Triple GAN, again achieving competitive results.

3.1.3 Capsule Networks

Geoffrey Hinton et al. [48] put forward Capsule Networks for the first time. A capsule network works a bit like inverse rendering [48]. It is able to recognise objects within an image and its associated instantiation parameters. It is meant to overcome the drawbacks of CNNs. Such network is able to capture spatial hierarchies (relationships between simpler features that make up a larger feature [51]) and needs less training data.

The architecture of a capsule network is typically split up in an encoder and a decoder network. The encoder encodes the image into its classes. The decoder tries to reconstruct the original image and is used as regulariser. It forces the capsule layers to capture features that are useful to reconstruct the image [51]. The decoder consists of: a convolutional layer, primary capsule layer and a digitCaps layer (capsule layer). The encoder consists of a three fully connected layers. A capsule layer consists of capsules.

A capsule captures an object. The output of each capsule is an activity vector. The length of the activity vector tells us the likelihood of finding that object. For example the probability of finding a circle in the image. The orientation of the activity vector tells us the instantiation parameters of that object, for example the color of the circle.

The ReLu activation function typically used in CNNs is replaced by a squashing function, along with max pooling that is replaced with routing by agreement [31]. It will take a weighted sum of the activation outputs of the previous layer and create a vector that is scaled between zero and one, while retaining direction [40]. The goal of the squashing function is to scale short vectors to almost zero and long ones to almost one [48]. Routing by agreement ensures that when activity vectors from a higher-level capsule have a big scalar product with predictions from lower-level capsules, that the output of these lower-level capsules are preferred to be send to these [48]. Interestingly, the digitCaps layer has the same amount of capsules as classes present.

The fact that they capture spatial hierarchies, recognise classes and are able to deal with less training data makes them suitable for HSI classification.

Luo et al. [52] used capsule networks for the first time in HSI classification, but could not find the benefits [53]. Later, Fei Deng et al. [53] and Mercedes E. Paoletti et al. [31] both released a paper around the same time applying a capsule network to the task of HSI classification. Both followed a similar architecture as described above. A patch of a HSI image is identified and used as input to the network. Just like a 3D CNN it is therefore able to capture spatial features.

Fei Deng et al. [54] did not use the decoder network and used the final capsule layer as the class identifier, where each capsule represented a unique class presented in the data. The highest output was the identified class. Notably, they also introduced a dropout layer in between the convolutional layer and the first capsule layer, in order to fight overfitting. They also found that their capsule network showed significantly higher confidence in predictions.

Mercedes E. Paoletti et al. [31] used an architecture and decoder similar to the one found by Geoffrey Hinton et al. [48]. They also highlighted that capsule networks had better border delineation (based on visual identification) within the classified HSI images compared to other methods.

Finally, both papers were able to achieve state-of-the-art results while dealing with limited training samples. Nonetheless, it is worth pointing out that capsule networks are still in their early days. Nevertheless, the results point out that capsule networks hold great promise within the field of HSI classification and deep learning in general.

3.2 Dimensionality

As with all learning applications, "The Curse of Dimensionality" also applies for HSI classification. As discussed in the previous section, good labelled data is scarce. Combined with the high dimensionality of the samples this can be detrimental for the accuracy of the classification [16]. In recent years, spatial information is used in conjunction with spectral information. These spatial-spectral methods suffer especially hard from high dimensionality [55]. Several methods have been developed for learning discriminative, uncorrelated features from this high-dimensional data.

3.2.1 General approaches

Most of the methods mentiones in the literature employ band selection, which is removing spectral bands that are do not contain discriminative power or contain outliers. Prime candidates to be cut are bands with low Signal to Noise Ratio (SNR), saturated bands and bands related to water absorption [5]. Methods to remove these outliers include PCA, mutual information maximization [56] or an entropy-based metric [57]. Paoletti et al. [42] propose a novel CNN-based deep network architecture based on ResNet specifically crafted to deal with large data cubes. The potential of available information on each residual block is better exploited than the traditional ResNet architecture by using a pyramidal structure: the proposed approach gradually increases the feature map dimension at all convolutional layers that make up the –now pyramidal– residual block. This gradually increases the diversity of high level spatial-spectral attributes across layers, enhancing the classification performance of the network when compared to state-of-the-art HSI classification methods.

More involved methods are the use of evolutionary-based optimisation of feature selection.

3.2.2 Auto-encoders

Another way to address the abundance of data in the hyperspectral images is making use of auto-encoders. Zhou et al. [55], Zhang et al. [59] and Tao et al. [58] use different implementations of s to reduce the dimensionality of the hyperspectral data and learn features in an unsupervised manner.

