# **Project Name: GreenLens**

# Group - 5 Members:

* Fatu Kromah
* Alfred Nyeswa
* Sulaiman Barry
* Jerome N. Tokpa
* Joseph S. Lablah Jr

**Problem Definition**

**Problem Statement**: Liberia faces significant deforestation due to logging, agriculture, and mining activities. Current monitoring methods have long delays, making it difficult to respond quickly to forest loss events.

**ML Problem Type**: **Classification Problem**

* **Classes**:
  + Forest (intact vegetation)
  + Non-forest (agriculture, urban, water)
  + Forest loss (recently cleared areas)

**Objective**: Build a machine learning model to classify satellite imagery pixels into these three categories, enabling near real-time forest monitoring.

## **Data**

**Primary Data Sources**:

* **Sentinel-2 satellite imagery** via Google Earth Engine (10-20m resolution, 13 spectral bands)
* **Landsat imagery** via Google Earth Engine (30m resolution, 11 spectral bands)
* **Global Forest Watch tree cover loss datasets** (ground truth labels)
* **Liberia administrative boundaries** from GADM

**Exploratory Data Analysis (EDA) Plan**:

1. **Statistical Summary**:  
   * Distribution of forest vs non-forest vs forest loss pixels
   * Spectral band statistics (mean, std, min, max) for each class
   * Temporal patterns in the data (seasonal variations)
2. **Missing Data Examination**:  
   * Cloud cover frequency and spatial distribution
   * Data gaps in satellite imagery time series
   * Quality of ground truth labels across different regions
3. **Spatial Analysis**:  
   * Geographic distribution of forest loss events
   * Correlation between deforestation and proximity to roads/settlements
   * Class imbalance assessment (forest loss is typically <5% of pixels)

## **Preprocessing**

**Data Cleaning Steps**:

1. **Cloud and Shadow Masking**:  
   * Remove cloudy pixels using QA bands
   * Apply shadow detection algorithms
   * Fill gaps using temporal interpolation
2. **Feature Engineering**:  
   * Calculate vegetation indices (NDVI, EVI)
   * Create seasonal composites to reduce noise
   * Extract texture features from infrared bands
3. **Missing Values**:  
   * Handle cloud-masked pixels through interpolation
   * Remove tiles with >30% missing data
   * Use multi-temporal composites for gap-filling
4. **Data Preparation**:  
   * Tile imagery into 512×512 pixel patches
   * Normalize spectral bands (0-1 scaling)
   * Balance classes using stratified sampling
   * Split into 70% training, 15% validation, 15% test

**Visualizations**:

* Spectral signatures for each class
* Correlation matrix of input features
* Geographic distribution maps
* Time series plots of vegetation indices

## **Methodology**

**Model Selection**:

1. **Baseline Model: Random Forest**
   * **Rationale**: Good for tabular data with mixed feature types, handles feature interactions well
   * **Input Features**: Spectral bands, vegetation indices, texture features (~20 features total)
   * **Hyperparameters**: Number of trees, max depth, min samples split
2. **Advanced Model: EfficientNet-Lite (CNN)**
   * **Rationale**: Can capture spatial patterns and relationships between neighboring pixels
   * **Input**: Multi-band satellite images (RGB + NIR + SWIR channels)
   * **Transfer Learning**: Pre-trained weights adapted for satellite imagery

**Cross-Validation Strategy**:

* **5-fold spatial cross-validation** to prevent data leakage
* Ensure folds are geographically separated
* Stratified sampling to maintain class balance in each fold

**Evaluation Metrics**:

* **Primary**: F1-score (handles class imbalance well)
* **Secondary**: Precision and Recall for each class
* **Additional**: Confusion matrix, ROC curves
* **Target Performance**: F1-score ≥ 85%

## **Experiments**

**Experiment Design**:

1. **Model Comparison**:  
   * Compare Random Forest vs EfficientNet-Lite performance
   * Analyze which features are most important for classification
   * Test different input band combinations
2. **Hyperparameter Tuning**:  
   * **Random Forest**: Grid search on n\_estimators (100, 300, 500), max\_depth (10, 20, None)
   * **EfficientNet**: Learning rate (1e-3, 1e-4), batch size (16, 32), dropout rate (0.2, 0.5)
3. **Feature Engineering Experiments**:  
   * Test different vegetation indices combinations
   * Compare seasonal vs annual composites
   * Evaluate texture features importance
4. **Data Augmentation** (for CNN):  
   * Test rotation, flipping, and brightness adjustments
   * Assess impact on model generalization

**Creative Approaches**:

* Ensemble methods combining Random Forest + CNN predictions
* Multi-temporal analysis using time series of images
* Active learning to identify most informative samples for labeling

## **Conclusion**

**Expected Results**:

* Achieve F1-score of at least 85% for forest loss detection
* Random Forest likely to perform well on spectral features
* CNN expected to excel at capturing spatial patterns
* Ensemble approach may provide best overall performance

**Performance Analysis Plan**:

* Compare models across different regions in Liberia
* Analyze false positives/negatives to understand model limitations
* Evaluate computational efficiency (training time, prediction speed)
* Assess model interpretability and feature importance

**Findings Discussion**:

* Identify which satellite bands are most informative
* Understand seasonal effects on model performance
* Document challenges with cloud cover and data quality
* Analyze geographic patterns in model accuracy

## **Future Work**

**Model Improvements**:

* **New Data Integration**:
  + Add radar data (Sentinel-1) for cloud-free monitoring
  + Incorporate higher resolution imagery for validation
  + Include socioeconomic data (roads, population density)

**Algorithm Enhancements**:

* Implement time series analysis for gradual forest degradation
* Develop change detection algorithms for monitoring forest recovery
* Experiment with newer architectures (Vision Transformers, U-Net for segmentation)

**Operational Considerations**:

* **Model Retraining**: Set up pipelines for regular model updates with new data
* **Drift Detection**: Monitor model performance over time and geographic regions
* **Scalability**: Optimize for processing larger areas efficiently

**Adaptation Strategy**:

* Plan for seasonal variations in model performance
* Prepare for new deforestation patterns (mining vs agriculture)
* Consider transfer learning to other West African countries

**Validation and Monitoring**:

* Establish ground truth collection protocols
* Set up continuous evaluation with field data
* Plan for stakeholder feedback integration

**Impact Measurement**:

* Track model usage by government agencies
* Monitor reduction in forest loss detection time
* Assess contribution to conservation decision-making