Neural Computing and Applications (2023) 35:7595–7609
https://doi.org/10.1007/s00521-022-08056-w

ORIGINAL ARTICLE

Multi-head attention with CNN and wavelet for classification of hyperspectral image

Harshula Tulapurkar[1] · Biplab Banerjee[1] · Krishna Mohan Buddhiraju[1]

Received: 20 October 2021 / Accepted: 11 November 2022 / Published online: 9 December 2022

Abstract

Hyperspectral Image (HSI) is characterized by large number of bands with a high spectral resolution where continuous spectrum is measured for each pixel. This high volume therefore leads to challenges in processing the dataset. Objective of Dimensionality Reduction (DR) algorithms is to identify and eliminate statistical redundancies of hyperspectral data while keeping as much spectral information as possible. Combining spectral and spatial information offers a more comprehensive classification approach. Convolutional neural network (CNN) has the potential to extract complex spatial and spectral features embedded in Hyperspectral data. Wavelet transform belongs to the family of multi-scale transformation where the input signal is analyzed at different levels of granularity. Attention mechanism is a method in neural networks to guide the algorithm to focus on the important information in the data. In this paper, we use Multi-head Transformer-based Attention (Vaswani et al. in Attention is all you Need, http://arxiv.org/abs/1706.03762 2017) technique for Channel attention which captures the long-range spectral dependencies. The experimental results show that the proposed algorithm MT-CW Band Selection-based multi-head transformer for dimensionality reduction and Wavelet CNN-based algorithm for feature extraction yields impressive results in terms of information conservation and class separability.

Keywords Transformer · Band attention · Convolutional neural network (CNN) · Hyperspectral (HSI) image classification · Wavelet · Dimensionality reduction · Multi-head channel attention

1 1 Introduction

Color images traditionally represented using three bands are an approximate representation of colors that are well received by the human eye. This is because the human eye is trichromatic, i.e., it has three color receptors. In reality, color can be represented as a spectral curve over a wide range across the electromagnetic spectrum. Multispectral and hyperspectral imaging capture image data within specific wavelength ranges across visible portion of the electromagnetic spectrum. The reflectance values for landscape features such as water, sand, roads, and forests across different wavelengths when plotted results in what is called "spectral response curves" or "spectral signatures." Each object has a unique spectral signature and can be used as a discriminating feature for analyzing the biophysical and biochemical properties of objects and classifying remotely sensed images into classes of landscape features. The multispectral and hyperspectral data are differentiated by the number of spectral bands and how narrow the bands are. For target classification, the more the spectral bands and the total wavelength range they cover, the better the differentiation between classes. Multispectral images are characterized by 3–10 bands of broad bandwidth, whereas hyperspectral images typically have hundreds of narrow contiguous bands. Images captured in narrow spectral bands over a continuous spectral range, producing the spectra of all pixels in the scenes are stacked together to produce a Hyperspectral (HSI) Cube. Hyperspectral imagery has wide applications from urban planning, defense and security, planetary exploration, and astrophysics to

✉ Harshula Tulapurkar
harshula@iitb.ac.in

Biplab Banerjee
bbanerjee@iitb.ac.in

Krishna Mohan Buddhiraju
bkmohan@iitb.ac.in

[1] Indian Institute of Technology Bombay, Centre of Studies in Resources Engineering, Mumbai 400076, India

monitoring and management of the environment, physical analysis of materials, biomedical imaging, food safety, detection of counterfeit objects (especially in pharmacology), and precision agriculture.

While a large number of narrow spectral bands helps in discriminating between land use categories accurately, it results in challenges like the complexity of computation, requirements of large memory, the need for a large number of training samples due to high dimensionality space, etc. A well-studied problem that arises when working with high-dimensional feature sets is the “Curse of dimensionality” because the volume of the feature space increases vastly and the available data becomes sparse. As a result, the number of samples required for training Machine Learning models increases and it becomes a problem when the size of the labeled training sample set is small. The goal of dimensionality reduction is to identify a subset of the spectral bands by eliminating statistical redundancies within the hyperspectral data while keeping as much spectral information as possible. Relatively few bands can be used to represent most of the information thereby overcoming the challenges of storage, transmission, classification, spectral unmixing, target detection, and visualization of Hyperspectral Imagery.

Conventional dimensionality reduction techniques like PCA [2] create new uncorrelated variables that successively maximize variance and thereby representing the HSI by smaller number of variables termed Principal Components. Independent component analysis (ICA) [3] separates the multivariate signal into additive subcomponents with the underlying assumption that the subcomponents are non-Gaussian signals and that they are statistically independent of each other. RNN’s/LSTM’s [4] are algorithms that are designed to handle sequential data and hence have been used for dimensionality reduction of HSI which have an inherent sequential relationship in the spectral dimension. However, RNN’s drawback of vanishing gradient problem limits its usage in high dimensional datasets. Squeeze operation in the Squeeze and Excitation [5] method aggregates the features across spatial domain and excitation operation in the form of gating that modulates channel-wise weights. Wavelet transforms [6–8] decompose the input signal into features at different scales with the help of a nonlinear wavelet that extracts features of hyperspectral data. CNN-based extraction techniques [9, 10] share the weights across layers and hence local spatial coherence in the images is considered. Any transformation applied to the input results in an equal shift or change in the output feature map because CNN feature maps take into consideration any affine transformations that are applied to the input image. Band Selection that uses the attention mechanism [11–23] generates an attention mask that has high values for the most informative bands for classification of the HSI image thereby giving less weightage to trivial bands. We can refer to [24, 25] for a comprehensive review of the various dimensionality reduction algorithms.

The sparse representation-based band selection (SpaBS) [27] method determines the histogram of the sparse coefficient matrix and selects the top K bands that appear more frequently than the others. It assumes that the larger coefficient means that more information resides in that band. In this way, the most important and representative bands are selected. However, it should be noted that the SpaBS method suffers from the very large computational complexity of the K-SVD process, due to a large number of pixels. The SpaBS method does not fully consider the interaction of different bands and hence it cannot capture the global structure information in the learned sparse coefficient matrix which is a limitation. Maximum-variance PCA (MVPCA) [28] uses band-prioritization along with a band-decorrelation approach for selecting discriminative bands. All bands are prioritized based on the information contained in them. Divergence is used to determine the correlation between the bands. A joint approach of band-prioritization and band-decorrelation is used for band selection. The computational time required in MVPCA is large due to the computation in the principal component analysis (PCA) transformation of the HSI dataset. Sparse Non-negative Matrix Factorization (SNMF) [29] is based on the Sparse representation (SR) theory which states that each band can be linearly represented by only a few other bands. The sparse coefficient matrix gives the relationships between different bands and also the geometrical structure information of the original dataset. Non-negativity of the coefficient ensures that factors do not cancel out each other. However, the drawback of the sparse matrix is that performing operations across this matrix may take a long time and the majority of computations performed will involve adding or multiplying zero values together. Also, though the majority of elements of the sparse matrix are zero, memory is required to be allocated and results in waste of resources. Recommender systems [33] are widely used in social media where users maintain profiles in multiple systems that reflect their preferences. Leveraging all the user preferences available in several systems or domains is used for generating more encompassing user models and better recommendations. In the context of Hyperspectral image processing these recommendation systems can be employed for the Band selection algorithm where Cross-Domain Collaborative Filtering (CDCF) algorithms can be used to address the sparsity problem where the spectral and spatial features are the different domains. Two-Sided CDCF model based on Selective Ensemble learning considering both Accuracy and Efficiency (TSSEAE) [33] addresses two important challenges, viz., how to select significant subsets from all the auxiliary domains to improve the recommendation accuracy and how to balance recommendation accuracy and efficiency because using all the auxiliary domains also lead to low efficiency. The problem of selecting a subset that bears two simultaneous objectives, optimal subset of bands and MAE value belongs to a bi-objective optimization problem. It uses the Non-negative Matrix Factorization (NMF), Pareto Ensemble Pruning and variable-depth search and Ensemble learning is then deployed to choose the best model. Ensemble methods have higher predictive accuracy, compared to the individual models, however, there is an overhead of computational cost.

Our proposed algorithm Multi-head Transformer-based Attention with CNN-Wavelet (MT-CW) for feature extraction consists of Multi-head Transformer for Channel Attention. Channel mask for selecting the informative bands and suppressing bands with trivial information is generated by analyzing weights for bands in all the training epochs. Multi-headed attention helps in richer representation of the input imagery. CNN and Wavelet transforms are used to generate spatial and spectral features. Spectral and spatial features provide a rich feature set of the input image. The results and analysis demonstrate that this network can achieve state-of-the-art classification performance on datasets that have spectrally similar classes.

2 MT-CE Multi-head transformer-based attention with CNN-Wavelet

While Hyperspectral imagery provides detailed spectral responses based on hundreds of spectral bands there is also increase in the cost of storage requirements, computational time to process, besides a need for a greater number of samples and highly redundant data. In order to remove redundant data, it is extremely important to select effective and representative spectral bands. Attention mechanism is a method in neural networks to guide the algorithm to focus on the important information in the data and thereby select the discriminative bands from the HSI. The attention mechanism adaptively selects bands from the HSI that are non-trivial, thereby addressing the issue of band redundancy. While doing so it considers not just the local but global information as well. The attention model is able to selectively focus on important bands of the input and learns the association between them. The proposed model uses Multi-headed Attention where it generates a channel mask or an attention vector that has high values for the bands that are important and low values for the trivial bands. The bands are ranked based on the mask value for the bands. The multi-headed attention together with the Band Ranking module forms the Band Selection, the output of which is the top ‘N’ non-trivial bands. ‘N’ is chosen empirically and is dependent on spectral similarity of classes in the imagery. More the spectral similarity in the classes, higher is the value of ‘N’. This multi-headed attention-based feature selection step results in minimizing the dimensionality of the input by excluding irrelevant features.