Tao et al. [58] propose a spectral-spatial feature learning framework and experimentally show improved accuracy over previously handcrafted features. They also demonstrate that their feature learning model can be shared among multiple related images. Many related images can thus be efficiently classified since the feature learning step only needs to be performed once. Tao et al. use a stacked sparse auto-encoder, a neural network consisting of multiple layers of basic shallow sparse auto-encoders (SSAE). In shallow sparse auto-encoder, the shallow means that only one hidden layer is present and sparse means that there is a sparsity constraint on the hidden layer. this encourages the network to activate a limited part of the network for a given input and deactivating the rest. The stack of auto-encoders can be constructed by wiring the output of the hidden layer from one sparse auto-encoder h_1 into the input of the hidden layer of another h_2. This is illustrated in Figure 7. The input x is used to learn the primary feature $h_1 = f_1(x)$ (Figure 7 a). This primary feature is then used as input for a second shallow sparse auto-encoder to learn the secondary feature $h_2 = f_2(h_1)$ (Figure 7 b). Finally, the auto-encoders are stacked together to form an SSAE (Figure 7 c). The resulting function transforms an input x to a deep feature representation $h_2 = f_2(f_1(x))$. In practice, they used two SSAE's, with two hidden layers, in parallel, one for learning from the spectral data, where every pixel is represented by a d-dimensional spectral vector. In the second, each pixel is represented as multiple image patches with different sizes. When both types of features are learned, they are combined in one feature vector and fed into a linear SVM for classification.

Zhang et al. [59] propose the use of recursive auto-encoders (RAE). A binary tree is built where every node represents an auto-encoder. the leaves of the tree have the HSI data of one pixel and other nodes take the combined output of one leaf node and one non-leaf node as input, except for one, which takes the output of two leaf nodes as input. To test their features they use SVM on features learned from the Indian Pines and Pavia datasets.

Zhou et al. [55] try to reduce the complexity of HSI classification without a drop in performance. They propose a compact and discriminative stacked auto-encoder (CDSAE) with as small a number of hidden neurons as possible. The proposed framework consists of two stages, first, one for low-dimensional feature mapping learning (feature extractor training) and the latter for the joint training of HSI classifier and feature extractor. Both spectral and spatial information is used for feature learning and they use PCA for dimensionality reduction. To increase the compactness and discriminative power of the classifier, each hidden layer of the CDSAE is fine-tuned with a local Fisher discriminant regularisation and a diversity regularisation.

3.2.3 Evolutionary-based optimisation of feature selection

Exhaustive feature (or band) selection approaches require huge amounts of computational power and memory. A new trend to select features is the use of (gradient-free) evolutionary-based optimisation approaches, such as genetic algorithms (GA) and particle swarm optimisation (PSO).

Genetic algorithms (GA) [45] are based on the idea of biology and evolution. First, the algorithm generates

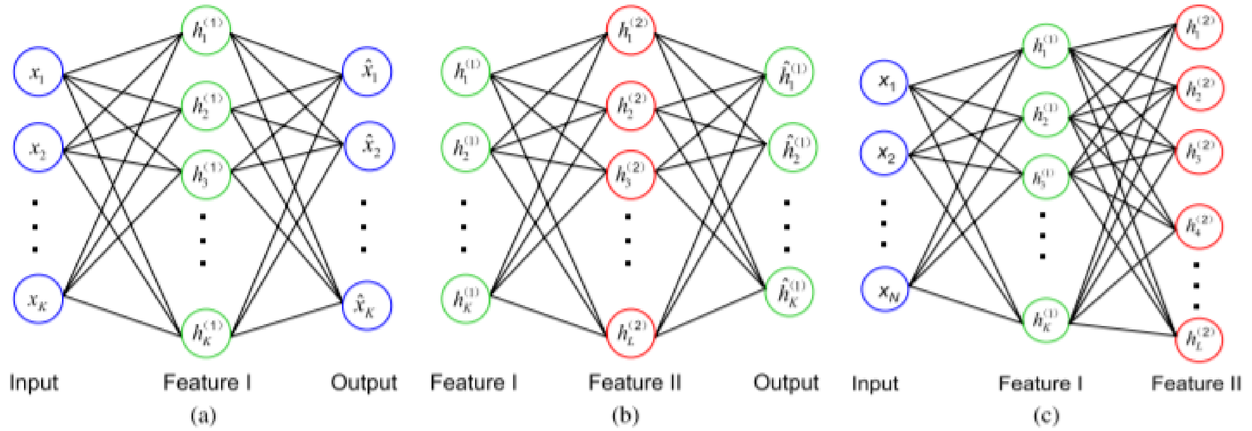

Figure 7: Structure of the stacked auto-encoder model [58]

many possible solutions that form a population. The solutions are scored using a fitness function (objective function) to decide which solutions are better than others. These candidate-solutions are then recombined so that the best solutions reproduce to form a new generation of solutions, with the best traits of the previous solution. This continues until improvement stops or until some maximum number of generations is reached. Ghamisi et al. use GA to tune hyperplane parameters of an SVM by selecting efficient features to be fed to the classifier.

