remote sensing

MDPI

Article

Hybrid Human–AI Model Fusion for Deforestation Detection in the Brazilian Legal Amazon

Rodrigo Schuller [1,*], Francisco Ganacim [1], Paulo Orenstein [1], Larissa S. V. B. Amorim [2], Raissa F. Ferreira [2], Manoela Athaide [2] and Carlos M. Souza Jr. [2]

[1] Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro 22460-320, RJ, Brazil; ganacim@impa.br (F.G.); pauloo@impa.br (P.O.)
[2] Amazon Institute for People and the Environment (Imazon), Travessa Dom Romualdo de Seixas 1698, Belém 66055-200, PA, Brazil; larissa.amorim@imazon.org.br (L.S.V.B.A.); raissa@imazon.org.br (R.F.F.); manoela@imazon.org.br (M.A.); souzajr@imazon.org.br (C.M.S.J.)
* Correspondence: rodrigo.loro@impa.br

Highlights

What are the main findings?

- A human-in-the-loop workflow was integrated into SAD, an established deforestation monitoring pipeline.
- Average annotation time per alert was 45% lower during deployment than in the historical comparison period.

What is the implication of the main findings?

- Spatial and temporal predictions can support expert alert verification while making each data source's contribution explicit.

Abstract

Frequent remote detection of deforestation is essential for rainforest conservation but remains challenging due to seasonal vegetation dynamics, biome heterogeneity, cloud cover, and anthropogenic disturbances. Traditional monitoring requires hundreds of expert hours monthly; deep learning reduces this workload but often lacks interpretability and expert-level reliability under strong class imbalance. We present a hybrid human–AI workflow integrating interpretable machine learning with expert validation to accelerate deforestation monitoring in the Brazilian Legal Amazon. The system was trained on 2.8 million manually labeled Sentinel-2 image pairs, each containing a candidate deforestation event preselected by automated change detection, plus vegetation-index time series and distance to the nearest water body. Our approach combines Convolutional Neural Networks for spatial features with a Recurrent Neural Network for temporal dynamics, fused through logistic regression to enable transparent estimation of feature-source contributions. During AI-assisted deployment, the model achieved F1 = 0.75 (precision 0.81, recall 0.70), ROC AUC 0.993, and PR AUC 0.781 against final analyst-validated labels. These AI predictions feed a graphical user interface that supports expert validation, cutting manual annotation time by 45 %. Deployed since October 2024 in Imazon's Sistema de Alerta ao Desmatamento (SAD), these results demonstrate that interpretable model fusion with expert oversight enables scalable, accurate, and transparent tropical forest monitoring.

Keywords: deforestation detection; Sentinel-2 time series; deep learning; human-in-the-loop; explainable artificial intelligence; Brazilian Legal Amazon

Check for updates

Academic Editor: Huaqiang Du

Received: 4 September 2026
Revised: 30 October 2026
Accepted: 13 November 2026
Published: 21 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The Brazilian Legal Amazon (BLA) spans about 5 million km², roughly two-thirds of Brazil's territory, and is critical for biodiversity conservation, climate regulation, and atmospheric moisture recycling [1]. About half of the rainfall in the region is returned to the atmosphere through evapotranspiration, underscoring the forest's essential role in maintaining regional hydrological cycles [2]. The BLA includes the entire Amazon biome within Brazil as well as portions of the Cerrado and Pantanal [3]. Despite its ecological importance, deforestation remains a major threat because it generates substantial carbon emissions and increases forest susceptibility to fire, thereby intensifying the effects of extreme droughts [4]. Although annual deforestation rates vary, they remain persistently high. The expansion of cattle ranching is a primary driver of deforestation and land-use change in the region. Between 1985 and 2022, an estimated 52.7 million hectares were deforested in the Brazilian Amazon biome, equivalent to approximately 12.5% of its total area [5]. As a result, ongoing land-use change has disrupted hydrological processes, increased greenhouse gas emissions, and moved the Amazon closer to an ecological tipping point with potentially irreversible consequences [1,6].

Accurate deforestation alerts are crucial preventive tools for legislative bodies and agencies responsible for overseeing natural resources. Three large-scale Brazilian programs currently monitor deforestation in the BLA: the Deforestation Alert System (SAD) [7], run by the non-governmental organization Amazon Institute for People and the Environment (Imazon); and two governmental initiatives: the Project for Satellite Monitoring of Deforestation in the Brazilian Amazon (PRODES) [8] and the Deforestation Tracking System in Real Time (DETER) [9]. International efforts include Global Land Analysis & Discovery (GLAD) [10] and MAAP [11]. Yearly programs such as PRODES use high-resolution imagery from the dry season (May to October) to produce what is widely considered the gold standard for annual deforestation estimates in the Amazon. Its lower monitoring frequency enables the use of more resource-intensive techniques that rely heavily on visual interpretation and expert analysis, resulting in greater accuracy. While such annual assessments are essential for measuring deforestation with high confidence, they are less suited to guiding immediate enforcement actions. This role is instead filled by higher-frequency alert systems.

Monthly deforestation monitoring initiatives generate thousands of alerts. These include monthly reports from SAD and DETER, as well as near real-time alerts from GLAD and MAAP. SAD and DETER rely on thousands of hours of visual interpretation each year. GLAD uses decision-tree models to generate alerts, while MAAP combines alert analysis with expert interpretation of satellite imagery. These reports provide authorities with timely insights that can inform short-term investigative strategies.

Despite these monitoring efforts, distinguishing deforestation from other changes in the forest remains challenging. Rainfall varies with the seasons and over longer periods as sea-surface temperatures change in the Pacific and Atlantic Oceans [6]. These changes can cause droughts and floods that damage or kill vegetation, creating natural canopy openings that are difficult to distinguish from deforestation in satellite images [12]. Disturbances associated with selective logging and road construction can produce signals that are also difficult to distinguish from forest clearing [7]. Further complicating detection, persistent cloud cover obscures optical satellite observations [13]. Figure 1 illustrates this ambiguity: before-and-after imagery suggests deforestation, whereas the longer time series reveals a seasonal pattern.

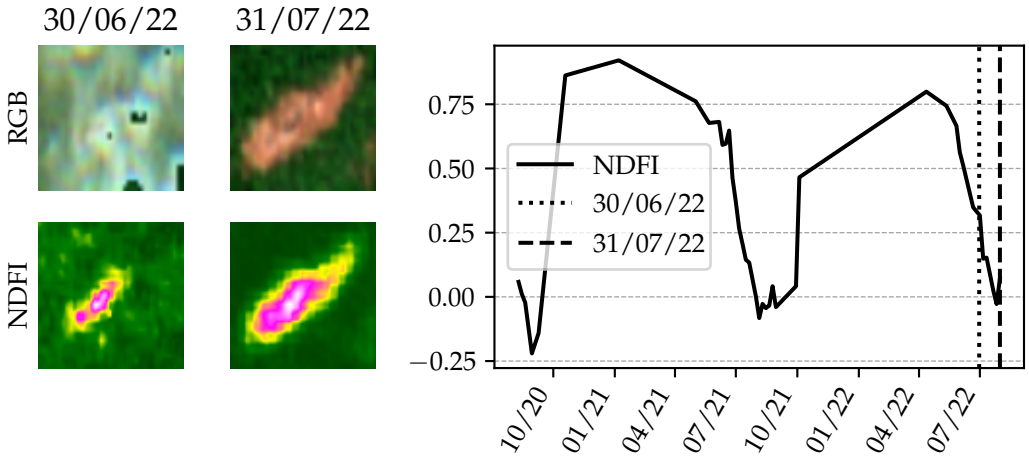

Figure 1. Importance of temporal context for distinguishing deforestation from seasonal canopy change in the Brazilian Legal Amazon (BLA). Before-and-after imagery shows an apparent canopy disturbance that could be interpreted as deforestation based on visual evidence alone (**left**). However, the corresponding two-year mean NDFI time series over the region (**right**) reveals a cyclical pattern consistent with seasonal vegetation dynamics rather than permanent forest loss. This example illustrates the limitations of image-based classification without temporal context and motivates the use of spatiotemporal information for deforestation detection. In the NDFI panels, green indicates high NDFI values (intact canopy), while yellow to magenta indicates progressively lower values (canopy disturbance).