Higher level spectral features are then extracted using Wavelet transform and the CNN module extracts the spatial features. These features are then combined resulting in a comprehensive feature set.

The proposed algorithm consists of six key modules.

- Multi-headed Transformer Attention module
- Band Ranking Module
- Patch Generation
- CNN Feature Extractor (CFE)
- Wavelet decomposer
- Classifier

Fig. 1 below is the block diagram for the proposed MT-CE algorithm.

2.1 Multi-headed transformer attention module

Attention in neural networks helps to focus on important parts of the inputs. In the context of the hyperspectral image, it is used to determine the important bands. There are different attention models namely Self-attention, Global and Local attention, Hard and Soft Attention. Global attention considers the current state of the neurons as well as prior states to compute the attention mask. All the hidden states of the network are concatenated into a matrix and multiplied with a weight matrix to get the final layer of the feedforward connection. As the input size increases, the matrix size also increases. Therefore, an increase in the number of nodes in the feedforward connection can increase computational complexity and cost. Like Global attention, Soft attention uses the weighted sum of the hidden states and is used to calculate the attention mask. In Hard attention instead of a weighted average of all hidden states, attention scores are used to select a single hidden state. A function like argmax can be used to make the selection, but since it is not differentiable and backpropagation training cannot change this selected hidden state, complex techniques are required for training. Local attention combines soft and hard attention concepts where instead of considering all the encoded inputs, only a part is considered for generating the attention mask. This results in computationally economical as well as ease of training the models. Transformer-based methods use the Self-attention technique where inputs interact with each other and therefore calculate the attention of all other inputs with respect to one input. Since the input sequence is passed in parallel, the GPU utilization is effective and the speed of training also increases.

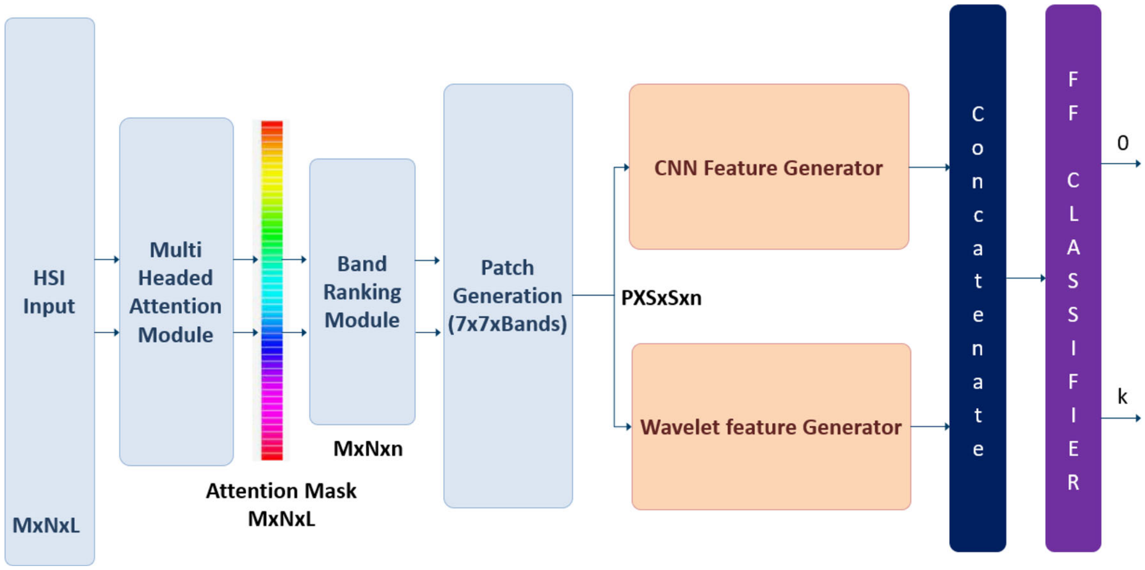

Fig. 1 Below is the block diagram for the proposed MT-CE algorithm

HSI intrinsically has a sequence-based data structure based on wavelength ranges of spectral features and conventionally recurrent neural network (RNN) and long short-term memory (LSTM) are the neural networks used to process sequential data. However, RNN and LSTM suffer from problems of vanishing gradient and slow training, respectively. The Transformer Neural Network [1] aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. The transformer model reduces the computation complexity involved in identifying the dependencies in the bands irrespective of their sequence. It generates a channel mask that helps in identifying the subset of bands from HSI that contain the most useful information for the classification of the given image without impacting the classification accuracy. Figure 2 shows the details of the Multi-headed transformer module.

The input to the model is HSI data cube $Xo \in \mathscr{R}^{\mathrm{MxNxL}}$ where M is the width, N is the height of image and L is number of spectral bands. The class set $Y = \{y_1, y_2 \ldots y_k\}$ where k denotes the total classes in Xo. Deep networks need a large amount of training data for good performance. The number of labeled samples available in HSI is very few and hence data augmentation layer has been added. Data augmentation is applied as part of the model. It is one layer in the model. The various image augmentation techniques applied include RandomFlip (“horizontal”), RandomRotation(factor = 0.02), RandomZoom(height_factor = 0.2, width_factor = 0.2). The advantage of having the augmentation layer as part of the model is that the model becomes portable and it helps reduce the training/serving skew. This is followed by patch generation which generates a 14*14*L sized patches. A patch size of 14 has been chosen empirically. Positional embedding is a representation of the location or “position” of Hyperspectral bands in a sequence. It avoids recursion, by processing data as a whole and by learning relationships between data using positional embedding. The encoder of the transformer network proposed in Vaswani et al. [1] consists of a multi-head self-attention network followed by a decoder layer. For our implementation, the decoder layer is replaced by a fully connected feedforward network. A residual connection from the input of the attention layer is followed by the normalization layer. Residual connection helps address two important aspects namely information preservation and vanishing gradient. When the input traverses through various layers in the forward propagation, they might get modified considerably and this might result in loss of information by the time they reach the final layer. Residual connections are like information highways that bypass a lot of intermediate layers and feed information into deeper layers. This helps the deeper layer retain information in the input layers. Normalization is done to standardize neuron activation along the axis of the feature. This is followed by a feedforward network and normalization layer. The proposed model consists of two layers of Encoder with four heads. The outputs of the final Transformer block are normalized, reshaped, and fed to a feedforward classifier. AdamW is used as the Optimizer function while the Loss function used is SparseCategoricalCrossentropy.

2.2 Band ranking module

Weights at the normalization layer are stored for every epoch. The normalization layer has the same number of neurons as the input bands. Hence, these weights reflect