PSO is similar to GA in that it creates a population of solutions (or *swarm*) for each iteration. Each solution (or *particle*) in the swarm has a direction and a velocity. At each iteration, the movement of the particle is determined by a mixture of the current direction, the direction of the best point it has found in the past, and the direction of the best point that the whole swarm has discovered. The idea is that more and more particles will eventually move to areas where better solutions are found, and that the swarm will eventually converge to the optimal value. Darwinian PSO (DPSO) also incorporates a natural selection process, or survival of the fittest, to enhance the ability to escape from local optima. To further improve on premature convergence of the swarm to a non-optimal point, the concept of fractional calculus is used to control the convergence rate of the DPSO, leading to a fractional-order Darwinian PSO (FODSPO). Ghamisi et al. propose a self-improving CNN (SICNN) that addresses the lack of available training samples by automatically selecting the best set of bands suitable for the defined network. They use FODPSO to select an optimal set of features and use the overall classification accuracy on validation samples as fitness value. Their results indicate that the method can significantly improve the classification accuracies of the CNN when there are only a limited number of training samples available.

4 Discussion

4.1 Evaluation

Paoletti et al. [42] showed that MLP is usually faster than CNN as shown in Table 2. However, CNN results in much better accuracies than MLP.
Differences between 1D CNN, 2D CNN and 3D CNN are also demonstrated. As 1D CNN only explores the spectral features and 2D CNN only the spatial features, it is self-evident that 3D CNN, which handles both spatial and spectral features, performs better accuracy-wise. The difference in overall accuracy between 1D CNN and 2D CNN is visibly smaller than the difference between 2D CNN and 3D CNN. However, the difference between the average accuracy between 1D CNN and 2D CNN is greater than the difference between 2D CNN and 3D CNN. These observations are also valid for the University of Pavia dataset as can be seen in Table 3.

In Figures 8, 9 and 10, the accuracy measurements of different methods on the Indian Pines, Salinas and University of Pavia datasets are shown. The graphs were compiled with data from [3], [7], [16], [22], [30]–[32], [38], [40], [44]–[46], [53]–[55], [59]–[63]. It is important to note here that these measurements can be taken from a very different environment e.g. the amount of classes can be cropped, the amount of test samples per class can vary, etc.
It is remarkable that the average accuracy does not always scale compared to the Kappa coefficient and average accuracy. This is the case for the method with recursive auto-encoders for the Indian Pines data set. In [59] it is however not mentioned what the cause of this behaviour is.
It can also be seen that all the accuracy measures of the most recent methods have values close to 100. Therefore, we can ask ourselves whether much improvement is still possible today.

Table 2: Accuracies and runtime obtained by different neural networks tested using the Indian Pines dataset

	Overall Accuracy	Average Accuracy	Runtime (sec.)
MLP	84.60	91.66	0.18
1D CNN	87.81	93.12	457.80
2D CNN	89.99	97.19	357.00
3D CNN	97.56	99.23	1675.20

Table 3: Accuracies and runtime obtained by different neural networks tested using the University of Pavia dataset

	Overall Accuracy	Average Accuracy	Runtime (sec.)
MLP	88.20	90.39	0.15
1D CNN	92.28	92.55	994.80
2D CNN	94.04	97.52	607.19
3D CNN	99.54	99.66	2769.00