Classical machine-learning methods remain in use in established forest-monitoring systems. Random forests have been used to classify forest changes from Landsat time-series features [14] and to combine change-detection outputs with environmental information [15]. More recently, Cué La Rosa et al. [16] found gradient boosting competitive with a convolutional neural network for forecasting deforestation across 17 tropical countries, with only a modest global performance advantage for the neural model. Benchmarks on medium-sized tabular datasets show that tree ensembles can outperform neural networks while requiring less training time [17].

Convolutional neural networks (CNNs) offer an alternative by learning spatial patterns directly from satellite imagery. De Bem et al. [18] found that they outperformed random forests in annual deforestation mapping in the Brazilian Amazon, producing fewer missed detections and cleaner boundaries. Irvin et al. [19] similarly found that ForestNet outperformed random forests when classifying deforestation drivers in Indonesia, combining image predictions with auxiliary environmental features. In both studies, CNNs that learn spatial context outperformed random forests trained on pixel values or regional summary statistics.

Vision transformers (ViTs) learn relationships between image patches through attention. Kaselimi et al. [20] introduced ForestViT for multilabel classification of Amazon satellite imagery, improving recognition of minority labels compared with the tested CNNs. Soto Vega et al. [21] compared a transformer-based model with a CNN for deforestation detection at three Brazilian sites. The transformer achieved higher F1 when trained and evaluated within each site, whereas the CNN performed better in four of six cross-site evaluations.

These approaches combine several kinds of information. ForestNet combines imagery with environmental variables [19], ForestViT integrates information across spectral bands and image patches [20], and Slagter et al. combine satellite imagery with disturbance masks and dates [22]. However, their predictions do not explicitly show how each input contributes to the decision. Showing how each data source influences the prediction would help analysts understand the model's decision.

Human-in-the-loop workflows involve experts in checking predictions and improving training data. Torres et al. [23] evaluated CNN-generated deforestation maps in Acre and Amazonas through expert auditing under the PRODES protocol. Their results supported using these maps as a starting point for analysts, while highlighting missed clearings and confusion between forest degradation and deforestation that required correction. Masolele et al. [24] used model uncertainty to select images for human annotation and retrained their model with the additional labels. Across two rounds of annotation and retraining, macro-F1 for classifying land uses following deforestation increased from 43% to 84%. Together, these studies show how model predictions can assist expert annotation and how expert annotations can, in turn, improve models through successive rounds of training.

The practical question is how to bring these advances into routine monitoring, where analysts must assess model predictions alongside the satellite evidence. Here, we integrate interpretable model fusion into SAD's established alert-verification workflow. Our aim is to help analysts use complementary spatial and temporal information while making each data source's contribution to the prediction explicit. We compare the model with random forests and gradient boosting to assess its predictive performance, then evaluate its use during a ten-month operational deployment. The study thus examines interpretable assistance in practice and at scale, considering both classification performance and analyst review time while retaining human responsibility for the final labels.

2. Materials and Methods

2.1. Study Area

The BLA is an administrative subregion that encompasses 60% of the Amazon rainforest, parts of the Cerrado and Pantanal biomes, and covers over 5 million km² (see Figure 2). This area is the subject of near-real-time deforestation monitoring by Imazon's SAD, which is of particular interest in this study. Figure 2 also shows the distributions of deforestation alerts (A) and the dominant class of false-positive detections (B) obtained in 2021 and 2022.

2.2. SAD (Deforestation Alert System)

The Deforestation Alert System (SAD) provides monthly monitoring of deforestation and forest degradation across the Brazilian Legal Amazon. It was first implemented in 2006 as a pilot program in the states of Pará and Mato Grosso, where deforestation alerts were generated from MODIS imagery. Since 2008, SAD has operated across the entire Legal Amazon, reporting both deforestation and forest degradation alerts. Currently, the SAD monitoring program relies on remote sensing and cloud-computing-based methods and runs on Google's geospatial analysis platform (Google Earth Engine; Google LLC, Mountain View, CA, USA). To support this monitoring, Imazon uses Landsat 8 and 9 imagery provided by the National Aeronautics and Space Administration (NASA), together with Sentinel-1 and Sentinel-2 imagery from the European Space Agency (ESA). Sentinel-1 microwave data are used to complement monitoring in areas affected by persistent cloud cover. When combined, these satellite sources allow the same area to be revisited every 5 to 8 days. Further details of the SAD change-detection algorithm are provided elsewhere [5]. Examples of SAD deforestation alerts and false-positive detections are shown in Figure 3.

The SAD methodology comprises three main stages: (i) pre-processing, (ii) change detection, and (iii) human validation. In the pre-processing stage, satellite imagery from Landsat 8 and 9, Sentinel-1, and Sentinel-2 is used to generate monthly mosaics. To construct them, Imazon uses proprietary cloud filtering techniques that preserve historical consistency and handle missing pixels. The SAD change-detection algorithm, which is based on the Normalized Difference Fraction Index (NDFI), is then applied to these monthly mosaics to identify changes in forest condition [5,25]. Temporal comparisons

between successive mosaics enable the detection and delineation of forest-cover changes associated with deforestation and forest degradation. In this study, we focus exclusively on deforestation detection. Change detection is performed by applying a threshold to the temporal differences between NDFI mosaics, generating deforestation alerts. The NDFI ranges from −1, indicating absence of vegetation, to 1, indicating full vegetation cover, and is more sensitive to canopy change than conventional vegetation indices [5,7]. This sensitivity is particularly important for detecting subtle canopy disturbances, including forest degradation, as monitored operationally by Imazon. Finally, each alert produced by the change-detection algorithm is reviewed by human experts, who classify it as either true deforestation or a false positive. Figure 3 shows examples of SAD alerts and false positives, including one caused by residual cloud contamination (panel E).

