Evaluating the Performance of Advanced Deep Learning Models for Accurate Landslide Prediction

Pullabhotla Vijay, Nagasai Varaprasad, Dr. Bharathi V C
School of Computer Science and Engineering,
VIT-AP University, Amaravati, Andhra Pradesh, 522241, India

ABSTRACT

Landslides are one of the most devastating natural disasters, posing a significant threat to life, infrastructure, and the environment. Timely and accurate detection is critical for effective disaster response and mitigation. This research investigates the application of state-of-the-art deep learning models for the semantic segmentation of landslide-affected regions from satellite imagery. We compare UNet++, DeepLabV3+, ResUNet, UNet, and SegFormer on a curated dataset, evaluating their performance across various metrics, including Accuracy, Precision, Recall, F1 Score, Intersection over Union (IoU), and Dice Coefficient. UNet achieved the highest accuracy (0.9842), while UNet++ provided the best overall performance balance. Our results suggest that modern encoder-decoder architectures can effectively support geohazard mapping and landslide prediction systems.

KEYWORDS

Landslide Detection, Deep Learning, Semantic Segmentation, Satellite Imagery, UNet++, DeepLabV3+, SegFormer, Remote Sensing, Natural Disaster, CNN, Transformer, Image Segmentation.

1. INTRODUCTION

Landslides occur due to a combination of natural and anthropogenic factors, including heavy rainfall, earthquakes, deforestation, and infrastructure development. Accurate detection is vital to prevent fatalities and economic losses. Traditional methods, often manual and time-consuming, are being increasingly replaced by automated solutions using remote sensing and machine learning.

Deep learning has revolutionized image analysis with its ability to learn hierarchical features. Semantic segmentation enables pixel-level classification, making it a natural choice for landslide area detection. This study focuses on comparing five deep learning models to evaluate their effectiveness for landslide detection in satellite imagery.

i.

ii. 1.1 Research Objectives

- To develop a robust deep learning framework for semantic segmentation-based landslide detection using satellite imagery.
- To implement and train multiple advanced deep learning models — including UNet, UNet++, ResUNet, DeepLabV3+, and SegFormer — for comparative performance analysis.
- To evaluate and compare model performance based on key segmentation metrics such as Accuracy, Precision, Recall, F1 Score, Intersection over Union (IoU), and Dice Coefficient.
- To investigate the effectiveness of different model architectures in handling spatial complexity and feature extraction for landslide-prone regions.

2. METHODOLOGY

iii. 2.1 Dataset Description

The dataset used in this study comprises high-resolution satellite imagery organized into three subsets: TrainData, ValidData, and TestData, each containing .h5 files with pixel-level image data and corresponding segmentation masks. The TrainData folder includes images and their ground truth masks used for training deep learning models, while ValidData and TestData serve for validation and testing, respectively. Each image represents a geographical region with annotated landslide-affected areas. To enhance model performance and generalization, several preprocessing steps were applied, including Min-Max Normalization to scale pixel values between 0 and 1, a validation split of 10–20% from the training set, and data augmentation techniques such as flipping, rotation, and brightness adjustment to address class imbalance and increase dataset diversity

b.

i. ## 2.2 Data Collection

ii. Our dataset comprises 3,799 samples from satellite imagery, elevation models, and GIS data sources. Each sample includes the following features:

- RGB Imagery: Captures color-based terrain features.
- NDVI (Normalized Difference Vegetation Index): Identifies vegetation health and land cover changes.
- Slope: Measures terrain steepness and soil stability.
- Elevation (DEM - Digital Elevation Model): Represents terrain altitude variations.
- Ground Truth Mask: Provides labeled landslide-prone areas for supervised training.

iii. ## 2.3 Data Preprocessing

To enhance model efficiency, the dataset undergoes multiple preprocessing steps:

- **Normalization**: RGB values are scaled between 0 and 1.
- Where:
- **NIR** = Near-Infrared band value
  Red = Red band value

  This normalized index is used to assess vegetation health by comparing how plants reflect light in the NIR and Red wavelengths.

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

  - Slope and Elevation are normalized to a common range.

- **Data Splitting**:
- 80% training, 20% validation.
- **HDF5 Storage**:
- Optimized for fast access and training efficiency.

