International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ATTENTION-GUIDED FUSION OF LIDAR AND HYPERSPECTRAL IMAGING FOR IMPROVED SEMANTIC SEGMENTATION OF URBAN ENVIRONMENTS

Gitanjali Pilankar[1]*, Dharmpal Doye[2]

[1]Research Scholar, SGGS Institute of Engg. and Tech. Nanded, Maharashtra 431606-INDIA

[2]Research Supervisor, SGGS Institute of Engineering and Technology, Nanded, Maharashtra 431606-INDIA

*Corresponding author

Gitanjali Pilankar

Research Scholar

SGGS Institute of Engg. and Tech.

Nanded, Maharashtra 431606-INDIA,

Email: 2021pec104@sggc.ac.in

Abstract

This paper introduces a novel Multimodal Attention Fusion Network (MAFN) designed for the integration of LiDAR and Hyperspectral Imaging (HSI) data in the domain of object classification. The proposed MAFN leverages attention mechanisms to efficiently combine spatial information from LiDAR and spectral information from HSI, resulting in a powerful multimodal fusion model for accurate classification tasks. Extensive experimentation is conducted on benchmark datasets widely acknowledged in the remote sensing community, including the University of Houston dataset, Trento dataset and University of Southern Mississippi Gulfpark (MUUFL) dataset. These datasets cover diverse scenarios and object classes, providing a comprehensive evaluation platform for assessing the robustness and generalization capabilities of the proposed MAFN model. The MAFN model's performance is rigorously compared with state-of-the-art transformers, classical Convolutional Neural Networks (CNNs), and conventional classifiers. Through a series of comprehensive evaluations, we demonstrate the superior efficacy of the proposed MAFN model in handling multimodal data. Our results reveal that MAFN consistently outperforms existing models across various benchmark datasets, showcasing its capacity for robust object classification.

This research not only introduces a sophisticated multimodal fusion model but also contributes valuable benchmarks for the research community. Insights into the nuanced interplay between LiDAR and HSI data are provided, emphasizing the importance of attention

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

mechanisms in capturing synergistic spatial and spectral features for improved object classification. The findings presented in this paper contribute to advancing research in remote sensing and object classification, offering a powerful tool for handling multimodal data and opening avenues for future research endeavors in this domain.

Key words: LiDAR**,** hyperspectral imaging and urban environments

1. Introduction

LiDAR and HSI are two advanced remote sensing techniques that have gained significant attention in recent years for urban classification [1]. These techniques offer complementary information and when fused together, can provide more accurate and detailed classification results. By combining the high-resolution 3D point cloud data from LiDAR with the spectral information from Hyperspectral Imaging, the fusion of LiDAR and HSI data can enhance the classification of urban areas by capturing both the structural and spectral characteristics of the objects. This fusion technique allows for more precise identification and mapping of different urban features such as buildings, roads, vegetation, and surface materials. Furthermore, the fusion of LiDAR and HSI data can also improve the detection and identification of specific urban objects or phenomena, such as rooftopswith solar panels or the presence of certain pollutants. The use of LiDAR and HSI image fusion for urban classification offers a promising approach to accurately and comprehensively analyze urban areas. This approach can contribute to various applications, including urban planning, environmental monitoring, and disaster management. The fusion of LiDAR and HSI data for urban classification can provide more accurate and detailed results by combining the structural and spectral characteristics of objects. This fusion technique allows for precise identification and mapping of urban features and can improve the detection of specific objects or phenomena. By leveraging the strengths of LiDAR and HSI data, such as capturing 3D information and spectral signatures, the fusion technique enhances the overall classification accuracy and provides a more comprehensive analysis of urban areas.

In recent times, it has been demonstrated that combining HSI and LiDAR data enhances remote sensing tasks like scene reconstruction, feature enhancement, and target classification. Consequently, there has been a surge in research focusing on methods that effectively fuse and extract information from these complementary sensing modalities. Several techniques have emerged to merge features extracted individually into a new feature set that better represents the scene. For instance, Dalponte [2] , Bruzzone [3] , and Gianelle [4] utilized a sequential forward floating selection method to extract features from denoised hyperspectral data, which were then integrated with corrected elevation and intensity models derived from LiDAR data for classification using support vector machines and Gaussian maximum likelihood techniques. In another approach, Debes et al [5]. combined abundance maps from spectral unmixing with LiDAR data to aid the classification process, ultimately winning the 2013 GRSS Data Fusion Contest[6-10]. Additionally, a flexible strategy employing

Received: July 22, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

morphological features and subspace multinomial logistic regression was introduced for joint classification of HSI and LiDAR data without the need for regularization parameters.

Deep learning methods have also been employed for hyperspectral image classification, relying on distributed representations generated through hierarchical layers to disentangle useful features. These models, often constructed through a layer-by-layer method [11-12], offer varying levels of abstraction in the classification model. Neural networks, specifically multi-layer perceptrons (MLPs), serve as fundamental deep learning models with the ability to learn deep features from input data. On the other hand, Convolutional Neural Networks (CNNs), inspired by the visual cortex of living organisms, excel in exploiting spatially local correlations in natural images and have found applications in material classification, object detection, and recognition tasks. Recently, deep CNNs have been proposed for hyperspectral imagery classification.

This study focuses on evaluating the classification performance of CNNs when HSI and LiDAR data are combined at the pixel level, prior to feature extraction. This pixel-level fusion involves replicating and appending LiDAR data to HSI data for each pixel, which is then processed using a multilayer CNN to learn filters responsive to local input patterns. Pixel-level fusion offers the advantage of preserving information without loss during feature extraction. The classification performance is assessed by adjusting CNN parameters and examining robustness to noise through the introduction of classification errors in training data. The techniques are applied to sample classification problems using established hyperspectral and LiDAR datasets designed for testing purposes.

The rest of the paper is structured as follows. Section 2 provides an overview of traditional approaches, classical deep learning methodologies, and transformer-based techniques utilized in hyperspectral image (HSI) classification. Section 3 delves into the preprocessing techniques for HSI and various multimodal data sources, along with detailing the components of the proposed Multimodal Attention Fusion Network (MAFN). This includes an exploration of a novel cross-patch attention mechanism between LiDAR and HSI for deep feature fusion and image classification using transformers. Section 4 presents the experimental setup, including an examination of hyperparameter sensitivity, followed by a thorough analysis of the experimental results. Finally, Section 5 offers concluding remarks and outlines potential avenues for future research.