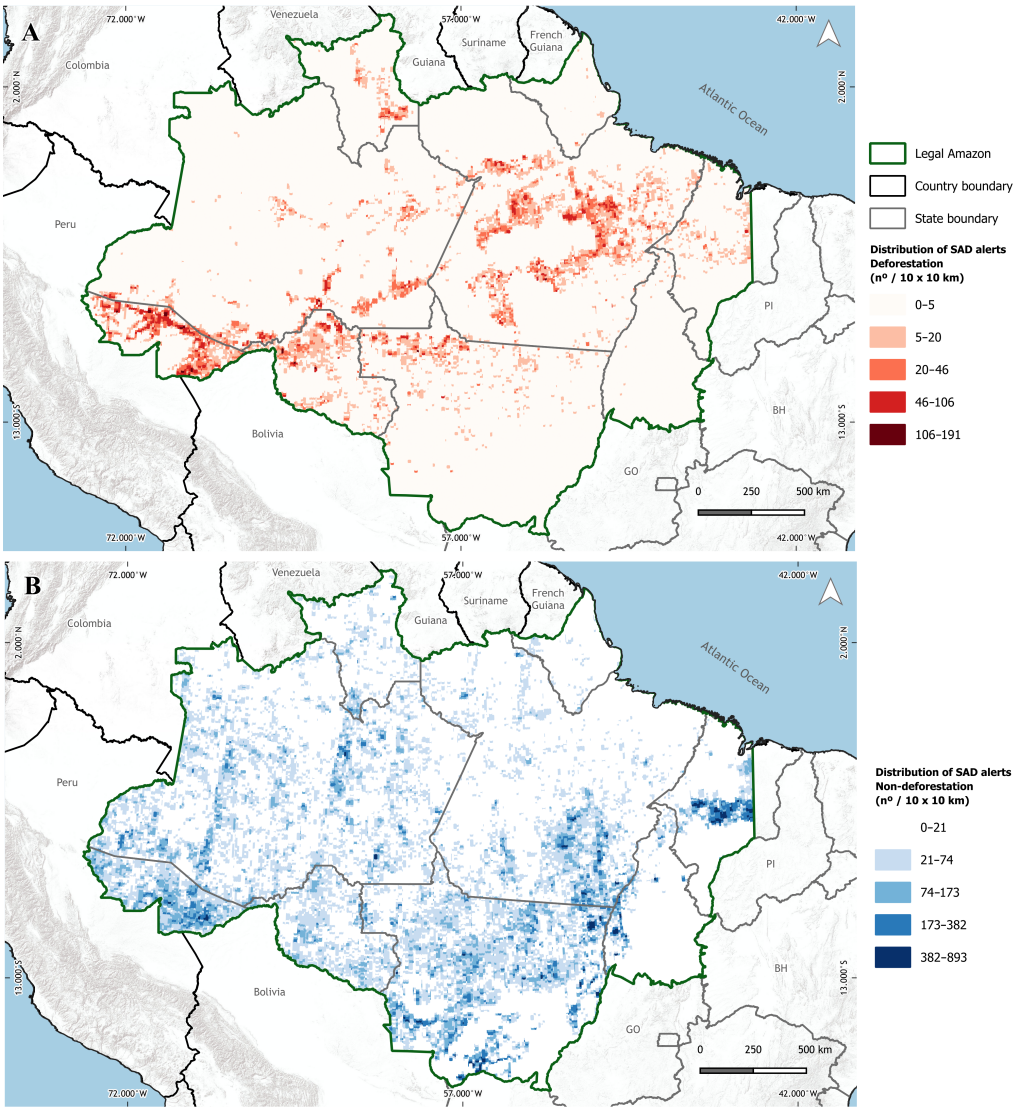

Figure 2. Study area covering the Brazilian Legal Amazon (BLA) and the spatial distribution of human-validated SAD alerts. (**A**) Density of SAD deforestation alerts issued between 2021 and 2022, aggregated to a 10 × 10 km grid and expressed as the number of alerts per cell. (**B**) Spatial distribution of alerts classified through human validation as non-deforestation, corresponding to false-positive detections.

Although deforestation detection is the primary focus of this study, distinguishing deforestation from forest degradation is essential for understanding both the complexity

of the classification task and the scope of the analysis. Deforestation refers to the complete removal of forest canopy cover, typically occurring over a relatively short period and often followed by land clearing for subsequent land-use change. Forest degradation, by contrast, involves a partial reduction in canopy cover caused by disturbances that do not fully remove the forest but alter its structure, composition, and ecological functioning [26]. Within the SAD system, degradation is primarily associated with two processes: (i) selective logging, which creates canopy openings, access trails, and log storage yards within otherwise forested areas; and (ii) forest fires, which can compromise vegetation structure, reduce environmental quality, and impair ecosystem health (see Figure 3D).

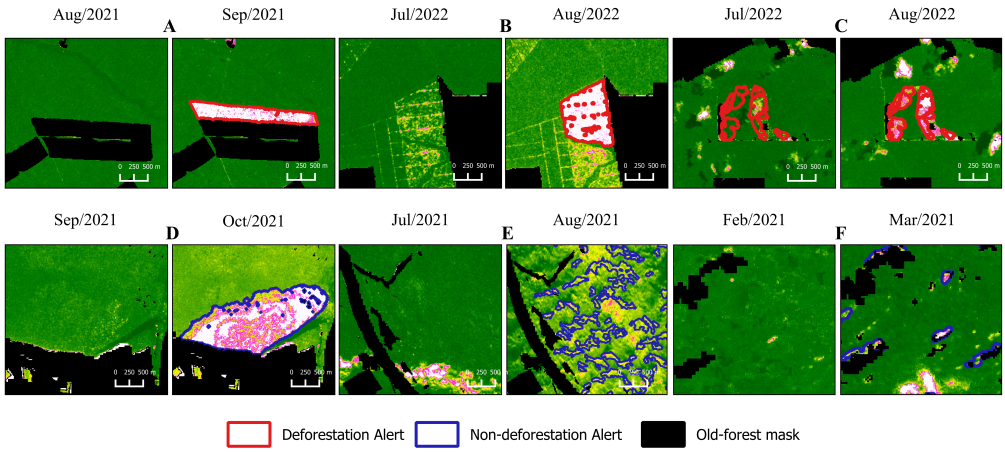

Figure 3. Examples of SAD deforestation alerts and false-positive detections. Recent deforestation alerts are shown in red, false positives in blue, old deforestation in black, remaining forest in green, and shadows as darker pixels. The panels show distinct cases: (**A**,**B**) automatically detected deforestation alerts; (**C**) deforestation visually identified by analysts despite fire smoke in the first image; (**D**) fire-associated forest degradation used as noise during AI model training; (**E**) cloud noise in the second image classified as noise; and (**F**) false alerts classified as non–forest loss caused by phenological changes associated with rocky outcrops.

2.3. The AI Model

2.3.1. Dataset Construction

To train the AI model, we assembled a primary dataset comprising 2.8 million manually labeled monthly image pairs delineating candidate deforestation events from January 2021 to December 2023. These polygons were generated through SAD's standard monthly processing pipeline and linked to Copernicus Harmonized Sentinel-2 MSI imagery, which provides 12 spectral channels at spatial resolutions ranging from 10 to 60 m per pixel [27].

On average, the polygons cover 0.05 km², with a standard deviation of 0.4 km² (see Figure 4). For each candidate deforestation polygon, an image patch of identical spatial extent is extracted from the previous month, producing a paired temporal sample. Image patches were resized to 64 × 64 pixels; their ground extent, and therefore the effective spatial resolution after resizing, varied with the candidate polygon. Each dataset entry therefore consists of two images: a baseline image from the preceding month, representing the forested area to be monitored (T0), and an image from the subsequent month containing a potential disturbance to be classified as either deforestation or non-deforestation. The latter category includes phenological variation, image noise, and forest degradation (see Figure 3). Monthly mosaics are constructed for both T0 and T1 to reduce cloud contamination and other visual artifacts.

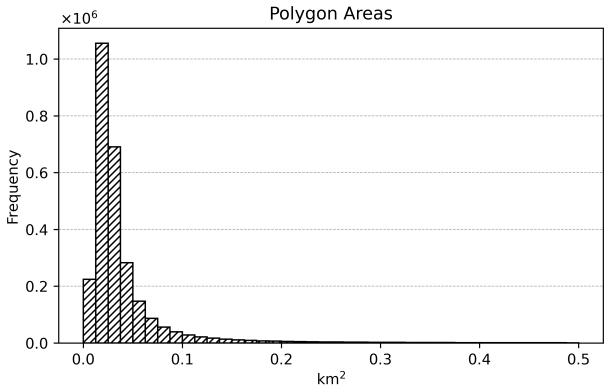

Figure 4. Distribution of polygon areas in the full dataset. Most candidate-alert polygons are relatively small, corresponding to compact spatial disturbances at the 10 m Sentinel-2 pixel resolution. The distribution is strongly right-skewed, with many small polygons and a very small number of larger alerts.

In addition to image pairs, our dataset includes two complementary features designed to be computationally efficient for near-real-time use:

- Water Proximity: A raster layer encoding the pointwise distance from each polygon to the nearest body of water, capped at 10 km and derived from global surface water maps [28].
- Vegetation Dynamics: A temporal sequence of NDFI statistics for each polygon, covering the two-year period ending at the T1 mosaic cutoff. At each observation date, the mean, standard deviation and selected quantiles summarize NDFI values spatially within the polygon. We retained the latest 128 observations, with shorter sequences left-padded.

These features provide additional context for distinguishing deforestation from flooding and seasonal vegetation changes. Figure 5 illustrates this for an example alert: although the Sentinel-2 image pair suggests deforestation, the NDFI series shows an earlier decline in January 2021 followed by recovery. This longer history helps assess whether the latest change represents new deforestation.

Each image pair and its associated features were labeled by a human expert as either deforestation or non-deforestation, providing the supervised labels for the AI model to learn. The full dataset of 2.8 million image pairs is strongly imbalanced: confirmed deforestation is the minority class, accounting for only 6% (approximately 180,000 cases). The remaining alerts, which experts judged did not contain a deforestation event, are primarily associated with cloud artifacts, natural forest degradation, rocky outcrops, legal logging, seasonal flooding, road construction, and other human activities.

