IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 361~373
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp361-373

Multi-scale features assisted knowledge distillation vision transformer for land cover segmentation and classification

Sujata Arjun Gaikwad, Vijaya Musande
Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, MGM University,
Chhatrapati Sambhaji Nagar, India

Article Info

Article history:

Received May 26, 2025
Revised Nov 17, 2025
Accepted Jan 10, 2026

Keywords:

Improved guided filter
Knowledge distillation
Multi scale segmentation
Pyramidal module
Remote sensing image
Vision transformer

ABSTRACT

The most significant problem in remote sensing interpretation is semantic segmentation, which attempts to give each pixel in the image a particular class. This research work follows the various steps, such as pre-processing, segmentation, and classification. Initially, high spatial resolution remote sensing images (RSI) are collected from the open-source dataset. In the pre-processing stage, an improved guided filter (Imp-GF) is used to remove various noises from images. Next, the segmentation is done by using a knowledge distillation-based vision transformer approach integrated with an atrous spatial multi-scale pyramidal module (KD-MuViTPy). Based on the segmented image, land cover classes such as vegetation, urban areas, forest, water bodies, and roads are classified. The proposed method outperformed the Bhuvan satellite dataset, achieving better accuracy, precision, recall, F1-score, Dice score, intersection over union (IoU), and Kappa score at values of 98.01%, 98.99%, 97.49%, 98.23%, 98.23%, 96.55%, and 95.91%, respectively.

Corresponding Author:

Sujata Arjun Gaikwad
Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, MGM University
Chhatrapati Sambhaji Nagar, Maharashtra, India
Email: sujatagaikwad414@gmail.com

1. INTRODUCTION

Land cover detection and classification are critical in economic assessment, resource management, and crop area analysis. This evaluation study greatly improved homeland security and national economic stability [1], [2]. Remote sensing is a powerful technology for earth land observation that employs sensors on satellites or airplanes to capture images of the planet's surface without requiring direct physical contact [3], [4]. One significant area of remote sensing is optical remote sensing, which has been used for a variety of purposes, such as super-resolution mapping of land cover and object detection [5]. Using the vast amount of remote sensing photos, researchers have focused on automatically classifying land cover using satellite pictures [6].

The examination of land coverage has become more visible in recent years due to the development of artificial intelligence (AI) [7]. AI-powered machine learning (ML) as well as deep learning (DL) models are employed to identify changes in the land [8]. Because of its deep layers, DL models outperform ML models in detection. ML models use statistical algorithms to detect patterns in data, but they may struggle with complicated, multidimensional datasets such as satellite images [9].

DL models, which use multi-layered neural networks, provide greater detection skills. These models can automatically extract detailed features from raw data using deeper layers, catching small changes more effectively [10]. DL's capacity to learn hierarchical representations allows it to perform well in tasks like image segmentation, classification, as well as change recognition [11]. As a result, DL-based techniques have emerged as the preferred method for detecting land cover changes, as they offer higher accuracy and resilience than standard ML models [12]. This makes them especially valuable in urban preparation, deforestation monitoring, as well as agricultural management [13].

Initially, land cover classifications are done using multi-scale convolutional neural networks (CNNs). However, issues like inadequate depth, constrained receptive fields, and the incapacity to accurately capture long-range interdependence in intricate situations hindered their performance [14]. They thus have trouble managing irregular patterns in large-scale remote sensing data and fine-grained classification. Furthermore, the segmentation and classification of overlapping or unclear land cover types were less accurately accomplished by the early multi-scale CNNs due to their absence of strong feature fusion processes [15].

Although CNN-based multi-scale models have been widely used for land cover segmentation, they still face some key challenges, such as limited receptive fields, ambiguous land cover types, insufficient depth to capture long-range dependencies, and weak feature fusion. Many existing models also suffered from high computational cost, memory inefficiency, and feature redundancy, which reduce the scalability for high-resolution satellite imagery. To bridge these gaps, this study introduces a novel knowledge distillation-based vision transformer approach integrated with multi-scale pyramidal module (KD-MuViTPy) model. The proposed model integrates noise reduction via an improved guided filter (Imp-GF), response-based knowledge distillation to transfer robust representation from teacher to student networks, and dynamic multi-scale pyramidal pooling to capture both global and fine-grained spatial features. This combination enhances segmentation accuracy, reduces computational overhead, and delivers state-of-the-art performance for land cover classification. The major contribution of this research work is given below: noise from the input images is removed by using an Imp-GF, which helps to enhance image quality. To segment the pre-processed images and identify land classes using atrous spatial KD-MuViTPy.

2. RELATED WORKS

Various previous studies have recommended a number of methods based on multi-scale and DL to improve the accuracy of extracting features and grouping images generated from remote sensing, such as:

- Cardama *et al.* [16] developed a consensus multi-scale binary alteration recognition approach for object-based feature extraction. Multiple detectors based on different segmentation approaches were utilized at different scales to exploit the very high resolution (VHR) pictures' high spatial resolution and capture changes at different granularity levels. The change vector analysis-segment anything model (CVA-SAM) was used on the segment level rather than the pixel level.
- Wang *et al.* [17] developed the parallel swin (P-Swin) transformer network, a transformer network based on parallel windows. The P-swin transformer block, which comprises the feed forward network (FFN) and window-based self-attention interaction (WSAI), is a critical component of P-Swin. WSAI can determine the link between windows as well as the relationship within windows. As a result, it increases the network's ability to collect feature context data.
- Ma *et al.* [18] developed a new CNN pixel-by-pixel classification approach with smaller sizes. The approach addresses the problem of inadequate multi-scale learning for classification by employing multi-scale networks to remove multi-scale contextual data at a fine-grained level.
- Jia *et al.* [19] developed a multi-attention semantic segmentation network for remote sensing images (RSI). The baseline model is U-Net, and the backbone network's capacity to remove fine-grained structures is improved by incorporating an organized attention-based residual network into the encoder. To increase network information extraction, the decoder's traditional up-sampling operator was replaced by a content-aware rearrangement module.
- Xu *et al.* [20] developed a two-branched supervised semantic segmentation system. A new symmetric attention module with enhanced strip pooling was developed. The multiple long accessible fields allow for the acquisition of more anisotropic contextual information and better visibility of irregular objects.
- Zhang *et al.* [21] developed the extended topology preserving segmentation (ETPS) model-based multi-scale as well as multi-feature method that employs convolutional neural segmentation. The suggested model splits the images into superpixels employing the ETPS method. However, the multi-resolution segmentation model retrieved features and mapped them to superpixels for multi-representation.
- Shi *et al.* [22] developed a land cover classification system based on multi-spectral light detection and ranging (LiDAR) with spatial multi-scale as well as spectral feature selection. Initially, k-nearest neighborhood was employed to choose neighborhood points from multi-spectral LiDAR data. Additionally, spatial as well as spectral information was retrieved from the multi-scale neighborhood.
- Li *et al.* [23] developed a multi-scale fully convolutional network (MSFCN) with a multi-scale convolutional kernel as well as a channel attention block (CAB) as well as a global pooling module for land cover categorization.