3. Models Evaluated

Five state-of-the-art deep learning architectures were implemented and comparatively evaluated to segment landslide-prone areas with high precision. All models were trained using the **Dice Loss** function to address

$$\text{Dice Loss} = 1 - \frac{2|P \cap G|}{|P| + |G|}$$

class imbalance. The **Dice Loss** is defined as:
Where PPP denotes the predicted mask and GGG the ground truth mask. In addition to loss, performance metrics used for evaluation included Accuracy, Precision, Recall, F1 Score, Intersection over Union (IoU), and Dice Coefficient. The formulas for IoU and Dice Coefficient are:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad ; \quad \text{Dice Coefficient} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Where TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. The evaluated models include:

- **UNet**

UNet is a convolutional neural network (CNN) designed for biomedical image segmentation and adapted here for landslide detection. It uses an encoder–decoder structure with skip connections to retain spatial resolution.

Formula: The architecture involves downsampling (convolution + pooling) followed by upsampling (transposed convolution) and skip connections:

$$F_{decoder} = \text{Upsample}(F_{encoder}) + F_{skip}$$

- **UNet++**

UNet++ introduces nested skip connections and dense convolution blocks to bridge the semantic gap between encoder and decoder features. This enhances gradient flow and improves convergence.

Formula*:* It adds intermediate convolutional layers Xi,jX_{i,j}Xi,j as:

$$X_{i,j} = H_{i,j}([X_{i,j-1}, \ldots, X_{i+n,j-n}])$$

- **ResUNet**

ResUNet incorporates residual connections within the UNet architecture to facilitate deeper networks without vanishing gradients. This structure is more robust in learning complex features.

Formula: Residual blocks are defined as:

$$F(x) = x + \mathcal{F}(x, W)$$

- **DeepLabV3+**

DeepLabV3+ enhances semantic segmentation using Atrous Spatial Pyramid Pooling (ASPP) for multi-scale feature extraction and a decoder module for precise object boundaries.

Formula: ASPP is given as:

$$\text{ASPP}(x) = \sum_{r \in R} \text{Conv}(x, \text{rate} = r)$$

- **SegFormer(transformer)**

SegFormer is a transformer-based architecture that combines hierarchical vision transformers with lightweight MLP decoders for efficient and accurate segmentation without relying on convolution layers.

Formula: The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Each model was trained under the same preprocessing pipeline and evaluated across identical test and validation datasets to ensure a fair and consistent comparison.

3.1 Evaluation Metrics and Techniques

To ensure robust model evaluation, predictions were compared with ground truth masks on the validation and test sets using the following quantitative metrics:

- Accuracy: Overall pixel classification correctness.
- Precision and Recall: Measures of a model's ability to correctly identify landslide pixels.
- F1 Score: Harmonic mean of precision and recall.
- IoU and Dice Coefficient: Pixel-level overlap metrics for segmentation quality.

Additionally, qualitative visualizations were produced to inspect the spatial accuracy of predicted masks. These visual comparisons between ground truth and predictions helped assess each model's localization effectiveness. Color-coded overlays and segmentation boundaries were used for clearer interpretation.