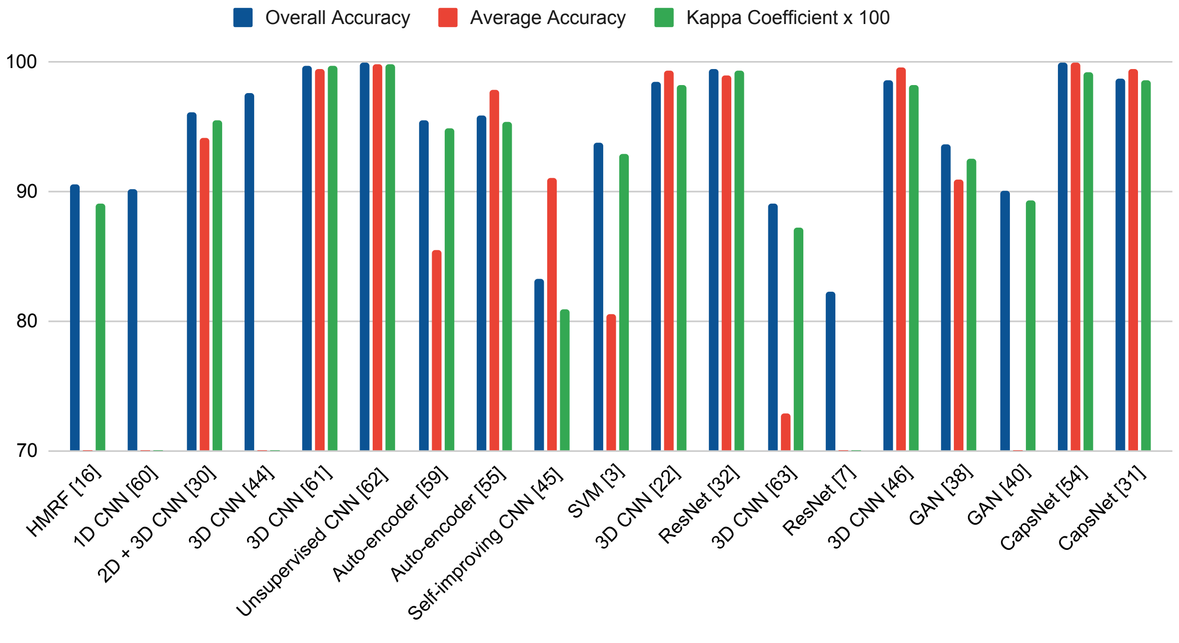

Figure 8: Accuracies of the Indian Pines dataset

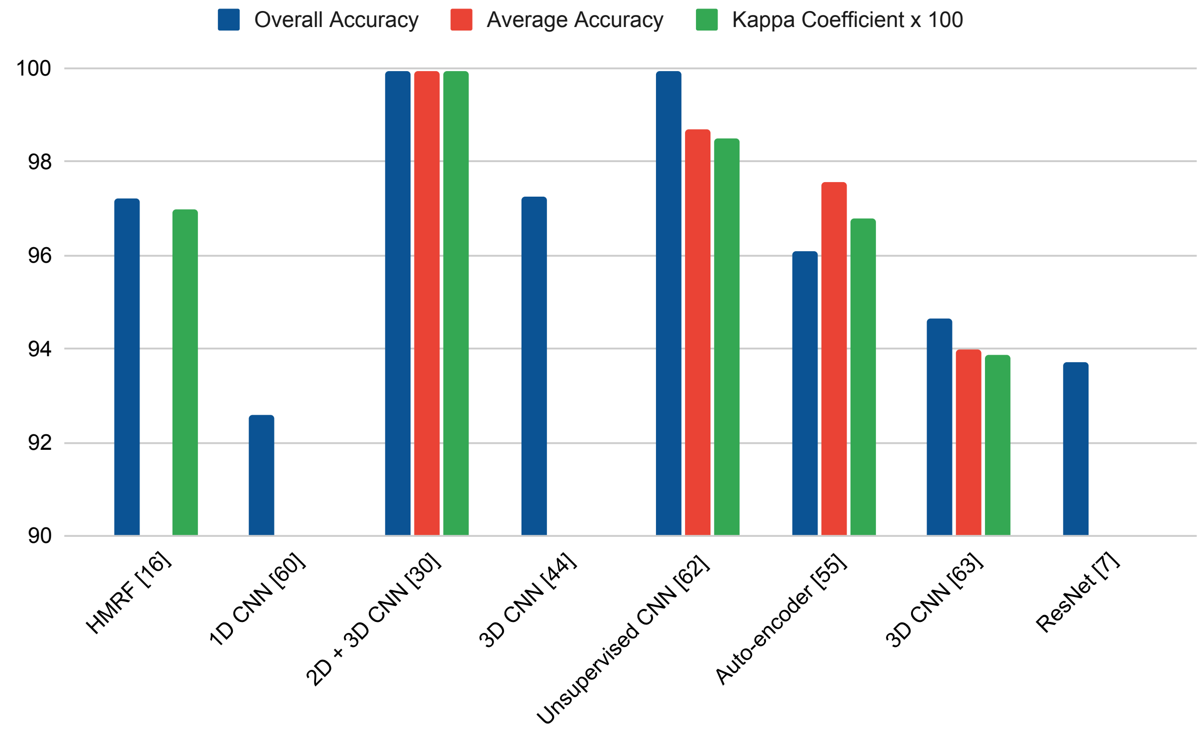

Figure 9: Accuracies of the Salinas dataset

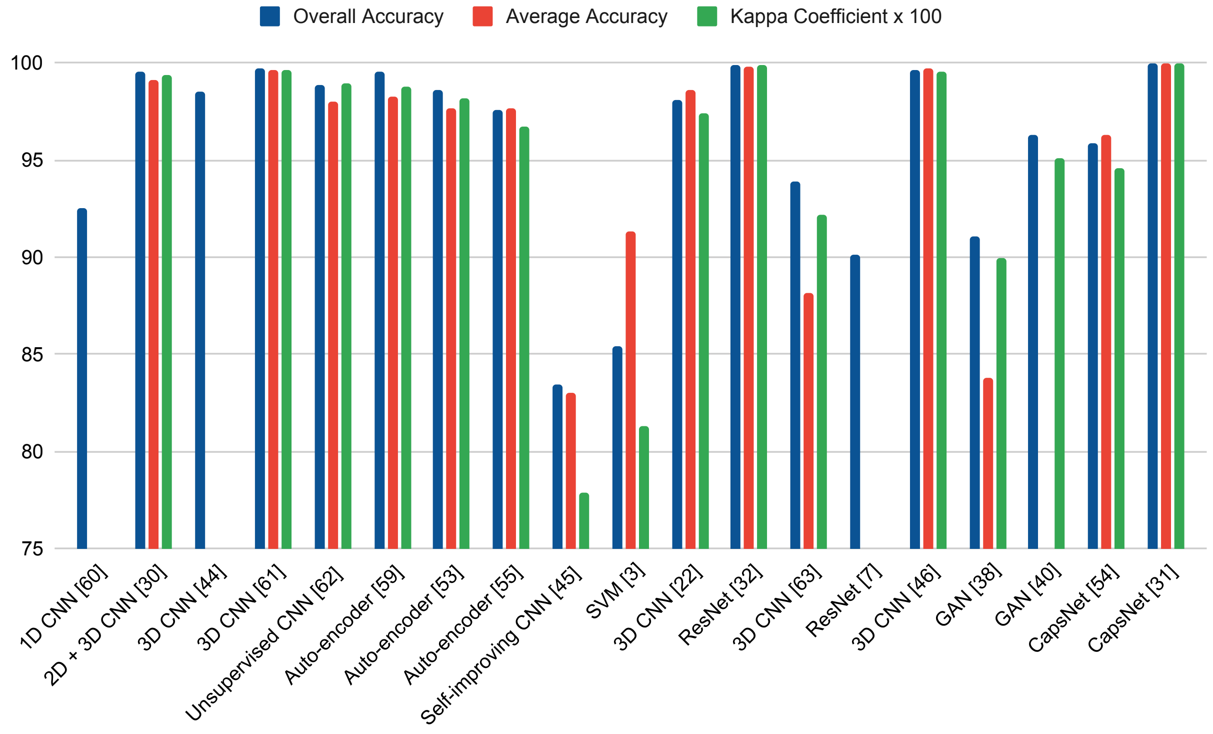

Figure 10: Accuracies of the University of Pavia dataset

4.2 A personal take

We are able to recognise 3 major items that have a big impact on the future of HSI classification these are:

- **New datasets.** Methods from past years are in general able to obtain an accuracy of 95% or more. This makes it difficult to distinguish different models. Furthermore, it illustrates that there is not much room to grow. New, more challenging datasets could differentiate promising methods better.

- **Train models on more general data than one specific set.** This way it can classify more diverse problems. Variability between datasets, spatial and spectral resolutions will pose a challenge.

- **Identify unknown classes.** Future work could focus on identifying unknown classes especially when they are scarce, to further improve HSI classification by rejecting unknown classes[7].

4.3 Trend

We can see that 3D CNN's are a true and tested classification method. A lot of research has been done on the subject and in recent years the accuracy is very high. A relative newcomer is the capsule network but it shows a lot of promise and will probably achieve a near perfect accuracy in the coming years. The addition of unknown classes will probably be explored further as well.

5 Conclusion

