# MixerSENet: A Lightweight Framework for Efficient Hyperspectral Image Classification

Mohammed Q. Alkhatib, *Senior Member, IEEE,* Swalpa Kumar Roy, *Senior Member, IEEE* and Ali Jamali

***Abstract*—In this paper, a novel framework, MixerSENet, is introduced for hyperspectral image (HSI) classification, designed to address the challenges of computational efficiency and limited labeled data. The proposed model processes hyperspectral image patches while maintaining consistent size and resolution throughout the network, effectively decoupling the mixing of spatial and channel dimensions. Notably, MixerSENet is lightweight and computationally efficient, requiring fewer parameters compared to traditional models, making it suitable for resource-constrained environments. A squeeze and excitation block is incorporated into the model to refine feature extraction, enhancing the network's ability to capture more informative features. Experimental results on two benchmark datasets demonstrate that MixerSENet achieves superior performance, reaching an overall accuracy (OA) of 82.47% on Houston13 dataset and 96.70% on the Qingyun dataset, outperforming state-of-the-art methods including 3D-CNN, HybridKAN, HSIFormer, SimPoolFormer, and MorphMamba. Furthermore, a detailed analysis of computational efficiency shows that MixerSENet achieves a favorable balance between accuracy and efficiency, with only 53,146 parameters and an low inference time, confirming its practicality for real-world applications. At publication, source code will be publicly available at https://github.com/mqalkhatib/MixerSENet.**

***Index Terms*—Hyperspectral Imaging (HSI), HSI Classification, Mixer Networks, Depth-Wise Convolution, Attention Block.**

## I. Introduction

Hyperspectral image (HSI) data has been available since the 1980s [1]. It offers an extensive amount of rich spectral and spatial information, spanning hundreds of narrow contiguous spectral bands from visible to infrared wavelengths. This rich data enables the execution of fine-grained remote sensing tasks that were previously challenging. HSI classification has become a prominent area of research in Remote Sensing (RS), given its wide range of applications in Earth Observation (EO), including land cover and land used mapping and environmental monitoring. The accuracy of HSI classification is a critical factor for the success of these applications. Achieving precise and reliable classification results necessitates the extraction of effective spatial and spectral features.

With the rapid advancements in Deep Learning (DL) technologies, researchers in computer vision have increasingly turned to DL methods, particularly Deep Convolutional Neural Networks (DCNNs), for classifying HSI data. DCNNs have proven to outperform traditional classification methods [2], [3]. In [4], a 1D-CNN method was introduced to extract spectral features. However, relying solely on spectral information is insufficient to achieve high classification accuracy. To address this, [5] proposed a 2D-CNN approach that incorporates spatial information for HSI classification. Despite this, these methods failed to fully exploit the three-dimensional characteristics of HSI data. To overcome this limitation, [6] introduced a 3D-CNN approach, which utilizes both spectral and spatial information to further enhance classification performance. Building on this, Roy *et al.* developed the Hybrid Spectral CNN (HybridSN) [3], which combines 2D-CNN and 3D-CNN where the 3D-CNN captures joint spatial–spectral features early in the process, and the 2D-CNN refines the extraction of higher-level spatial features. Numerous other studies have proposed similar approaches [2], [7].

Building on the success of transformer models in natural language processing, researchers are increasingly exploring their applications in computer vision and Earth observation [8], [9]. These advanced models have demonstrated significant potential in various fields, including hyperspectral imagery analysis [10]. However, one major challenge with transformers is their higher requirement for large training datasets compared to Convolutional Neural Networks (CNNs). This makes their application in remote sensing, particularly for hyperspectral imaging with limited labeled data, more difficult. To overcome this challenge, the HSIFormer model [11] was introduced. This model employs an efficient Vision Transformer (ViT) with local window attention (LWA) for precise hyperspectral image classification. The aim is to improve classification accuracy while addressing the limitations posed by limited labeled data in remote sensing. However, the major drawback is the significant computational cost and hardware resources required compared to standard CNN classifiers.