- Martins *et al*. [24] suggested the multi-scale object-based convolutional neural networks (OCNN) model. However, the suggested model classifies the large-scale land area with a 145,750 Km^2 resolution. Moreover, the suggested method consists of three phases: image segmentation, skeleton based algorithm for object analysis, and the application of multiple CNNs for ultimate classification.
- Chen *et al*. [25] developed the multi-level feature aggregation network (MFANet) model. The suggested approach had improved two phases, such as deep feature extraction as well as up-sampling feature fusion.

The literature search identified several constraints, including the model's high computational complexity, need for more memory to process, and high computational cost; also, the model has scaling concerns [16]–[18]. In addition, conventional models require more time for training. Furthermore, increased feature redundancy renders the model inefficient [19]–[23]. Regardless of how many layers are present in the network, network difficulties may arise, and future fusion approaches may enhance complexity [24], [25]. These are the overall limitations mentioned in the existing survey.

To address these current challenges, a unique knowledge distillation-based vision transformer approach was presented, which is coupled with an atrous spatial multi-scale pyramidal module to divide and classify land cover effectively. This study found that a knowledge distillation-based vision transformer model enhances accuracy while lowering computing costs. In addition, the unique module includes a multi-scale pyramidal module layer that processes and extracts the multi-scale aspects of the input image while also assisting in the acquisition of high-level contextual information. Furthermore, the multi-scale pyramidal module improves computational efficiency and is appropriate for real-time applications. The KD-based vision transformer consistently improves classification performance to boost semantic segmentation transformers, resulting in higher accuracy, lower computational cost, and improved pre-training needs.

3. PROPOSED METHOD

This study presented a multi-scale feature assisted semantic segmentation method for land segmentation and classification. The working flow of research is represented in Figure 1. The architecture shows the flow of the suggested methodology; initially, data were acquired from the Bhuvan satellite image dataset. The pre-processing stages use the Imp-GF method, which effectively eliminates the undesired noise and artifacts from the image. After pre-processing is done, segmentation is performed by using the KD-MuViTPy approach. Therefore, the segmented images are used to classify the land cover into numerous kinds, such as roads, urban areas, forests, water bodies, as well as vegetation.

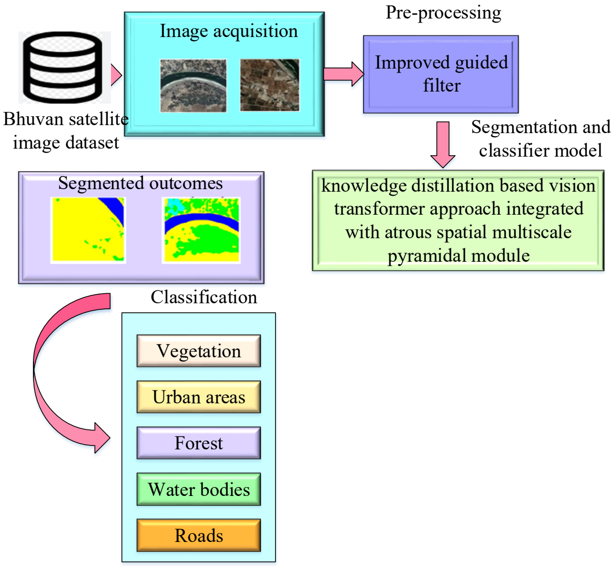


Image acquisition
Pre-processing
Improved guided filter
Bhuvan satellite image dataset
Segmentation and classifier model
Segmented outcomes
knowledge distillation based vision transformer approach integrated with atrous spatial multiscale pyramidal module
Classification
Vegetation
Urban areas
Forest
Water bodies
Roads

Figure 1. Architecture of the proposed methodology

3.1. Pre-processing by using an improved guided filter

In this work, images are obtained from a satellite image dataset; nevertheless, the collected images contain some artifacts and undesired noises such as speckle noise and salt and pepper noise. These noises were removed using the Imp-GF method. However, this filter can be used to eliminate the noise from the sub-band images [26]. The noise eliminated from the image by using Imp-GF is formulated in (1).

$$p_i = b_l I_l + a_l, \forall_i \in \omega_l \tag{1}$$

Here ω_l denotes the window and linear variables are represented as b_l and a_l. Next, the (2) removes the unwanted noise to determine the linear variables.

$$p_i = \vartheta_1 - m_l \tag{2}$$

Here, the input image is specified as ϑ_l, and a noise element is represented asm_l, however, minimizing the difference between ϑ_l and p_i is formulated in (3).

$$A(b_l, a_l) = \sum_{i\in\omega_l}((b_l I_l + a_l - \vartheta_l)^2 + \alpha b_l^2) \tag{3}$$

Where α represent the normalization variable. Imp-GF introduced the Jaccard similarity, which is used for identifying images near the edges, and is formulated in (4) and (5). Here Jaccard similarity is represented as E_{js}, thus, the process of Imp-GF effectively eliminates the noise and enhances the edge and contrast accuracy of the degraded image.

$$A(b_l, a_l) = \sum_{i\in\omega_l}\left((b_l I_l + a_l - \vartheta_l)^2 + \alpha b_l^2\right) + E_{js} \tag{4}$$

$$E_{js} = 1 - \frac{|a\cap b|}{|a\cup b|} \tag{5}$$

3.2. Segmentation and land cover classification by KD-MuViTPy

This study introduced a KD-MuViTPy model for picture segmentation and classification. This technique addresses the challenges that exist in the usual approach. In general, distillation methodologies, knowledge type, and teacher student structure all have an impact on student model learning. Knowledge can be categorized as feature, response, or relationship-based, employing the teacher model's knowledge sets. For picture classification, response-based knowledge serves as a classification problem, often known as hard labels. In classification, real labels are utilized as hard labels, and the probability dissemination of the model is fed into the SoftMax function. Furthermore, the result of the SoftMax function is immediately matched with the hard label to establish the specific category.

3.2.1. Knowledge distillation-based vision transformer

This study used the KD-based vision transformer, which consistently improves classification performance and semantic segmentation transformers, resulting in higher accuracy, lower processing costs, and improved pre-training needs [27]. In KD, knowledge is transferred from teacher to student using a model that is extensively utilized in computer vision. However, the main purpose of this research is to create a KD structure for segmentation using transformer-based models. Meanwhile, this study used a response-based KD technique to train a teacher and student model. Compared to previous KD models, the response-based KD method offers the advantages of minimizing error, distributing data adaptively based on diverse types, enhancing model interpretability, and significantly improving model robustness. Figure 2 depicts the architecture of the knowledge distillation-based vision transformer.