This overview paper first introduced hyperspectral imaging and classification of these images. With the understanding of HSI classification we established a structured overview of the different deep learning techniques available. These were structured by way of a tree representation, making the current landscape interpretable. Each technique was shortly covered and a timeline was able to further visualise the evolution. Additionally the most commonly used datasets and their characteristics were covered.

Subsequently two of the major challenges within hyperspectral imaging were covered. Namely, limited training samples and dimensionality reduction Both stemming from *Hughes Phenomenon*: the imbalance between the high dimensionality of the data and the limited number of training samples available [42]. The first challenge covered GANs and Capsule networks more in depth, highlighting the influence of augmentation and using a model that uses less training data. The second challenge covered auto-encoders in detail as an answer for dimensionality reduction.

Finally a discussion surrounding the current state of the art was made. An evaluation of the accuracy of different deep learning approaches was given. Next, our personal take was expressed. Lastly, a general trend within the field was recognised, confirming that 3D CNNs generally offer the highest accuracy. We could recognise that techniques such as Capsule Networks that mitigate the shortcomings of CNNs further improve accuracies, pushing the field in that direction.

References

[1] A. F. Goetz, "Three decades of hyperspectral remote sensing of the Earth: A personal view," *Remote Sensing of Environment*, volume 113, number SUPPL. 1, S5–S16, 2009, ISSN: 00344257. DOI: 10.1016/j.rse.2007.12.014.

[2] NASA, *NASA Landsat Science — Landsat 1.* [Online]. Available: https://landsat.gsfc.nasa.gov/landsat-1-3/landsat-1 (visited on 12/20/2020).