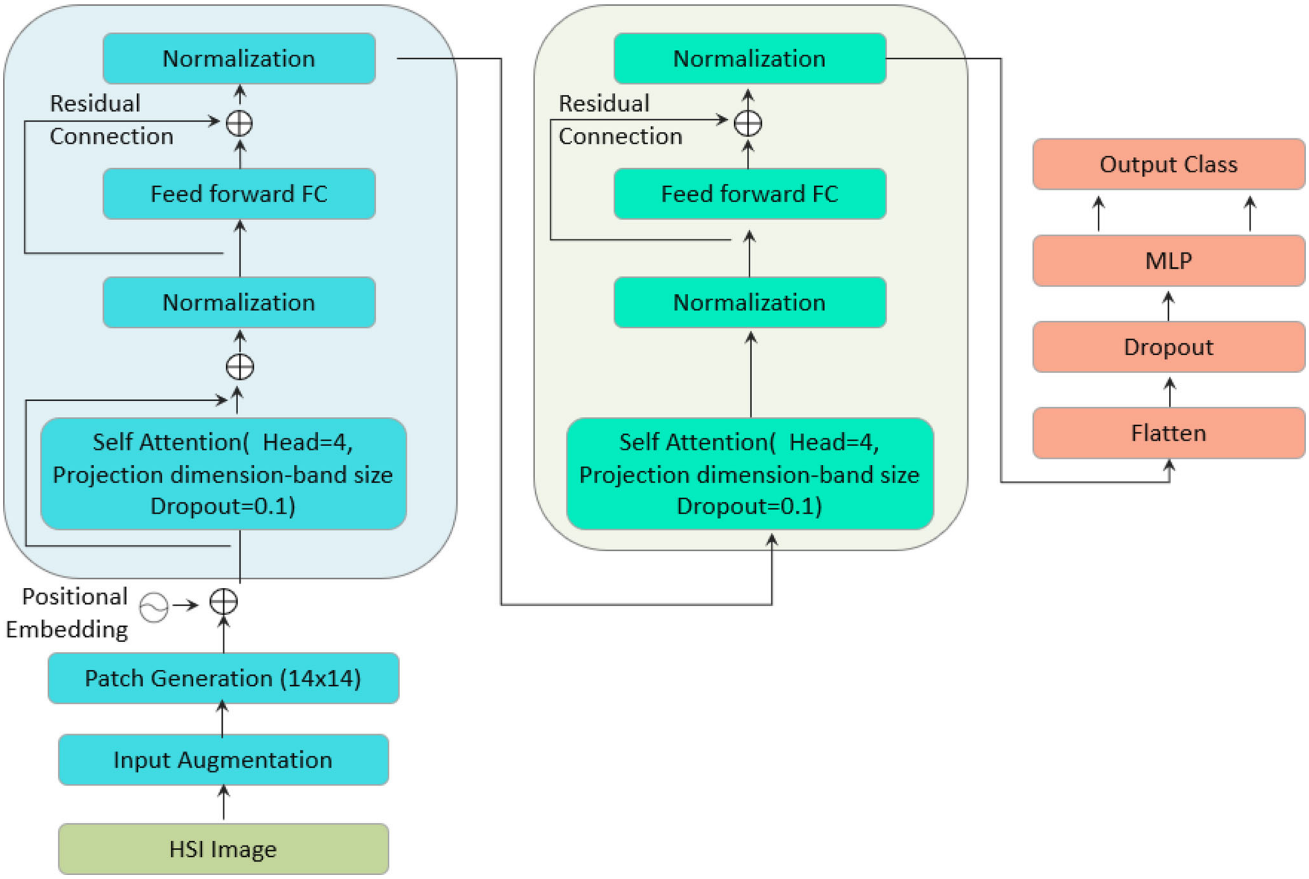

Fig. 2 Multi-headed transformer module

trivial and non-trivial bands. At the end of the training, all the epochs where classification accuracy is less than 95% are filtered out from the attention mask generation process. For every epoch, bands are ranked based on their weight. Band with the highest weight is ranked 1. Average accuracy ‘A’ for the band across all the subset of epochs is available from model training history. A threshold ‘K’ is defined to identify bands that are consistently in Top K bands. ‘K’ is empirically set to 40% of the total bands in the dataset. The number of times the bands appeared in Top K is determined. Let this be represented by ‘P’. The final score for the band is determined by the product of P and A. The resultant vector is the attention mask.

2.3 Patch generation

The reduced dimension ‘*n*’ is dependent on the dataset. If the classes have a high similarity index, the number of bands required ‘*n*’ to classify them is slightly higher. Number ‘n’ used was 25 for Indian Pines and 15 for the Pavia dataset so that it can be compared on the same parameters against the other models used in this study. The Attention mask has high values for informative bands and low values for trivial bands. Top ‘*n*’ bands from the HSI input are selected based on the attention mask. The dimensionality reduced input vector $X_{dr} \in \mathscr{R}^{MxNxn}$ where *M* is the width, *N* is the height of the image and *n* consists of only top n bands and is a sparse representation of the given dataset.

To ensure that there is no loss of contextual information, HSI data cube X_{dr} is first divided into overlapping cubic patches of size S x S x n where S < M and S < N and ‘n’ is the number of sparse bands of the given dataset. The spatial resolution of the Indian Pine dataset using the Aviris sensor is 20 m and so if the patches are very large then it interferes with the localization results and if the patch size is small it results in loss of contextual information so a patch size of 7 is chosen. Each cubic patch is centered at the spatial location (i, j), $Z_{i,j} \in \mathscr{R}^{SxSxn}$. The cubic patch is assigned the class of the center pixel. The transformed input is thus represented by $X_p \in \mathscr{R}^{PxSxSxn}$ where P represents the number of patches, ‘S’ the patch size, and ‘n’ the number of bands.

2.4 CNN feature extractor (CFE)

3D-CNN and 2D-CNN are used to extract the Spatial and Spectral features. Since the data set has spectrally similar classes additional features extracted from wavelet transform enhance the classification accuracy. The input to CFE is $X_p \in \mathscr{R}^{PxSxSxn}$ and the output is F_{cl} spatial-spectral feature map. CFE consists of two layers of 3D convolutions with 8 and 32 filters with kernel size (3,3,3) followed by two layers of 2D convolution with 64 and 128 filters and kernel size (3,3). The 2D convolution layer feeds into the Average pool layer with a kernel size of (3, 3). The output of CFE is the spatial-spectral feature map for the Dimensionality-reduced input image patches. Sigmoid is used as an activation function for the first 3D convolution while ReLu is used for subsequent layers.

2.5 Wavelet decomposition module (WM)

Wavelets provide spectral data features in a hierarchical framework. The image is decomposed using wavelet transform into various approximation and detail components, as a result it helps in separating finer details in a signal. By changing the scaling function and wavelet function finer details and coarse details of the imagery can be extracted. This is effective when the classes are spectrally similar and adding wavelet features enhances discriminative capabilities. Wavelet transform uses a mother wavelet function to analyze input data at different resolutions by decomposing it into a series of shifted and scaled versions of the mother wavelet. The scaling and shifting function of the wavelet transform results in a rich set of features that represents the local spectral variation of the dataset. This adds to discriminative features that can improve classification accuracy. Coiflet was used as the wavelet for both the data sets. The second level of decomposition was employed to ensure that the number of features does not overshadow the features provided by the CFE module. These features are fed to a 2D Average Pooling Layer and reshaped to give the wavelet feature set Fw.

Table 1 Summary of Indian Pines and Pavia University datasets

Data set	Indian Pines	Pavia University
Pixels	145*145	610*340
Channels	200	103
Classes	16	9
Labeled Pixels	10,249	42,776
Sensor	AVIRIS	ROSIS
Spatial resolution	20 m	1.3 m

2.6 Classifier

The features from the CFE module (Fcl) and WM (Fw) are concatenated to provide a rich feature set. These comprehensive features are fed to a fully connected layer followed by a softmax layer for classification. Softmax is used to calculate the class probability of the center pixel, and the Class with the highest probability is selected as the final Class result of the given input block.

3 Results and discussion

3.1 3.1 Dataset

Two variations of the model are proposed first one is MT-CNN which consists of the Multi-head attention module for band selection followed by CNN for features selection. The second model MT-CW consists of CNN and Wavelet for feature selection. Both the models were tested on two HSI data sets: Indian Pines and Pavia University. The summary of the 2 data sets is shown in Table 1.

The Indian Pines test site in Northwestern Indiana consists of 145 × 145 pixels and 224 spectral reflectance bands in the wavelength range of 0.4–2.5 microns. The ground truth consists of 16 classes. The number of bands is 200 where bands covering the region of water absorption: [104–108], [150–163], and 220 have been removed. Pavia University data set was acquired by the ROSIS sensor during a flight campaign over Pavia, Northern Italy. This scene is a 103 spectral band 610 × 340 pixels image. The

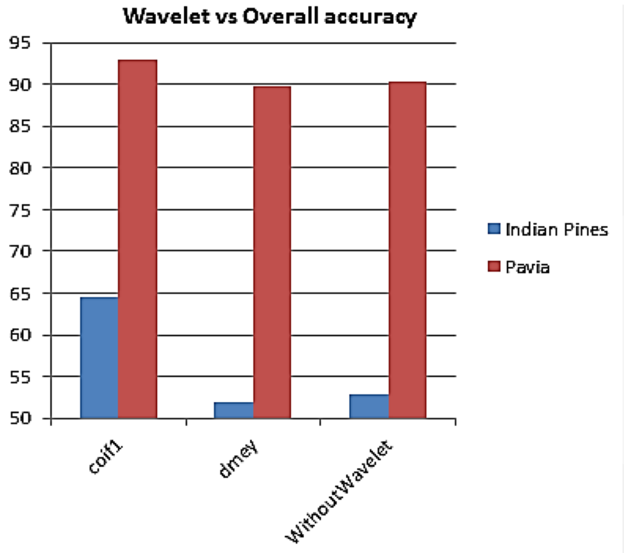

Fig. 3 Overall accuracy Performance of different MT-CW for Pavia Data set and Indian Pines Dataset

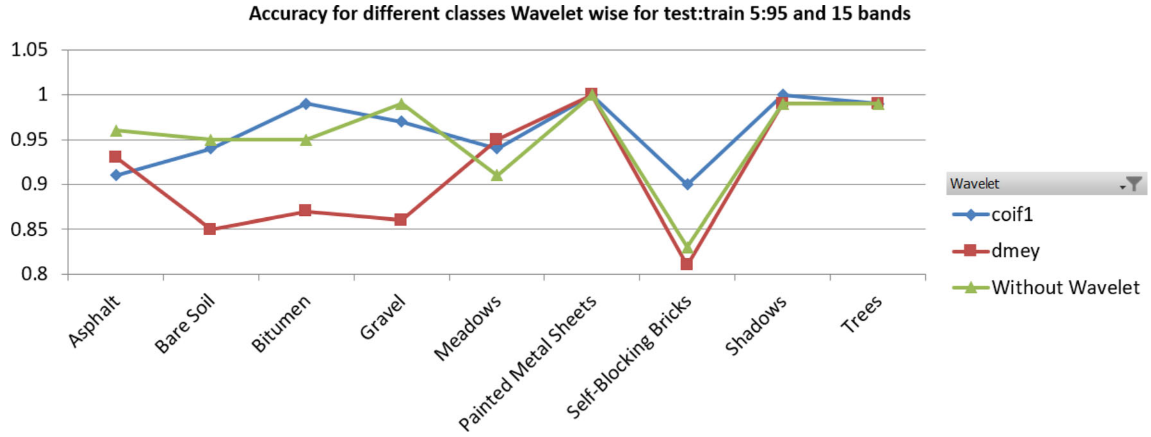

Fig. 4 Performance of MT-CW for Pavia Data set for different wavelet with 5% training data, with 15 bands