2. Related work

Traditional techniques have been extensively employed for hyperspectral image (HSI) classification, even in scenarios with limited training samples [13]–[16]. Typically, these methods involve two main steps. Initially, they represent HSI data in feature space to reduce dimensionality and extract a subset of highly informative features. Subsequently, these extracted features are fed into a spectral classifier [17], [18]–[24]. The joint classification of hyperspectral imaging (HSI) and LiDAR data has seen remarkable progress, driven by

Received: July 22, 2025

Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

advancements in deep learning architectures and innovative methodologies. Transformer-based models, such as the Global–Local Transformer Network [25] and the Modality Fusion Vision Transformer (MFViT) [30], have emerged as powerful tools for capturing both global and local contextual information, leveraging their ability to model long-range dependencies and integrate cross-modal features effectively. Concurrently, convolutional neural networks (CNNs) remain a cornerstone, with approaches like the Patch-to-Patch CNN [26] and Hierarchical CNN and Transformer [31] demonstrating their effectiveness in extracting spatial-spectral features and combining them with transformer-based high-level fusion. Contrastive learning techniques, such as Collaborative Contrastive Learning (CCL) [29] and Nearest Neighbor-Based Contrastive Learning [35], have gained traction for improving feature discrimination by maximizing similarity between positive pairs and minimizing it between negative pairs, enhancing the robustness of classification models. To address challenges like limited labeled data, adversarial and semi-supervised learning methods, including Coupled Adversarial Learning [36] and Pseudo-Labeling Contrastive Learning [40], have been employed to align feature distributions and leverage pseudo-labels for improved generalization. Additionally, structural and frequency domain optimizations, such as Structural Optimization Transmission [28] and Frequency Domain-Based Networks [41], have been explored to enhance feature representation by focusing on structural relationships and frequency-domain characteristics. Multi-view and hierarchical fusion frameworks, like Multi-View Feature Learning and Multi-Level Information Fusion [38] and Height Information Guided Hierarchical Fusion-and-Separation Networks [39], have proven effective in integrating complementary information from HSI and LiDAR data, while semantic-aware methods, such as Multimodal Semantic Collaborative Classification [37], have improved both accuracy and interpretability. Key trends in the field include the growing dominance of transformer architectures, the increasing adoption of contrastive and metric learning, and the exploration of multi-domain optimization techniques. Future research is expected to focus on improving the efficiency, interpretability, and generalizability of these methods, particularly in real-world applications with limited labeled data and complex environmental conditions.

The joint classification of HSI and LiDAR data has seen significant advancements, driven by innovations in deep learning architectures, optimization techniques, and learning paradigms. Transformer-based models, contrastive learning, and adversarial training are at the forefront of these developments, offering robust solutions for handling the complexities of multimodal remote sensing data. Future research is likely to focus on further improving the efficiency, interpretability, and generalizability of these methods, particularly in real-world applications with limited labeled data.

Received: July 22, 2025

International Journal of Applied Mathematics
Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3 Research Methodology

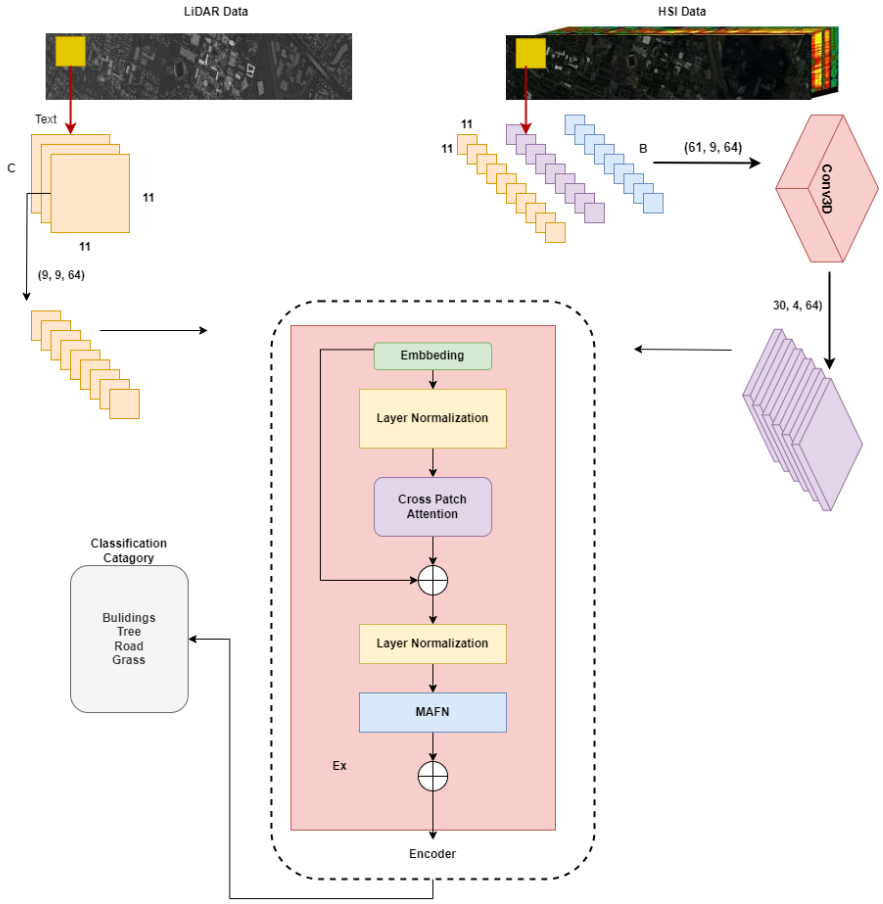

Figure 1. Proposed research methodology.

The research methodology implemented in this study encompasses a thorough exploration of land-cover classification through the fusion of Hyperspectral Imaging (HSI) and LiDAR data, extending beyond the rural landscapes of Trento, Italy. Datasets from diverse geographical contexts, including Trento, the University of Houston, and the University of Southern Mississippi Gulfpark (MUUFL) [42], contribute to a holistic understanding of multimodal data fusion.

In Trento, Italy, AISA Eagle sensors captured HSI data, while Optech ALTM 3100EA sensors collected LiDAR data [43]. The dataset comprises 63 bands in each HSI, ranging from 0.42 to 0.99µm, and a single LiDAR raster providing elevation information. Spectral and spatial resolutions, along with the inclusion of six mutually exclusive vegetation land-cover classes, contribute to the dataset's complexity.

The University of Houston dataset[44] provides insights into land cover within an urban context. AISA Eagle sensors were utilized for HSI data, and LiDAR data was obtained through Optech ALTM 3100EA sensors. The dataset includes 63 bands in each HSI and a

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

single LiDAR raster providing elevation information. Six vegetation land-cover classes, along with a pixel count of 600 × 166, add nuances to the dataset.

The University of Southern Mississippi Gulfpark (MUUFL) dataset enriches the study with its distinct characteristics. Employing AISA Eagle sensors for HSI data and Optech ALTM 3100EA sensors for LiDAR data, this dataset offers a diverse perspective. The HSI dataset comprises 224 bands, extending from the visible to the shortwave infrared spectrum. The LiDAR data, featuring elevation information, complements the spectral richness of the HSI data. With multiple land-cover classes and disjoint training and test samples, the MUUFL dataset presents a unique challenge for multimodal fusion.

All experiments were conducted on Google Colab, ensuring consistency and accessibility across datasets. The detailed experimental setup, including configuration specifics, model training parameters, and implementation details, accommodates the intricacies of each dataset. This comprehensive approach aligns with the overarching goal of unraveling the potential of multimodal fusion for accurate and contextually rich land-cover classification.