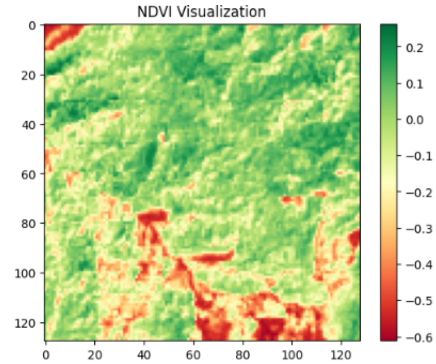

Fig.1: NDVI Visualization

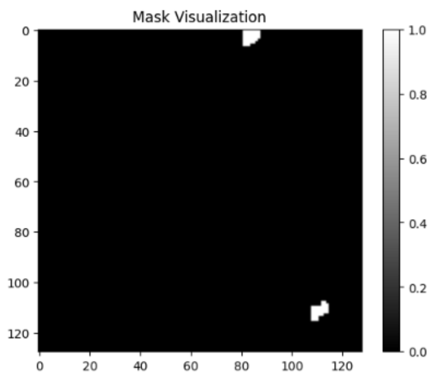

Fig.2: Mask Visualization

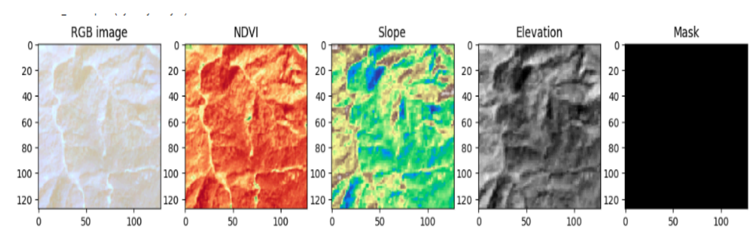

Fig. 3: RGB, NDVI, Slope, Elevation, Mask Images.

3.2 Model Performance

To evaluate the effectiveness of the proposed deep learning models in segmenting landslide-prone areas, a comprehensive performance analysis was conducted using six key metrics: Accuracy, Precision, Recall, F1 Score, Intersection over Union (IoU), and Dice Coefficient. These metrics offer a balanced view of each model's ability to identify true positives while mitigating the effects of class imbalance, which is common in geospatial image segmentation tasks. The models were trained and tested on consistent datasets with standardized preprocessing, and their performance was measured across identical conditions to ensure fair comparison.