To evaluate a broader range of predictors, we also constructed a smaller balanced dataset consisting of 216,000 image pairs, evenly split between deforestation and non-deforestation. This subset incorporates Sentinel-1 SAR [27], Landsat-7, and Landsat-8 optical imagery [29], NDFI measurements, digital elevation data [30], and surface water maps.

Both datasets were randomly divided into training, validation and test sets using 80/10/10 proportions. Each fold of the balanced dataset was drawn exclusively from the corresponding fold of the full dataset. For the ablation, this split allowed models to be compared across all three seasonal cycles. Splitting chronologically would keep training observations earlier than evaluation observations, but would limit evaluation to a shorter period that might miss important seasonal and interannual changes. We therefore chose to

evaluate across the full period, accepting that later observations could inform predictions for earlier ones.

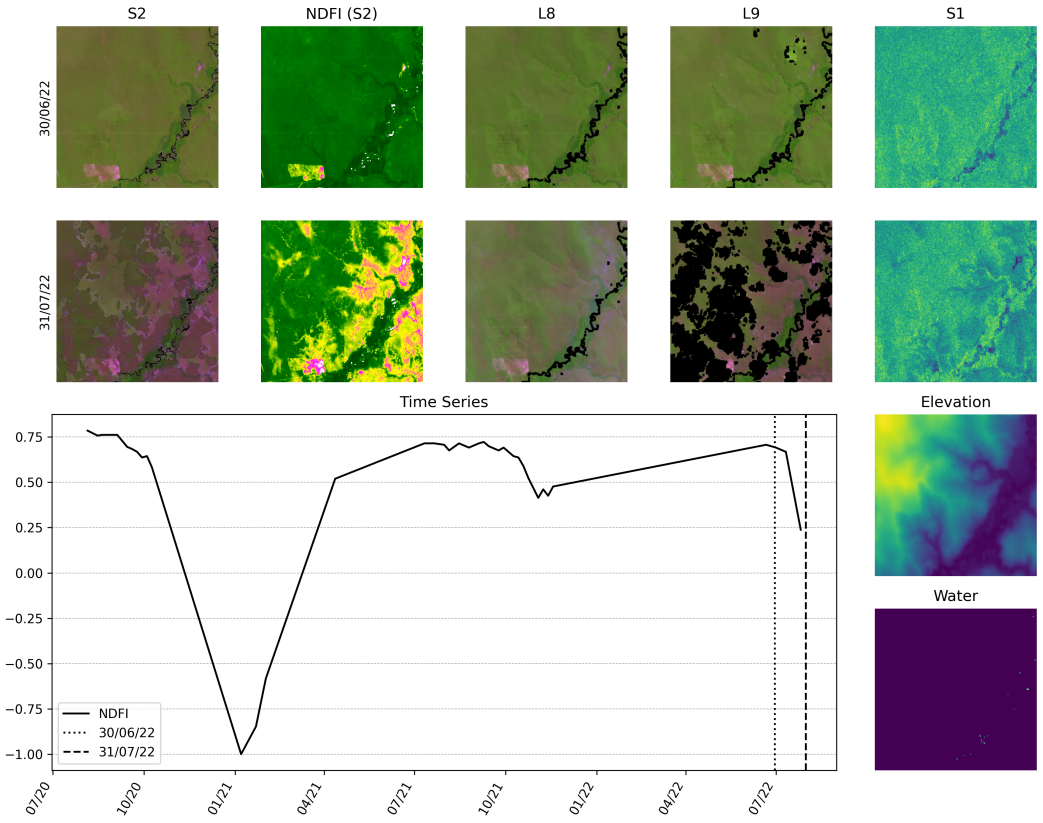

Figure 5. Example dataset sample combining multi-sensor imagery and vegetation-index time series. The figure shows satellite image patches from multiple sensors together with the corresponding mean Normalized Difference Fraction Index (NDFI) time series over the region. Although the Sentinel-2 before-and-after image pair may visually resemble a deforestation event, the NDFI time series indicates a previous episode of soil exposure in January 2021, suggesting that the observed canopy change is unlikely to represent new deforestation. The optical panels (S2, L8 and L9) are color composites in which black pixels indicate masked or missing data; the NDFI panel uses the same color scale as Figure 1; and the Sentinel-1 (S1), elevation and water panels use a color scale ranging from dark purple (low values) to yellow (high values).

We did not separate geographical areas across folds because the model is intended to monitor the same BLA region over time. Recurring local patterns can be useful in this setting: learning that alerts near a particular river are often not deforestation may help classify later alerts there. The ablation therefore compares models for monitoring the same region, not for generalizing to new regions.

Section 2.5 describes the separate deployment evaluation used to assess performance on future alerts. The training data covered 2021–2023, while deployment evaluation began in October 2024. This evaluation included a human-only group whose analysts reviewed alerts without access to model predictions.

2.3.2. Model Architecture and Training

Because processing features for the entire BLA is computationally costly, we used the balanced subset of 216,000 samples only to identify the most predictive feature combinations. This analysis indicated that the strongest predictors were: (i) Sentinel-2 image pairs from two consecutive months; (ii) summary statistics of the NDFI time series, including

mean, standard deviation and selected quantiles; and (iii) the minimum distance to rivers (see Figure 6).

To combine these heterogeneous data sources, we evaluated several feature-composition strategies. The best-performing approach processed spatial and temporal inputs using specialized submodules: Long Short-Term Memory networks (LSTMs) for temporal features [31] and Residual Networks (ResNets) for image-based features [32]. The LSTM used two stacked recurrent layers with 20 hidden units each, followed by a linear layer producing one logit. The ResNets received 64×64 inputs, with 22 channels for Sentinel-2 and one for distance to rivers. Following the architecture shown in Figure 6, both used three groups of four residual blocks, with 64, 128, and 256 channels, respectively. Each submodel produced one scalar logit. Logistic regression combined three logits in the final model, corresponding to Sentinel-2 imagery, NDFI time series and distance to rivers. The eight-source ablation combined eight logits. For each sample, the resulting model estimates the probability that the image pair contains a deforestation event. This probability is converted into a binary prediction by applying a user-defined threshold, typically set to 0.5. The submodels were trained separately on the training set of the full dataset using binary cross-entropy loss and the Adam optimizer [33] with a learning rate of 3×10^{-4}, without class weighting or resampling, for up to 100 epochs with a batch size of 32. Additional ablation results are presented in Section 3.1.2.

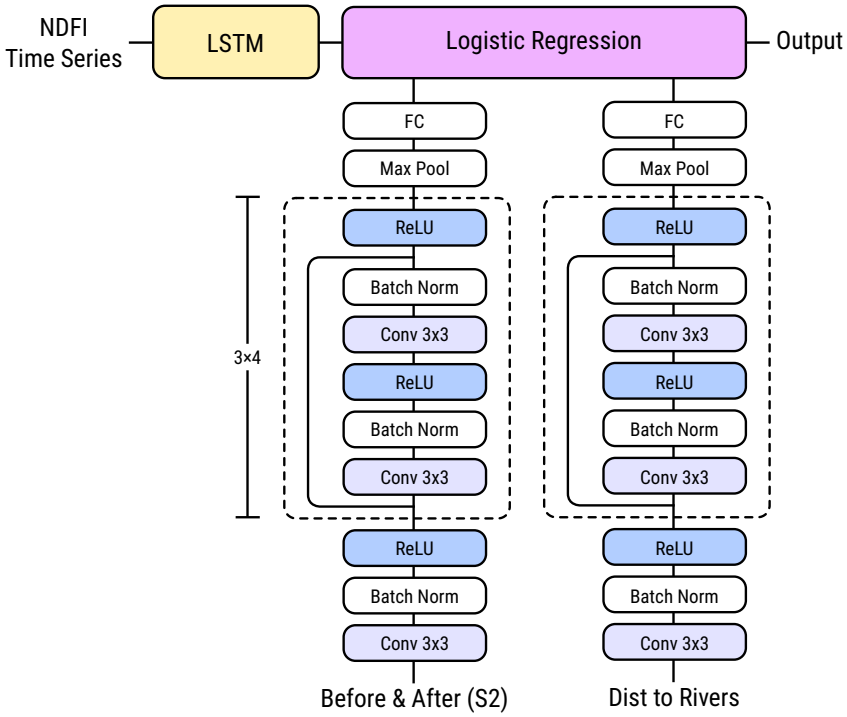

Figure 6. Architecture of the AI model. The model combines three complementary inputs: before-and-after Sentinel-2 mosaics, minimum distance to rivers, and NDFI time-series summaries. Spatial information from the Sentinel-2 image pair and temporal information from the NDFI sequence are processed through separate specialized modules, while the river-distance feature provides additional contextual information. These input-specific representations are fused only in the final linear layer, allowing the predicted deforestation probability to be decomposed into source-level contributions and thereby supporting model interpretability.