[3] Y. Tarabalka, J. Chanussot, and J. A. Benediktsson, "Segmentation and classification of hyperspectral images using watershed transformation," *Pattern Recognition*, volume 43, number 7, pages 2367–2379, 2010, ISSN: 00313203. DOI: 10.1016/j.patcog.2010.01.016. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2010.01.016.

[4] California Institute of Technology, *NASA Jet Propulsion Lab.* [Online]. Available: https://www.jpl.nasa.gov/imagepolicy/ (visited on 12/20/2020).

[5] N. Audebert, B. Le Saux, and S. Lefevre, "Deep Learning for Classification of Hyperspectral Data: A Comparative Review," *IEEE Geoscience and Remote Sensing Magazine*, volume 7, number 2, pages 159–173, Jun. 2019, ISSN: 2168-6831. DOI: 10.1109/MGRS.2019.2912563. arXiv: 1904.10674.

[6] A. Mohammadpour, O. Feron, and A. Mohammad-Djafari, "Bayesian segmentation of hyperspectral images," English, in *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING*, volume 735, AMER INST PHYSICS, 2004, 541–548, ISBN: 0-7354-0217-5. DOI: 10.1063/v757.frontmatter.

[7] S. Liu, Q. Shi, and L. Zhang, "Few-Shot Hyperspectral Image Classification With Unknown Classes Using Multitask Deep Learning," *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–18, 2020, ISSN: 0196-2892. DOI: 10.1109/tgrs.2020.3018879.

[8] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, volume 43, pages 1351–1362, Jul. 2005. DOI: 10.1109/TGRS.2005.846154.

[9] J. Haut and M. Paoletti, "Cloud implementation of multinomial logistic regression for uav hyperspectral images," *IEEE Journal on Miniaturization for Air and Space Systems*, volume PP, pages 1–1, Aug. 2020. DOI: 10.1109/JMASS.2020.3019669.

[10] J. Ham, Y. Chen, M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, volume 43, pages 492–501, Apr. 2005. DOI: 10.1109/TGRS.2004.842481.

[11] Y. Bazi and F. Melgani, "Gaussian process approach to remote sensing image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, volume 48, pages 186–197, Feb. 2010. DOI: 10.1109/TGRS.2009.2023983.

[12] L. Chapel, T. Burger, N. Courty, and S. Lefèvre, "Perturbo manifold learning algorithm for weakly labeled hyperspectral image classification," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, volume 7, pages 1070–1078, Apr. 2014. DOI: 10.1109/JSTARS.2014.2304304.

[13] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced Spectral Classifiers for Hyperspectral Images: A review," *IEEE Geoscience and Remote Sensing Magazine*, volume 5, number 1, pages 8–32, Mar. 2017, ISSN: 2168-6831. DOI: 10.1109/MGRS.2016.2616418. [Online]. Available: https://ieeexplore.ieee.org/document/7882742/.

[14] L. Parra, C. Spence, P. Sajda, A. Ziehe, and K. Müller, "Unmixing hyperspectral data," English, in *Advances in Neural Information Processing Systems 12 - Proceedings of the 1999 Conference, NIPS 1999*, series Advances in Neural Information Processing Systems, 13th Annual Neural Information Processing Systems Conference, NIPS 1999 ; Conference date: 29-11-1999 Through 04-12-1999, Neural information processing systems foundation, 2000, pages 942–948, ISBN: 0262194503.

[15] J. Haut, M. Paoletti, J. Plaza, and A. Plaza, "Cloud implementation of the k-means algorithm for hyperspectral image analysis," *The Journal of Supercomputing*, volume 73, Jan. 2017. DOI: 10.1007/s11227-016-1896-3.

[16] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, volume 52, number 5, pages 2565–2574, 2014, ISSN: 01962892. DOI: 10.1109/TGRS.2013.2263282.

[17] P. Goel, S. Prasher, R. Patel, J. Landry, R. Bonnell, and A. Viau, "Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn," *Computers and Electronics in Agriculture*, volume 39, number 2, pages 67–93, 2003, ISSN: 0168-1699. DOI: https://doi.org/10.1016/S0168-1699(03)00020-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169903000206.

[18] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised Neural Networks for Efficient Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 48, number 5, pages 2271–2282, May 2010, ISSN: 0196-2892. DOI: 10.1109/TGRS.2009.2037898.

[19] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 55, number 7, pages 3639–3655, 2017. DOI: 10.1109/TGRS.2016.2636241.

[20] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Transactions on Image Processing*, volume 27, number 3, pages 1259–1270, 2018. DOI: 10.1109/TIP.2017.2772836.

[21] M. He, X. Li, Y. Zhang, J. Zhang, and W. Wang, "Hyperspectral image classification based on deep stacking network," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pages 3286–3289. DOI: 10.1109/IGARSS.2016.7729850.