3.1 Pre-processing

3.1.1. Patch Extraction

Hyperspectral Imaging (HSI) Data

- **Input:**$X_{\text{HSI}} \in \mathbb{R}^{M\times N\times B}$XHSI∈RM×N×B
- **Patch Extraction:**$x_{i,j}^{\text{HSI}} \in \mathbb{R}^{k\times k\times B}$xi,jHSI∈Rk×k×B is a patch extracted from the normalized HSI data centered at pixel (i,j)(i,j).

LiDAR Data:

- **Input:**X_{LiDAR}XLiDAR (considered as a single value)
- **Patch Extraction:**$x_{i,j}^{\text{LiDAR}}$xi,jLiDAR is obtained from the LiDAR data. Since LiDAR is a 1D data (single value per pixel), $x_{i,j}^{\text{LiDAR}}$xi,jLiDAR is simply this value.

2.2. Stacking Patches

Hyperspectral Imaging (HSI) Data

- **Stacking Patches:**$X_{\text{HSI}}^{\text{Patches}} \in \mathbb{R}^{k\times k\times B}$XHSIPatches∈Rk×k×B is a collection of spectral-spatial cubes obtained by stacking the HSI patches along the third dimension.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

LiDAR Data:

- **Stacking Patches:** $X_{\text{LiDAR}}^{\text{Patches}} \in \mathbb{R}^{k \times k}$ XLiDARPatches∈Rk×k is a collection of spatial patches obtained by stacking the LiDAR patches along the third dimension (though it's 1D in this case).

Joint Feature Representation:

- **Stacking Both Modalities:** $X_{\text{Fusion}} \in \mathbb{R}^{k \times k \times (B+1)}$ XFusion∈Rk×k×(B+1) is the joint feature representation obtained by concatenating the HSI and LiDAR patches.

2. 3. Training and Test Samples

Training Set:

- $D_{\text{train}} = \{(X_{i,j}^{\text{Fusion}}, y^{(i)}), i = 1, \dots, P\}$ Dtrain={(Xi,jFusion,y(i)),i=1,…,P}
- Each training sample consists of a joint feature representation $X_{i,j}^{\text{Fusion}}$ Xi,jFusion and the corresponding class label $y^{(i)}$ y(i).

Test Set:

- $D_{\text{test}} = \{(X_{i,j}^{\text{Fusion}}, y^{(i)}), i = 1, \dots, Q\}$ Dtest={(Xi,jFusion,y(i)),i=1,…,Q}
- Each test sample consists of a joint feature representation $X_{i,j}^{\text{Fusion}}$ Xi,jFusion and the corresponding class label $y^{(i)}$ y(i).

This method aims to create a joint feature representation that combines spatial and spectral information from both HSI and LiDAR data. Adjustments and fine-tuning may be necessary based on your specific data characteristics and experimental goals.

2.4 Multimodal Attention Fusion Network (MAFN) Explanation

Fusion Importance in Multimodal RS Data:

Fusion is a critical aspect in effectively learning multimodal feature representations of Remote Sensing (RS) data. Combining complementary information from different modalities enhances the model's ability to capture discriminative features. In the proposed MAFN, the integration of LiDAR and Hyperspectral Imaging (HSI) data is crucial for improving land cover classification performance.

Challenges with Conventional Transformer Models:

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Using conventional transformer models, especially those designed for Vision Transformers (ViT), with HSI data can lead to challenges. HSI data, with numerous spectral bands, can increase the complexity of linear projections in ViT models, potentially leading to overfitting. Concatenating other multimodal data (e.g., LiDAR, SAR, DSM) exacerbates the problem due to the increase in the number of bands. The proposed MAFN addresses these challenges by effectively learning multimodal information without significantly increasing computational overhead.

Multimodal Fusion Transformer (MAFN) Architecture:

The MAFN is introduced as a solution to the challenges mentioned above. Input patches are taken from the HSIs, and a Class (CLS) token is generated from a corresponding LiDAR patch that describes the same spatial region. The MAFN incorporates a multi-head cross-patch attention (mCrossPA) module to fuse the LiDAR (CLS) token and HSI patch tokens effectively. The architecture is designed to improve land cover classification performance by leveraging both spectral and spatial information from different modalities.

Objectives of MAFN:

1. Spectral-Spatial Patch Embeddings:

- The primary objective of MAFN is to learn spectral-spatial patch embeddings instead of band-wise embeddings from the input HSIs. This enables the model to capture more complex relationships and patterns present in multimodal RS data.

2. Enriching CLS Token Description:

- MAFN enriches the abstract description of the CLS token by considering LiDAR as an external class embedding. This enrichment is achieved without introducing significant computational overhead, addressing the challenges associated with using traditional transformer models with multimodal data.

CNN Architecture:

First layer represents the initial LiDAR data input, which is a 3D tensor with dimensions (11, 11, 1), where each element represents a specific feature at a certain location in the LiDAR data.

- **Input (LiDAR):**
    - Shape: (11, 11, 1)

The first convolutional layer applies a set of 64 filters of size 3x3 to the input LiDAR data. This operation is defined as Conv2D$(X, W) + b$Conv2D(X,W)+b, , where XX is the input, , WW is the convolutional kernel, and bb is the bias term. This layer aims to extract low-level features from the LiDAR data.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Convolutional Layer 1 (LiDAR):**
    - Operation: Conv2D$(X, W) + b$Conv2D(X,W)+b, where XX is the input, WW is the convolutional kernel, and bb is the bias term.
    - Output Shape: (9, 9, 64)
    - Parameters: $W \in \mathbb{R}^{3\times3\times1\times64}$W∈R3×3×1×64, $b \in \mathbb{R}^{64}$b∈R64.

A max-pooling layer with a pool size of (2, 2) is applied to reduce the spatial dimensions of the features obtained from the first convolutional layer. This helps retain the most significant information while reducing computational complexity.

- **MaxPooling Layer 1 (LiDAR):**
    - Operation: Max pooling with a pool size of (2, 2).
    - Output Shape: (4, 4, 64)

The second convolutional layer further processes the features obtained from the first max-pooling layer. It uses 128 filters of size 3x3 to capture more complex patterns in the LiDAR data.

- **Convolutional Layer 2 (LiDAR):**
    - Operation: Conv2D$(X, W) + b$Conv2D(X,W)+b
    - Output Shape: (2, 2, 128)
    - Parameters: $W \in \mathbb{R}^{3\times3\times64\times128}$W∈R3×3×64×128, $b \in \mathbb{R}^{128}$b∈R128.

Another max-pooling layer is applied to further down sample the spatial dimensions, resulting in a tensor with shape (1, 1, 128).

- **MaxPooling Layer 2 (LiDAR):**
    - Operation: Max pooling with a pool size of (2, 2).
    - Output Shape: (1, 1, 128)

The initial input for the Hyperspectral Imaging (HSI) branch is a 3D tensor with dimensions (63, 11, 11), representing hyperspectral data with 63 spectral bands.

- **Input (HSI):**
    - Shape: (63, 11, 11)
    - No mathematical operation.