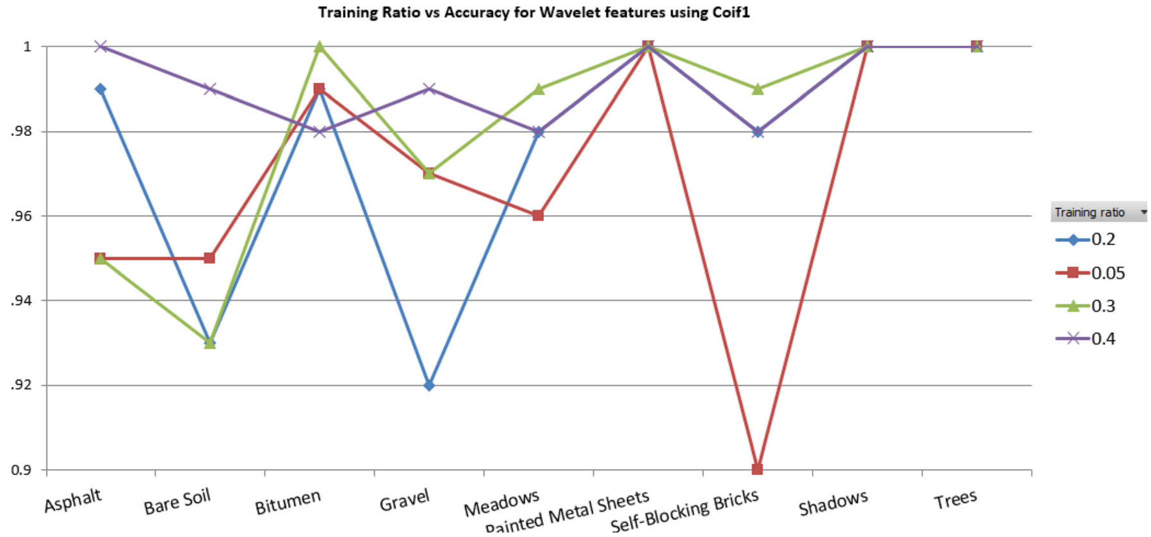

Fig. 5 Performance of MT-CW for Pavia Data set for different training ratios, 15 bands and 'Coiflet1' wavelet

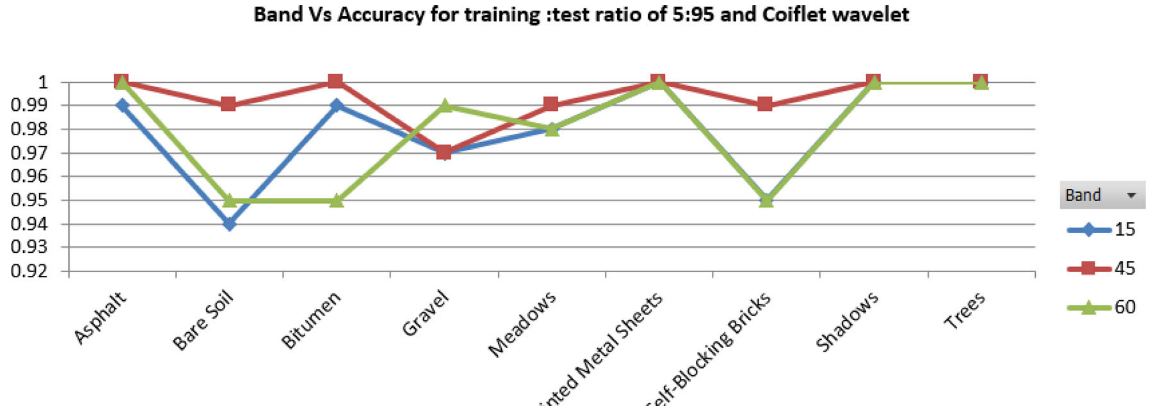

Fig. 6 Performance of MT-CW for Pavia Data set with different band size, training: test ratio of 5:95 and 'Coiflet1' wavelet

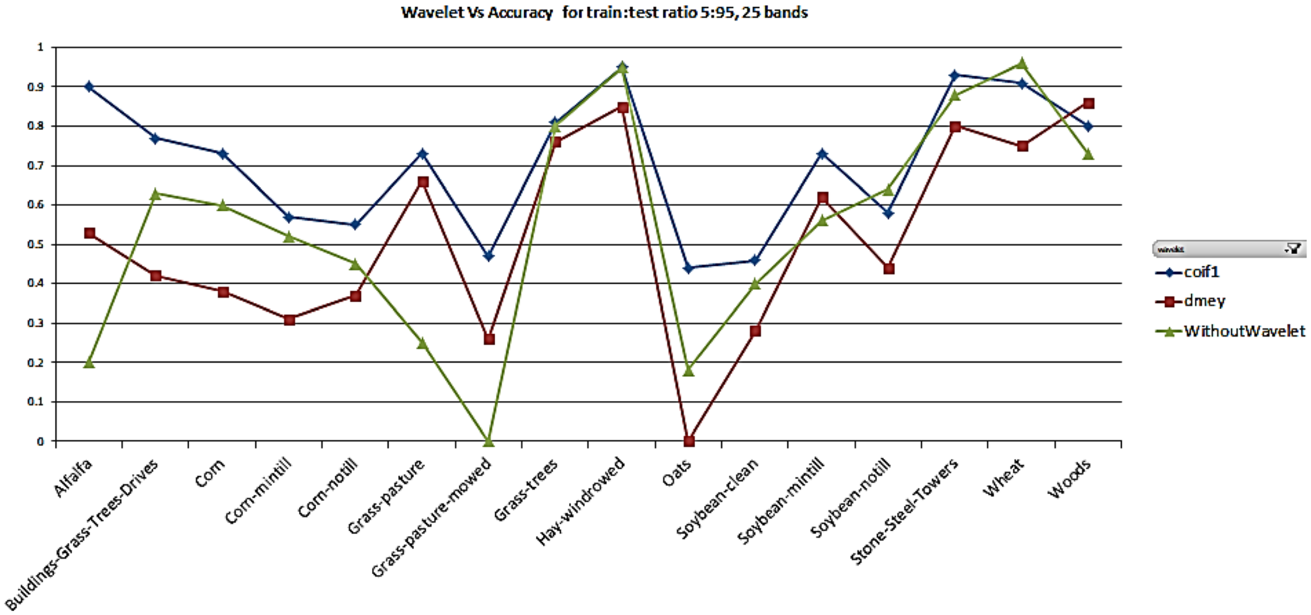

Fig. 7 Performance of MT-CW for Indian Pines Data set for different wavelet with 5% training data, with 25 bands

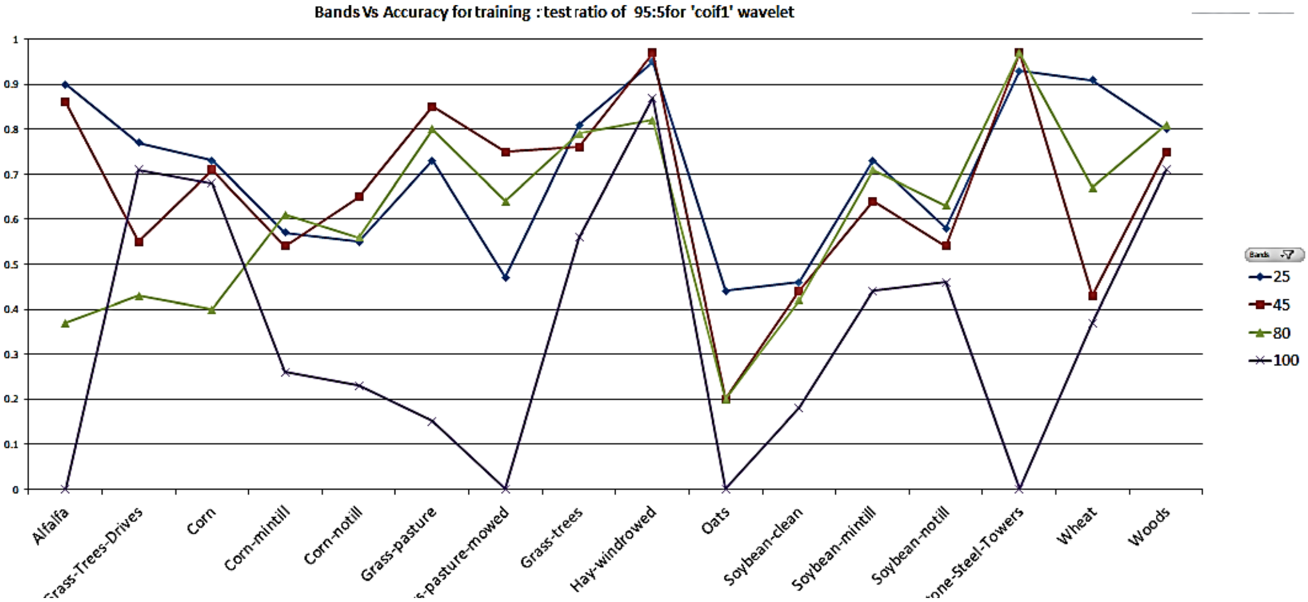

Fig. 8 Performance of MT-CW for Indian Pines Data set for different wavelet with 5% training data, with ‘Coiflet1’

geometric resolution is 1.3 m. There are nine classes in this dataset.

