MAIS 202 - PROJECT PROPOSAL

1. Choice of Dataset

Dataset selected: https://www.kaggle.com/datasets/akhilchibber/deforestation-detection-dataset/data

I have chosen the Cloud-Free Dataset for Semantic Segmentation of Deforestation, which includes high-resolution satellite images with PRODES Ground Truth Data (2020) labeling deforested areas. Sentinel-1 captures images through cloud cover, while Sentinel-2 helps distinguish vegetation from cleared land. The dataset's labeled ground truth enables effective training without manual labeling, making it well-suited for deep learning-based segmentation and accurate deforestation detection.

2. Methodology

Machine Learning Model

This project aims to classify deforested areas from satellite imagery, producing segmentation masks that highlight affected regions. To achieve this, I will use DeepLabV3+, a convolutional neural network (CNN) architecture designed for semantic segmentation. I believe DeepLabV3+ to be adequate for this task because it processes images at multiple scales, allowing it to capture both large and small deforested patches with high accuracy and efficiency. Other models such as SegNet and U-Net were considered, but SegNet struggles with high-resolution imagery, while U-Net is generally used for medical imaging.

Data Preprocessing

Before training, the dataset will undergo preprocessing to ensure consistency. All images will be resized to a fixed resolution, and pixel values will be normalized to a [0,1] range for better model performance. Since deep learning models require diverse training samples, data augmentation techniques such as rotations and flips will be applied to improve generalization. The dataset will be split into 80% training and 20% testing to judge model performance. Since the data is labeled, no additional labeling is needed.

Evaluation Metrics

Model performance will be assessed using Intersection over Union (IoU), which measures how well predicted deforestation areas overlap with ground truth. A baseline of IoU ≥ 0.7 (70%) is considered acceptable for this task. Additionally, the Dice Coefficient will be used to balance precision and recall, with a target of ≥ 0.75 (75%). Pixel accuracy will also be reported, though it can be misleading if the dataset is imbalanced. A Confusion Matrix will help visualize misclassified pixels and refine the model.

3. Application & Future Extensions

The final model will be integrated into a web-based tool where users can enter a location, and a Sentinel-2 satellite image will be automatically fetched from Google Earth Engine (GEE) for deforestation classification. If the selected location differs significantly from the model's training data, a warning will be displayed to indicate potential inaccuracies. In the future, I plan to implement historical trend analysis, allowing users to track deforestation changes over time by comparing past and present images. A Deforestation Likelihood Score may also be added to quantify risk based on detected patterns.