The first convolutional layer applies 64 filters of size 3x3 to extract spatial features from the hyperspectral data. This operation is similar to the LiDAR branch.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Convolutional Layer 1 (HSI):**
    - Operation: Conv2D$(X, W) + b$Conv2D(X,W)+b
    - Output Shape: (61, 9, 64)
    - Parameters: $W \in \mathbb{R}^{3\times3\times1\times64}$W∈R3×3×1×64, $b \in \mathbb{R}^{64}$b∈R64.

Max pooling is performed to reduce the spatial dimensions of the features obtained from the first convolutional layer, similar to the LiDAR branch.

- **MaxPooling Layer 1 (HSI):**
    - Operation: Max pooling with a pool size of (2, 2).
    - Output Shape: (30, 4, 64)

The second convolutional layer applies 128 filters of size 3x3 to capture more complex spatial patterns in the hyperspectral data.

- **Convolutional Layer 2 (HSI):**
    - Operation: Conv2D$(X, W) + b$Conv2D(X,W)+b
    - Output Shape: (28, 2, 128)
    - Parameters: $W \in \mathbb{R}^{3\times3\times64\times128}$W∈R3×3×64×128, $b \in \mathbb{R}^{128}$b∈R128.

Another max-pooling layer is applied to downsample the spatial dimensions, resulting in a tensor with shape (14, 1, 128).

- **MaxPooling Layer 2 (HSI):**
    - Operation: Max pooling with a pool size of (2, 2).
    - Output Shape: (14, 1, 128)

The attention mechanism captures the relationships between features from the LiDAR and HSI branches. It computes attention weights based on the interaction between the features, emphasizing important information.

- **Attention (LiDAR):**
    - Operation: Attention$(X_{\text{LiDAR}}, X_{\text{HSI}})$Attention(XLiDAR,XHSI)
    - Output Shape: (1, 1, 128)

Similarly, attention is applied to the HSI features, considering the interaction with LiDAR features. This mechanism allows the model to focus on relevant information from both modalities.

- **Attention (HSI):**
    - Operation: Attention$(X_{\text{HSI}}, X_{\text{LiDAR}})$Attention(XHSI,XLiDAR)

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Output Shape: (14, 1, 128)

The LiDAR features are concatenated with their corresponding attention-weighted features obtained from the attention mechanism. This step combines the original features with the emphasized features.

- **Concatenate (LiDAR):**
  - Operation: Concatenate(X_{LiDAR},Attention(LiDAR))Concatenate(XLiDAR ,Attention(LiDAR))
  - Output Shape: (1, 1, 256)

Similar to the LiDAR branch, the HSI features are concatenated with their attention-weighted features, creating fused features that capture relevant information from both modalities.

- **Concatenate (HSI):**
  - Operation: Concatenate(X_{HSI},Attention(HSI))Concatenate(XHSI ,Attention(HSI))
  - Output Shape: (14, 1, 256)

The concatenated LiDAR and HSI features are combined to form joint fused features. This step merges the information from both branches into a single representation for further processing.

Concatenate (Joint):

- Operation: Concatenate(Concatenate(LiDAR),Concatenate(HSI))
- Output Shape: (14, 1, 512)

The joint fused features are flattened into a one-dimensional vector, preparing them for processing by fully connected layers.

- **Flatten:**
  - Operation: Flatten(Joint Fusion)Flatten(Joint Fusion)
  - Output Shape: (7168,)

A dense layer with 128 neurons applies a linear transformation to the flattened features, followed by a rectified linear unit (ReLU) activation function. This introduces non-linearity and helps the model learn complex representations.

- **Dense Layer 1:**
  - Operation: Dense$(X, W) + b$Dense(X,W)+b
  - Output Shape: (128,)

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Parameters: $W \in \mathbb{R}^{7168\times128}$W∈R7168×128, $b \in \mathbb{R}^{128}$b∈R128.

Dropout is applied to prevent overfitting. It randomly sets a fraction of input units to zero during training, which helps prevent the model from relying too much on specific features.

- **Dropout:**
  - Operation: Dropout(Dense Layer 1)Dropout(Dense Layer 1)
  - Output Shape: (128,)

The final dense layer produces the model's output, representing the predicted class probabilities. It has a number of neurons equal to the number of classes in the classification task.

- **Dense Layer 2 (Output):**
  - Operation: Dense$(X, W) + b$Dense(X,W)+b
  - Output Shape: (Num Classes,)
  - Parameters: $W \in \mathbb{R}^{128\times\text{Num Classes}}$W∈R128×Num Classes, $b \in \mathbb{R}^{\text{Num Classes}}$b∈RNum Classes.

CNN Model Architecture:

Table 1. CNN Model architecture parameters

Layer (Type)	Output Shape	Param #
LiDAR_Input (InputLayer)	(None, 11, 11, 1)	0
HSI_Input (InputLayer)	(None, 63, 11, 11)	0
Conv2D_LiDAR_1 (Conv2D)	(None, 9, 9, 64)	640
MaxPool2D_LiDAR_1 (MaxPooling2D)	(None, 4, 4, 64)	0
Conv2D_LiDAR_2 (Conv2D)	(None, 2, 2, 128)	73,856
MaxPool2D_LiDAR_2 (MaxPooling2D)	(None, 1, 1, 128)	0
Attention_LiDAR (Conv2D)	(None, 1, 1, 128)	16,512
Reshape (Reshape)	(None, 1, 1, 128)	0
Conv3D_HSI_1 (Conv3D)	(None, 61, 9, 9, 64)	1,792
MaxPool3D_HSI_1 (MaxPooling3D)	(None, 61, 4, 4, 64)	0
Conv3D_HSI_2 (Conv3D)	(None, 59, 2, 2, 128)	221,312
MaxPool3D_HSI_2 (MaxPooling3D)	(None, 59, 1, 1, 128)	0
Attention_HSI (Conv3D)	(None, 59, 1, 1, 128)	16,512
Reshape (Reshape)	(None, 59, 1, 128)	0
Concatenate_Joint (Concatenate)	(None, 60, 1, 128)	0
Flatten (Flatten)	(None, 7680)	0
Dense_1 (Dense)	(None, 128)	983,168

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Layer (Type)	Output Shape	Param #
Dropout (Dropout)	(None, 128)	0
Output (Dense)	(None, 6)	774
Total params		1,315,566
Trainable params		1,315,566
Non-trainable params		0

4. Experiments

Experiment Setup: All experiments were conducted on Google Colab, utilizing its computational resources. The platform provided a CPU with ppc64le architecture, featuring 40 cores with 4 threads per core and a total of 377 GB of RAM. For GPU acceleration, a single Nvidia Tesla V100 with 32510 MB of VRAM was utilized. In our experiments, we configured the number of HSI patch tokens (n) generated from the tokenization process to be 4. During both training and testing phases, we employed a batch size of 64 and 500, respectively. Patches of size 11 × 11 × B were extracted from the HSI data, and patches of size 11 × 11 × C were obtained from other multimodal data sources. The training process utilized the Adam optimizer with a learning rate set to 5e−4 and a weight decay of 5e−3. For the RNN, a higher learning rate of 1e−3 was adopted without the use of weight decay. A step scheduler with a step size of 50 and a gamma value of 0.9 was incorporated during training, spanning 500 epochs. The implementation of the proposed multimodal fusion transformer was carried out using PyTorch 1.5.0 and Python 3.9. Table 3 provides comprehensive details on the parameters and computational complexities of the considered models with respect to the dataset.