3.2 Evaluation criteria

For each data set, we first evaluate the model performance by analyzing the classification accuracy for different dimensionality reduction band sizes ranging from 15 to 60 for the Pavia dataset and 25–100 for the Indian pines dataset. To ensure a reliable classification result, we used the fivefold cross-validation. We also compare the classification performance [14] for MT-CW with many existing BS methods, i.e., ISSC [26], SpaBS [27], MVPCA [28], SNMF [29], MOBS [30], OPBS [31], BS-NET-FC [32] and BS-NET-CONV [32]. Three indices used to evaluate performance are.

1. Overall Accuracy (OA): This index is the ratio of the correctly classified hyperspectral pixels to the number of test samples.

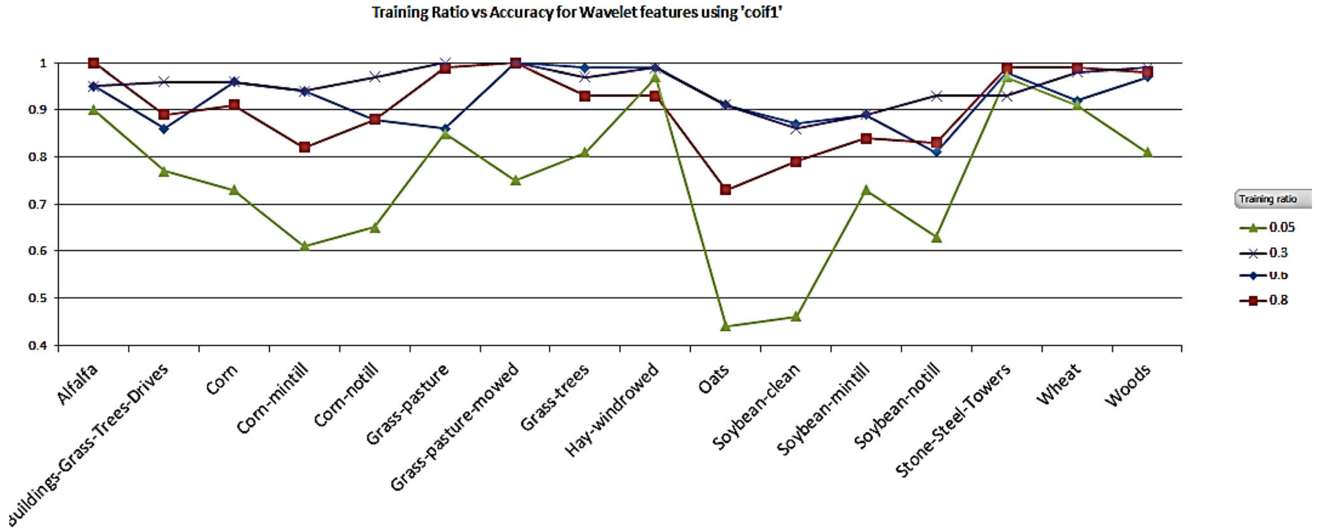

Fig. 9 Performance of MT-CW for Indian Pines Data set for different training ratio, 25 bands and 'Coiflet1' wavelet

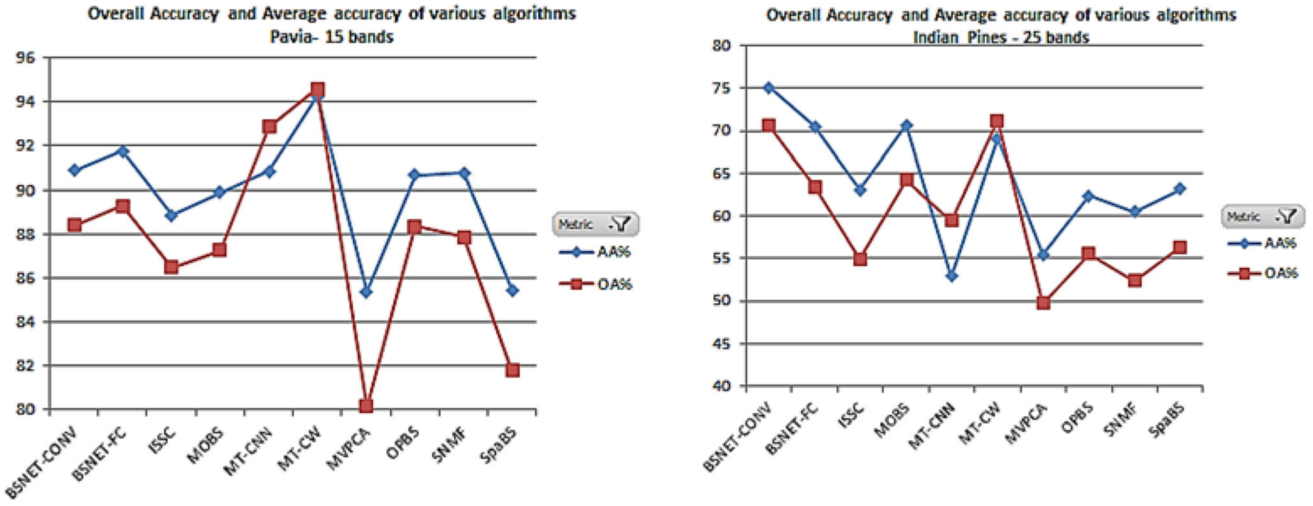

Fig. 10 Overall Accuracy (OA) and Average Accuracy (AA) for the different methods [14]

Table 2 Tabulation of Accuracy with different combination of Transformer layers and Number of Headers

Data set / Overall Accuracy	# of Header = 4, Transformer layers = 1	# of Header = 4, Transformer layers = 3	# of Header = 2, Transformer layers = 2	# of Header = 4, Transformer layers = 2
Indian Pines	48.25%	59.61%	51.33%	65.54%
Pavia	89.36%	89.92%	89.91%	94.62%

2. Average Accuracy (AA): This measure is the average accuracy per class (sum of accuracy for each class predicted/number of classes).
3. Kappa Coefficient: This index is more robust and is a measure of agreement between the final classification map and the ground-truth map.

Figure 3 shows the overall accuracy of MT-CW for different wavelets for Pavia Data and Indian Pines datasets and we can infer that for both the datasets, use of the 'Coiflet1' wavelet feature yields good results. Figures 4, 5, and 6 show the performance of MT-CW for different bands sizes, percentages of training samples, and different wavelets for the Pavia Data set.

Table 3 Classification performance of different methods using 25 Bands and 5% training data on Indian Pines dataset