The soft label is the key point of response-based KD; in this case, the soft label has obtained a smoothed representation of the SoftMax role from the probability of output dissemination. However, the softening output is calculated employing the teacher training model, as represented in (6).

$$p_i = \frac{exp\left(\frac{x_i}{T}\right)}{\sum_i exp\left(\frac{x_j}{T}\right)} \tag{6}$$

Here, the soft label is denoted as p_i, which is engaged to guide the student model as well as x_i signifies the likelihood of posterior dissemination, which is performed before SoftMax. The x_j denotes the value of the class; here j is the posterior probability. Moreover, one significant restriction is the temperature system T, which permits smoothing the posterior probability dissemination. When the temperature T is 1, it is functional to the SoftMax model; if T is larger than 1, the probability distribution of the output becomes smoother and serves to preserve similar information, while T is infinite, it resembles a uniform dissemination.

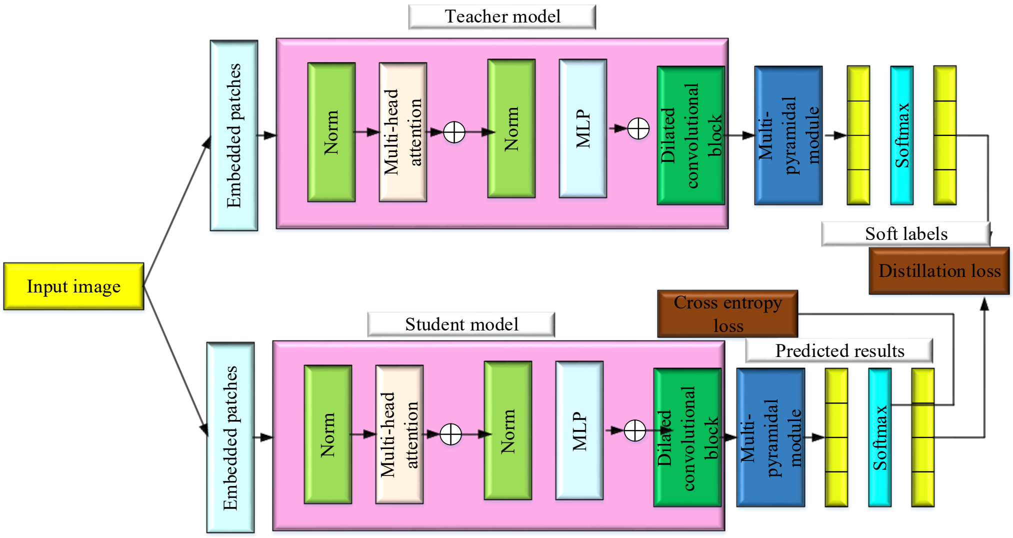


Figure 2. Architecture of knowledge distillation-based vision transformer

3.2.2. Multi-scale pyramidal module

Based on computer vision, the multi-scale pyramidal module is introduced. The term multi-scale refers to sampling of features at different stages. However, performing a specific task requires different features at different scale conditions; therefore, the multi-scale attention network (MSANet) must be introduced. The multi-scale pyramidal pooling module on the encoder and the attention mechanism on the decoder, which primarily reflect the multi-scale feature into two portions, such as feature fusion and feature map, are the foundation of the MSANet network's design. Figure 3 represents the architecture of the multi-scale pyramidal module model.

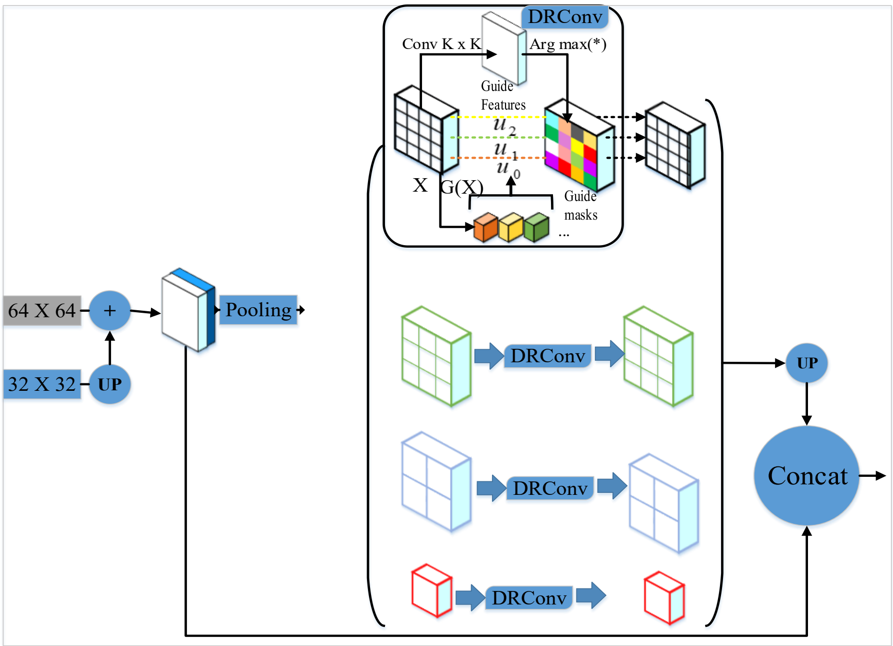


Figure 3. Architecture of multi-scale pyramidal module

The use of a multi-scale input feature map increases fusions between low- and high-resolution depths as well as shallow features, enriching detailed and semantic segmentation. However, global average pooling is used in the multi-scale pooling pyramidal module, which allows for a better grasp of contextual information. The pooling module adapts a four-parallel layer structure, although it has been adjusted to allow feature maps at different scales. Dynamic region-aware convolution (DRConv) is applied under each pooling layer. The multi-scale pooling module introduces the concept of DRConv, which is based on the pyramidal pooling modules. Pyramidal pooling modules can use contextual information at various scales, and DRConv improves the spatial interaction of changing semantic data. The DRConv separates the spatial dimensions into numerous regions using the learnable guided mask, allowing it to combine a set of semantically similar features into a single region, as formulated in (7). Similarly, the multi-scale pyramidal pooling module added to DRConv can effectively simulate semantic data from multiple locations in high-level structures, where the region assignment is determined by the guided mask defined in (8), hence improving the spatial interaction of structures at different scales.

$$X_{u,v,g} = \sum_{b=1}^{B} Y_{u,v,a} * U_{t,a}^{(o)}(u,v) \in A_t \tag{7}$$

$$N_{u,v} = argmax\left(G_{u,v}^{0}, G_{u,v}^{1}, \ldots., G_{u,v}^{n-1}\right) \tag{8}$$

Here defines the mask $N = A_0, \ldots, A_{n-1}$, where n denotes the regions in the spatial dimension, based on the input image feature extraction of the mask is performed. Similarly, semantic structures are grouped into uniform regions. A significant measure of DRConv, the mask N regulates the dissemination of convolutional kernels in the spatial dimension. In each region, only one convolutional kernel $A_t(t \in [0, m-1])$ is shared. The convolutional kernels are defined as $U = [U_0, \ldots. U_{n-1}]$, $U_t \in R^B$ corresponding to the region A_t. The corresponding convolutional kernel, which convolves on the feature map, is altered every time based on its mask. According to the shared area, the convolutional kernel will automatically adjust the substantial structures of the input image. At the end of the module, a feature map is implemented using bilinear interpolation. Further, at each level, features follow the superposition operation, which finally provides output to the decoder. Therefore, a multi-scale pyramidal module allows better contextual data across various regions and improves its ability to seize global data. Therefore, the multi-scale features accurately identify and classify the land cover types effectively.