In this paper, a novel and lightweight framework, MixerSENet, is presented for HSI classification to address these challenges. The framework processes hyperspectral image patches as input, maintaining consistent size and resolution across the network while effectively decoupling the mixing of spatial and channel dimensions. Notably, MixerSENet is designed to be computationally efficient, requiring fewer parameters compared to conventional models. This design is inspired by the MLPMixer [12] and PolSARconMixer [13] models, respectively. Additionally, a squeeze and excitation block is incorporated for feature refinement, enhancing the network's ability to capture more informative and robust features.

The rest of the paper is organized as follows: Section II

M. Q. Alkhatib is with the College of Engineering and IT, University of Dubai, Dubai, 14143, UAE. (e-mail: mqalkhatib@ieee.org).

S. K. Roy is with the Department of Computer Science and Engineering, Alipurduar Government Engineering and Management College, West Bengal 736206, India (e-mail: swalpa@agemc.ac.in).

A. Jamali is with the Department of Geography, Simon Fraser University, British Columbia 8888, Canada (e-mail: alij@sfu.ca).

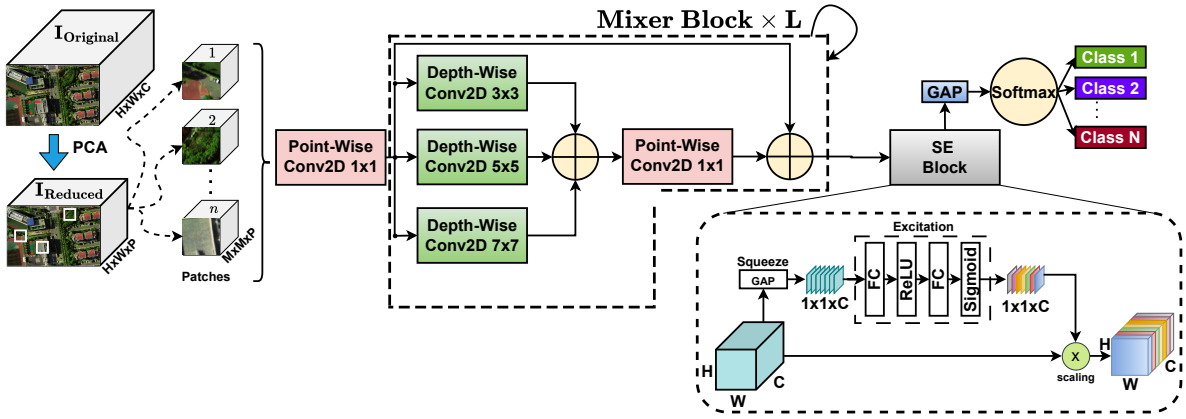


Fig. 1. Architecture of the proposed Model

explains the architecture and building blocks of the model used in the paper, experimental results and comparisons against state-of-the-art models are discussed in Section III, and finally, Section IV summarizes the paper and states the future direction of this research.

## II. NETWORK ARCHITECTURE

The architecture of the proposed model, shown in Fig. 1, starts with the input image ($\mathbf{I_{Original}}$) of size ($\mathbf{H} \times \mathbf{W} \times \mathbf{C}$). Dimensionality reduction is performed using PCA to reduce the number of spectral channels, yielding an output image ($\mathbf{I_{Reduced}}$) with a reduced spectral dimension $\mathbf{P}$, where $\mathbf{P} \ll \mathbf{C}$. The reduced data is then divided into equally sized patches, with each patch processed through point-wise convolution to extract channel features. Depth-wise convolutions with kernel sizes of $3 \times 3$, $5 \times 5$, and $7 \times 7$ are applied to capture spatial features at various scales, followed by a $1 \times 1$ point-wise convolution to mix channel information. This process is repeated $\mathbf{L}$ times to iteratively mix spatial and channel features. A squeeze and excitation (SE) block is introduced for feature refinement, where global average pooling (GAP) compresses the spatial dimensions, followed by fully connected layers with ReLU activation for excitation, and a sigmoid function to calculate the rescaling weights for the input features. Finally, the output is passed through a softmax layer to classify the image patches into one of the predefined classes.

### A. Depthwise Convolution

Unlike regular 2D convolution, which mixes information across all channels, depthwise convolution applies a separate filter to each channel independently. This greatly reduces parameters and computational cost while preserving the number of channels in the output. By decoupling spatial and channel operations, depthwise convolution extracts channel-specific spatial features efficiently, making it well-suited for resource-constrained environments. The process of depthwise convolution is illustrated in Fig. 2.