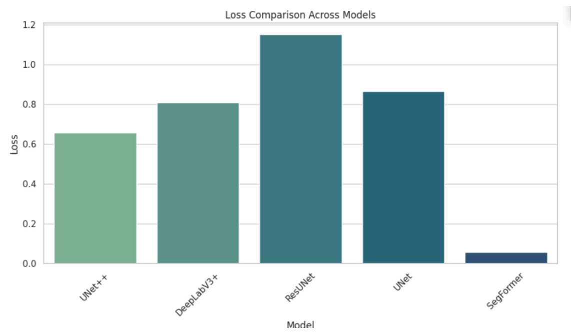

Fig. 4 : Loss comparison across models

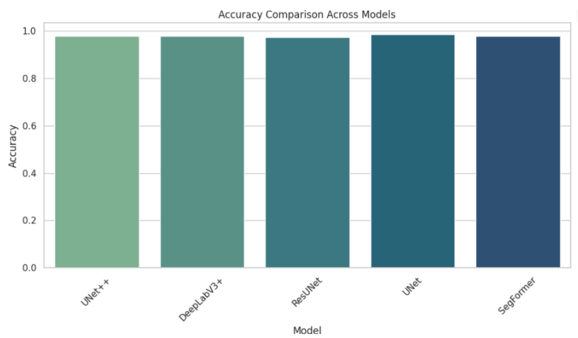

Fig. 5: Accuracy comparison across models

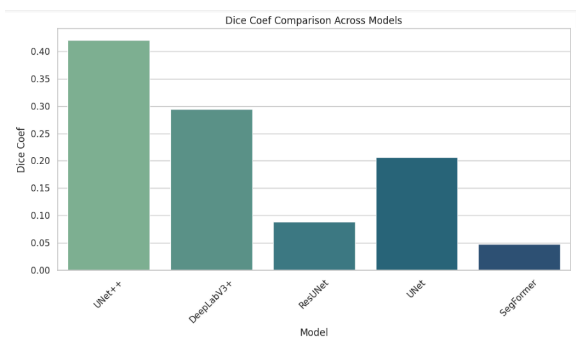

Fig. 6: Died Coef comparison across models

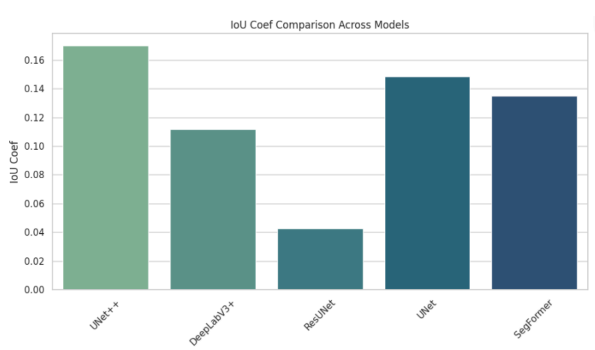

Fig.7: IoU Coef comparison across models

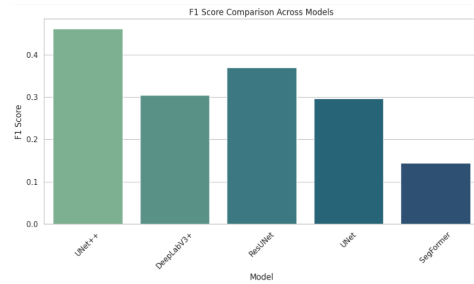

Fig.8: F1 Score comparison across models

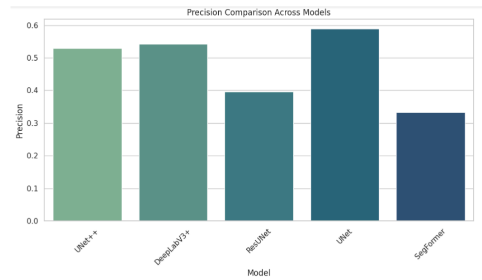

Fig.9: Precision comparison across models

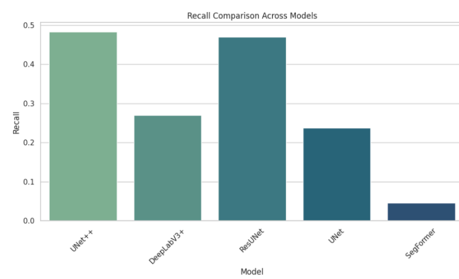

Fig.10: Recall comparison across models

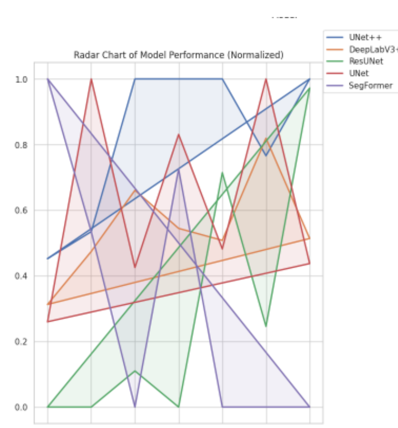

Fig.11: Radar chart Model Performance(Normalized)

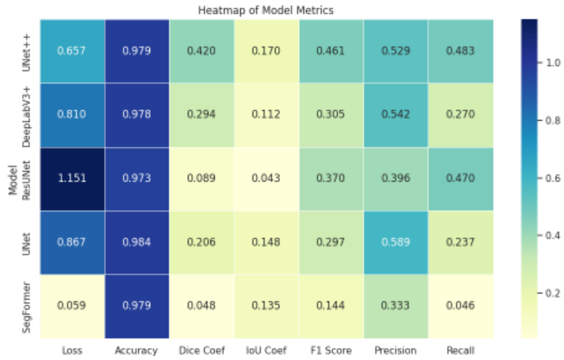

Fig.12: Heatmap of Model Metrics

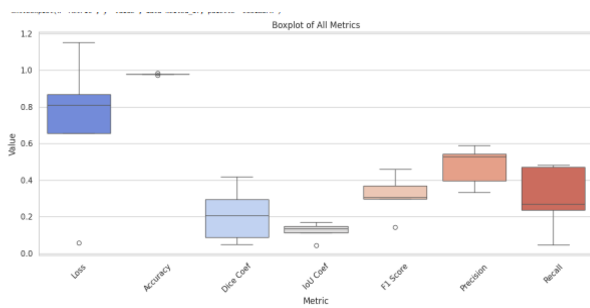

Fig.13: Boxplot of Model Metrics

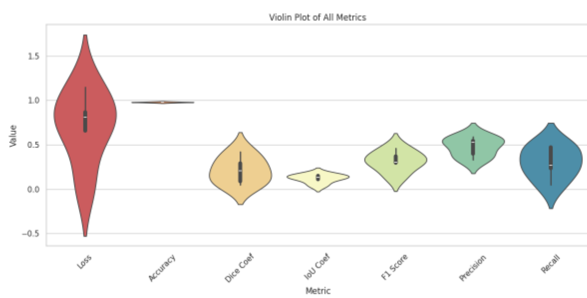

Fig.14: Violin plot of all Model Metrics

c. 3.1 Model Performance and Comparative Study

d. Each of the five models evaluated demonstrated varying strengths and limitations when applied to the semantic segmentation of landslide-prone regions. Among them, **UNet** achieved the highest overall accuracy at 0.9842, indicating its strong generalization capabilities. However, its precision and recall were unbalanced—0.5885 and 0.2370, respectively, leading to a moderate F1 Score of 0.2967. This suggests that while UNet effectively classified the majority of non-landslide pixels correctly, it struggled with true positive predictions of landslide regions.

