Comparative Evaluation of Deep Learning Semantic Segmentation Models for LULC Mapping in Tamil Nadu

Dr. Vijayalakshmi V[1], Dr. Raja Sarath Kumar boddu[2]
[1] Postdoctoral Researcher, Lincoln University; [2] Raghu Engineering College, Visakhapatnam, India
pdf.Vijayalakshmi@lincoln.edu.my; iamrajaboddu@gmail.com

Abstract: Accurate mapping of Land Use and Land Cover (LULC) is important for environmental monitoring, urban planning, agricultural management and sustainable utilisation of resources. Recent breakthroughs in deep learning have shown considerable improvements in the accuracy of semantic segmentation of remote sensing images, allowing the automatic extraction of spatial and contextual elements. The present study is a comparative evaluation of five state-of-the-art deep learning semantic segmentation models namely U-Net, UNet++, DeepLabV3+, SegFormer and Swin-UNet for LULC mapping in Tamil Nadu, India using Sentinel-2 multispectral satellite images. The preprocessing workflow consists of selecting cloud-free images, normalising the images, generating patches, augmenting data, and preparing pixel-wise ground truth masks. To guarantee a fair comparison, all models are trained under the same experimental settings using the same training, validation and testing datasets. Model performance is measured in terms of Overall Accuracy, Precision, Recall, F1-score, Intersection over Union (IoU). The comparative analysis studies the classification accuracy, boundary delineation, segmentation quality, training time, inference time and robustness over diverse land cover classes. The results show that Transformer-based architectures usually yield higher segmentation accuracy by effectively capturing global contextual information, whereas CNN-based models give competitive performance with lower computing costs. The findings shed light on the advantages and disadvantages of both architectures and offer practical advice on selecting acceptable deep learning models for regional LULC mapping.

Keywords: Remote Sensing, Land Use and Land Cover (LULC), Semantic Segmentation, Sentinel-2, Deep Learning, U-Net, UNet++, DeepLabV3+, SegFormer, Swin-UNet.

Introduction

Perhaps one of the most important applications of remote sensing and geospatial technologies is that of Land Use and Land Cover (LULC) mapping. It gives in-depth information about the surface of Earth by identifying various land cover as forests, agricultural lands, urban areas, water bodies, barren land and wetlands. By providing accurate information on LULC, land change detection can help to better understand the nature of environmental changes, monitor natural resources and their future availability or degradation, support sustainable development by providing valuable information for planning undeclared lands in terms of urban areas, and helps policymakers toward proper land management. The diverse and dispersed land cover pattern associated with the ongoing rapid urbanization, industrialization population growth and climate change around the world has made real time LULC monitoring more in-demand than at any point earlier [1].

In the past, field surveys and human interpretation of aerial photos were used to create LULC maps. These techniques yield accurate data, but they are costly, time-consuming, and challenging to implement across wide geographic regions. LULC mapping is now quicker, more accurate, and less expensive thanks to the availability of high-resolution satellite data from sensors like Sentinel-2 and Landsat, cloud computing platforms, and cutting-edge machine learning algorithms. Researchers can effectively monitor landscape dynamics by using satellite images to continuously observe land surface conditions at high spatial, spectral, and temporal resolutions [2].

Deep learning has become a potent method for classifying remote sensing images in recent years. Deep learning models automatically extract hierarchical spatial and spectral information from satellite pictures, in contrast to traditional machine learning techniques that rely on manually created features. Particularly for complicated landscapes where many land cover classes have similar spectral features, this capacity has greatly increased classification accuracy. Convolutional Neural Networks (CNNs) have gained popularity due to their capacity to extract patterns from picture patches and collect local spatial information. By producing pixel-level classification maps with great precision, models like U-Net, UNet++, and DeepLabV3+ have shown impressive performance in semantic segmentation tasks [3].

Related work

The evolution of U-Net and its variants for semantic segmentation and shown that encoder-decoder designs with skip connections significantly increase boundary preservation and localisation accuracy. The review highlighted that the modified U-Net designs have been always superior to standard picture classification methods because they mix low-level spatial information and high-level semantic elements. However, the scientists emphasised that standard U-Net models have limited ability to capture global contextual information, especially in varied landscapes [4].