### B. Squeeze and Excitation

The Squeeze and Excitation (SE) block [14] is designed to improve feature representation by adaptively recalibrating

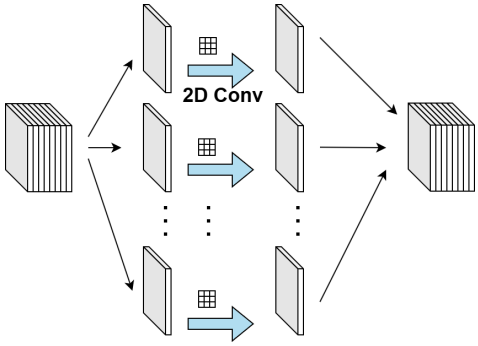


Fig. 2. Depthwise Separable Convolution: Input channels are separated, and each is convolved with a spatial filter. The split channels are then concatenated.

channel-wise responses. This mechanism is particularly beneficial for hyperspectral imagery, where channels correspond to spectral bands that often contain redundant or highly correlated information. By selectively emphasizing more discriminative features while suppressing less informative ones, the SE block enhances the network's ability to capture meaningful patterns for classification.

Formally, given a transformation $F_{tr}$ that maps the input $X$ to feature maps $u_c \in \mathbb{R}^{H \times W \times C}$, where $u_c$ is the $c$-th channel, the *squeeze* operation aggregates global spatial information into a compact channel descriptor. This is achieved using global average pooling, which reduces the representation from $H \times W \times C$ to $1 \times 1 \times C$ as:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{1}$$

The resulting vector $z$ encodes a global view of channel statistics. The *excitation* step then models channel interdependencies and learns nonlinear relationships between them. It generates a set of weights that highlight the importance of each channel, defined as:

$$s = F_{ex}(z, W) = \sigma(W_2 \, \mathrm{ReLU}(W_1 z)), \tag{2}$$

where $\sigma$ denotes the Sigmoid activation, and $W_1$ and $W_2$ are the weights of fully connected layers with a reduction ratio $r$ to control complexity. The resulting channel-wise attention vector $s$ is used to rescale the feature maps $u$, thereby enhancing informative responses while diminishing irrelevant

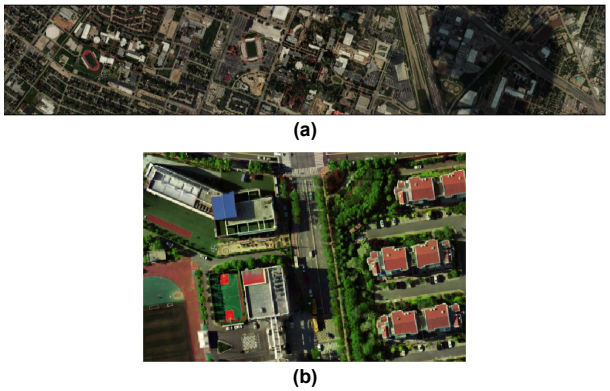


Fig. 3. RGB Composites: (a) Houston13; (b) QUH-Qingyun.

TABLE I
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS FOR THE HOUSTON13 DATASET.