Crucially, using a logistic regression layer rather than a black-box ensemble allows analysts to interpret the model's estimated deforestation probability by examining the relative contribution of each data source. For example, Figure 7 shows a case in which the NDFI time series contributed negatively to the final prediction, suggesting seasonal vegetation

dynamics, whereas the Sentinel-2 imagery provided strong evidence of canopy clearing. This decomposition of the model output helps clarify cases in which spatial evidence alone may resemble deforestation but temporal dynamics indicate a non-deforestation process. By making the contribution of each input modality explicit, the fusion strategy supports interpretability and helps build the trust required for operational deployment.

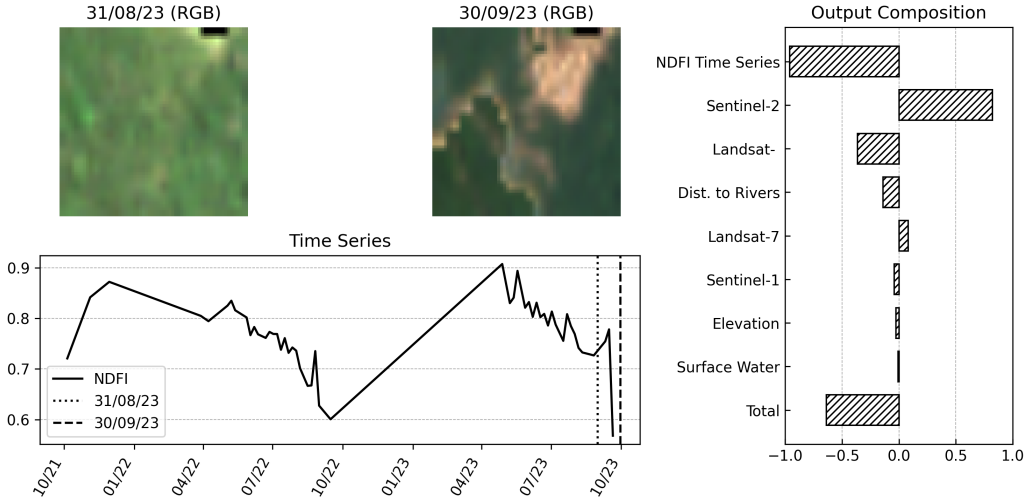

Figure 7. Source-level attribution enabled by logistic-regression model fusion. The panel on the right decomposes the model prediction into spatial and temporal logit contributions, which are combined to produce the final deforestation probability. In this example, the Sentinel-2 image component provides evidence consistent with deforestation, whereas the NDFI time-series component contributes negatively to the final prediction because it captures seasonal vegetation dynamics. The combined model therefore correctly classifies the alert as non-deforestation, demonstrating how temporal context can override visually ambiguous spatial evidence.

For the classical baselines, we trained random forests and histogram-based gradient boosting using the same labels and balanced dataset splits as the neural models. Each image channel was summarized by its mean, standard deviation, minimum and maximum over the 64×64 input, yielding 88 features for the 22-channel Sentinel-2 input. For the NDFI time series, we flattened the LSTM's 11×128 input into 1408 features, retaining the latest 128 observations and left-padding shorter sequences with -2. We evaluated each method on Sentinel-2 and NDFI time-series inputs separately, and on concatenated features from either the three selected sources (Sentinel-2, NDFI time series and distance to rivers) or all eight sources of the balanced dataset. The same four statistics per image channel were used for the additional sources.

The random forests used 200 trees, a maximum depth of 20, at least five samples per leaf and the square root of the feature count as the number of candidate features at each split. Gradient boosting used 200 iterations, a learning rate of 0.1, at most 31 leaves per tree, at least 20 samples per leaf and no early stopping. Both methods used fixed settings without validation-set hyperparameter tuning or class weighting, a random seed of zero and a probability threshold of 0.5 for classification.

We also evaluated an image-pair NDFI-threshold baseline. Its score was the mean before-minus-after NDFI difference within the SAD alert mask, computed at the native image resolution. We selected a nonnegative threshold to maximize accuracy on the balanced validation set and fixed it before evaluating the test set. Alerts with scores strictly above this threshold were classified as deforestation.

2.4. Operational Deployment and User Interface

The model was integrated into the existing SAD pipeline as a lightweight module running on modest hardware: six CPU cores, 16 GB of RAM, and a single NVIDIA Tesla T4 GPU (NVIDIA Corporation, Santa Clara, CA, USA). We leveraged the Google Earth Engine (GEE) infrastructure already used by Imazon to download image pairs, process time-series data, and generate the prediction assets displayed in the user interface. Maintaining the workflow within the GEE ecosystem reduces computational overhead and minimizes potential reprojection errors.

Integration into the GEE application involves three concurrent steps executed across different spatial tiles:

1. Retrieval of raw alerts, imagery, and time series data via GEE's RESTful API.
2. Generation of deforestation predictions using the AI model.
3. Uploading the resulting predictions as assets that can be accessed in real time through the user interface.

This workflow produces predictions within a few hours, a latency that is small relative to the several days typically required for analysts to complete the monthly validation process.

The updated interface was designed to support human-in-the-loop review by displaying polygon contours according to the model prediction at a threshold of $t = 0.5$: red for predicted deforestation and blue for predicted non-deforestation (see Figure 8). Keyboard shortcuts and dedicated interface buttons further allow analysts to move rapidly between polygons predicted as deforestation. Because true deforestation typically represents less than 10% of all candidate polygons, even modest reductions in the review burden associated with non-deforestation cases can substantially improve analyst throughput. The model's high agreement with analysts on non-deforestation cases therefore has direct operational value, reducing time spent rejecting obvious false positives while allowing analysts to focus on the smaller subset of likely forest-loss events. We also tested a four-color visualization scheme in which prediction confidence was represented through luminance variation. However, feedback from analysts indicated that the binary red/blue system provided the most effective balance between decision support and visual simplicity.

2.5. Deployment Evaluation Protocol

We evaluated the model's accuracy on future candidate alerts within the BLA during a 10-month operational deployment beginning in October 2024. The model was trained on data from the entire BLA covering 2021–2023, so the training data preceded the deployment period. We randomly selected 25% of deployment alerts (approximately 377 thousand alerts) for human review without access to model predictions. This blinded control prevented model outputs from influencing the reference labels for these alerts.

2.6. Metrics

A standard metric for evaluating binary classifiers is predictive accuracy, defined as the proportion of correctly classified samples. Although accuracy can be useful for comparing models on a shared and balanced dataset, it can be misleading under strong class imbalance [34]. In such settings, a naïve classifier that always predicts the majority class (in this case, non-deforestation) may achieve high accuracy without learning to distinguish deforestation from non-deforestation events.