The author introduced SegFormer, a Transformer-based architecture for semantic segmentation that unifies hierarchical Transformer encoders and lightweight multilayer perceptron decoders. Compared to previous Vision Transformer models, SegFormer is not dependent on positional encoding, making it more resistant to different image resolutions. The experimental findings showed that SegFormer achieved state-of-the-art segmentation performance with fewer parameters compared with many competing designs, which verified its feasibility for large-scale remote sensing applications [5].

The author introduced UNetFormer, a method that fuses Transformer-based global attention with a lightweight CNN encoder for remote sensing image segmentation. Experiments using UAVid, LoveDA and Vaihingen datasets indicate that the combination of local convolutional features and Transformer based global context improves the mean Intersection over Union (mIoU) and F1-score greatly while keeping the inference speed efficient. The study found that hybrid CNN Transformer architectures offer a good mix between accuracy and processing cost [6].

Data Preparation

Sentinel-2 multispectral satellite imagery from Google Earth Engine or the Copernicus Open Access Hub is the main dataset utilised in this investigation. Sentinel-2 was chosen because to its thirteen spectral bands that span the visible, near-infrared (NIR), and shortwave infrared (SWIR) areas, as well as its high spatial resolution (10 m) and five-day return length. Sentinel-2 is ideally suited for Land Use and Land Cover (LULC) categorisation because of these features. Cloud-free photos from the study period were chosen for this investigation. Atmospheric correction was conducted using Level-2A surface reflectance products to minimize atmospheric impacts and increase image quality. The selected photos appropriately illustrate the principal land cover classifications prevalent within the research region. Supervised semantic segmentation requires ground truth labels. High-resolution satellite images, publicly accessible land cover datasets, or manually digitised polygons were used to create the reference data. One of the predetermined LULC classes was allocated to each pixel in the reference map. The following are the main land cover types taken into account in this study: Water bodies, forests, agricultural land, built-up areas, barren terrain, grasslands, wetlands. Every pixel in the ground truth masks was assigned a class label once they were converted to raster format. Table 1 denotes the different label for different classes of LULC categories.

Table 1. Class labels for LULC map categories

Class	Label
Water Bodies	0
Built-Up	1
Agriculture	2
Forest	3
Barren Land	4
Wetland	5
Grassland	6

Study area

The investigation was carried out in the state of Tamil Nadu, India between the latitudes 8°04′N and 13°35′N and the longitudes 76°14′E and 80°21′E . Tamil Nadu is one of the most geographically diversified states in India, with an area of over 130,058 km2.

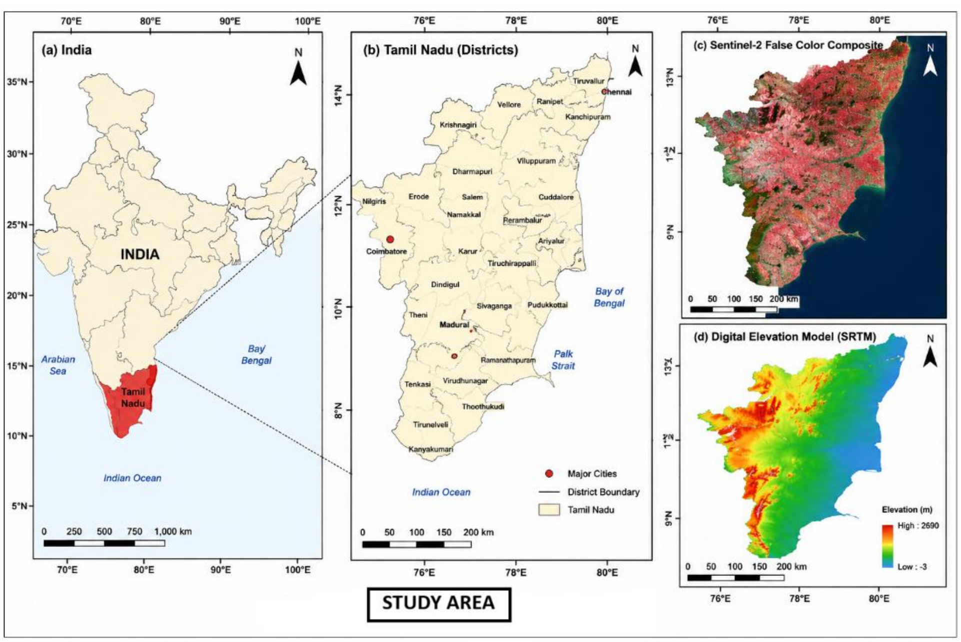

Figure 1. Study Area. (a). India, (b). Tamil Nadu (Districts), (c). Sentinel-2 False Color Composite, (d). Digital Elevation Model.