4. RESULTS AND DISCUSSION

The suggested method has been related to various prevailing methods like U-Net, HRNet, DeepLabV3, ResNet50+DeepLabV3, and Xception+DeepLabV3. Analytical performance of this method, such as accuracy, precision, recall, F1-score, Dice score, intersection over union (IoU), Kappa score, and error metrics, is evaluated as well as compared with recent research works. The suggested method is implemented in MATLAB R2021b, utilizing Windows 10 OS and 8 GB RAM.

4.1. Bhuvan satellite image dataset description

For tasks involving the analysis and segmentation of land cover, this dataset is considered a valuable resource [28]. A set of satellite photos with a high spatial resolution is part of the dataset. The dataset includes satellite 2D images of Varanasi, a city located in the northern part of India, in the state of Uttar Pradesh, organized alternating from 25.3° to 25.5° N latitude as well as 83° to 83.2° Elongitude. It consists of a series of high-resolution images of the Earth's surface. These images were captured by the Indian remote sensing satellite (IRS), processed, and made accessible through the Bhuvan geo platform, which is operated by the Indian space research organization (ISRO).

4.2. Experimental analysis

Figure 4 depicts an experimental investigation of the suggested strategy using the Bhuvan satellite image dataset. The input image is given in Figure 4(a). Figure 4(b) shows the pre-processed image acquired through the application of an Imp-GF. Figure 4(c) shows the segmented image that the suggested KD-MuViTPy model obtains.

4.3. Performance analysis in both training and testing

Training based performance analysis of the proposed model, assessed in terms of the metrics such as accuracy, precision, recall, as well as F1-score using the Bhuvan satellite image dataset, is depicted in Figure 5. The performance analysis of the recommended model using the Bhuvan satellite image dataset is shown in Figure 6. This analysis is assessed in terms of metrics such as accuracy, precision, recall, F1-score, Dice score, IoU, and Kappa score.

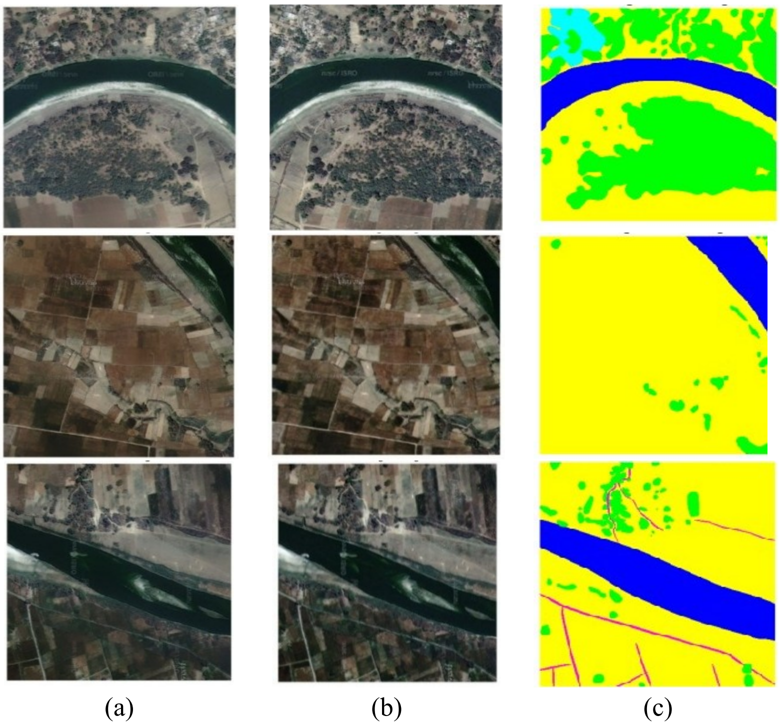


Figure 4. Experimental analysis using the Bhuvan satellite dataset in terms of (a) original image, (b) pre-processed image, and (c) segmented outcomes