Class	ISSC	SpaBS	MVPCA	SNMF	MOBS	OPBS	BSNET-FC	BSNET-CONV	MT-CNN	MT-CW Accuracy	σ
1	26.28 ± 16.72	36.56 ± 20.05	19.77 ± 172.26	21.88 ± 17.02	45.65 ± 22.02	37.26 ± 24.51	36.89 ± 24.09	59.83 ± 28.50	21.43	59.89	0.23
2	63.28 ± 3.97	59.11 ± 5.08	52.26 ± 2.69	58.42 ± 3.84	70.5 ± 3.49	59.71 ± 4.41	69.56 ± 3.23	74.34 ± 3.07	45.29	48.67	0.09
3	49.26 ± 6.91	48.64 ± 2.93	37.3 ± 3.95	40.13 ± 3.61	61.95 ± 5.00	48.14 ± 4.00	60.13 ± 4.96	62.38 ± 4.49	52.77	45.68	0.06
4	34.81 ± 8.84	38.56 ± 9.78	13.43 ± 3.38	34.74 ± 9.90	43.32 ± 13.31	28.46 ± 11.24	42.15 ± 7.69	54.48 ± 14.08	61.25	47.60	0.20
5	65.56 ± 8.52	82.34 ± 4.01	64.29 ± 6.32	73.5 ± 6.44	78.88 ± 7.18	78.16 ± 6.24	80.34 ± 4.98	83.17 ± 5.56	25.65	67.84	0.10
6	83.58 ± 3.14	82.67 ± 4.96	81.44 ± 4.17	87.62 ± 4.28	86.36 ± 3.76	87.62 ± 4.66	92.15 ± 3.44	94.63 ± 2.51	80.11	73.24	0.12
7	21.09 ± 19.21	26.4 ± 27.63	33.91 ± 24.45	16.36 ± 18.59	36.35 ± 33.38	30.94 ± 31.70	44.95 ± 32.58	52.18 ± 40.35	57	60.29	0.27
8	87.1 ± 4.10	86.45 ± 7.78	86.6 ± 8.29	83.5 ± 6.87	95.96 ± 1.91	89.09 ± 7.15	92.8 ± 6.51	96.32 ± 2.41	95.76	84.87	0.07
9	7.24 ± 9.77	4.64 ± 6.76	2.66 ± 3.96	7.76 ± 9.66	16.86 ± 21.22	13.75 ± 15.83	19.84 ± 24.25	28.35 ± 26.93	18.23	47.83	0.32
10	54.28 ± 7.40	52 ± 5.81	43.33 ± 5.70	46.27 ± 7.32	63.55 ± 6.34	49.44 ± 4.85	61.94 ± 4.71	69.05 ± 5.38	64.26	52.43	0.24
11	62.3 ± 3.02	58.06 ± 3.05	49.13 ± 3.53	56.63 ± 4.18	65.44 ± 6.34	60.99 ± 3.36	68.93 ± 3.96	71.39 ± 3.76	56.18	59.63	0.09
12	35.06 ± 5.41	40.61 ± 7.02	32.35 ± 8.25	33.93 ± 6.98	60.1 ± 10.04	30.29 ± 5.88	43.72 ± 6.47	64.99 ± 9.31	42.72	38.21	0.07
13	87.69 ± 8.86	89.42 ± 6.71	80.2 ± 13.33	75.94 ± 10.84	90.86 ± 6.93	79.2 ± 9.12	86.64 ± 10.70	95.24 ± 2.94	96.13	69.44	0.20
14	83.56 ± 4.74	88.61 ± 3.71	85.26 ± 4.20	88.8 ± 3.23	89.88 ± 3.68	86.74 ± 5.04	89.22 ± 2.97	89.03 ± 3.89	73.41	77.83	0.06
15	33.29 ± 9.82	41.14 ± 8.54	32.1 ± 6.37	40.56 ± 8.65	34.38 ± 6.19	30.76 ± 6.73	42.03 ± 7.37	49.17 ± 7.54	63.84	62.67	0.16
16	84.7 ± 4.74	66.69 ± 19.80	81.94 ± 6.44	72.54 ± 25.44	87.47 ± 5.03	78.89 ± 16.62	82.94 ± 82.94	85.16 ± 5.15	88.29	92.17	0.07
OA%	54.95 ± 2.01	56.37 ± 3.12	49.75 ± 2.00	52.41 ± 2.33	64.22 ± 3.46	55.59 ± 2.96	63.39 ± 3.05	70.61 ± 2.56	59.45	65.54	4.64
AA%	63.06 ± 1.07	63.16 ± 1.23	55.47 ± 1.04	60.5 + 1.07	70.67 ± 1.23	62.41 ± 1.12	70.51 ± 1.35	75.15 ± 1.17	52.92	57.95	1.17
Kappa	0.579 ± 0.012	0.58 ± 0.014	0.494 ± 0.011	0.55 ± 0.12	0.66 ± 0.014	0.572 ± 0.013	0.664 ± 0.015	0.717 ± 0.013	0.5442	0.608	.0950

Table 4 Classification Performance of different methods using 15 Bands and 5% training data on Pavia dataset [14]

Class	ISSC	SpaBS	MVPCA	SNMF	MOBS	OPBS	BSNET- FC	BSNET-CONV	MT-CNN	MT-CW	Accuracy σ
1	90.82 ± 1.23	88.79 ± 1.37	89.91 ± 1.18	90.21 ± 1.11	91.17 ± 1.33	90.72 ± 1.32	91.53 ± 0.80	91.15 ± 0.88	96.12	93.81	0.015
2	95.07 ± 0.36	95.88 ± 0.63	94.99 ± 0.41	96.36 ± 0.37	95.68 ± 0.45	95.61 ± 0.55	96.64 ± 0.35	95.72 ± 0.37	91.83	95.42	0.013
3	69.65 ± 3.30	62.98 ± 3.05	57.06 ± 4.59	72.77 ± 1.49	70.97 ± 2.88	73.6 ± 3.13	72.3 ± 2.94	71.2 ± 3.01	99.01	87.63	0.091
4	89.29 ± 2.07	85.1 ± 1.66	79.86 ± 2.38	89.37 ± 0.39	89 ± 1.68	89.72 ± 1.97	91.47 ± 1.78	90.6 ± 1.78	99.09	99.2	0.004
5	98.89 ± 0.35	99.1 ± 0.27	98.39 ± 0.67	99.18 ± 1.44	99.44 ± 0.20	99.18 ± 0.50	99.31 ± 0.30	99.36 ± 0.28	99.94	99.89	0.004
6	74.49 ± 1.64	52.0 ± 3.01	72.76 ± 1.87	85.53 ± 4.23	78.74 ± 1.69	84.64 ± 1.49	86.55 ± 1.65	84.31 ± 1.54	95.13	94.78	0.047
7	78.89 ± 4.01	70.63 ± 6.14	55.44 ± 6.72	73.47 ± 4.23	76.62 ± 4.08	78.9 ± 3.16	80.66 ± 2.80	80.01 ± 3.21	95.24	99.89	0.057
8	81.28 ± 1.83	81.87 ± 2.45	76.49 ± 3.61	84.11 ± 0.17	83.56 ± 2.18	82.92 ± 1.60	85.16 ± 2.30	83.37 ± 1.96	83.12	91.34	0.021
9	99.77 ± 0.16	99.71 ± 0.14	99.96 ± 0.15	99.81 ± 0.17	99.97 ± 0.07	99.81 ± 0.14	99.97 ± 0.05	99.92 ± 0.08	99.89	99.12	0.012
OA%	86.46 ± 0.54	81.78 ± 0.89	80.15 ± 0.68	87.87 ± 0.58	87.24 ± 0.59	88.34 ± 0.62	89.29 ± 0.47	88.4 ± 0.51	92.87	94.62	1.346
AA%	88.86 ± 0.28	85.42 ± 0.33	85.35 ± 0.22	90.76 ± 0.26	89.87 ± 0.27	90.65 ± 0.37	91.77 ± 0.30	90.87 ± 0.28	90.84	94.25	0.992
Kappa	0.852 ± 0.004	0.803 ± 0.005	0.804 ± 0.003	0.877 ± 0.003	0.865 ± 0.004	0.876 ± 0.005	0.891 ± 0.004	0.879 ± 0.004	0.904	0.929	.0014

From Fig. 4 it is seen that when 'dmey' is used as the wavelet the performance degrades and is lesser than the accuracy achieved by the model without wavelet features. Therefore, the use of the right wavelet is very important. MT-CNN is represented by results the under heading 'Without wavelet'. Figure 5 indicates that 30% training data yields the best response. From Fig. 6 we deduce that the Pavia dataset gives good results with a reduced subset of 15 bands. However, best performance is with 45 bands.

Figures 7, 8 and 9 show the performance of MT-CW for different wavelets, band sizes, and percentages of training samples for Indian Pines. From Fig. 7 it is seen that classification results are good with 'Coiflet1' as a wavelet for the Indian Pines dataset as well. From Fig. 8 it is observed that the classification accuracy drops drastically for 5% training data when bands are increased to 100. This is due to the "Curse of dimensionality". As the dimensions increase, the amount of training data required increases. This is also evident from Fig. 5 where the classification accuracy is higher for 20% of training data as against 5% of training data. From Fig. 9 it is clear that best results are for 60% training data. Figure 10 shows the comparative analysis of the different methods.

The results from experiment with different combinations of number of headers and number of transformer layers are tabulated below in Table 2. It is evident that number of header if chosen as 4 and transformer layers as 2 gives best performance for both data sets. Band selection Model converge faster for Pavia dataset than the Indian Pines dataset. Hence, epochs is set as 80 for Pavia whereas it is set as 150 for Indian Pines dataset.

Table 3 below tabulates the classification performance of different methods [14] for the best 25 bands for Indian Pines using 5% training samples. It is observed that the MT-CW method provides OA (65.54%), AA (57.95%), and Kappa (0.608). For class 7 and class 9 which have comparatively lesser training samples classification accuracy of MT-CW is better than all other models. The same can be said about class 15 and Class 16. It is also observed that while the overall accuracy of classification of MT-CW is more than MT-CNN, adding wavelet features reduces the accuracy of certain classes like corn and Hay-windrowed.

In Table 4, we show the detailed classification performance of different methods by selecting the best 15 bands for Pavia data set using 5% training samples. As can be seen from Table 4, MT-CW achieves the best OA (94.62%), AA (94.25%) and Kappa (0.929) and performs better than all other methods. Table 4 lists the selected band indices determined by different BS methods [14] for Indian Pines, Pavia University data sets. The numbers of selected bands of the Indian Pines and Pavia data sets are 25 and 15, respectively.