| Class | Train / Val | Test | 3D-CNN | Hybrid Kan | HSI-Former | Simpool-Former | Morph-Mamba | Mixer-Net | Mixer-SENet |
|---|---|---|---|---|---|---|---|---|---|
| Healthy Grass | 99 | 1,053 | 81.39 | 80.15 | 80.72 | 80.25 | 81.10 | 81.77 | 81.58 |
| Stressed Grass | 95 | 1,064 | 85.15 | 76.41 | 85.15 | 85.15 | 83.74 | 81.58 | 83.83 |
| Synthetic Grass | 96 | 505 | 94.46 | 81.19 | 79.80 | 73.07 | 70.69 | 98.42 | 99.80 |
| Tree | 94 | 1,056 | 92.71 | 88.92 | 88.92 | 90.34 | 83.52 | 91.29 | 89.02 |
| Soil | 93 | 1,056 | 99.81 | 94.89 | 99.15 | 98.39 | 99.81 | 99.91 | 100.00 |
| Water | 91 | 143 | 97.90 | 86.71 | 76.92 | 98.60 | 74.13 | 95.80 | 95.10 |
| Residential | 98 | 1,072 | 91.88 | 79.85 | 90.95 | 72.01 | 83.12 | 88.71 | 94.31 |
| Commercial | 96 | 1,053 | 69.61 | 63.34 | 66.95 | 68.47 | 69.52 | 73.60 | 78.35 |
| Road | 97 | 1,059 | 70.07 | 73.65 | 74.22 | 73.18 | 71.29 | 76.02 | 77.43 |
| Highway | 96 | 1,035 | 55.50 | 42.47 | 44.31 | 49.52 | 40.35 | 38.42 | 44.69 |
| Railway | 90 | 1,054 | 72.68 | 69.26 | 74.00 | 66.32 | 66.22 | 70.59 | 66.98 |
| Parking Lot1 | 96 | 1,041 | 56.58 | 72.91 | 57.25 | 80.40 | 73.87 | 85.01 | 86.07 |
| Parking Lot2 | 92 | 285 | 92.28 | 83.51 | 86.32 | 87.72 | 72.28 | 86.32 | 82.11 |
| Tennis Court | 90 | 247 | 100.00 | 76.11 | 100.00 | 100.00 | 93.12 | 98.38 | 100.00 |
| Running Track | 93 | 473 | 95.35 | 79.49 | 83.30 | 89.43 | 90.91 | 100.00 | 99.37 |
| OA (%) | | | 80.13 ±0.52 | 75.27 ±0.63 | 77.38 ±0.41 | 77.82 ±0.46 | 76.04 ±0.74 | 81.23 ±0.28 | 82.47 ±0.25 |
| AA (%) | | | 83.72 ±0.58 | 76.64 ±0.71 | 79.18 ±0.46 | 80.93 ±0.49 | 76.87 ±0.83 | 84.42 ±0.37 | 85.21 ±0.33 |
| Kappa x 100 | | | 78.42 ±0.39 | 73.12 ±0.64 | 75.48 ±0.43 | 76.13 ±0.48 | 74.08 ±0.69 | 79.63 ±0.31 | 81.03 ±0.27 |

or noisy channels. This recalibration mechanism improves the network's ability to exploit subtle variations that are critical in hyperspectral image classification.

## III. EXPERIMENTS AND ANALYSIS

To evaluate the performance of the proposed approach shown in Fig. 1, we compare it against several state-of-the-art methods, including 3D-CNN [6], the Kolmogorov Arnold Network (HybridKAN) [15], HSIFormer [11], SimPoolFormer [16] and MorphMamba [17]. To assess the impact of attention mechanisms, the model is tested both with and without the attention module. In this context, "MixerNet" denotes the proposed framework without the SE block, whereas "MixerSENet" incorporates the SE block. This distinction effectively serves as an ablation study, isolating the contribution of the SE component to the overall model performance. Performance is measured using Overall Accuracy (OA), Average Accuracy (AA), and the Kappa statistic (Kappa), with the classification accuracy for each class also reported. The evaluation is performed on two widely used hyperspectral datasets: Houston13 and QUH-Qingyun. Fig. 3 shows the RGB composite for both datasets,while Figures 4(a) and 5(a) show the available reference class maps. Full description of both datasets is available in [18] and [15], respectively.

For the Houston13 dataset, we followed the partitioning scheme published in the 2013 IEEE GRSS Data Fusion Contest. The available training data were evenly divided into training and validation sets, while the provided testing set

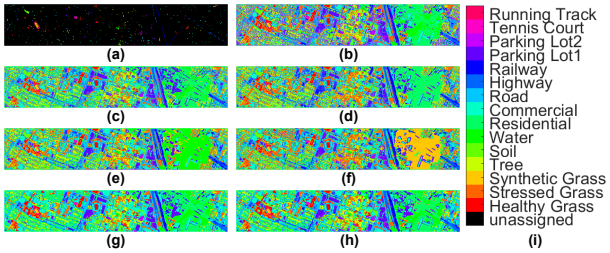


Fig. 4. Classification maps of Houston13 Dataset. (a) Reference Data; (b) 3D-CNN; (c) HybridKAN; (d) HSIFormer; (e) SimPoolFormer; (f) MorphMamba; (g) MixerNet; (h) MixerSENet; (i) Class Labels.