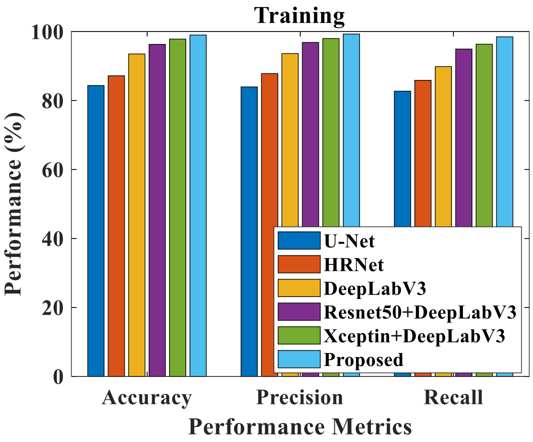
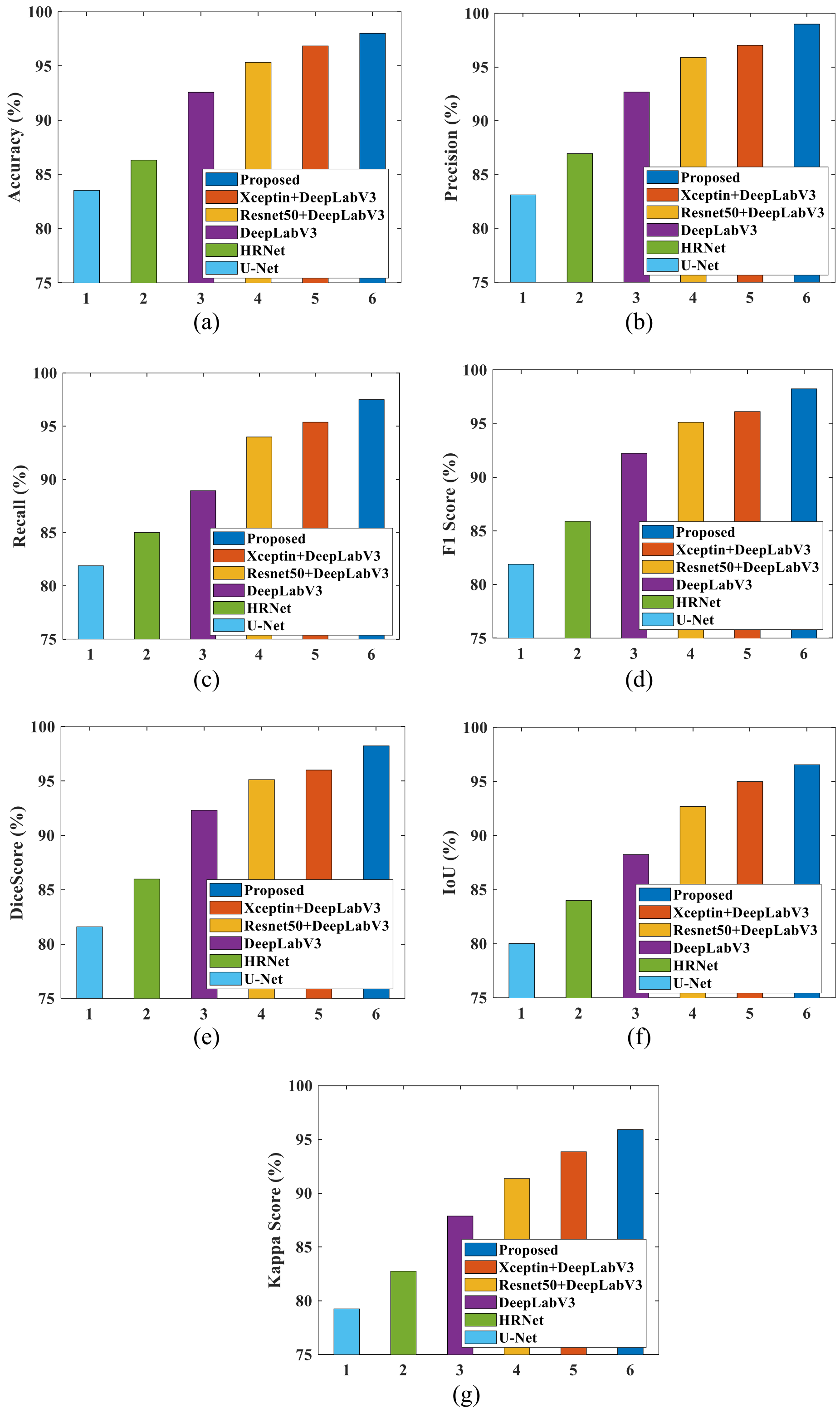


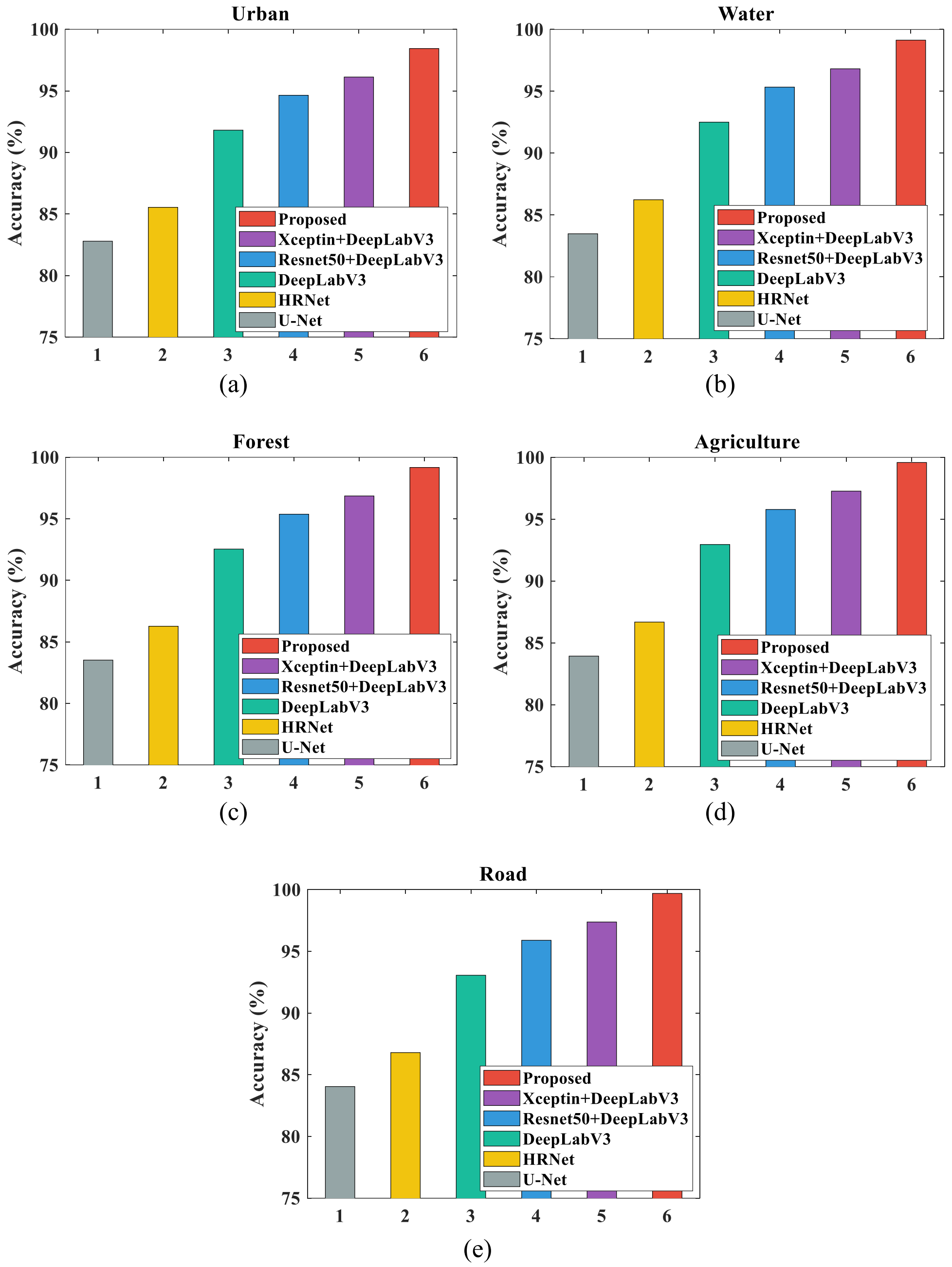
Figure 5. Training based performance analysis of segmentation using Bhuvan Satellite dataset