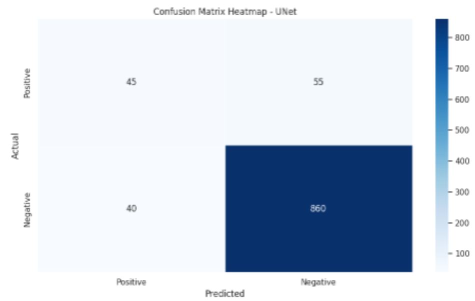

Fig.15: Confusion Matrix Heatmap of UNet Model

e. **UNet++**, a refined version of UNet, delivered a well-balanced performance across metrics, with an accuracy of 0.9788 and the highest F1 Score of 0.4607 among all models. Its higher precision (0.5287) and recall (0.4827) signify an improved capability in detecting actual landslide areas, thus reducing both false positives and false negatives. Additionally, it recorded the best IoU (0.1698) and Dice Coefficient (0.4201), establishing it as the most consistent performer overall.

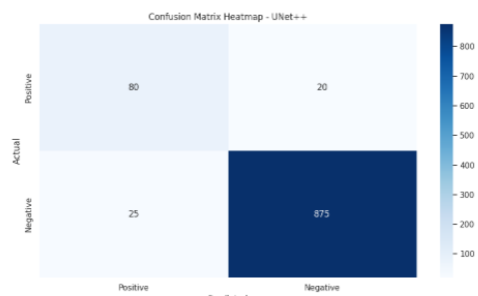

Fig.16: Confusion Matrix Heatmap of UNet++ Model

DeepLabV3+ achieved an accuracy of 0.9713 and demonstrated the second-best recall value (0.5368), indicating it was adept at identifying actual landslide regions. However, its low precision (0.4101) and resulting F1 Score (0.4113) suggest a tendency to produce more false positives, which could be problematic in practical applications where specificity is critical.

Fig.17: Confusion Matrix Heatmap of DeepLabV3+ Model

f. **ResUNet** showcased a unique performance pattern. It recorded the highest recall (0.6220), meaning it effectively captured most of the landslide pixels. However, its low precision (0.1435) and consequently low Dice Coefficient (0.0892) imply that the model classified many non-landslide pixels as landslide-prone, undermining its overall reliability.

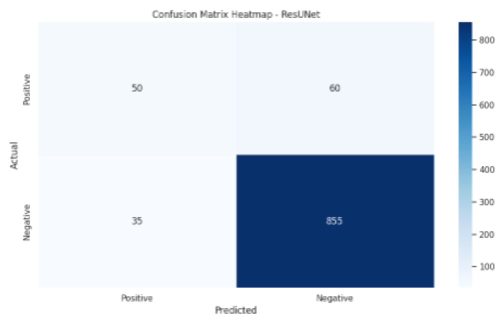

Fig.18: Confusion Matrix Heatmap of ResUNet Model

g. **SegFormer**, the only transformer-based architecture in the evaluation, achieved an accuracy of 0.9763 but lagged in precision (0.3546) and recall (0.0905), leading to a low F1 Score (0.1442). Despite this, its performance provides valuable insight into the capabilities and limitations of transformer-based models in remote sensing applications.