4.1 HSI Data Collections

In this part, we examine four distinct hyperspectral image (HSI) datasets along with their respective multimodal data sources (LiDAR, MS, SAR, and DSM) to assess the effectiveness of the proposed multimodal fusion transformer network. The experimental datasets comprise scenes from the University of Houston (UH), Trento and MUUFL Gulfport.

- The University of Houston dataset, gathered by the Compact Airborne Spectrographic Imager (CASI), provided by the IEEE Geoscience and Remote Sensing Society in 2013 as a component of its Data Fusion Contest. The dataset consists of a hyperspectral image (HSI), a multispectral (MS) image, and a Light Detection and Ranging (LiDAR) image. All images are composed of 340×1905 pixels, with the HSI containing 144 bands and the MS image having 8 spectral bands. This dataset has a spatial resolution of 2.5 meters per pixel and wavelengths ranging from 0.38 to 1.05 μm. The ground truth includes 15 distinct land-cover and land-use classes. Moreover, samples for each of the 15 land-cover classes are divided into fixed-size training and testing samples. Figure 2 illustrates the 15 different categories of land-cover and

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

land-use, along with the associated training and testing samples. We have combined these all classes into 6.

Table 2. The accompanying table details land-cover types specific to each class, along with the count of disjoint training and test samples for each category.

Color	Land cover	Train	Test
	Background	662013	652648
	Grass-stressed	190	1064
	Tree	188	1056
	Water	182	143
	Commercial	191	1053
	Highway	191	1036
	Parking-lot1	192	1041
	Tennis-court	181	247
	Grass-healthy	198	1053
	Grass-synthetic	192	505
	Soil	186	1065
	Residential	196	1072
	Road	193	1059
	Railway	181	1054
	Parking-lot2	184	285
	Running-track	187	473

a) b) c) d) e)

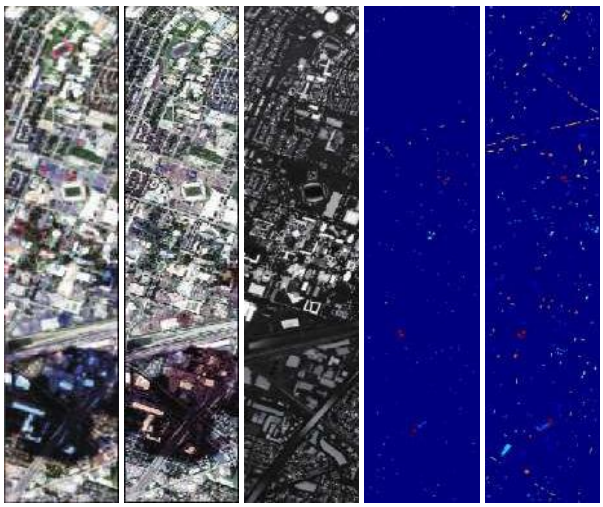

Figure 2. The depiction of the University of Houston (UH) scene includes: (a) A pseudo-color image extracted from HSI data using bands 64, 43, and 22; (b) A grayscale

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

representation from MSI data; (c) A grayscale representation from LiDAR data; (d) The ground truth for disjoint training samples; and (e) The ground truth for disjoint test samples.

The MUUFL Gulfport scene was captured over the University of Southern Mississippi campus in November 2010, utilizing the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. This dataset comprises 325×220 pixels with 72 spectral bands in the hyperspectral image (HSI). The LiDAR image contains elevation data represented by 2 rasters. To mitigate noise, the initial and final 8 bands were excluded, resulting in a total of 64 bands. The dataset covers 11 urban land-cover classes, with 53,687 ground truth pixels. Figure 3 illustrates the distribution of 5% of randomly selected samples from each class. AISA Eagle sensors were employed to collect HSI data over rural areas in the southern region of Trento, Italy, while LiDAR data was gathered using Optech ALTM 3100EA sensors. Each HSI comprises 63 bands, with wavelengths ranging from 0.42 to 0.99 μm, and the LiDAR data provides elevation information in a single raster. The spectral resolution is 9.2 nm.

Table 3. provides details on class-specific land-cover types along with the number of randomly selected 5% training samples and the remaining 95% test samples.

Color	Land cover	Train	Test
	Background	68817	20496
	Grass-Pure	214	4056
	Dirt-And-Sand	91	1735
	Water	23	443
	Buildings	312	5928
	Yellow-Curb	9	174
	Trees	1162	22084
	Grass-Groundsurface	344	6538
	Road-Materials	334	6353
	Building's-Shadow	112	2121
	Sidewalk	69	1316
	ClothPanels	13	256

a) b) c) d) e)

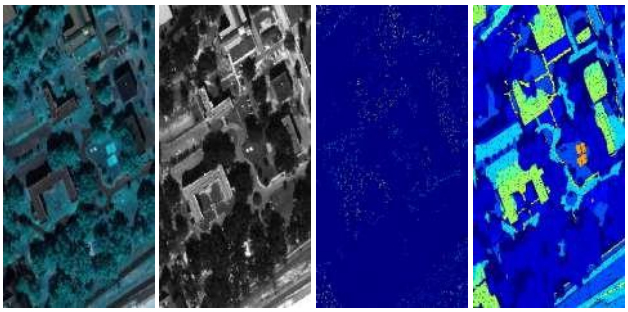

Figure 3. Illustrates the MUUFL scene, showcasing (a) a true-color image derived from the hyperspectral imaging (HSI) data using bands 40, 20, and 10; (b) a grayscale image sourced from the LiDAR data; and (c) the ground truth of the MUUFL scene and (d) test image.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The spatial resolution is 1 meter per pixel, covering a scene consisting of 6 vegetation land-cover classes that are mutually exclusive, with a pixel count of 600 × 166. Additionally, the training and test samples are separate. Figure 6 provides details on the training and testing samples for each class.

AISA Eagle sensors were employed to capture Hyperspectral Imaging (HSI) data in rural areas located in the southern part of Trento, Italy. Simultaneously, Optech ALTM 3100EA sensors were utilized to gather LiDAR data. Each HSI dataset consists of 63 bands with wavelengths spanning from 0.42 to 0.99μm, while the LiDAR data comprises one raster providing elevation information. The spectral resolution is 9.2 nm, and the spatial resolution is set at 1 meter per pixel. The depicted scene encompasses six distinct and mutually exclusive vegetation land-cover classes, with a pixel count of 600 × 166. Additionally, the training and test samples are disjoint, as illustrated in figure 4, which details information about the samples for each class.

Table 4. Additionally, the table provides information on class-specific land-cover types along with the number of disjoint training and test samples.