was kept unchanged. This resulted in approximately 9% for training, 9% for validation, and the remaining 82% for testing. This setup was adopted to ensure consistency with the competition standards. For the Qingyun datasets, the image patches were randomly split into 5% for training, 5% for validation, and 90% for testing and evaluation, the Train/Val/Test splits are shown in Tables I and II, it highlights the difference in the amount of samples used for training. A patch size of $9 \times 9$ was used, and the number of principal components was set to 15. To ensure consistency in classification outcomes and minimize the impact of random sample selection, the experiments were repeated 10 times. The final result was determined by averaging the outcomes of these experiments, where recorded values are represented in terms of mean and standard deviation. Additionally, detailed classification results for each category of the best performing iteration were provided. The model was trained for 100 epochs with a batch size of 32. An early stopping strategy was employed, where validation accuracy was evaluated at the end of each epoch. If the validation accuracy did not improve for 10 consecutive epochs, training was halted, and the model was restored to the weights that achieved the highest validation accuracy. The Adam optimizer was used with a learning rate of $1 \times 10^{-3}$. All models were implemented in Python using the Keras framework with TensorFlow as the backend. To ensure a fair comparison, all models were trained under identical conditions.

As shown in Table I, MixerSENet attains the highest overall accuracy (82.47% $\pm$ 0.25), surpassing MixerNet, 3D-CNN, and transformer-based methods by notable margins, while also yielding the best AA and Kappa values. Class-wise, MixerSENet performs particularly well on Synthetic Grass (99.80%), Soil (100%), Residential (94.31%), Commercial (78.35%), Road (77.43%), Parking Lot 1 (86.07%), and Running Track (99.37%), indicating strong capability in urban and man-made categories where fine boundaries are critical. Although slightly less competitive for Tree and Water, its performance remains comparable to the strongest baselines. The classification maps in Fig. 4 confirm these trends, showing smoother and more continuous labeling in residential and commercial regions when compared against the RGB image, direct comparisons with the reference map remain challenging due to the large proportion of unassigned pixels. Furthermore, the cloudy region observed in the RGB image (Fig. 3(a)) adds difficulty in distinguishing the classes covered by it,

TABLE II
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS FOR THE QINGYUN DATASET.

| Class | Train / Val | Test | 3D-CNN | Hybrid-KAN | HSI-Former | SimPool-Former | Morph-Mamba | Mixer-Net | Mixer-SENet |
|---|---|---|---|---|---|---|---|---|---|
| Trees | 13,907 | 250,335 | 96.95 | 95.78 | 95.60 | 97.17 | 95.25 | 98.25 | 97.25 |
| Concrete building | 8,976 | 161,561 | 96.85 | 93.28 | 96.87 | 93.52 | 84.91 | 98.91 | 93.92 |
| Car | 689 | 12,405 | 61.53 | 53.59 | 68.03 | 70.91 | 50.80 | 85.63 | 83.32 |
| Ironhide building | 489 | 8,790 | 98.59 | 96.84 | 96.92 | 97.98 | 98.40 | 99.12 | 99.81 |
| Plastic playground | 10,886 | 195,962 | 97.22 | 95.74 | 97.18 | 96.74 | 96.44 | 97.63 | 96.86 |
| Asphalt road | 12,797 | 230,351 | 94.74 | 91.37 | 92.92 | 95.88 | 90.29 | 93.19 | 96.06 |
| OA (%) | | | 95.91 ±0.46 | 93.52 ±0.61 | 95.10 ±0.39 | 95.67 ±0.44 | 91.64 ±0.72 | 96.04 ±0.33 | 96.70 ±0.21 |
| AA (%) | | | 90.98 ±0.53 | 87.77 ±0.68 | 91.25 ±0.41 | 92.03 ±0.48 | 86.02 ±0.76 | 94.54 ±0.36 | 95.46 ±0.24 |
| Kappa × 100 | | | 94.58 ±0.44 | 91.42 ±0.59 | 93.51 ±0.37 | 94.26 ±0.42 | 88.90 ±0.69 | 94.76 ±0.31 | 95.64 ±0.20 |