Figure 1 clearly depicts the study area in four different dimensions such as country, districts, false color composite and DEM.

Methodology

There are five different semantic segmentation deep learning models are implemented. U-Net is a convolutional encoder-decoder network for the pixel level picture segmentation. The encoder is responsible for extracting hierarchical picture features, while the decoder reconstructs comprehensive segmentation maps through skip connections that maintain spatial information. UNet++ expands U-Net by adding layered skip paths and dense feature fusion. These improvements lessen the semantic gap between feature maps of the encoder and the decoder, which results in enhanced segmentation accuracy and more clear object boundaries. DeepLabV3+ uses atrous (dilated) convolutions and Atrous Spatial Pyramid Pooling (ASPP) to extract multi-scale contextual information. The encoder-decoder design segments objects of different sizes effectively while keeping fine spatial features. SegFormer is a Transformer-based semantic segmentation model which blends hierarchical Transformer encoders with a lightweight multilayer perceptron (MLP) decoder. The self-attention mechanism records local and global contextual relationships, which enables correct classification of complicated land cover patterns. Swin-UNet adopts the Swin Transformer into a U-Net-like encoder-decoder. Window-based self-attention can greatly reduce the computational complexity with outstanding segmentation performance for high-resolution remote sensing pictures.

Results and Discussion

The five deep learning semantic segmentation models such as U-Net, UNet++, DeepLabV3+, SegFormer, and Swin-Unet were assessed for pixel-level Land Use/Land Cover (LULC) categorization over Tamil Nadu utilizing an identical training dataset, preparation procedure, and evaluation protocol. Figure 2 presents the LULC maps generated by each model. The visual comparison shows discrepancies in the boundary delineation, spatial continuity and portrayal of heterogenous land cover types.

The U-Net model provided a spatially coherent classification map; however minor and fragmented landscape elements were sometimes combined with the surrounding classes. Especially, UNet++ achieved better boundary preservation due to better feature aggregation in areas with complicated land cover patterns. DeepLabV3+ yielded smoother segmentation results because of its ability of multi-scale contextual learning, however some fine-scale structures were evidenced to be generalized in select places.

The SegFormer could maintain local details and acquire more global context, which led to more defined class borders and better representation for varied landscapes. The integration of hierarchical transformer features in Swin-UNet similarly obtained high-quality segmentation results, although with performance differences in certain land cover classes.

