Contextually Appropriate Dataset

The Congo Basin was selected as the target region for deforestation detection due to its extensive rainforest coverage, its role as a significant global carbon sink, and its high ecological significance. The dataset consists of Sentinel-2 satellite imagery tiles paired with binary forest–non-forest masks representing forested and deforested areas.

The dataset was generated in Nkoteng, Cameroon in the Congo Basin. This allows the model to learn spatial and spectral differences between intact forest, large-scale industrial deforestation, smaller fragmented clearings, and areas undergoing natural regeneration. The structure and purpose of the dataset closely mirror those used in the original paper, making it appropriate for replicating and evaluating the Attention U-Net methodology in a new geographical context.

Data Collection, Access Processing and Ethical Consideration

Existing curated datasets suitable for deep-learning-based deforestation detection in the Congo Basin are extremely limited. Many regions in the basin remain understudied, and although Cameroon's Ministry of Forestry and Wildlife (MINFOF) is the primary authority responsible for forest monitoring and protected areas, much of the relevant spatial data is confidential and unavailable for individual research use. Additionally, a significant proportion of logging activity in Cameroon is illegal, making it difficult to capture accurate deforestation records through official datasets alone, as these activities are deliberately concealed.

As a solution, the dataset was created using Google Earth Engine (GEE). Sentinel-2 Surface Reflectance imagery from 2021 was collected for a defined area of interest containing intact forest, large agro-industrial clearings, smaller deforested patches, and regenerating zones. The ground truth masks were generated using the ESA WorldCover 10 m land-cover product, where the "tree cover" class was extracted to create a binary forest/non-forest mask aligned spatially and temporally with the Sentinel-2 imagery.

From an ethical standpoint, the data used are openly accessible, satellite-derived products that do not contain personal or community-identifiable information. However, ethical considerations remain critical, as spatial datasets can influence land governance and resource control. The generated maps must therefore be interpreted carefully to avoid legitimising coercive land acquisition or disproportionately targeting subsistence land use. Transparency in data generation and limitations is essential, and the dataset is intended strictly for research and policy-support purposes rather than enforcement without local consultation. Ensuring responsible use requires acknowledging uncertainties in the ground truth data and avoiding conclusions that could exacerbate existing power imbalances in land governance.

Data Preprocessing Timeline

The Sentinel-2 imagery and corresponding ground truth masks were processed using a structured preprocessing pipeline (found in file ‘Generate_tiles_and_masks’) to ensure compatibility with the Attention U-Net model. First, the four-band Sentinel-2 composite (Red, Green, Blue, and Near-Infrared) and the binary forest mask were exported from Google Earth Engine at 10 m resolution using a consistent coordinate reference system. The raster data were then read using Rasterio and converted into NumPy arrays. Pixel values were optionally normalised to a 0–1 range using percentile-based clipping to reduce the influence of outliers.

Both imagery and masks were split into identical 512 × 512 pixel tiles. Tiles containing NaN values, near-zero pixels, or uninformative masks (almost entirely forest or non-forest) were automatically filtered out to prevent the model from learning from non-representative data. This step was crucial for removing edge tiles and areas outside the effective region of interest. Finally, the valid tiles were randomly split into training, validation, and test sets using a 60%, 24%, 5% ratio, respectively. Images and masks were stored in separate directories following a consistent naming convention to ensure correct pairing during model training. This preprocessing pipeline ensured data quality, reduced noise, and produced a balanced dataset suitable for the model and segmentation task.