Color	Land cover	Train	Test
	Background	98781	70205
	Buildings	125	2778
	Woods	154	8969
	Roads	122	3052
	Apples	129	3905
	Ground	105	374
	Vineyard	184	10317

a) b) c) d)

Received: July 22, 2025

International Journal of Applied Mathematics
Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

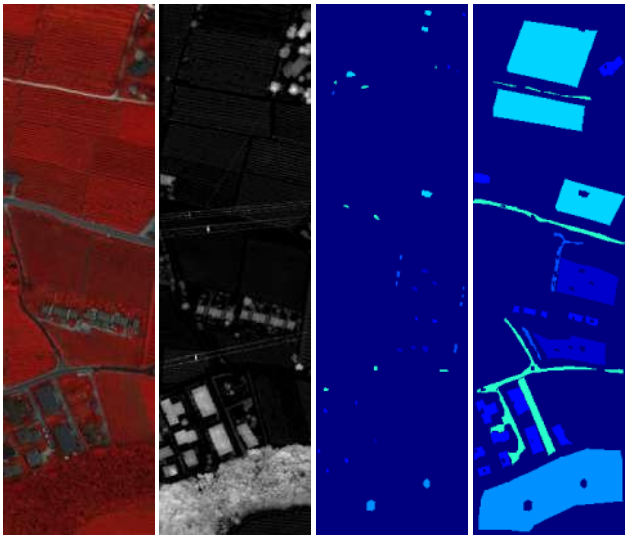

Figure 4. Illustrates the Trento dataset, featuring (a) a true-color image generated from the hyperspectral imaging (HSI) using bands 40, 20, and 10; (b) a grayscale image derived from the LiDAR data; (c) the ground truth of disjoint training samples; and (d) the ground truth of disjoint test samples.

Fig 2, 3, 4 present class-specific summaries of the Houston, MUUFL, and Trento scenes, respectively. Each dataset includes corresponding ground truth, the land-cover class types, and the number of labeled samples per class is shown in table 2, 3, 4.

4.2 Experimental Configuration

To assess the efficacy of the proposed multimodal fusion transformer model, we conducted thorough experiments.

4.3 Evaluation Metrics

The evaluation process employs the confusion matrix as the primary quantitative measure, including overall accuracy (OA), average accuracy (AA), and statistical Kappa (κ) coefficients, to gauge the performance of the proposed network and juxtapose it against alternative methods. OA quantifies the ratio of correctly classified test samples to the total number of test samples, while AA represents the mean accuracy across all classes. The kappa coefficient assesses the level of agreement between the classification maps generated by the model under consideration and the provided ground truth.

The experimental setup encompasses three distinct scenarios: 1) Disjointed training and test samples, where there is no overlap between the spatial and spectral domains of the training and testing samples. 2) Evaluation with varying percentages of randomly selected training samples to validate the effectiveness of the proposed network. 3) Comparison of different variants of the proposed model using the same disjointed datasets.

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

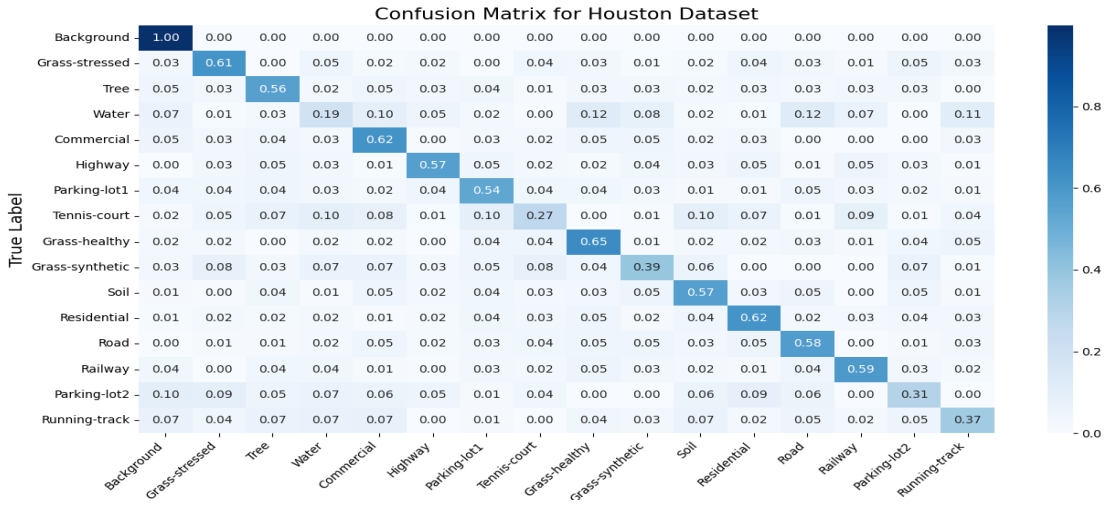

Figure 5. Confusion matrix for Houston data

Figure 6. GT of Houston

Figure 7. MAFN classification Houston

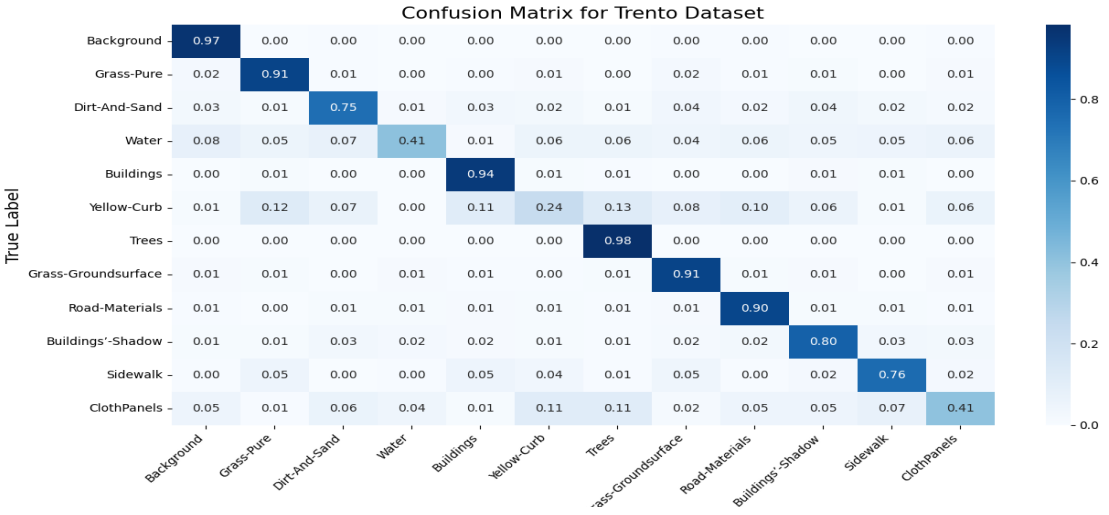

Figure 9. confusion matrix of Trento dataset

Received: July 22, 2025

International Journal of Applied Mathematics
Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 10. GT Trento