[22] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, volume 145, Dec. 2017. DOI: 10.1016/j.isprsjprs.2017.11.021.

[23] P. Fisher, "The pixel: A snare and a delusion," *International Journal of Remote Sensing*, volume 18, Feb. 1997. DOI: 10.1080/014311697219015.

[24] V. Slavkovikj, S. Verstockt, W. De Neve, S. Hoecke, and R. Van de Walle, "Unsupervised spectral sub-feature learning for hyperspectral image classification," *International Journal of Remote Sensing*, volume 37, pages 309–326, Jan. 2016. DOI: 10.1080/01431161.2015.1125554.

[25] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pages 4959–4962. DOI: 10.1109/IGARSS.2015.7326945.

[26] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, volume 54, number 8, pages 4544–4554, 2016. DOI: 10.1109/TGRS.2016.2543748.

[27] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, volume 6, number 6, pages 468–477, 2015. DOI: 10.1080/2150704X.2015.1047045.

[28] A. B. Hamida, A. Benoît, P. Lambert, and C. Benamar, "Deep learning approach for remote sensing image analysis," 2016. DOI: 10.1109/TGRS.2018.2818945.

[29] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sensing*, volume 9, number 1, 2017, ISSN: 20724292. DOI: 10.3390/rs9010067.

[30] Z. Ge, G. Cao, X. Li, and P. Fu, "Hyperspectral Image Classification Method Based on 2D–3D CNN and Multibranch Feature Fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, volume 13, pages 5776–5788, 2020, ISSN: 1939-1404. DOI: 10.1109/jstars.2020.3024841.

[31] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. Plaza, J. Li, and F. Pla, "Capsule Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 57, number 4, pages 2145–2160, 2019, ISSN: 01962892. DOI: 10.1109/TGRS.2018.2871782.

[32] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 57, number 2, pages 740–754, Feb. 2019, ISSN: 01962892. DOI: 10.1109/TGRS.2018.2860125.

[33] H. Lee and H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification," *IEEE Transactions on Image Processing*, volume 26, pages 1–1, Jul. 2017. DOI: 10.1109/TIP.2017.2725580.

[34] X. Ma, H. Wang, and J. Geng, "Spectral–spatial classification of hyperspectral image based on deep auto-encoder," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, volume 9, pages 1–13, Feb. 2016. DOI: 10.1109/JSTARS.2016.2517204.

[35] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning*, series ICML '08, Helsinki, Finland: Association for Computing Machinery, 2008, pages 536–543, ISBN: 9781605582054. DOI: 10.1145/1390156.1390224.

[36] T. Li, J. Zhang, and Y. Zhang, "Classification of hyperspectral image based on deep belief networks," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pages 5132–5136. DOI: 10.1109/ICIP.2014.7026039.

[37] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 54, number 3, pages 1349–1362, 2016. DOI: 10.1109/TGRS.2015.2478379.

[38] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative Adversarial Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 56, number 9, pages 5046–5063, 2018, ISSN: 01962892. DOI: 10.1109/TGRS.2018.2805286.

[39] Y. Lv, K. Wang, Y. Liu, and Q. Zhao, "Improved triple generative adversarial nets," *International Journal of Computer Applications in Technology*, volume 59, number 2, page 114, 2019, ISSN: 0952-8091. DOI: 10.1504/ijcat.2019.10019437.

[40] Y. Wang, L. Y. B, and S. Wang, *Multi-scale Densely 3D CNN for Hyperspectral Image Classification.* 2019, pages 150–160, ISBN: 9783030317232. DOI: 10.1007/978-3-030-31723-2.

[41] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlögl, and G. E. Birch, "Comparison of evaluation metrics in classification applications with imbalanced datasets," *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, pages 777–782, 2008. DOI: 10.1109/ICMLA.2008.34.

[42] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, “A new deep convolutional neural network for fast hyperspectral image classification,” *ISPRS Journal of Photogrammetry and Remote Sensing*, volume 145, pages 120–147, 2018, ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2017.11.021.

[43] M. Alonso, J. Malpica, and A. Martinez-Agirre, “Consequences of the hughes phenomenon on some classification techniques,” May 2011.

[44] H. C. Mingyi He, Bo Li, “Multi-Scale 3D Deep Convolutional Neural Network For Hyperspectral Image Classification,” pages 3904–3908, 2017. DOI: 10.1109/ICIP.2017.8297014.

[45] P. Ghamisi, Y. Chen, and X. X. Zhu, “A Self-Improving Convolution Neural Network for the Classification of Hyperspectral Data,” *IEEE Geoscience and Remote Sensing Letters*, volume 13, number 10, pages 1537–1541, 2016, ISSN: 1545598X. DOI: 10.1109/LGRS.2016.2595108.