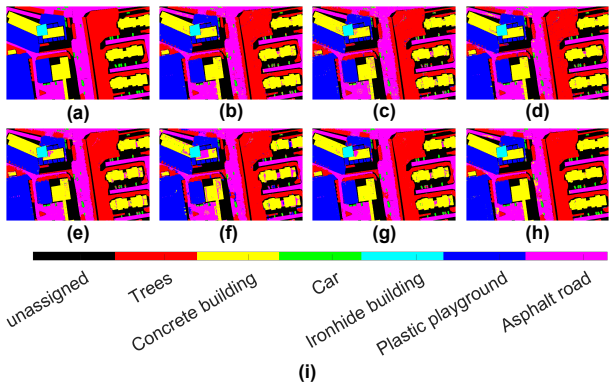


Fig. 5. Classification maps of Qingyun Dataset. (a) Reference Data; (b) 3D-CNN; (c) HybridKAN; (d) HSIFormer; (e) SimPoolFormer; (f) MorphMamba; (g) MixerNet; (h) MixerSENet; (i) Class Labels.

contributing to several misclassifications.

For the Qingyun dataset, the detailed results are summarized in Table II, which reports the classification performance of different models, while Fig. 5 illustrates the corresponding classification maps. In terms of Kappa, MixerSENet achieves 95.64, higher than 3D-CNN (94.58), HybridKAN (91.42), SimPoolFormer (94.26), and HSIFormer (93.51). Although MixerNet obtains a comparable Kappa (94.76), MixerSENet achieves superior OA (96.70% vs. 96.04%) and AA (95.46% vs. 94.54%), indicating more reliable overall performance. The class-wise results also show MixerSENet performing best in challenging categories such as "Car" (83.32% compared to 61.53% for 3D-CNN and 50.80% for MorphMamba).

To assess the impact of the depth value $L$, several tests were conducted on the Qingyun dataset while training on 1% of the data, with results reported in Table III. Increasing the depth from $\times 1$ to $\times 5$ improved all metrics, with OA, AA, and Kappa peaking at 93.96%, 89.45%, and 92.18%, respectively, before slightly degrading at $\times 6$, suggesting mild overfitting at higher depth. Here, $\times$L denotes the number of repeated Mixer blocks, where each block performs multi-scale spatial mixing with depth-wise convolutions followed by spectral mixing via point-wise convolution. Parameter growth is linear with depth, as each block adds a fixed number of parameters; for instance, x6 has 61,190 parameters compared to 51,270 for x5, an increase of 9,920.

*1) Model Size and Computational Efficiency:* Table IV provides a detailed breakdown the efficiency of each model in terms of parameters, floating point operations (FLOPs), multiply accumulate operations (MACs), and measured inference time. MixerSENet achieves a balanced trade off with

TABLE III
IMPACT OF DEPTH VALUE (L) ON CLASSIFICATION PERFORMANCE.

| Depth (L) | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| OA | 92.80 | 93.42 | 93.48 | 93.68 | 93.96 | 93.85 |
| AA | 86.08 | 87.92 | 88.10 | 88.35 | 89.45 | 89.18 |
| Kappa | 90.46 | 91.27 | 91.36 | 91.63 | 92.18 | 91.93 |
| Parameters | 11,590 | 21,510 | 31,430 | 41,350 | 51,270 | 61,190 |

TABLE IV
PARAMETERS, FLOPS, AND MACS OF EACH MODEL USED IN THE RESEARCH

| Model | Parameters | FLOPs ($\times 10^6$) | MACs ($\times 10^6$) | Inference (m:s) |
|---|---|---|---|---|
| 3D-CNN | 397,586 | 0.682 | 0.341 | 1:37 |
| HybridKAN | 142,690 | 20.200 | 10.027 | 03:25 |
| HSIFormer | 1,373,084 | 18.392 | 9.192 | 11:00 |
| SimPoolFormer | 771,122 | 57.497 | 28.423 | 4:52 |
| MorphMamba | 67,650 | 8.592 | 4.294 | 7:32 |
| MixerNet | 52,050 | 7.887 | 3.889 | 2:30 |
| MixerSENet | 53,146 | 7.894 | 3.890 | 2:32 |