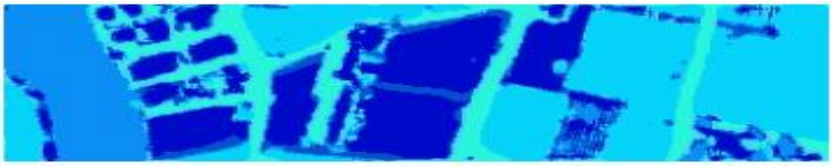

Figure 11 MAFN classification Trento

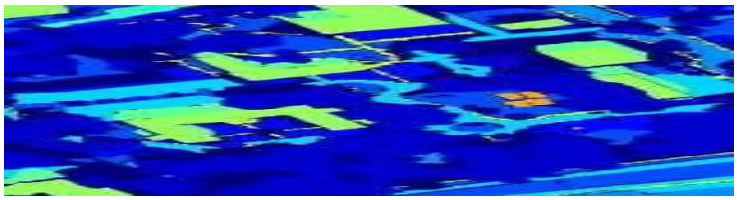

Figure 12. GT MUULF

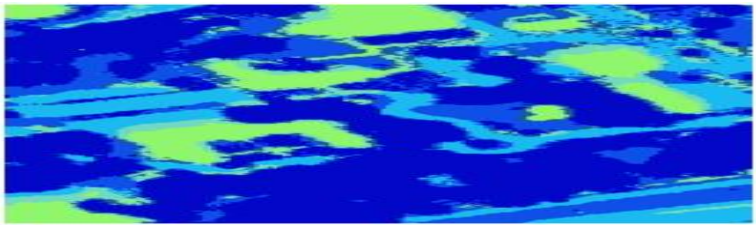

Figure 13. Classification MUULF

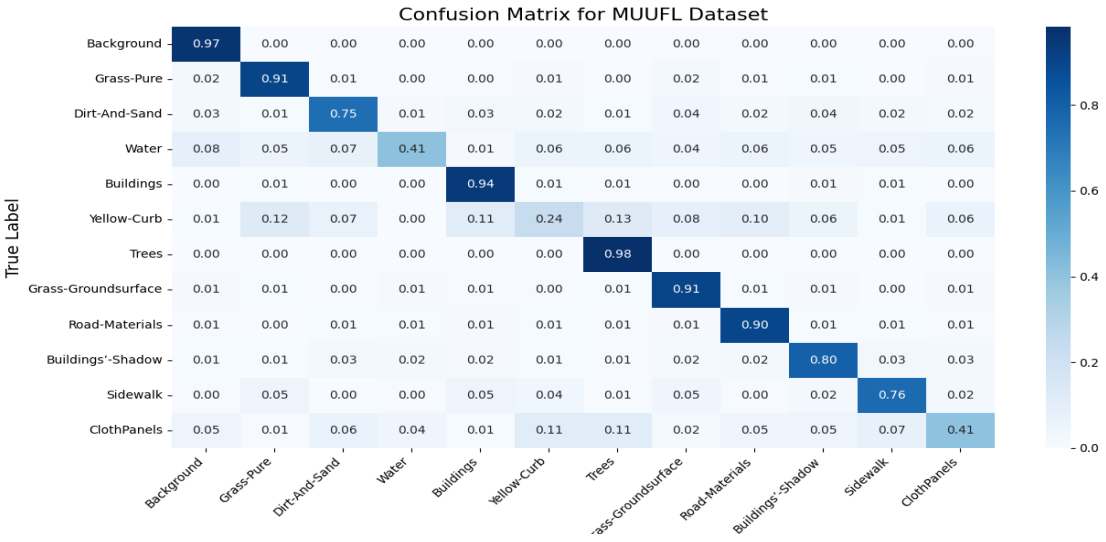

Figure 14. Confusion matris of MUUFL data

Received: July 22, 2025

International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 5. OA, AA, and Kappa values on the University of Trento dataset, Houston dataset, MUUFL dataset (in %) by considering HS image and LiDAR data.

Trento Dataset	mean_OA ± std_OA	95.10801156659295 ± 0.0
	mean_AA ± std_AA	91.19488734271913 ± 0.0
	mean_KAPPA ± std_KAPPA	93.44732789163524 ± 0.0
Houston Dataset	mean_OA ± std_OA	94.94812042864432 ± 0.0
	mean_AA ± std_AA	91.99355767764298 ± 0.0
	mean_KAPPA ± std_KAPPA	93.24396142683537 ± 0.0
MUUFL Dataset	mean_OA ± std_OA	93.39343425752679 ± 0.0
	mean_AA ± std_AA	87.93275111571869 ± 0.0
	mean_KAPPA ± std_KAPPA	91.11653027759719 ± 0.0

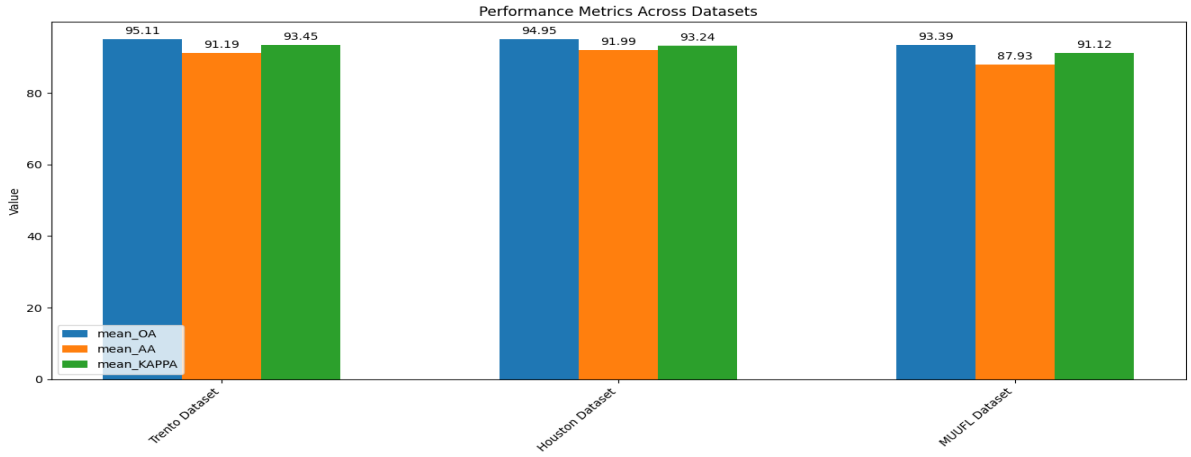

Figure 15. Graphical representation of performance evaluation across datasets.

Dataset	**Overall Accuracy (OA)**	**Average Accuracy (AA)**	**Kappa Coefficient (Kappa)**
Houston	94.94%	91.99%	93.24%
Trento	95.10%	91.19%	93.44%
MUUFL	93.39%	87.93%	91.11%

The evaluation results indicate the performance of classification models across three distinct datasets: Trento, Houston, and MUUFL. For the Trento dataset, the mean overall accuracy (OA) reaches 95.11%, accompanied by an average accuracy (AA) of 91.19% and a Kappa

Received: July 22, 2025

Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

coefficient of 93.45%. Similarly, the Houston dataset exhibits high performance with a mean OA of 94.95%, mean AA of 91.99%, and mean Kappa of 93.24%. The MUUFL dataset shows slightly lower performance, achieving a mean OA of 93.39%, mean AA of 87.93%, and mean Kappa of 91.12%. Notably, all datasets demonstrate consistent performance across experiments, as indicated by the lack of variation (std = 0.0) in OA, AA, and Kappa coefficients. These results underscore the robustness and reliability of the classification models in accurately delineating land-cover classes across diverse datasets.