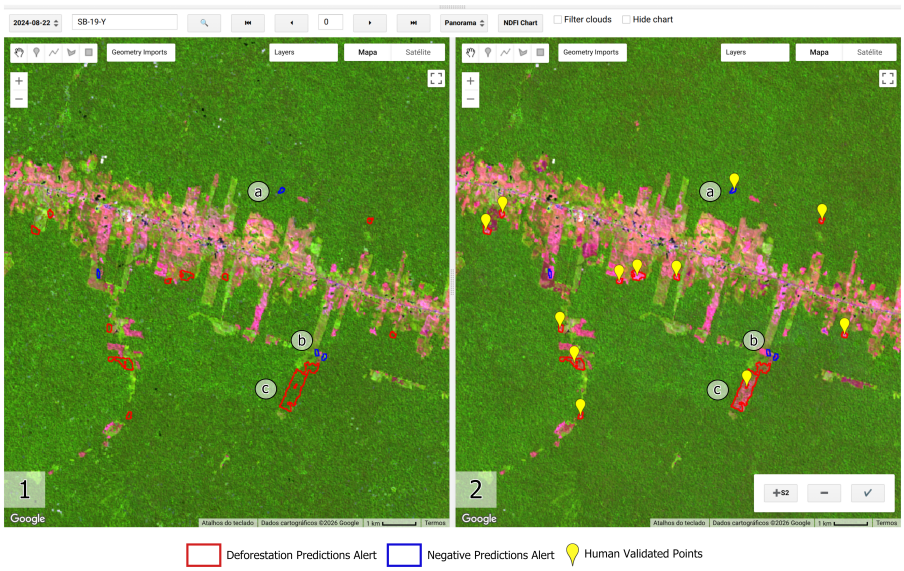

Figure 8. Human-in-the-loop visualization interface for reviewing model predictions in Google Earth Engine. The interface displays SAD alert polygons with contours colored according to the model prediction: red indicates predicted deforestation and blue indicates predicted non-deforestation. (1) View of the previous month's data, without analyst validation marks. (2) View of the current validation month, in which analyst validations are marked with yellow pins. The examples illustrate three representative review outcomes: (a) a false negative, in which the model predicts non-deforestation for an alert validated by analysts as deforestation; (b) a true negative, in which both the model and analyst classify the alert as non-deforestation; and (c) a true positive, in which both identify the alert as deforestation. This visualization supports rapid analyst screening while preserving expert control over the final validation decision. Some Google Maps interface labels appear in Portuguese because of the browser language settings: "Mapa" (Map), "Satélite" (Satellite), "Atalhos do teclado" (Keyboard shortcuts), "Dados cartográficos" (Map data) and "Termos" (Terms).

To address this limitation, we evaluated model performance using metrics that are more informative under class imbalance. First, we report the F1-score, defined as the harmonic mean of precision and recall. Precision, also referred to as User's Accuracy, is the proportion of predicted deforestation cases that are correctly classified as deforestation. It therefore measures the reliability of positive model predictions. Recall, also referred to as Producer's Accuracy, is the proportion of true deforestation cases that are correctly identified by the model. It therefore measures the model's sensitivity to deforestation events. The F1-score combines these two quantities into a single metric and is particularly useful when performance on the minority class is of primary interest.

We also report Cohen's kappa, which measures agreement between model predictions and reference labels while accounting for agreement expected by chance. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed probability of agreement and p_e is the expected probability of agreement under chance assignment, estimated from the marginal class frequencies. A value of $\kappa = 1$ indicates perfect agreement, $\kappa = 0$ indicates agreement no better than chance, and negative values indicate systematic disagreement. In addition, we report balanced accuracy, defined as the average of the true positive rate, or recall, and the true negative rate, or specificity. Balanced accuracy provides a complementary summary of performance by giving equal weight to both classes.

Finally, because the model outputs continuous probability scores rather than only binary classifications, we evaluated its performance across a range of decision thresholds. Varying the threshold allows the model to be tuned toward higher sensitivity or higher precision, depending on operational priorities. We summarize this threshold-dependent behavior using two curves: The precision–recall (PR) curve plots precision against recall across all classification thresholds and is especially informative when the positive class is rare. The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate across thresholds. The corresponding areas under these curves, PR-AUC and ROC-AUC, provide threshold-independent summaries of model performance.

3. Results

3.1. Model Evaluation

In this subsection, we evaluate the model's performance as a binary deforestation classifier using the test fold of the dataset. We first assess its consistency across different regions of the Brazilian Legal Amazon (BLA) and between the dry and rainy seasons. We then present an ablation study comparing alternative strategies for integrating signals derived from different data sources.

3.1.1. Model Performance

The logistic-regression fusion model, which combines Sentinel-2 imagery, NDFI time-series statistics, and river proximity, achieved an F1-score of 0.77, precision of 0.77, recall of 0.77, and balanced accuracy of 88.0% on the test fold of the full dataset (Table 1). Its ROC-AUC of 0.989 (see Figure 9) indicates strong discriminative performance across classification thresholds. This enables the model to be calibrated toward either lower false-negative rates or lower false-positive rates, depending on operational priorities, while maintaining high overall performance. Overall accuracy was 97.3% on the full dataset and 96.1% on the balanced subset.

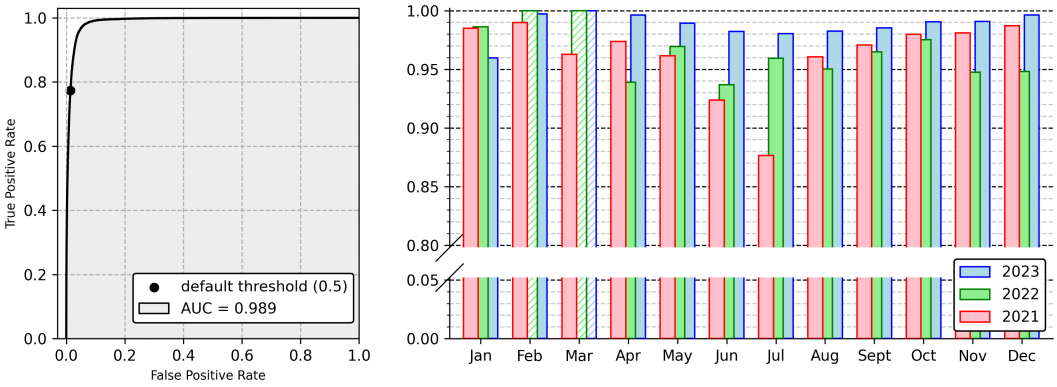

Figure 9. Test-set discrimination performance and temporal consistency of model accuracy. The left panel shows the receiver operating characteristic (ROC) curve for the held-out test dataset. The right panel shows monthly test accuracy, with hatched bars indicating months for which no alerts were present in the test fold. Accuracy remained consistently high across the evaluation period.

Table 1. Confusion matrix of model predictions on the held-out test subset. The table reports absolute counts and their proportions relative to the full test subset. Rows correspond to analyst-validated reference labels, and columns correspond to model predictions. The model correctly classified 12,676 deforestation alerts and 256,821 non-deforestation alerts, while producing 3740 false positives and 3708 false negatives.

	Positive Prediction	**Negative Prediction**
Positive target	12,676 (4.6%)	3708 (1.3%)
Negative target	3740 (1.4%)	256,821 (92.7%)

A key advantage of the proposed architecture is the interpretability provided by the logistic-regression fusion layer, which allows the spatial and temporal contributions to the final prediction to be examined separately (see Figure 7). This decomposition proved useful for identifying edge cases in which the Sentinel-2 before-and-after imagery appeared consistent with deforestation, but the NDFI time series indicated a seasonal fluctuation rather than permanent forest loss.

Because the BLA is a vast geographic area encompassing distinct biomes and two primary seasons (the wet season from November to April and the dry season from May to October), high aggregate accuracy is insufficient for operational deployment. Robust performance across both space and time is essential. From August 2021 onward, the model maintained consistently high performance across the evaluation period. (The lower accuracy observed before August 2021 is likely attributable to historical changes in the SAD pipeline, including modifications to the temporal windows used for mosaic construction and to the floating-point precision applied during cloud filtering.)

The model also exhibited strong spatial robustness. As shown in Figure 10, the spatial distribution of errors was broadly proportional to the distribution of positive labels, indicating that errors were not concentrated in a small number of regions. This pattern is expected, as non-deforestation cases are generally easier to classify than true deforestation events, which are rarer and often more visually heterogeneous.