The segmentation accuracy evaluated by existing methods, such as U-Net, HRNet, DeepLabV3, ResNet50+DeepLabV3, Xceptin+DeepLabV3, achieves values of 84.35727%, 87.18534%, 93.50809%, 96.2957%, and 97.8209%, respectively, as depicted in Figure 6(a). Similarly, the precision evaluated by existing methods is illustrated in Figure 6(b), which achieved values of 83.95326%, 87.82165%, 93.60909%, 96.85127%, and 97.9926%, respectively. The recall was assessed using existing methods, such as 82.71093%, 85.8622%, 93.15458%, 96.07355%, and 97.08358%, respectively. The F1-score evaluated by the suggested method is 99.22397 %, as depicted in Figures 6(c) and 6(d). The Dice score was evaluated by existing methods, such as 81.59%, 85.98%, 92.31%, 95.13%, and 96.01%, respectively. The Dice score assessed by the suggested method is 98.23%, as shown in Figure 6(e). The IoU evaluated by existing methods is 80.02%, 83.99%, 88.25%, 92.68%, and 94.99%. The proposed method obtained an IoU of 96.55%, which is shown in Figure 6(f). The Kappa score was evaluated by existing methods at values of 79.25%, 82.75%, 87.88%, 91.35%, and 93.86%. The suggested method obtained a Kappa score of 95.91%, which is depicted in Figure 6(g). The suggested method produced the best results when compared to other techniques. Figure 7 shows the performance analysis of the recommended technique across including many classes; Figure 7(a) for urban, Figure 7(b) for water, Figure 7(c) for forest, Figure 7 (d) for agriculture, and Figure 7(e) for road; using the Bhuvan Satellite image dataset.



(a) (b) (c) (d) (e) (f) (g)

Figure 6. Testing based performance analysis of segmentation using Bhuvan satellite dataset in terms of (a) accuracy, (b) precision, (c) recall, (d) F1-score, (e) Dice score, (f) IoU, and (g) Kappa score



(a) (b)

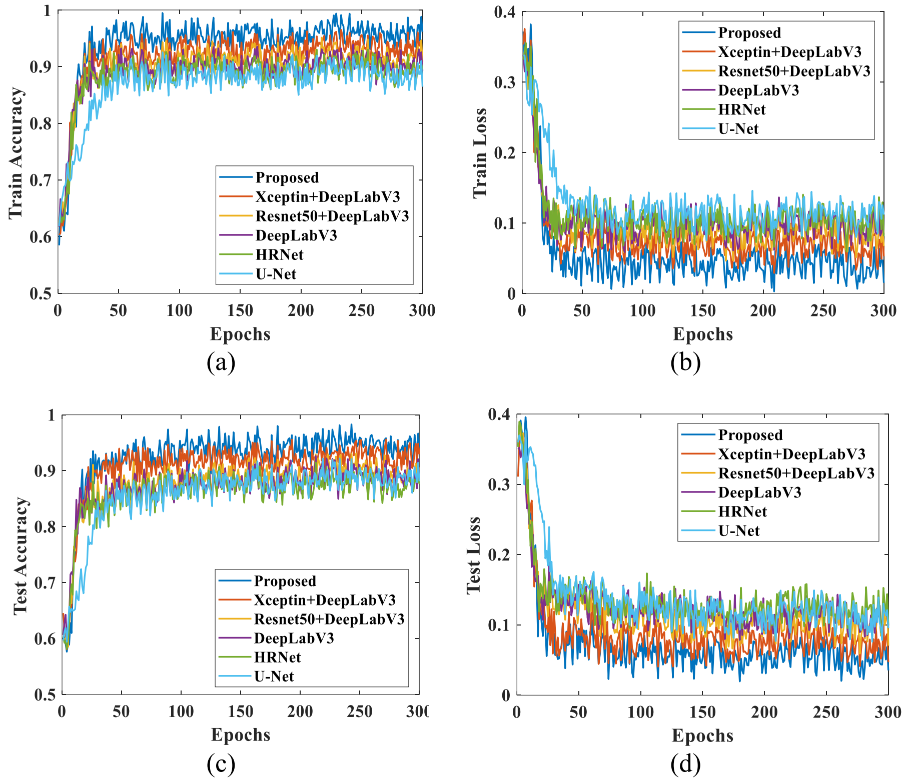
(c) (d)

(e)

Figure 7. Land cover classification accuracy of (a) urban, (b) water, (c) forest, (d) agriculture, and (e) road

4.4. Comparative analysis

In this analysis, the suggested model is related to conventional models such as U-Net, HRNet, DeepLabV3, ResNet50+DeepLabV3, and Xceptin+DeepLabV3. The proposed model has obtained a training phase accuracy of 99% with 300 epochs and 80% of the training data. The suggested model has a training loss value of less than 1 with 300 epochs, with 80% of the training data. The suggested model achieved better outcomes compared with other existing models. The suggested model has obtained a testing phase accuracy of 98% with 300 epochs. The suggested model has a testing loss value of less than 1 with 300 epochs. The suggested model achieved better outcomes compared with other existing methods. Both training as well as testing accuracy and loss are shown in Figure 8. Specifically, Figure 8(a) illustrates the training accuracy, Figure 8(b) shows the training loss, Figure 8(c) presents the testing accuracy, and Figure 8(d) depicts the testing loss. Figure 8 shows the training-based performance analysis for accuracy, precision, recall, as well as the F1-score. When compared to current models, the suggested model attained an accuracy of 99.00%. Similarly, analyses of precision, recall, as well as F1-score revealed superior performance values of 99.28%, 98.47%, and 99.22%, respectively. Table 1 shows a comparison of the existing and proposed models.



(a) (b)

(c) (d)

Figure 8. Training and testing accuracy as well as loss using Bhuvan satellite dataset of (a) training accuracy, (b) training loss, (c) testing accuracy, and (d) testing loss

Table 1. Training based comparative analysis of segmentation using Bhuvan satellite dataset

Metrics	U-Net	HRNet	DeepLabV3	Resnet50+DeepLabV3	Xceptin+DeepLabV3	Proposed model
Accuracy (%)	84.35727	87.18534	93.50809	96.29576	97.8209	99.00064
Precision (%)	83.95326	87.82165	93.60909	96.85127	97.9926	99.28672
Recall (%)	82.71093	85.8622	89.8417	94.93223	96.33616	98.47489
F1-score (%)	82.71093	86.75103	93.15458	96.07355	97.08358	99.22397

Testing based comparative analysis of the proposed model using Bhuvan satellite image data is shown in Table 2. The suggested model attained the maximal accuracy of 98.01 %, which is 14.49%, 11.69%, 5.43%, 2.67%, and 1.16% superior to the existing U-Net, HRNet, DeepLabV3, ResNet50+DeepLabV3, and Xceptin+DeepLabV3 models. Similarly, the suggested model obtained the precision value of 98.99%, which is 15.87%, 12.04%, 6.31%, 3.1%, and 1.97% superior to the traditional models. The suggested model obtained a recall value of 97.49%, which is 15.6%, 12.48%, 8.54%, 3.5%, and 2.11% superior to the existing models. The suggested model obtained an F1-score of 98.23%, which is 16.34%, 12.34%, 6.00%, 3.11%, and 2.11% superior to the existing models. However, the proposed model obtained a Dice score value of 98.23 %, which is 16.64%, 12.25%, 5.92%, 3.1%, and 2.22% more than the existing models. The proposed model obtained an IoU value of 96.55 %, which is 16.53%, 12.56%, 8.3%, 3.87%, and 1.56% superior to the conventional model. The proposed model obtained a Kappa score value of 95.91 %, which is 16.66%, 13.16%, 8.03%, 4.56%, and 2.05% better than the current models.

