Comparison between Baseline and Adapted Model

The adapted Attention U-Net demonstrates a consistent and statistically significant improvement over the baseline model across nearly all relevant segmentation metrics. Overall accuracy increased from 0.8776 ± 0.0935 to 0.9585 ± 0.0286, indicating more stable and reliable pixel-level predictions. Additionally, task-relevant metrics show substantial gains. Recall improved from 0.8590 to 0.9763, highlighting the adapted model's superior ability to detect deforested areas, particularly small or fragmented regions that were frequently missed by the baseline. This improvement is critical in deforestation monitoring, where false negatives represent undetected forest loss. The F1-score increased from 0.9139 to 0.9735, reflecting a better balance between precision and recall, while Intersection-over-Union (IoU) improved from 0.8494 to 0.9490, indicating a better spatial overlap between predicted and ground-truth deforestation masks. Although precision decreased slightly (from 0.9831 to 0.9717), this change is minor and not statistically significant, suggesting that the increase in recall did not occur because of an increased number of false positives. Overall, the adapted model prioritises correct detection of deforestation while maintaining high spatial accuracy.

Appropriateness of Evaluation Metrics

The selected metrics are well suited to the binary semantic segmentation task and the class imbalance inherent in deforestation detection. The dominance of forest pixels makes accuracy insufficient on its own, however when paired with precision, recall, F1-score, and IoU, it provides useful context. Recall is particularly important, as missing deforested areas can undermine conservation and policy interventions. F1-score captures the trade-off between false positives and false negatives, while IoU directly measures spatial agreement, making it a standard and reliable metric for segmentation quality.

Statistical Significance Testing

A paired t-tests was conducted between the baseline and the adapted model across all metrics, to determine if the observed difference between the model's was due to chance. Statistically significant improvements were observed for accuracy ($p = 6.78\times10^{-5}$), recall ($p = 2.04\times10^{-6}$), F1-score ($p = 1.55\times10^{-4}$), and IoU ($p = 4.83\times10^{-5}$). These results confirm that the adapted model's improvements are robust and systematic, rather than arising from random variation. The lack of a statistically significant difference in precision ($p = 0.226$) further supports the conclusion that recall improvements were not achieved by disproportionately increasing false positives. This reinforces the effectiveness of the combined weighted BCE and focal Tversky loss in guiding learning towards under-represented deforestation pixels.

Failure Case Analysis

The qualitative analysis of the lowest-performing mask shows that the improved baseline model captures the large-scale spatial structure of the region correctly, with errors mostly along boundaries and within small, fragmented regions. The model occasionally misses small isolated components, and smooth irregular ground truth boundaries. This behaviour indicates that the model optimises well for global region overlap but struggles with fine-grained spatial details, which explains why relatively high F1 scores are maintained despite visible differences between the predicted and ground truth masks.

A consistent pattern across these failure cases is systematic under-segmentation, where false negatives occur more frequently than false positives. Predicted masks are typically cleaner and less fragmented than the corresponding ground truth, suggesting the model has a precision-biased that favours conservative predictions. In areas with dense texture and low foreground–background contrast, the model appears to rely more on semantic context than high-frequency spatial information, leading to the suppression of small regions present in the ground truth.