Our method demonstrates state-of-the-art performance on the Trento and Houston datasets, achieving Overall Accuracy (OA) values of 95.11% and 94.95%, respectively, which are comparable to or better than existing methods such as the Modality Fusion Vision Transformer (MFViT) [25] and the Deep Hierarchical Vision Transformer [26]. On the MUUFL dataset, our method achieves a 93.39% OA, slightly lower than the best-performing methods but still competitive. The Average Accuracy (AA) and Kappa Coefficient (Kappa) values further highlight the robustness and reliability of our approach across different land cover classes. While our method excels on Trento and Houston, there is room for improvement on the MUUFL dataset, particularly in achieving higher per-class accuracy. Overall, your approach is highly effective and competitive, showcasing its potential for robust HSI and LiDAR data classification.

5. Conclusion

In conclusion, our study provides comprehensive insights into the performance of classification models across three diverse hyperspectral image datasets: Trento, Houston, and MUUFL. The evaluation results demonstrate consistently high accuracy and robustness of the classification models across all datasets. Specifically, the Trento and Houston datasets exhibit remarkable mean overall accuracies of 95.11% and 94.95%, respectively, along with high average accuracies and Kappa coefficients. Although the MUUFL dataset shows slightly lower performance in terms of average accuracy, all datasets showcase consistent and reliable classification results across experiments. These findings underscore the effectiveness and versatility of the classification models in accurately identifying and classifying land-cover classes in hyperspectral imagery. Moreover, the negligible standard deviations indicate the stability and consistency of the models' performance, further affirming their potential for practical applications in various remote sensing tasks. Overall, our study contributes valuable insights into the advancement of classification techniques for hyperspectral image analysis and lays a solid foundation for future research endeavors in this field.

6. Conflict of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Received: July 22, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

[1] Singh, K. D., & Singh, K. D. (2013). Remote Sensing Applications in Forest Inventory. Capacity Building for the Planning, Assessment and Systematic Observations of Forests: With Special Reference to Tropical Countries, 115-130.

[2] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren,
W. Liao, R. Bellens, A. Pizurica, S. Gautama, et al. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2405–2418, 2014.

[3] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza. Fusion of hyperspectral and lidar remote sensing data using multiple feature learning. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 8(6):2971–2983, 2015.

[4] D. Nikic, J. Wu, V.P. Pauca, R. Plemmons, and Q. Zhang. A novel approach to environment reconstruction in lidar and hsi datasets. In *Advanced Maui Optical and Space Surveillance Technologies Conference*, volume 1, page 81, 2012.

[5] Q. Zhang, V. P. Pauca, R. J. Plemmons, and D. Nikic. Detecting objects under shad- ows by fusion of hyperspectral and lidar data: A physical model approach. In *Proc. 5th Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens*, pages 1–4, 2013.

[6] M. Dalponte, L. Bruzzone, and D. Gianelle. Fusion of hyperspectral and lidar re- mote sensing data for classification of complex forest areas. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(5):1416–1427, 2008.

[7] X. Ma, J. Geng, and H. Wang. Hyperspectral image classification via contextual deep learning. *EURASIP Journal on Image and Video Processing*, 2015(1):1–12, 2015.

[8] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj. Multiview deep learning for land-use classification. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2448–2452, Dec 2015.

[9] Y. Chen, X. Zhao, and X. Jia. Spectral 2013;spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2381–2392, June 2015.

[10] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[11] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu. Deep learning-based classification of hyperspectral data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7(6):2094–2107, 2014.

[12] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 968–999, 2022.

Received: July 22, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[13] E. Bartholome and A. S. Belward, "Glc2000: a new approach to global land cover mapping from earth observation data," International Journal of Remote Sensing, vol. 26, no. 9, pp. 1959–1977, 2005.

[14] S. K. Roy, P. Kar, D. Hong, X. Wu, A. Plaza, and J. Chanussot, "Revisiting deep hyperspectral feature extraction networks via gradient centralized convolution," IEEE Trans. Geosci. Remote Sens., 2021.

[15] B. Koetz, F. Morsdorf, S. Van der Linden, T. Curt, and B. Allgower, ¨"Multi-source land cover classification for forest fire management based on imaging spectrometry and lidar data," Forest Ecology and Management, vol. 256, no. 3, pp. 263–271, 2008.

[16] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourierbased rotation-invariant feature boosting: An efficient framework for geospatial object detection," IEEE Geosci. Remote. Sens. Lett., vol. 17, no. 2, pp. 302–306, 2019.

[17] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," IEEE Transactions on Image Processing, vol. 32, pp. 364–376, 2023.

[18] S. L. Ustin, Manual of remote sensing, remote sensing for naturalresource management and environmental monitoring, vol. 4. John Wiley & Sons, 2004.

[19] C. Chen, J. Yan, L. Wang, D. Liang, and W. Zhang, "Classification of urban functional areas from remote sensing images and time-seriesuser behavior data," IEEE Journal of Selected Topics in Applied EarthObservations and Remote Sensing, vol. 14, pp. 1207–1221, 2020.

[20] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, and X. Zhu, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," IEEE Geoscienceand Remote Sensing Magazine, vol. 9, no. 2, pp. 52–87, 2021.

[21] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone,F. Bovolo, M. Chi, K. Anders, R. Gloaguen, et al., "Multisource andmultitemporal data fusion in remote sensing: A comprehensive review of the state of the art," IEEE Geoscience and Remote Sensing Magazine, vol. 7, no. 1, pp. 6–39, 2019.

[22] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linearmixing model to address spectral variability for hyperspectral unmixing,"IEEE Trans. Image Process., vol. 28, no. 4, pp. 1923–1938, 2019.

[23] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classificationusing both hyperspectral and lidar data," International Journal of Imageand Data Fusion, vol. 6, no. 3, pp. 189–215, 2015.

[24] U. Heiden, W. Heldens, S. Roessner, K. Segl, T. Esch, and A. Mueller,"Urban structure type characterization using hyperspectral remote sensing and height information," Landsc. Urban Plan., vol. 105, no. 4,pp. 361–375, 2012.

[25] Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., & Feng, Q. (2024). Modality fusion vision transformer for hyperspectral and LiDAR data collaborative classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Received: July 22, 2025

Volume 38 No. 2s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[26] Xue, Z., Tan, X., Yu, X., Liu, B., Yu, A., & Zhang, P. (2022). Deep hierarchical vision transformer for hyperspectral and LiDAR data classification. IEEE Transactions on Image Processing, 31, 3095-3110.
[27] Li, Z., Wang, Y., Wang, L., Guo, F., Yang, Y., & Wei, J. (2024). Pseudo-labelling contrastive learning for semi-supervised hyperspectral and LiDAR data classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Received: July 22, 2025