[46] “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” *IEEE Transactions on Geoscience and Remote Sensing*, volume 54, number 10, pages 6232–6251, 2016, ISSN: 01962892. DOI: 10.1109/TGRS.2016.2584107.

[47] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, volume 3, 2014. DOI: 10.3156/jsoft.29.5_177_2.

[48] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems*, volume 2017-Decem, 2017. [Online]. Available: http://arxiv.org/abs/1710.09829.

[49] A. Odena, “Semi-Supervised Learning with Generative Adversarial Networks,” pages 1–3, 2016. arXiv: 1606.01583. [Online]. Available: http://arxiv.org/abs/1606.01583.

[50] H. Xu, H. Zhang, W. He, and L. Zhang, “Superpixel-based spatial-spectral dimension reduction for hyperspectral imagery classification,” *Neurocomputing*, volume 360, pages 138–150, 2019, ISSN: 18728286. DOI: 10.1016/j.neucom.2019.06.023.

[51] M. Pechyonkin, *Understanding hinton’s capsule networks. part 1. intuition.* Nov. 2017. [Online]. Available: https://pechyonkin.me/capsules-1/.

[52] Y. Luo, J. Zou, C. Yao, X. Zhao, T. Li, and G. Bai, “HSI-CNN: A Novel Convolution Neural Network for Hyperspectral Image,” in *ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing*, 2018. DOI: 10.1109/ICALIP.2018.8455251.

[53] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, “Active Transfer Learning Network: A Unified Deep Joint Spectral-Spatial Feature Learning Model for Hyperspectral Image Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, volume 57, number 3, pages 1741–1754, 2019, ISSN: 01962892. DOI: 10.1109/TGRS.2018.2868851. arXiv: 1904.02454.

[54] F. Deng, S. Pu, X. Chen, Y. Shi, T. Yuan, and S. Pu, “Hyperspectral Image Classification with Capsule Network Using Limited Training Samples,” *Sensors*, volume 18, number 9, page 3153, Sep. 2018, ISSN: 1424-8220. DOI: 10.3390/s18093153.

[55] P. Zhou, J. Han, G. Cheng, and B. Zhang, “Learning Compact and Discriminative Stacked Autoencoder for Hyperspectral Image Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, volume 57, number 7, pages 4823–4833, 2019, ISSN: 15580644. DOI: 10.1109/TGRS.2019.2893180.

[56] A. Marinoni and P. Gamba, “Unsupervised Data Driven Feature Extraction by Means of Mutual Information Maximization,” *IEEE Transactions on Computational Imaging*, volume 3, number 2, pages 243–253, 2017, ISSN: 2573-0436. DOI: 10.1109/tci.2017.2669731.

[57] A. Marinoni, G. C. Iannelli, and P. Gamba, “An Information Theory-Based Scheme for Efficient Classification of Remote Sensing Data,” *IEEE Transactions on Geoscience and Remote Sensing*, volume 55, number 10, pages 5864–5876, 2017, ISSN: 01962892. DOI: 10.1109/TGRS.2017.2716187.

[58] C. Tao, H. Pan, Y. Li, and Z. Zou, “Unsupervised Spectral-Spatial Feature Learning With Stacked Sparse Autoencoder for Hyperspectral Imagery Classification,” *IEEE Geoscience and Remote Sensing Letters*, volume 12, number 12, pages 2438–2442, 2015, ISSN: 1545598X. DOI: 10.1109/LGRS.2015.2482520.

[59] X. Zhang, Y. Liang, C. Li, N. Huyan, L. Jiao, and H. Zhou, “Recursive Autoencoders-Based Unsupervised Feature Learning for Hyperspectral Image Classification,” *IEEE Geoscience and Remote Sensing Letters*, volume 14, number 11, pages 1928–1932, 2017, ISSN: 1545598X. DOI: 10.1109/LGRS.2017.2737823.

[60] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep Convolutional Neural Networks for Hyperspectral Image Classification," *Journal of Sensors*, volume 2015, pages 1–12, 2015, ISSN: 1687-725X. DOI: 10.1155/2015/258619.

[61] Q. Xu, Y. Xiao, D. Wang, and B. Luo, "CSA-MSO3DCNN: Multiscale octave 3D CNN with channel and spatial attention for hyperspectral image classification," *Remote Sensing*, volume 12, number 1, 2020, ISSN: 20724292. DOI: 10.3390/RS12010188.

[62] R. Vaddi and P. Manoharan, "CNN based hyperspectral image classification using unsupervised band selection and structure-preserving spatial features," *Infrared Physics and Technology*, volume 110, number May, page 103 457, 2020, ISSN: 13504495. DOI: 10.1016/j.infrared.2020.103457.

[63] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, volume 57, number 9, 2019, ISSN: 15580644. DOI: 10.1109/TGRS.2019.2910603.