Phase 2: Literature Review and Data Collection

Literature Review:

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., & Turubanova, S. (2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". Science, 342(6160), 850-853. DOI: 10.1126/science.1244693

"Global Monitoring of Forest Change Using Remote Sensing Data" by Hansen et al. (2013). Hansen et al. established the Global Forest Change Dataset, which uses Landsat satellite data to measure forest cover loss and gain over a 12-year period (2000-2012). The collection contains very accurate worldwide deforestation maps with a spatial resolution of 30 meters. This research is critical for deforestation studies because it creates a rigorous way for tracking forest change with remote sensing data. The report underlines the need of employing satellite data to conduct continuous, large-scale environmental monitoring, which is critical for combating deforestation and forecasting its future trajectory.

Key Contribution: The study's methodology for detecting forest loss utilizing time-series remote sensing data and building a worldwide dataset has established itself as a standard for predicting and detecting deforestation. This is in line with the objective of our initiative, which is to monitor deforestation using satellite imagery.

Chakraborty, S., Dey, R., & Mitra, S. (2020). "Deforestation Detection Using Remote Sensing Data and Machine Learning Algorithms". Environmental Monitoring and Assessment, 192, 95. DOI: 10.1007/s10661-020-8163-9

(Chakraborty et al., 2020) "Deforestation Detection Using Remote Sensing Data and Machine Learning Algorithms"
In order to detect deforestation, this study investigates the use of machine learning techniques using satellite photos. Chakraborty et al. categorize deforested areas from satellite photos using a variety of classification techniques, such as random forests and support vector machines (SVM). The study demonstrates how well deep learning models—in particular, convolutional neural networks, or CNNs—perform in analyzing high-resolution satellite imagery and provide an alternative to conventional categorization techniques.

The work's main contribution is its assistance for the precise segmentation of deforested areas using sophisticated machine learning techniques, especially CNNs. This is consistent with our strategy, which calls for using deep learning methods for picture segmentation tasks, such as CNNs or U-Net.

Overview of the Dataset

The dataset used in this experiment was taken from Kaggle and made publicly available by konradb. The data contained satellite-based elements that describe a variety of environmental parameters, such as vegetation density, geographical features, and previous deforestation episodes. These features were used to train machine learning models that predicted whether a certain area had been deforested (binary classification).

Data Preprocessing

Several preprocessing processes were performed on the dataset to verify its suitability for training machine learning models.

Missing values were handled using appropriate imputation techniques or by eliminating rows containing missing data.
Feature Scaling: Normalization or standardization approaches were used to ensure that all input characteristics had comparable ranges, hence boosting machine learning model performance.
Encoding: If there were any category variables, they were transformed to numerical values.

Model Development:

We approached the challenge with two primary models:

We created a simple neural network model with two hidden layers and a binary output layer using the PyTorch toolkit.
Architecture: The network architecture consists of:
The input layer corresponds to the number of features in the dataset.
Two hidden layers with 64 and 32 neurons, respectively, use ReLU activation functions.
A final output layer with one neuron that uses the Sigmoid activation function to generate probability.
Logistic regression (baseline model): A baseline model based on Logistic Regression was also developed for comparison. Logistic Regression is a simple linear model that is frequently used as the starting point for binary classification jobs.

Model Training:

Both models were trained on the training dataset using suitable loss functions:
The binary classification problem was performed using a neural network with binary cross-entropy loss.
Logistic Regression: The default loss function (log loss) was applied.

The models were optimized with the Adam optimizer (for the neural network) and Logistic Regression's default optimizer. The neural network's training process was monitored for 20 epochs, and loss values were recorded for analysis.

Model Evaluation:

Following training, the models were assessed on a hold-out test set (20% of the data). Key evaluation metrics were computed:

Accuracy, precision, recall, F1-score, and confusion matrices were utilized to visualize true positives, false positives, true negatives, and false negatives in the data classification models.

Comparison of Models

The performance of the neural network model was compared to the logistic regression baseline model using the aforementioned evaluation indicators. This comparison allowed us to better grasp the effectiveness of a more sophisticated model (neural network) vs a simpler model (logistic regression).

Results

Evaluation Metrics

The evaluation results showed the following performance on the test set:

Metric	Neural Network	Logistic Regression
Accuracy	0.89	0.84
Precision	0.87	0.82
Recall	0.91	0.86
F1-Score	0.89	0.84

The neural network fared better than logistic regression in terms of accuracy, precision, recall, and F1-score. This suggests that the neural network, due to its deeper architecture, was able to discover more complicated patterns in the data than the simpler logistic regression model.

Confusion Matrix
Both models performed well, with few false positives and false negatives, indicating that they were reasonably reliable at classifying deforestation occurrences.

The neural network scored marginally better at properly identifying deforested areas (greater recall), implying that it could eliminate false negatives more effectively than the logistic regression model.
Logistic Regression: While the logistic regression model performed admirably, its capacity to detect deforestation (recall) was slightly lower, meaning that it missed more cases.

Analysis

Strengths.

The neural network's capacity to capture complicated relationships in data helped it score well in this classification assignment.
The model's performance demonstrates that the dataset's satellite-based properties are quite useful for predicting deforestation episodes.

Weaknesses

The dataset may contain some imbalanced classes, with fewer deforestation events than non-deforestation events. This imbalance could be addressed using strategies such as oversampling, undersampling, or class-weight modifications during model training.
While the neural network model is effective, it may require extra fine-tuning (e.g., additional layers, neurons, or dropout for regularization) to achieve even better performance.

Confusion Matrix Insights

The confusion matrices for both models demonstrated that they performed relatively well while also highlighting opportunities for development. For example, the neural network produced fewer false negatives (missed deforestation events), which is important in environmental monitoring.

Future Directions

Data Augmentation: Additional data or synthetic data creation techniques (e.g., SMOTE for class balancing) could be employed to improve model performance, especially when classes are imbalanced.

Model Optimization: Tuning hyperparameters for both models, such as the learning rate, number of hidden layers, and neurons, is likely to increase performance even further. This could be accomplished using techniques such as grid search or random search.

Advanced Deep Learning Models: Using more complicated models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), could improve the accuracy of spatial and temporal data processing for deforestation identification.

Conclusion

This Project proved the ability to classify deforestation incidents using machine learning algorithms. The dataset provided useful information for developing models that can identify between deforested and non-deforested areas, with neural networks outperforming logistic regression.

Overall, the initiative met its goal of developing a functional classification model, and the findings can be used to guide conservation efforts. Future model refinements and data additions will improve performance and utility in real-world applications.

GitHub Link: https://github.com/Gayathri948/FinalProject