Machine Learning Project Documentation

# Group 5 – GreenLens

Team Members:  
- Fatu Kromah  
- Alfred Nyeswa  
- Sulaiman Barry  
- Jerome N. Tokpa  
- Joseph S. Lablah Jr

# Model Refinement

## 1. Overview

The model refinement phase focused on enhancing the predictive performance of the GreenLens forest monitoring model. After initial exploration using Random Forest and EfficientNet-Lite, refinement targeted improved accuracy, generalization across regions, and reduced misclassification in cloud-prone and dry-season imagery.

## 2. Model Evaluation

Initial evaluations showed Random Forest achieving an F1-score of approximately 82%, and EfficientNet-Lite achieving around 87%. Misclassifications primarily occurred in regions with mixed vegetation or cloud artifacts. Improvement areas included handling data imbalance, hyperparameter tuning, and enhanced feature selection.

## 3. Refinement Techniques

- Hyperparameter tuning through grid search and learning rate optimization.  
- Integration of Sentinel-1 radar data for cloud-independent monitoring.  
- Ensemble combination of Random Forest and CNN outputs.  
- Class reweighting to address rare forest loss events.  
- Spatial cross-validation to assess regional robustness.

## 4. Hyperparameter Tuning

Random Forest tuning increased the number of estimators from 500 to 800 and set max depth to 25, resulting in an improved F1-score of 85%. EfficientNet-Lite tuning adjusted the learning rate to 0.0008 and batch size to 32, improving validation accuracy to 89%. Dropout was raised from 0.3 to 0.4 to reduce overfitting.

## 5. Cross-Validation

Spatial cross-validation was adopted using Liberia’s counties (Lofa, Nimba, Sinoe, etc.) as folds, replacing random splits. This improved the reliability of model performance by ensuring evaluation on unseen geographic areas.

## 6. Feature Selection

Permutation importance and SHAP analysis identified NDVI drop, NBR, SWIR1, and proximity to roads as the top predictors. Bands B2 and B12 were removed due to low contribution, leading to slightly faster training and improved model stability.

# Test Submission

## 1. Overview

The test submission phase validated the refined model’s performance on unseen Sentinel-2 and GFW data representing new forest loss events across Liberia.

## 2. Data Preparation for Testing

The test dataset consisted of 10% holdout samples covering Grand Gedeh and Bomi counties. Cloud masking, NDVI, NBR, and EVI indices were recomputed to maintain consistency with training data.

## 3. Model Application

The trained ensemble model (Random Forest + CNN) was applied to test tiles, generating probabilities for each class: Forest, Non-Forest, and Forest Loss. Predictions were mapped and validated visually using Google Earth Engine overlays.

## 4. Test Metrics

Performance metrics for the final test evaluation:

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Train | Validation | Test |
| Accuracy | 0.92 | 0.89 | 0.88 |
| F1 (macro) | 0.90 | 0.87 | 0.86 |
| Recall (Forest Loss) | 0.83 | 0.81 | 0.80 |
| ROC-AUC | 0.94 | 0.92 | 0.91 |

## 5. Model Deployment

The CNN model was converted to TensorFlow Lite for deployment on mobile or Raspberry Pi devices. Integration with Google Earth Engine enabled real-time visualization and forest loss alerts for Liberia’s EPA.

## 6. Code Implementation

Key refinement and testing code examples:  
  
***Random Forest Grid Search***from sklearn.model\_selection import GridSearchCV  
params = {'n\_estimators':[500,800,1000], 'max\_depth':[15,20,25]}  
grid = GridSearchCV(rf, param\_grid=params, scoring='f1\_macro', cv=3, n\_jobs=-1)  
grid.fit(X\_train, y\_train)  
best\_rf = grid.best\_estimator\_  
  
***CNN Refinement***model.compile(optimizer=tf.keras.optimizers.Adam(learning\_rate=8e-4),  
 loss='sparse\_categorical\_crossentropy', metrics=['accuracy'])  
model.fit(train\_ds, validation\_data=val\_ds, epochs=15, callbacks=[early\_stop])

# Conclusion

The GreenLens refined model achieved 88% test accuracy and an 86% macro F1-score, demonstrating strong potential for near real-time forest monitoring in Liberia. Key challenges included data imbalance, limited ground truth, and cloud contamination. The refined system shows readiness for field deployment and policy integration.

# References

- Hansen, M. C., et al. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. Science, 342(6160), 850–853.  
- Reiche, J., et al. (2018). A Framework for Near Real-Time Forest Monitoring with Sentinel-1. Remote Sensing of Environment, 215, 12–25.  
- Belgiu, M., & Csillik, O. (2018). Sentinel-2 for Land Cover Classification Using Random Forest. Remote Sensing, 10(2), 247.  
- Zhu, X., et al. (2020). Deep Learning in Remote Sensing: A Review. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 156–177.  
- Google Earth Engine (2023). Planetary-Scale Geospatial Analysis. https://earthengine.google.com