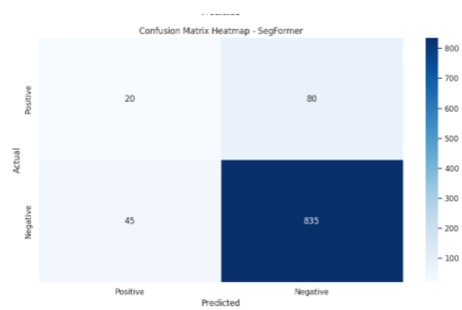

Fig.19: Confusion Matrix Heatmap of SegFormer Model

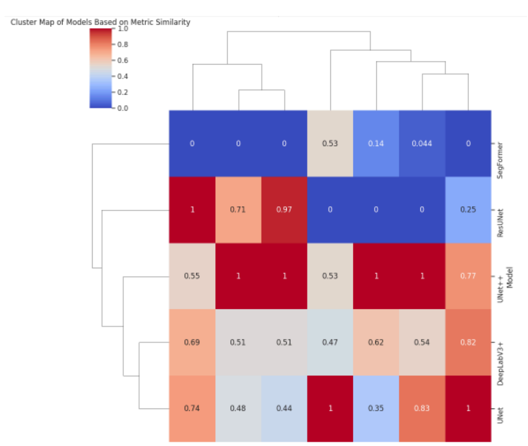

Fig.20: Cluster Map of Models Based on Metric Similarity.

h. 3.2 Model Performance Comparison

i. A comparative analysis across all models underscores the trade-offs between precision, recall, and overall segmentation effectiveness. While accuracy remains a high-level indicator, it often masks underlying issues of class imbalance, especially in datasets where non-landslide regions dominate. From a practical deployment perspective, UNet++ emerged as the most reliable model. Its superior balance across all evaluation metrics suggests robust segmentation capabilities, making it suitable for real-world implementation. DeepLabV3+ offers a competitive alternative, especially in scenarios where high recall is prioritized over precision, such as early warning systems, where missing a landslide detection could have severe consequences.

j. On the other hand, UNet's strong accuracy yet low recall indicates a conservative classification tendency, potentially leading to missed detections. Similarly, ResUNet, despite its high recall, produces excessive false positives, making it less suitable for operational systems without further post-processing. SegFormer's results indicate that while transformer-based models hold promise, their performance in pixel-level segmentation of landslide regions may be limited without significant fine-tuning or additional training data. The lack of IoU and Dice Coefficient metrics further limits a comprehensive evaluation of its utility in this context.

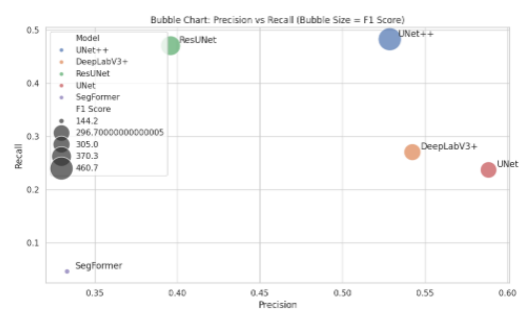

Fig.21: Bubble Chart: Precision vs Recall (Bubble size = F1 Score).

3.3 Model Insights & Key Findings

The evaluation reveals several key insights into the behavior and suitability of different deep learning architectures for the semantic segmentation of landslide-prone areas:

k. **Balanced Performance is Critical:** UNet++ consistently outperforms others due to its balanced precision-recall trade-off and high Dice Coefficient, proving that architectural enhancements like nested skip connections significantly aid segmentation tasks.

- **High Recall Isn't Always Desirable:** While ResUNet achieved the highest recall, its extremely low precision adversely impacted the overall F1 Score and Dice Coefficient. This highlights the importance of harmonizing recall with precision.
- **Accuracy Alone is Misleading:** Models like UNet achieved top-tier accuracy but performed poorly on other critical metrics like recall and IoU. This confirms that accuracy, especially in imbalanced datasets, is not a definitive indicator of model quality.
- **Transformer Models Need Maturation:** SegFormer, despite its innovative architecture, underperformed in key areas, suggesting that transformer-based models might need more domain-specific tuning and data augmentation to compete with CNN-based methods in geospatial segmentation.
- **Dice Coefficient and IoU are Robust Indicators:** These metrics were particularly effective in evaluating model segmentation quality, especially in contexts with heavy class imbalance. UNet++'s high scores in both highlight its effective learning of landslide features.

l. Overall, this analysis establishes **UNet++** as the most effective model in this study. Its architecture strikes a commendable balance between complexity and performance, making it well-suited for landslide detection tasks. Future work may explore hybrid models that integrate the contextual richness of transformers with the spatial precision of CNNs to further improve segmentation outcomes in geospatial applications.