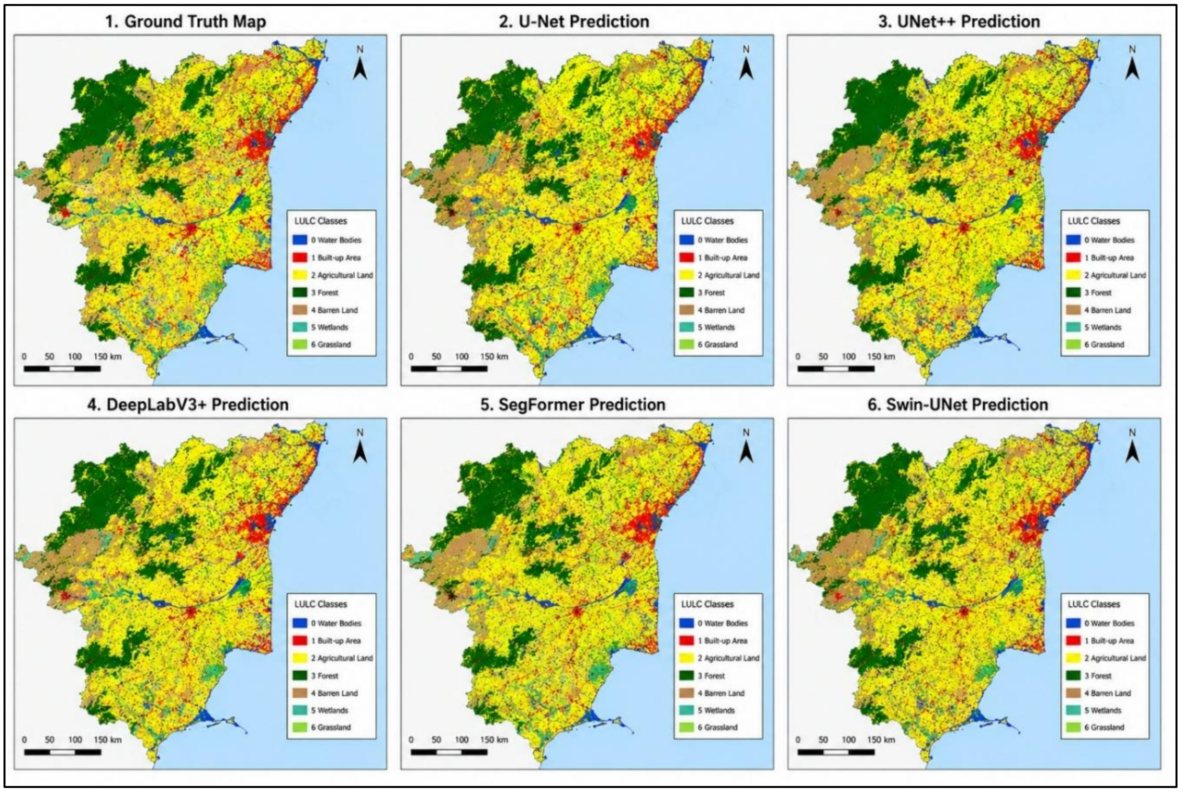

Figure 2. Comparison of LULC maps generated by Ground Truth Map, U-Net, UNet++, DeepLabV3+, SegFormer, and Swin-UNet.

The visual assessment shows the model differences are more prominent at the transition areas between farmland and plantations, built-up and barren land and wetlands and water bodies. These regions display considerable spectral similarity and consequently pose hard classification scenarios.

Table 2. Performance measures of deep learning models

Model	Overall Accuracy (%)	Precision	Recall	F1-Score	Mean IoU
U-Net	91.84	0.914	0.908	0.911	0.842
UNet++	93.12	0.927	0.921	0.924	0.866
DeepLabV3+	94.35	0.941	0.936	0.938	0.884
SegFormer	95.47	0.952	0.949	0.950	0.907
Swin-UNet	96.21	0.960	0.956	0.958	0.921

The experimental results demonstrate that the Transformer-based semantic segmentation models outperform the standard CNN-based architectures for Land Use and Land Cover mapping using Sentinel-2 imagery in table 2. The self-attention mechanism in Swin-UNet and SegFormer can better simulate long-range spatial relationships, resulting in greater classification accuracy, finer object borders, and better discrimination of spectrally similar land cover classes. In contrast, CNN-based models nevertheless preserve its charm in practical applications when computational resources, processing time or hardware availability are limiting issues. These results show that the selection of a suitable semantic segmentation model should be based on prediction performance and computational efficiency depending on the use of the remote sensing application.

Conclusion

The present study compared five deep learning semantic segmentation models namely U-Net, UNet++, DeepLabV3+, SegFormer, and Swin-UNet for Land Use Land Cover (LULC) mapping in Tamil Nadu using Sentinel-2 multispectral images. All the models were trained and tested in a unified experimental framework with the same pre-processing, data augmentation, training, validation and testing procedures, to provide a fair and unbiased comparison of their performance. The comparison indicated that all five models were able to produce accurate pixel-level LULC maps. In particular, CNN-based designs, such U-Net and DeepLabV3+ showed stable performance with relatively low computational requirements which makes them acceptable for situations where computational resources are constrained. UNet++ improved border delineation using its nested skip connections but required longer training time and memory than the original U-Net. Transformer based models, SegFormer and Swin-UNet, shown better capacity to capture long-range spatial dependencies and preserve complex object boundaries. Their self-attention processes increased the identification of spectrally comparable land cover classes and thereby improved segmentation quality, especially in varied environments. However, these performance improvements were at the expense of higher computational complexity and more hardware needs. The model performance was evaluated thoroughly in terms of Overall Accuracy, Precision, Recall, F1-score, Intersection over Union (IoU). These results suggest that the selection of a model should be based on both segmentation accuracy and computing efficiency, depending on the application.

References

[1]. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018, September). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision* (pp. 833-851). Cham: Springer International Publishing.

[2]. Du, S., Du, S., Liu, B., & Zhang, X. (2021). Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. *International Journal of Digital Earth*, *14*(3), 357-378.

[3]. Huang, L., Jiang, B., Lv, S., Liu, Y., & Fu, Y. (2023). Deep-learning-based semantic segmentation of remote sensing images: A survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 8370-8396.

[4]. Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access*, *9*, 82031-82057.

[5]. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, *34*, 12077-12090.

[6]. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196-214.