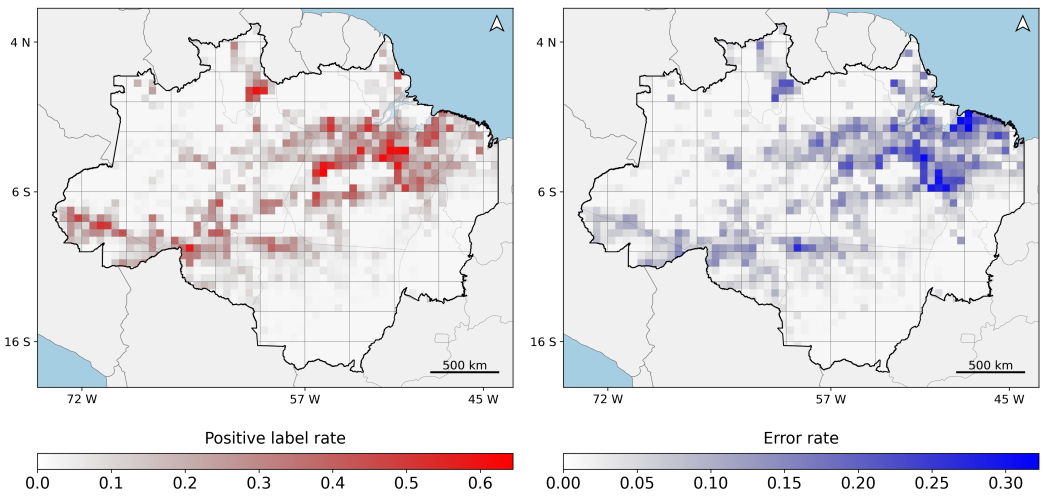

Figure 10. Spatial distribution of positive-label prevalence and model prediction errors in the test dataset. The left panel shows the spatial distribution of the proportion of positive labels, corresponding to confirmed deforestation alerts, while the right panel shows the spatial distribution of model prediction errors. The two distributions are strongly correlated, with a Pearson correlation coefficient of 0.8, indicating that areas with higher deforestation prevalence also tend to exhibit higher error rates. This pattern suggests that prediction errors are primarily associated with the local frequency of positive cases rather than systematic geographic bias.

3.1.2. Model Ablation

As shown in Table 2, the best-performing single-source model on the balanced dataset of 216,000 samples was the ResNet trained on before-and-after Sentinel-2 mosaics, which achieved an accuracy of 95.9%. This result is consistent with operational practice, as Sentinel-2 mosaics constitute the primary visual evidence used by human analysts during SAD validation. Using Sentinel-2 inputs, random forests and gradient boosting achieved test accuracies of 91.6% and 93.1%, respectively, while the image-pair NDFI-threshold baseline reached 77.8%.

The LSTM trained exclusively on the NDFI time series performed nearly as well, reaching an accuracy of 94.9%, indicating that temporal vegetation dynamics provide highly informative signals for distinguishing deforestation from non-deforestation alerts. For this source, the gap to the classical methods was smaller: random forests achieved 93.3% accuracy and gradient boosting reached 94.2%.

To assess whether combining complementary data sources could improve the model's ability to distinguish deforestation from seasonal vegetation dynamics, flooding, and atmospheric artifacts, we evaluated two strategies for combining neural models and compared them with random forests and gradient boosting trained on concatenated input features (see Table 3).

- Logistic regression: We fit a logistic regression on the logit outputs of individual pre-trained single-source models.
- Fully Connected Adapter (FCA): We substitute the last layer of each model by a single linear layer and train the composite network from scratch. Note that the resulting architecture is similar to the aforementioned Logistic Regression, differing only by the collapse of redundant hyperparameters (e.g., subsequent linear layers).

Table 2. Accuracy of single-source models on the balanced dataset. The results indicate that before-and-after Sentinel-2 imagery and NDFI time series are the most predictive individual data sources for deforestation classification, capturing distinct sources of signal in the features.

Model	Data Source	Accuracy
ResNet	Sentinel-1	0.865
ResNet	**Sentinel-2**	**0.959**
Random forest	Sentinel-2	0.916
Gradient boosting	Sentinel-2	0.931
NDFI threshold	Sentinel-2	0.778
ResNet	Landsat 7	0.838
ResNet	Landsat 8	0.880
ResNet	DEM	0.663
ResNet	JRC	0.569
ResNet	Dist to Rivers	0.567
LSTM	NDFI Time series	0.949
Random forest	NDFI Time series	0.933
Gradient boosting	NDFI Time series	0.942

Bold indicates the highest accuracy.

Combining the predictions of all eight neural models with logistic regression achieved the highest accuracy on the balanced test set, reaching 96.2%. Using the same eight sources, random forests and gradient boosting achieved 93.9% and 95.4%, respectively. With Sentinel-2, NDFI time series and distance to rivers, the corresponding tree-model accuracies were 93.8% and 95.1%.

Table 3. Accuracy on the balanced test set. Combining neural-model predictions with logistic regression achieved the highest accuracy, outperforming random forests and gradient boosting using the same eight data sources.

Accuracy	Method	Sources
0.957	FCA	All eight
0.954	FCA	All except NDFI time-series
0.962	**Logistic Regression**	**All eight**
0.938	Random forest	S2, NDFI time-series, Dist. to Rivers
0.951	Gradient boosting	S2, NDFI time-series, Dist. to Rivers
0.939	Random forest	All eight
0.954	Gradient boosting	All eight

Bold indicates the highest accuracy.

We first evaluated all 28 two-source combinations and then all 56 three-source combinations of the eight data sources, ranking them by accuracy on the balanced validation set. Through best-subset selection among the features listed in Table 2, we identified before-and-after Sentinel-2 imagery and NDFI time series as the strongest predictors. We additionally incorporated distance to rivers based on domain-expert guidance, as errors from non-composite models were frequently concentrated near water bodies. Together, these three features achieved a validation accuracy of 96.1% on the balanced dataset. When retrained and evaluated on the full dataset, the final model reached an accuracy of 97.3%.

From a model-complexity perspective, it is notable that the least flexible composition strategy was also the most accurate. This result may be related to the training instability often observed in deep neural networks under noisy or weakly aligned supervision [35]. Although the deforestation labels are not typically considered corrupted, inconsistencies can arise from factors such as different satellite revisit times, viewing geometries, atmospheric conditions, and weak feature–label alignment across sensors. For example, a deforestation event may be clearly visible in a Sentinel-2 mosaic but obscured in Landsat-8 imagery because of cloud cover or differences in acquisition timing. In highly flexible, jointly trained models, such inconsistencies may produce large gradients in spurious directions, reducing the stability of stochastic optimization. By contrast, logistic regression fusion constrains the final composition step, allowing the model to combine reliable source-level evidence while reducing the risk of overfitting to sensor-specific artifacts.

3.2. Deployment Performance

Over a 10-month operational deployment from October 2024 to July 2025, the model analyzed 1.5 million alerts and maintained consistently strong performance despite a substantial covariate shift in live data. During deployment, the system achieved a ROC-AUC of 0.993, a PR-AUC of 0.781 (see Figure 11), an F1-score of 0.75, a balanced accuracy of 84.8%, and a Cohen's kappa of 0.76 (see Table 4). These results indicate robust classification performance across a range of operational thresholds.

Notably, the deployment F1-score of 0.75 closely matched the F1-score of 0.77 obtained on the test fold, even though the prevalence of confirmed deforestation alerts declined from 5.9% in the test set to only 0.9% in the operational setting. This stability is particularly important because political, economic, environmental, and seasonal factors can substantially alter the distribution of alerts encountered during real-world monitoring. The model's ability to preserve performance under this shift suggests that it generalizes well to operational conditions and does not require frequent retraining or recalibration to remain effective.

Table 4. Performance of the proposed model and validation workflows under test and operational conditions. The table compares model performance on the held-out test set with live deployment results from October 2024 to July 2025. During deployment, 25% of alerts were randomly assigned to a control workflow in which analysts reviewed alerts without access to model predictions, while the remaining alerts were reviewed using the AI-assisted interface. Metrics for the deployment columns were computed using the final analyst-validated labels as the reference standard. Always-negative accuracy equals the proportion of non-deforestation labels; this classifier has zero deforestation recall. The similar F1-scores between the test set and AI-assisted deployment indicate robustness under substantial distribution shift, while the higher agreement metrics in the AI-assisted workflow suggest that model support may help standardize analyst interpretation.