Table 2. Testing based comparative analysis of segmentation using Bhuvan satellite dataset [29]

Metrics	U-Net	HRNet	DeepLabV3	Resnet50+DeepLabV3	Xceptin+DeepLabV3	Proposed model
Accuracy (%)	83.52	86.32	92.58	95.34	96.85	98.01
Precision (%)	83.12	86.95	92.68	95.89	97.02	98.99
Recall (%)	81.89	85.01	88.95	93.99	95.38	97.49
F1-score (%)	81.89	85.89	92.23	95.12	96.12	98.23
Dice score (%)	81.59	85.98	92.31	95.13	96.01	98.23
IoU (%)	80.02	83.99	88.25	92.68	94.99	96.55
Kappa score (%)	79.25	82.75	87.88	91.35	93.86	95.91

The comparative study of the suggested model evaluated in terms of every class, such as urban area, water, forest area, agriculture part, and road, using the Bhuvan satellite image dataset is shown in Table 3. The suggested method obtained the maximal classification accuracy of 98.43%, 99.13%, 99.16%, 99.58%, and 99.69% to classify the urban land cover, water, forest, agriculture, and road, respectively. The proposed model outperformed the conventional models in classifying various land covers accurately. Table 4 represents some other comparative analysis for the various existing and proposed models.

The inclusion of references [16]–[25] highlights that prior models achieved accuracies ranging from 80.2% to 93.49% and IoU values up to 86.45%, whereas the proposed model attains 99.00% accuracy and 96.55% IoU values. These comparisons with existing literature clearly justify the superior performance of the proposed approach. This contextual evidence supports the robustness and effectiveness of the proposed model.

Table 3. Comparative study of classification accuracy of numerous land cover types using the Bhuvan satellite image dataset

Models	Urban (%)	Water (%)	Forest (%)	Agriculture (%)	Road (%)
U-Net	82.78	83.48	83.51	83.93	84.04
HRNet	85.53	86.23	86.26	86.68	86.79
DeepLabV3	91.80	92.50	92.53	92.95	93.06
ResNet50+DeepLabV3	94.64	95.33	95.37	95.78	95.89
Xception+DeepLabv3	96.123	96.82	96.85	97.26	97.38
Proposed model	98.43	99.13	99.16	99.58	99.69

Table 4. Comparative analysis for state-of-the-art models with the proposed approach

References	Performances	Percentage (%)
[16]	Accuracy	93.49
[17]	IoU	64.61
[18]	mIOU	63.27
[19]	Accuracy	80.2
[20]	Accuracy	87.63
	IoU	68.53
[21]	Accuracy	87.54
[22]	Accuracy	91.99
[23]	Accuracy	83.78
[24]	Accuracy	87.2
[25]	IoU	86.45
	Accuracy	86.54
Proposed	Accuracy	99.00
	IoU	96.55

5. CONCLUSION

This research introduced a novel technique using an atrous spatial assisted knowledge distillation-based vision transformer approach for land cover segmentation and classification. Here, effective segmentation and classification are employed through the introduction of the proposed model, in which the image segmentation has been improved by KD-MuViTPy, and obtains a more accurate segmentation. Using the SoftMax layer, the various types of land covers are classified accurately. The proposed model is assessed using the Bhuvan satellite dataset, which yields better accuracy, precision, recall, F1-score, Dice score, IoU, and Kappa score at values of 98.01%, 98.99%, 97.49%, 98.23%, 98.23%, 96.55%, and 95.91%, respectively. However, its error analysis shows a minimum error of less than 1%. The proposed model can be functional in managing land resources, monitoring the urban environment, identifying land cover changes, and protecting the environment. Future work will explore a more elaborate model for land cover segmentation and classification of high spatial resolution RSI to enhance the representation capacity of the network and more effectively differentiate targets that are prone to confusion.

ACKNOWLEDGMENTS

The authors are grateful to the management of TPCT’s College of Engineering, Dharashiv, Maharashtra, India, for their support. The authors also acknowledge the management of the Research Centre, Jawaharlal Nehru Engineering College, Chhatrapati Sambhaji Nagar, Maharashtra, India, for their support and cooperation.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sujata Arjun Gaikwad	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓			
Vijaya Musande	✓	✓			✓	✓	✓	✓		✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

[1] C. Chen, X. He, Z. Liu, W. Sun, H. Dong, and Y. Chu, “Analysis of regional economic development based on land use and land cover change information derived from Landsat imagery,” *Scientific Reports*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-69716-2.

[2] L. T. Ramos and A. D. Sappa, “Leveraging U-Net and selective feature extraction for land cover classification using remote sensing imagery,” *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-024-84795-1.

[3] A. Vali, S. Comai, and M. Matteucci, “Deep learning for land use and land cover classification based on hyperspectral and multi-spectral earth observation data: a review,” *Remote Sensing*, vol. 12, no. 15, 2020, doi: 10.3390/rs12152495.

[4] S. Talukdar, P. Singha, S. Mahato, S. Pal, Y. A. Liou, and A. Rahman, “Land-use land-cover classification by machine learning classifiers for satellite observations—a review,” *Remote Sensing*, vol. 12, no. 7, 2020, doi: 10.3390/rs12071135.

[5] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, “Super resolution guided deep network for land cover classification from remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021, doi: 10.1109/TGRS.2021.3120891.

[6] S. Basheer *et al.*, “Comparison of land use land cover classifiers using different satellite imagery and machine learning techniques,” *Remote Sensing*, vol. 14, no. 19, 2022, doi: 10.3390/rs14194978.

[7] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, “Change detection based on artificial intelligence: state-of-the-art and challenges,” *Remote Sensing*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101688.

[8] M. Beak *et al.*, “Land cover classification for Siberia leveraging diverse global land cover datasets,” *Progress in Earth and Planetary Science*, vol. 12, no. 1, 2025, doi: 10.1186/s40645-024-00672-5.