53,146 parameters, $7.894\times10^6$ FLOPs, $3.890\times10^6$ MACs, and an inference time of 2 minutes (m) and 32 seconds (s), which is almost identical to MixerNet (2:30) despite the inclusion of the SE block. In contrast, transformer based models are considerably heavier. SimPoolFormer requires $57.497\times10^6$ FLOPs and $28.423\times10^6$ MACs with an inference time of 4:52, while HSIFormer includes more than 1.3 million parameters and takes 11:00 to complete inference, highlighting the computational burden of transformer architectures. MorphMamba, though lightweight in parameters (67,650) and moderate in FLOPs and MACs (8.592 and $4.294\times10^6$), still exhibits a long inference time (7:32), consistent with the use of computationally intensive morphological operations and token processing. Other models such as 3D CNN and HybridKAN span the middle range, with 3D CNN showing very low FLOPs and MACs (0.682 and $0.341\times10^6$) and the shortest inference time (1:37), yet its accuracy remains lower than MixerSENet. HybridKAN, on the other hand, requires $20.200\times10^6$ FLOPs and $10.027\times10^6$ MACs and runs in 3:25. Overall, MixerSENet demonstrates strong computational efficiency without compromising accuracy, making it practical for deployment. All inference times were measured on a Windows 10 machine with 64 GB RAM and an NVIDIA GeForce RTX 2080 GPU with 8 GB VRAM.

*2) Effect of training data:* To evaluate the impact of training data size on the classification performance of MixerSENet, the model's performance on the Qingyun benchmark dataset was analyzed across various training ratios, ranging from 1% to 5%. The results are presented in Fig. 6. It is observed that MixerSENet achieves a high overall accuracy of 94.10% with a relatively small amount of training data, outperforming other classifiers such as 3D-CNN (92.59%), HybridKAN (90.26%), HSIFormer (91.26%), SimPoolFormer (90.10%), MorphMamba (87.31%), and the Mixer-only network (93.35%). Notably, a training proportion of 1% was considered from the reference data. This highlights the effectiveness of MixerSENet in classifying hyperspectral image (HSI) data with limited labeled data, compared to other developed classification algorithms. Furthermore, across all training ratios from 1% to 5%, MixerSENet consistently achieves the highest performance in terms of OA, AA, and Kappa. For example, at 5% training data, MixerSENet obtains 96.70% OA,

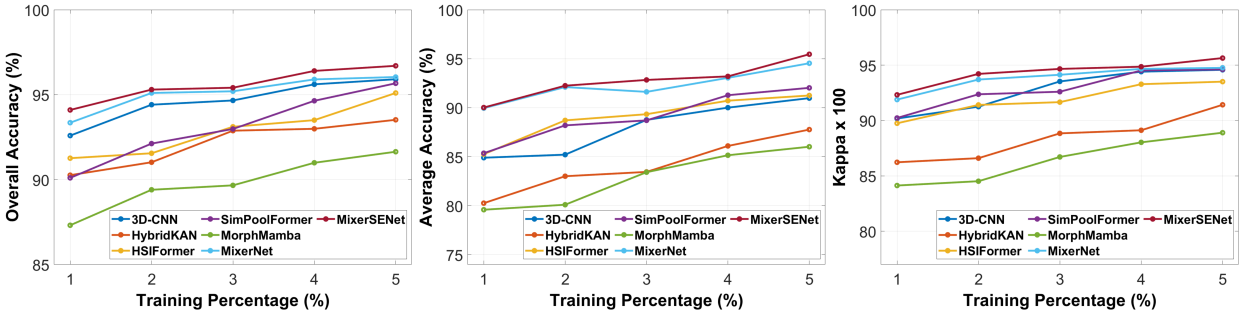


Fig. 6. Classification accuracy of Qingyun dataset at different percentages of training data (left) OA (center) AA and (right) Kappa index.