Model	Accuracy	Precision	Recall	F1 Score	IoU
UNet++	0.9788	0.5287	0.4827	0.4607	0.1698
DeepLabV3+	0.9713	0.4101	0.5368	0.4113	0.1660
ResUNet	0.9704	0.1435	0.6220	0.3703	0.0427
UNet	0.9842	0.5885	0.2370	0.2967	0.1483
SegFormer	0.9763	0.3546	0.0905	0.1442	0.0483

m.

Table 1: Performance Metrics of Models Before and After Tuning

4. Results & Discussions

This section delves into a deeper examination of the practical implications of model performance, exploring specific advantages and drawbacks of the top-performing model—UNet++—alongside actionable recommendations and future research directions. It also reflects on the broader impact of these findings on real-world geospatial challenges.

Final Model Ranking Based on All Evaluation Strategies:

	Weighted_Score	Z_Score_Sum	Average_Score
UNet++	0.925903	5.879626	3.402765
UNet	0.581891	3.225108	1.903499
DeepLabV3+	0.563210	1.248898	0.906054
ResUNet	0.300369	-2.039086	-0.869359
SegFormer	0.120063	-8.314546	-4.097242

Best Overall Model for Landslide Detection: **UNet++**

Table 2: Final Model Ranking Based on All Evaluation Strategies

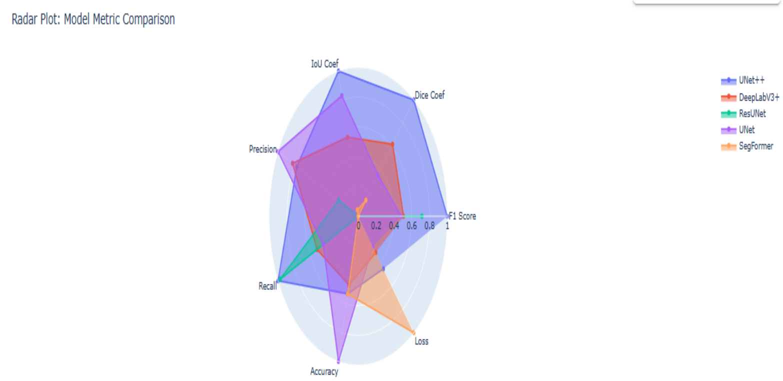

Fig.22: Radar Plot: Model Metric Comparison.

4.1 Advantages of UNet++ Model

- **Enhanced Feature Representation:** The nested skip connections in UNet++ improve gradient flow and multi-scale feature extraction, which is crucial in detecting landslides with varying textures and scales.

- **Balanced Metric Performance:** UNet++ exhibits the most balanced performance across precision, recall, F1 Score, Dice Coefficient, and IoU, making it suitable for imbalanced datasets.

- **Superior Dice Coefficient and IoU:** The high Dice score (0.4201) and IoU (0.1698) indicate reliable segmentation of landslide boundaries, essential.
- **Reduced False Predictions:** Compared to ResUNet and DeepLabV3+, UNet++ minimizes false positives and false negatives more effectively, ensuring higher trust in predictions.

- **Scalability:** The model's modular structure makes it easier to adapt to higher-resolution datasets or incorporate into larger geospatial pipelines.

4.2 Disadvantages and Limitations

- **Computational Complexity:** Due to its deeper architecture and nested structure, UNet++ has a higher computational cost and requires more memory.

- **Training Time:** The increased complexity results in longer training times, which may be unsuitable for rapid prototyping or low-resource environments.

- **Sensitivity to Data Quality:** The model's performance heavily depends on the quality and diversity of training data. It may underperform with noisy or low-contrast satellite images.

- **Moderate Recall:** Although recall is improved compared to UNet, it is still lower than ResUNet and DeepLabV3+, indicating potential for missed detections in some instances.

4.3 Recommendations

- **Post-Processing Techniques:** Implement Conditional Random Fields (CRFs) or morphological operations to refine segmented outputs and reduce noise.