[9] K. Karra, C. Kontgis, Z. S. -Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, “Global land use/land cover with Sentinel 2 and deep learning,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Brussels, Belgium: IEEE, 2021, pp. 4704–4707, doi: 10.1109/IGARSS47720.2021.9553499.

[10] R. Swathika, N. Radha, S. P. Sasirekha, and S. Dhanabal, “Flood susceptibility in urban environment using multi-layered neural network model from satellite imagery sources,” *Global Nest Journal*, vol. 25, no. 8, pp. 60–73, 2023, doi: 10.30955/gnj.005139.

[11] H. Jiang *et al.*, “A survey on deep learning-based change detection from high-resolution remote sensing images,” *Remote Sensing*, vol. 14, no. 7, 2022, doi: 10.3390/rs14071552.

[12] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, “Deep learning-based change detection in remote sensing images: a review,” *Remote Sensing*, vol. 14, no. 4, 2022, doi: 10.3390/rs14040871.

[13] C. Persello *et al.*, “Deep learning and earth observation to support the sustainable development goals: current approaches, open challenges, and future opportunities,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 172–200, 2022, doi: 10.1109/MGRS.2021.3136100.

[14] Z. Zhao *et al.*, “Comparison of three machine learning algorithms using Google Earth Engine for land use land cover classification,” *Rangeland Ecology and Management*, vol. 92, pp. 129–137, 2024, doi: 10.1016/j.rama.2023.10.007.

[15] C. Liu, D. Zeng, H. Wu, Y. Wang, S. Jia, and L. Xin, “Urban land cover classification of high-resolution aerial imagery using a relation-enhanced multi-scale convolutional network,” *Remote Sensing*, vol. 12, no. 2, 2020, doi: 10.3390/rs12020311.

[16] F. J. Cardama, D. B. Heras, and F. Argüello, “Consensus techniques for unsupervised binary change detection using multi-scale segmentation detectors for land cover vegetation images,” *Remote Sensing*, vol. 15, no. 11, 2023, doi: 10.3390/rs15112889.

[17] D. Wang *et al.*, “P-Swin: parallel swin transformer multi-scale semantic segmentation network for land cover classification,” *Computers and Geosciences*, vol. 175, 2023, doi: 10.1016/j.cageo.2023.105340.

[18] Y. Ma, X. Deng, and J. Wei, “Land use classification of high-resolution multi-spectral satellite images with fine-grained multi-scale networks and superpixel postprocessing,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3264–3278, 2023, doi: 10.1109/JSTARS.2023.3260448.

[19] J. Jia, J. Song, Q. Kong, H. Yang, Y. Teng, and X. Song, “Multi-attention-based semantic segmentation network for land cover remote sensing images,” *Electronics*, vol. 12, no. 6, 2023, doi: 10.3390/electronics12061347.

[20] D. Xu, Z. Li, H. Feng, F. Wu, and Y. Wang, “Multi-scale feature fusion network with symmetric attention for land cover classification using SAR and optical images,” *Remote Sensing*, vol. 16, no. 6, 2024, doi: 10.3390/rs16060957.

[21] S. Zhang *et al.*, “EMMCNN: an ETPS-based multi-scale and multi-feature method using CNN for high spatial resolution image land-cover classification,” *Remote Sensing*, vol. 12, no. 1, 2019, doi: 10.3390/rs12010066.

[22] S. Shi *et al.*, “Land cover classification with multi-spectral LiDAR based on multi-scale spatial and spectral feature selection,” *Remote Sensing*, vol. 13, no. 20, 2021, doi: 10.3390/rs13204118.

[23] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, “Land cover classification from remote sensing images based on multi-scale fully convolutional network,” *Geo-spatial Information Science*, vol. 25, no. 2, pp. 278–294, 2022, doi: 10.1080/10095020.2021.2017237.

[24] V. S. Martins, A. L. Kaleita, B. K. Gelder, H. L. da Silveira, and C. A. Abe, “Exploring multi-scale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 56–73, 2020, doi: 10.1016/j.isprsjprs.2020.08.004.

[25] B. Chen, M. Xia, and J. Huang, “MFANet: a multi-level feature aggregation network for semantic segmentation of land cover,” *Remote Sensing*, vol. 13, no. 4, 2021, doi: 10.3390/rs13040731.

[26] H. Singh, S. V. R. Kommuri, A. Kumar, and V. Bajaj, “A new technique for guided filter based image denoising using modified cuckoo search optimization,” *Expert Systems with Applications*, vol. 176, 2021, doi: 10.1016/j.eswa.2021.114884.

[27] Q. Huang, K. Yang, Y. Zhu, L. Chen, and L. Cao, “Knowledge distillation for enhancing a lightweight magnet tile target detection model: leveraging spatial attention and multi-scale output features,” *Electronics*, vol. 12, no. 22, 2023, doi: 10.3390/electronics12224589.

[28] K. Patni, “Satellite image and mask,” *Kaggle*. Accessed: Feb 20, 2025. [Online]. Available: https://www.kaggle.com/datasets/khushiipatni/satellite-image-and-mask

[29] J. Wang *et al.*, “A transformer-based knowledge distillation network for cortical cataract grading,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1089–1101, 2023, doi: 10.1109/TMI.2023.3327274.

BIOGRAPHIES OF AUTHORS

Sujata Arjun Gaikwad is working as an assistant professor, head of the Department of Computer Science and Engineering, at TPCT’s College of Engineering, Osmanabad (Dharashiv), Maharashtra, India. She received her B.E. degree (Computer Science and Engineering) and M.E. degree (Computer Science and Engineering) from Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhaji Nagar, India. She is pursuing Ph. D. from Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhaji Nagar, India. She has more than 15 years of teaching experience. Her area of research is remote sensing and GIS, deep learning, image processing, and artificial intelligence. She has publications in international conferences and journals. She can be contacted at email: sujatagaikwad414@gmail.com.

Dr. Vijaya Musande is a professor and the principal of Jawaharlal Nehru Engineering College at MGM University, Chhatrapati, Sambhajinagar, Maharashtra, India. She holds a Ph.D. in Computer Science and Engineering for her work on "Development of feature extraction technique for remote sensing images: a temporal fuzzy approach." Her research expertise spans remote sensing and GIS, digital image processing, soft computing, and artificial intelligence. She has an extensive publication record, including 24 journal papers (majority Scopus-indexed), 15 conference papers, and 4 patents granted in India, Australia, and Germany. She has successfully guided 11 M.Tech. and 5 Ph.D. students to completion, with 6 more currently pursuing their Ph.D. under her supervision. With over 28 years of academic experience and numerous awards, including the Outstanding Engineers Award for Sir M. Visvesvaraya, she is a distinguished personality in her field. She can be contacted at email: vijayamusande@gmail.com.