95.46% AA, and 95.64 Kappa, surpassing MixerNet (96.04 % OA, 94.54% AA,and 94.76 Kappa) and significantly outperforming models such as HSIFormer (95.10% OA, 91.25% AA, and 93.51 Kappa) and MorphMamba (91.64% OA, 86.02% AA, and 88.90 Kappa). These results further confirm not only the robustness of MixerSENet to scarce training data but also its superiority in stability and generalization compared to transformer-based and Mamba-based architectures.

## IV. Conclusions

In this paper, MixerSENet, a lightweight and efficient framework for hyperspectral image classification, is introduced. The proposed model utilizes depth-wise convolutions and a squeeze and excitation (SE) block to enhance feature extraction while maintaining computational efficiency. Experimental results demonstrate that MixerSENet outperforms several state-of-the-art models in terms of overall accuracy, average accuracy, and Kappa, with a significant improvement in class-wise performance. Moreover, competitive performance is achieved with fewer parameters, making it an ideal choice for resource-constrained environments. These findings highlight the potential of MixerSENet as a reliable and efficient solution for hyperspectral image classification tasks.

Future work will focus on further optimizing the model and exploring its applicability to other remote sensing datasets.In particular, while the inclusion of point-wise convolution and the SE block helps mitigate the tendency of depth-wise convolution to struggle with highly correlated spectral bands, this limitation remains and motivates future exploration of alternative spectral mixing strategies. In addition, more sophisticated techniques such as transformers will be considered to further enhance performance in complex classification scenarios.

## References


[1] S. Qu, X. Li, and Z. Gan, "A Review of Hyperspectral Image Classification Based on Joint Spatial-spectral Features," in *Journal of Physics: Conference Series*, vol. 2203. IOP Publishing, 2022, p. 012040.

[2] M. Q. Alkhatib, M. Al-Saad, N. Aburaed, S. Almansoori, J. Zabalza, S. Marshall, and H. Al-Ahmad, "Tri-CNN: a three branch model for hyperspectral image classification," *Remote Sensing*, vol. 15, no. 2, p. 316, 2023.

[3] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2019.

[4] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, pp. 1–12, 2015.

[5] 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)*. IEEE, 2015, pp. 4959–4962.

[6] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Transactions on geoscience and remote sensing*, vol. 56, no. 8, pp. 4420–4434, 2018.

[7] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2485–2501, 2020.

[8] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[9] D. Cai and P. Zhang, "T$^3$SR: Texture Transfer Transformer for Remote Sensing Image Superresolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7346–7358, 2022.

[10] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[11] M. Q. Alkhatib and A. Jamali, "HSIFormer: An Efficient Vision Transformer Framework for Enhanced Hyperspectral Image Classification Using Local Window Attention," in *2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2024, pp. 1–5.

[12] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "MLP-mixer: An all-MLP architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.

[13] A. Jamali, S. K. Roy, B. Lu, A. Bhattacharya, and P. Ghamisi, "PolSAR-ConvMixer: A Channel and Spatial Mixing Convolutional Algorithm for PolSAR Data Classification," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 11 248–11 251.

[14] J. Zhang, P. Ma, T. Jiang, X. Zhao, W. Tan, J. Zhang, S. Zou, X. Huang, M. Grzegorzek, and C. Li, "SEM-RCNN: a squeeze-and-excitation-based mask region convolutional neural network for multi-class environmental microorganism detection," *Applied Sciences*, vol. 12, no. 19, p. 9902, 2022.

[15] A. Jamali, S. K. Roy, D. Hong, B. Lu, and P. Ghamisi, "How to learn more? Exploring Kolmogorov–Arnold networks for hyperspectral image classification," *Remote Sensing*, vol. 16, no. 21, p. 4015, 2024.

[16] S. K. Roy, A. Jamali, J. Chanussot, P. Ghamisi, E. Ghaderpour, and H. Shahabi, "SimPoolFormer: A two-stream vision transformer for hyperspectral image classification," *Remote Sensing Applications: Society and Environment*, p. 101478, 2025.

[17] M. Ahmad, M. H. F. Butt, A. M. Khan, M. Mazzara, S. Distefano, M. Usama, S. K. Roy, J. Chanussot, and D. Hong, "Spatial–spectral morphological mamba for hyperspectral image classification," *Neurocomputing*, vol. 636, p. 129995, 2025.

[18] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. Van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.