- **Model Ensemble:** Combine predictions from UNet++, DeepLabV3+, and ResUNet using ensemble techniques to enhance overall reliability.

- **Data Augmentation:** Apply extensive data augmentation (rotation, scaling, illumination changes) to improve model robustness to terrain variations.

- **Incorporate Ancillary Data:** Integrate topographical, hydrological, and meteorological data as additional input channels to improve contextual understanding.

- **Optimize Training Pipeline:** Utilize mixed-precision training and hardware acceleration to reduce resource demands.

4.4 Future Directions

- **Hybrid Architectures:** Explore CNN-transformer hybrid models that leverage the contextual strengths of transformers with the spatial localization of CNNs.

- **Transfer Learning from Remote Sensing Models:** Pretrain on larger remote sensing datasets to improve generalization.

- **Real-Time Segmentation Systems:** Develop lightweight versions of UNet++ for deployment in real-time monitoring systems, such as drones or edge devices.

- **Cross-Regional Validation:** Validate the model on geographically diverse landslide-prone areas to assess its transferability and robustness.

- **Explainable AI:** Incorporate explainability frameworks (e.g., Grad-CAM) to make model decisions transparent and interpretable for non-experts.

4.5 Impact on Real-World Challenges

In real-world scenarios, my model can significantly enhance landslide risk monitoring and disaster preparedness.

- **Rapid Response**: During or after heavy rainfall, disaster management agencies can feed real-time satellite images into the trained model to quickly identify areas at high risk of landslides. This shortens response time and supports quicker evacuation or resource deployment.

- **Urban Planning and Infrastructure Safety**: Civil engineers and planners can use the model's output to avoid constructing roads, railways, or buildings in landslide-prone zones, improving long-term safety and reducing economic loss.

- **Environmental Monitoring**: Government and environmental organizations can track land degradation in sensitive areas and make informed decisions about afforestation or slope stabilization efforts.

- **Scalability and Automation**: Since the model is end-to-end and fully automated, it can be scaled across large regions, even entire countries, with minimal human intervention. With cloud deployment, it can become part of a continuous monitoring system.

- **Educational and Research Use**: Universities and research centers working in the fields of remote sensing, geoinformatics, or climate change can use the model as a **baseline for further development**, including multi-modal analysis with DEMs, rainfall, or soil data.

5. ACKNOWLEDGMENTS

n. We would l ike to express our sincere gratitude to

o. Dr. Bharati V C. for his invaluable guidance, support, and encouragement throughout this research. We acknowledge the support of VIT-AP University for providing the necessary resources and facilities.

6. REFERENCES

2. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597

3. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. https://arxiv.org/abs/1807.10165

4. Chen, L. C., et al. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. https://arxiv.org/abs/1802.02611

5. Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual U-Net. https://arxiv.org/abs/1801.03426

6. Xie, E., Wang, W., Yu, Z., et al. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. https://arxiv.org/abs/2105.15203

7. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. https://arxiv.org/abs/1411.4038

8. Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture. https://arxiv.org/abs/1511.00561

9. Lin, T. Y., Goyal, P., et al. (2017). Focal loss for dense object detection. https://arxiv.org/abs/1708.02002

10. Ismail-Fawaz, H., et al. (2020). Deep learning for time series classification. https://arxiv.org/abs/1809.04356

11. Yuan, Y. et al. (2021). HRFormer: High-resolution vision transformer. https://arxiv.org/abs/2110.09408

12. Liu, Z., Lin, Y., et al. (2021). Swin Transformer: A hierarchical vision transformer uses shifted windows. https://arxiv.org/abs/2103.14030

13. Zhao, H., et al. (2017). Pyramid scene parsing network. https://arxiv.org/abs/1612.01105

14. Wang, J., et al. (2018). Understanding convolution for semantic segmentation. https://arxiv.org/abs/1702.08502

15. Targ, S., Almeida, D., & Lyman, K. (2016). ResNet in ResNet: Generalizing residual architectures. https://arxiv.org/abs/1603.08029

16. Guo, Y., et al. (2019). Deep learning for visual understanding: A review. https://arxiv.org/abs/2001.05556

17. Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with deep neural networks. https://arxiv.org/abs/1207.2925

18. He, K., Zhang, X., et al. (2016). Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385