Table 5 Best band indices for Pavia (15 bands) and Indian Pines (25 bands) by different methods [14]

Indian Pines	MT-CNN / MT-CW	[2,6,11,17,36,44,47,51,61,77,78,89,93,104,116,136,140,141,146,149,161,167,175,182,196]
	BSNET-FC	[165.38,51,65,12,100,0,71,5,60,88,26,164,75,74,52,22,94,35,11,184,179,34,160,46]
	BSNET-CONV	[46,33,140,161,80,35,178,44,126,36,138,71,180,66,192,16,53,152,185,119,24,28,26,156, 83]
	ISSC	[171,130,67,85,182,183,47,143,138,90,129,141,25,142,21,181,3,34,13,14,11,10,190,187, 189]
	SpaBS	[7.96,52,171,53,3,76,75,74,95,77,73,78,54,81,92,88,91,71,72,79,80,55,92,56]
	MVPCA	[167,74,168,9,147,165,161,162,152,19,160,119,164,159,157,163,158,156,20,154,118,148, 153,149,155]
	SNMF	[23,197,198,94,76,2,87,105,143,145,11,84,132,108,28,104,144,34,44,74,71,96,75,171,162]
	MOBS	[8,14,21,22,30,33,39,51,52,64,79,80,86,87,106,114,124,128,135,141,160,168,169,182,196]
	OPBS	[28,41,60,0,74,34,88,19,17,33,56,87,22,31,73,12,18,32,90,39,11,89,75,24,70]
Pavia U	MT-CNN / MT-CW	[3,9,16,19,24,37,38,40,51,52,57,80,90,94,98]
	BSNET-FC	[38,78,17,20,85,98,65,81,79,90,95,74,66,62,92]
	BSNET-CONV	[90,42,16,48,71,3,78,38,80,53,7,31,4,99,98]
	ISSC	[51,76,7,64,31,8,0,24,40,30,5,3,6,27,2]
	SpaBS	[50,48,16,22,4,102,21,25,23,47,24,20,31,26,41]
	MVPCA	[48,22,51,16,52,21,65,17,20,53,18,54,19,55,76]
	SNMF	[92,53,43,66,22,89,82,30,51,5,83,77,80,2,48]
	MOBS	[4,15,23,25,33,35,42,53,58,61,62,64,67,73,101]
	OPBS	[90,62,14,0,2,72,102,4,33,1,6,84,45,82,8]

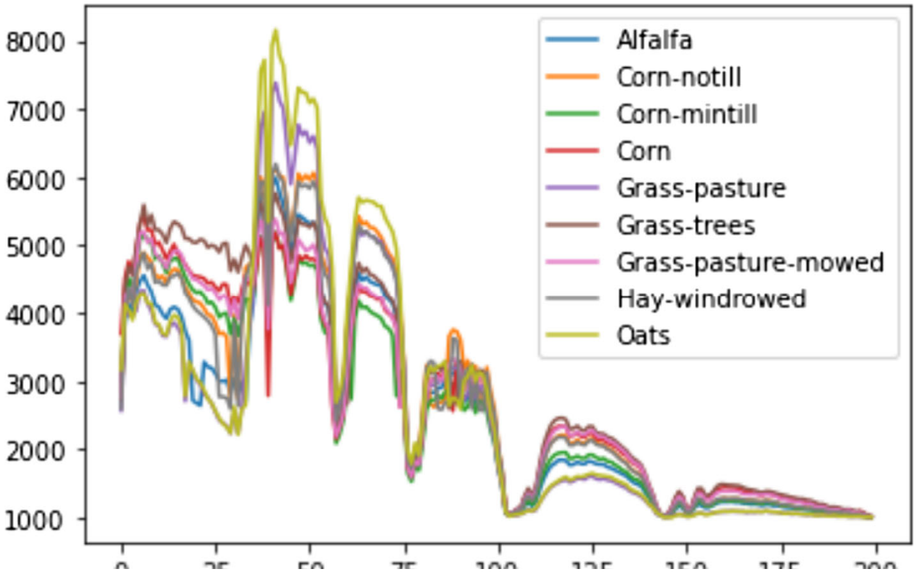

Fig. 11 Full band Spectra Plot of Indian Pines

Figures 11 and 12 show the spectra plot for the Indian Pines data set with all bands and Top15 bands, respectively. Similarly, Figs. 13 and 14 show the spectra plot for the Pavia data set with all bands and top 25 bands, respectively. From the plots, it is evident that the reduced bands provide good class separability for both the data sets.

4 Conclusion

This article presents an approach for Hyperspectral image classification with dimensionality reduction. It uses the state-of-art Transformer [1] algorithm for generating the attention mask. The attention masks help in identifying bands that contain discriminative information and eliminate trivial bands. The attention masks consider the average weight for the bands during the training process along with the frequency of its occurrence in top bands. This

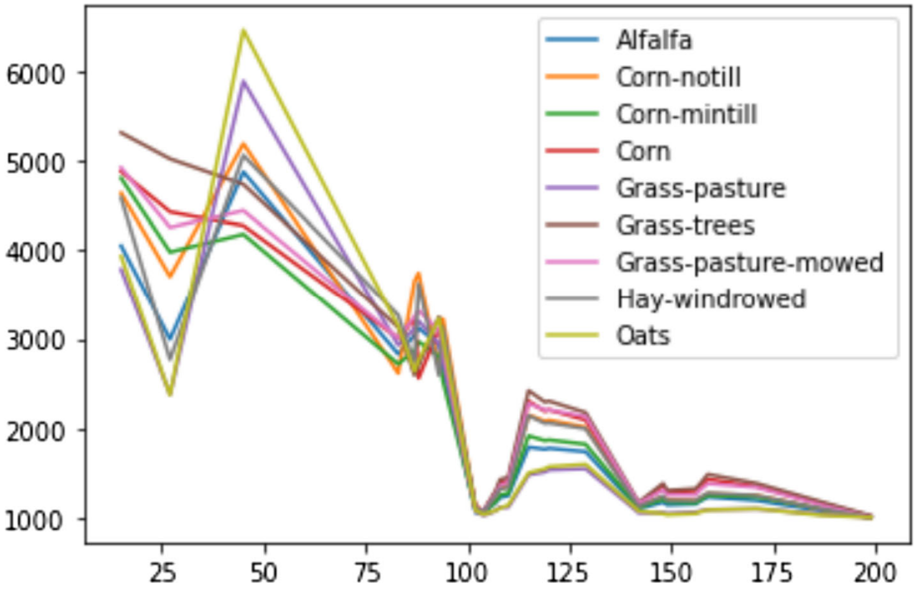

Fig. 12 Spectra Plot of Indian Pines with 25 Best bands from MT-CW

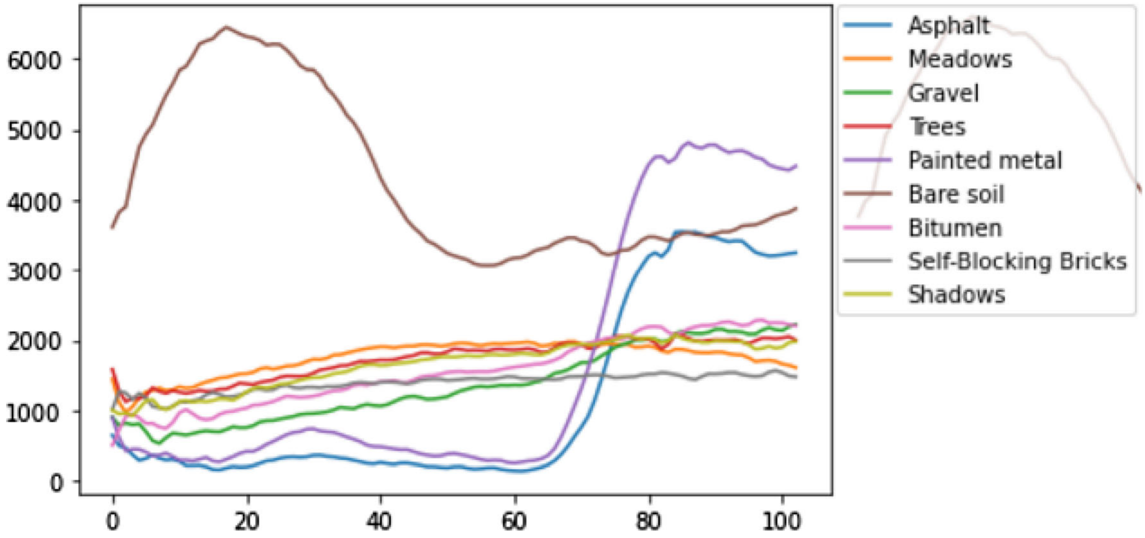

Fig. 13 Full band Spectra Plot of Pavia

eliminates the outliers and results in a better sparse representation of the subset of bands.

Experiments show that when the CNN features are coupled with wavelet features it improves the classification accuracy for both Indian Pines and Pavia datasets across different training sets and subsets of bands. The choice of wavelet depends on the type of classes in any given imagery. Experiments with different wavelets led to the conclusion that 'Coiflet1' wavelet results in discriminating features for classes in Indian Pines and Pavia University dataset. Since 'Coiflet1' uses six scaling and wavelet function coefficients, an increase in pixel averaging and differencing leads to a smoother wavelet. The effect of wavelet features is not so visible in Pavia dataset because the classes have distinct signatures. However, the impact of wavelet features is more prominent in Indian pines because it contains classes that have very high similarity index, and adding wavelet features to CNN features further enhances the classification accuracy.

For Pavia University dataset we conclude that since the classes in this dataset are sufficiently distinct a lesser number of bands are required to classify them with high accuracy. The experimental results show that for the Pavia dataset best results are achieved with 45 bands, 30% training data, and Coiflet1 for extracting wavelet features.

Similarly, for the Indian Pines dataset, we conclude that since classes have a high similarity index the number of bands required to classify them is slightly higher than Pavia

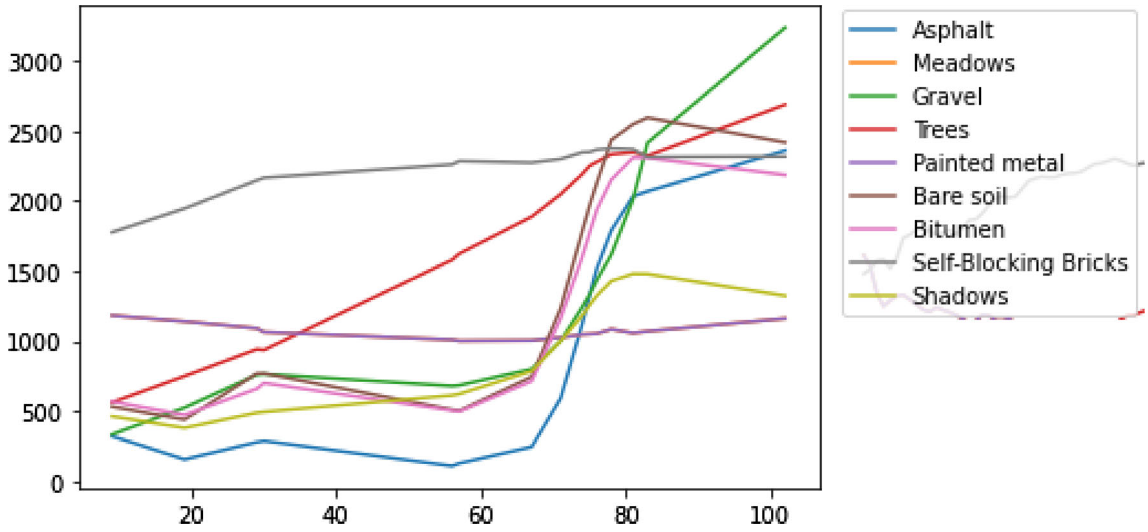

Fig. 14 Spectra Plot of Pavia with 15 Best bands from MT-CW

University. It was also observed that for the training set of 5% certain classes like 'Oats' which have a lesser number of labeled pixels result in a drop in classification accuracy. This is because of the 'Curse of Dimensionality' effect. Augmenting of samples may help address this issue. The experimental results show that for Indian Pines best results are achieved with 80 bands, 30% training data, and Coiflet1 for extracting wavelet features. Except for class 'Oats' all classes give an accuracy of above 90% for this setup. CNN feature extractor hyperparameters remain the same to classify these two different datasets.

The proposed solution has been tested on Anand and Surendra Nagar, India, datasets as well provides accuracy to the tune of 98% for almost all classes. This dataset was acquired using the Aviris NG sensor that has a Spectral resolution of 5 nm ± 0.5 nm and a spatial resolution of 4 m. The field data were collected for this on March 18, 2018. However, since there is no comparative study available for these data sets, the result for them are not presented in this paper.

The experimental results show that the implemented MT-CW attention module has the ability to efficiently and effectively produce a sparse representation of data. It performs well when compared to some of the state-of-art techniques. The proposed solution employs 3D convolutions and can be computationally expensive if the subset bands are high. Future work will include enhancing the band selection algorithm to eliminate bands that have high similarities.

Acknowledgements The authors would like to thank the authors of paper 'BS-Nets: An End-to-End Framework for Band Selection of Hyperspectral Image' [14] for providing a comprehensive comparative analysis of different methods. The authors have used their analysis as a basis of comparison. The data that support the findings of this study is publicly available.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, & Polosukhin I (2017). Attention is all you need. ArXiv, abs/1706.03762
2. Vidal R, Ma Yi, Sastry S (2005) Generalized principal component analysis (GPCA). IEEE Trans Pattern Anal Mach Intell 27(12):1945–1959
3. Wang J, Chang C-I (2006) Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. IEEE Trans Geosci Remote Sens 44(6):1586–1600. https://doi.org/10.1109/TGRS.2005.863297
4. Mou L, Ghamisi P, Zhu XX (2017) Deep recurrent neural networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens 55(7):3639–3655. https://doi.org/10.1109/TGRS.2016.2636241
5. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. IEEE/CVF Conf Comput Vision Pattern Recogn 2018:7132–7141. https://doi.org/10.1109/CVPR.2018.00745
6. Abdolmaleki M, Fathianpour N, Tabaei M (2018) Evaluating the performance of the wavelet transform in extracting spectral alteration features from hyperspectral images. Int J Remote Sens 39(19):6076–6094. https://doi.org/10.1080/01431161.2018.1434324
7. Cao X, Yao J, Fu X, Bi H, Hong D (2020) An Enhanced 3-D discrete wavelet transform for hyperspectral image classification. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2020.2990407
8. Wang Y, Cui S (2014) Hyperspectral image feature classification using stationary wavelet transform. Int Conf Wavelet Anal Pattern Recogn 2014:104–108. https://doi.org/10.1109/ICWAPR.2014.6961299

9. Hsu, P-H & Yang, H-H (2007). Hyperspectral image classification using wavelet networks. IEEE International Geoscience and Remote Sensing Symposium. 1767 - 1770. https://doi.org/10.1109/IGARSS.2007.4423162.
10. Paoletti ME et al (2018) Deep & dense convolutional neural network for hyperspectral image classification. Remote Sens 10:1454
11. Paoletti ME et al (2017) A new deep convolutional neural network for fast hyperspectral image classification. Isprs J Photogramm Remote Sens 145:120–147
12. Yang H, Du Q, Chen G (2011) Unsupervised hyperspectral band selection using graphics processing units. IEEE J Sel Topics Appl Earth Observ Remote Sens 4(3):660–668
13. Chen, Long et al. "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning." In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 6298–6306
14. Cai Y, Liu X, Cai Z (2020) BS-nets: an end-to-end framework for band selection of hyperspectral image. IEEE Trans Geosci Remote Sens 58(3):1969–1984. https://doi.org/10.1109/TGRS.2019.2951433
15. Wang J, Zhou J, Huang W (2019) Attend in bands: hyperspectral band weighting and selection for image classification. IEEE J Select Topics Appl Earth Obs Remote Sens 12(12):4712–4727. https://doi.org/10.1109/JSTARS.2019.2955097
16. P. Ribalta Lorenzo, L. Tulczyjew, M. Marcinkiewicz and J. Nalepa, "Hyperspectral Band Selection Using Attention-Based Convolutional Neural Networks,“ in IEEE Access, vol. 8, pp. 42384–42403, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2977454.
17. Li S, Qiu J, Yang X, Liu H, Wan D, Zhu Y (2014) 'A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search.' Eng Appl Artif Intell 27:241–250
18. Li F, Zhang P, Huchuan L (2018) 'Unsupervised band selection of hyperspectral images via multi-dictionary sparse representation.' IEEE Access 6:71632–71643
19. S. Jia, Z. Ji, Y. Qian, and L. Shen, ''Unsupervised band selection for hyperspectral imagery classification without manual band removal,'' IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 5, no. 2, pp. 531–543, Apr. 2012.
20. Martinez-Uso A, Pla F, Sotoca JM, Garcia-Sevilla P (2007) Clustering-based hyperspectral band selection using information measures. IEEE Trans Geosci Remote Sens 45(12):4158–4171. https://doi.org/10.1109/TGRS.2007.904951
21. Yang R, Su L, Zhao X, Wan H, Sun J (2017) 'Representative band selection for hyperspectral image classification.' J Vis Commun Image Represent 48:396–403
22. Guo B, Gunn SR, Damper RI, Nelson JDB (2006) 'Band selection for hyperspectral image classification using mutual information.' IEEE Geosci Remote Sens Lett 3(4):522–526
23. Du Q, Yang H (2008) 'Similarity-based unsupervised band selection for hyperspectral image analysis.' IEEE Geosci Remote Sens Lett 5(4):564–568
24. Van der Maaten L, Postma E, Van Den Herik H (2009) Dimensionality reduction: A comparative review. J Mach Learn Res 10:1–41
25. Ghamisi P, Yokoya N, Li J, Liao W, Liu S, Plaza J, Rasti B, Plaza A (2017) Advances in hyperspectral image and signal processing: a comprehensive overview of the state of the art. IEEE Geosci Remote Sens Mag 5:37–78
26. Zhai H, Zhang H, Zhang L, Li P (2019) Laplacian-regularized lowrank subspace clustering for hyperspectral image band selection. IEEE Trans Geosci Remote Sens 57(3):1723–1740
27. W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,vol. 8, no. 6, pp. 2784–2797, Jun. 2015.
28. Sun K, Geng X, Ji L (2015) A new sparsity-based band selection method for target detection of hyperspectral image. IEEE Geosci Remote Sens Lett 12(2):329–333
29. Chang C-I, Du Q, Sun T-L, Althouse MLG (1999) A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. IEEE Trans Geosci Remote Sens 37(6):2631–2641
30. Li J-M, Qian Y-T (2011) Clustering-based hyperspectral band selection using sparse nonnegative matrix factorization. J Zhejiang Univ Sci C 12(7):542–549
31. Gong M, Zhang M, Yuan Y (2016) Unsupervised band selection based on evolutionary multi-objective optimization for hyperspectral images. IEEE Trans Geosci Remote Sens 54(1):544–557
32. Zhang W, Li X, Dou Y, Zhao L (2018) A geometry-based band selection approach for hyperspectral image analysis. IEEE Trans Geosci Remote Sens 56(8):4318–4333
33. Yu, Xu, et al. "A selective ensemble learning based two-sided cross-domain collaborative filtering algorithm." Information Processing & Management 58.6 (2021): 102691.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.