	Training Phase	**Deployment Phase**	
Metric	**Test Fold**	**AI-Assisted**	**Human-Only**
Accuracy	97.3%	99.6%	99.4%
Always-negative accuracy	94.1%	99.1%	99.2%
F1-Score	0.77	0.75	0.65
Precision	0.77	0.81	0.66
Recall	0.77	0.70	0.65
Balanced accuracy	88.0%	84.8%	81.7%
Cohen's kappa	0.76	0.75	0.6471
Positive labels	5.9 %	0.9%	0.8%

To assess the operational effect of AI assistance, 25% of live alerts were randomly assigned to a control workflow in which analysts reviewed polygons without access to model predictions (see Table 4, column 3). The remaining alerts were reviewed through the AI-assisted interface. Model agreement with analyst labels was higher in the AI-assisted group than in the human-only group. This difference may partly reflect the influence of model predictions on analysts. The human-only group, reviewed without access to predictions, therefore provides a more independent assessment of model performance.

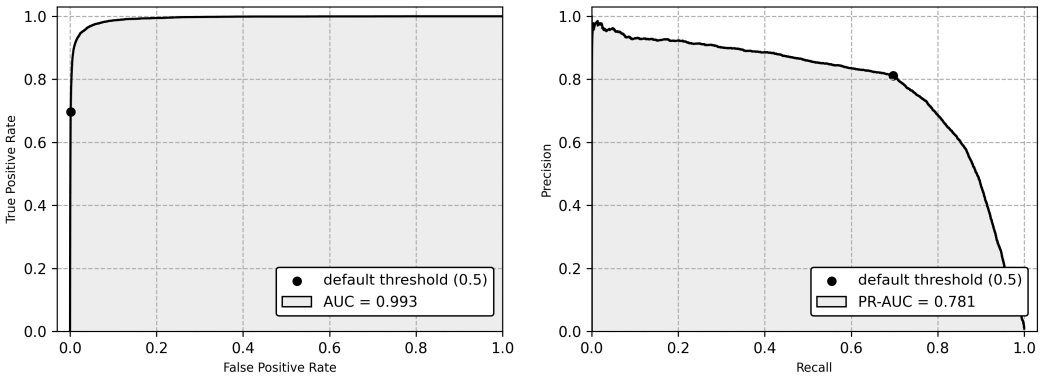

Figure 11. Deployment performance across probability thresholds. The figure shows the receiver operating characteristic (ROC) curve on the left and the precision–recall (PR) curve on the right for the model during operational deployment. Final labels from the hybrid human–AI validation workflow were used as the reference standard. The ROC-AUC of 0.993 and PR-AUC of 0.781 indicate strong discrimination under the deployment distribution and demonstrate that the model can be calibrated across a range of operating thresholds.

3.3. Operational Impact

Deployment of the AI model in Imazon's operational SAD pipeline began in August 2024, with the objective of reducing the manual workload associated with visual inspection of deforestation alerts. In the standard SAD workflow, analysts visually inspect alerts to determine whether they correspond to true deforestation or to false-positive detections caused

by clouds, shadows, fire scars, seasonal vegetation dynamics, or other non-deforestation processes. The AI model was therefore introduced as part of a human-in-the-loop workflow designed to accelerate alert review by prioritizing candidate detections while preserving expert control over final validation decisions.

We compared the historical workflow (October 2022–July 2023) with the deployment workflow (October 2024–July 2025). The deployment workflow included both alerts shown with AI prediction colors and the 25% shown without them. Analysts reviewed 695,251 alerts during the historical period and 1,508,800 during deployment, with total recorded review times of 915 h 43 min and 1085 h 7 min, respectively (Table 5). Consequently, the average review time per alert decreased from 4.742 s to 2.589 s, corresponding to a 45% reduction. These results indicate that the deployment workflow improved operational efficiency, allowing analysts to review more alerts with a lower average review time per alert.

The agreement analysis further shows that the AI model was particularly effective at identifying alerts that did not correspond to deforestation. From August 2024 to July 2025, agreement between AI predictions and final SAD analyst validation averaged 99.77% for non-deforestation alerts and 69.06% for deforestation alerts (see Table 6). The very high agreement for non-deforestation cases is especially important operationally, because false-positive alerts constitute a major source of analyst workload in large-scale monitoring systems. By reliably identifying likely non-deforestation cases, the model can reduce the burden of detailed visual inspection for low-priority alerts and allow analysts to focus attention on the smaller subset of likely forest-loss events.

Agreement was lower and more variable for deforestation alerts, indicating that the model should be used as a decision-support and triage tool rather than as a replacement for expert validation. Human analysts remain essential for confirming forest-loss events, adjudicating ambiguous cases, and maintaining the integrity of the final SAD classifications.

Table 5. Operational efficiency of SAD alert validation during the historical and deployment periods. The table reports total analyst time, number of alerts reviewed, and average review time per alert. Deployment totals include alerts shown both with and without AI prediction colors. Average review time per alert decreased by 45%, from 4.742s to 2.589s.

	Historical Workflow October 2022–July 2023		**Deployment Workflow October 2024–July 2025**	
	Time	**# of Alerts**	**Time**	**# of Alerts**
October	130 h 00 m	51,606	173 h 40 m	301,799
November	119 h 47 m	19,126	94 h 00 m	220,711
December	69 h 50 m	5291	80 h 45 m	136,919
January	75 h 11 m	4887	87 h 45 m	109,968
February	67 h 40 m	108,067	80 h 33 m	127,712
March	85 h 12 m	156,270	86 h 05 m	123,660
April	86 h 58 m	117,628	97 h 54 m	90,026
May	93 h 10 m	53,656	106 h 14 m	145,192
June	98 h 10 m	63,656	145 h 49 m	103,212
July	89 h 45 m	115,064	132 h 22 m	149,601
Total	915 h 43 m	695,251	1085 h 07 m	1,508,800
Time per alert	4.742 s		2.589 s	

Together, these results show that the AI-assisted workflow improved operational efficiency while preserving analyst oversight. The model reduced the time required to evaluate the alerts and enabled the system to manage a substantially larger volume of detections, demonstrating its potential to increase the responsiveness and scalability of near-real-time deforestation monitoring.

Table 6. Monthly agreement between SAD analyst validation and AI model predictions from August 2024 to July 2025. Agreement rates are reported separately for alerts classified as non-deforestation and deforestation. The AI model showed consistently high agreement with analysts for non-deforestation alerts, averaging 99.77%, while agreement for deforestation alerts was lower and more variable, averaging 69.06%. These results indicate that the model is especially useful for filtering likely false positives, while expert review remains important for confirming deforestation events.

Month	No Deforestation	Deforestation
August 2024	99.46%	74.42%
September 2024	99.46%	69.51%
October 2024	99.76%	56.05%
November 2024	99.86%	65.95%
December 2024	99.93%	65.99%
January 2025	99.88%	71.65%
February 2025	99.95%	72.56%
March 2025	99.96%	60.62%
April 2025	99.85%	72.90%
May 2025	99.76%	71.88%
June 2025	99.66%	61.51%
July 2025	99.75%	85.72%
Average	99.77%	69.06%

3.4. Threshold Calibration and Operational Alert Filtering

During deployment, the threshold of $t = 0.5$ determined alert colors, without excluding alerts from review. Here, we evaluate the potential of alternative thresholds for alert filtering. Different monitoring contexts may prioritize minimizing missed deforestation events or reducing false alarms, requiring different trade-offs between recall and precision.

Beyond binary classification, threshold calibration is particularly useful for human-in-the-loop alert filtering. Near-real-time change-detection algorithms often generate large numbers of false-positive alerts caused by clouds, fire scars, seasonal vegetation dynamics, and other transient surface changes. In this setting, the model can serve as a screening layer that filters low-probability alerts before they enter the human validation queue.

In this filtering setting, the model's positive prediction rate corresponds to the proportion of alerts forwarded to analysts for expert review and can therefore be interpreted as a direct proxy for human workload. For instance, applying a low threshold of $t = 0.0025$ excludes alerts with extremely low predicted probabilities of deforestation while retaining 99% of polygons containing true deforestation. At this threshold, only 13% of all alerts would be forwarded to analysts (see Figure 12). In other words, the model could substantially reduce the number of alerts requiring visual inspection while preserving nearly all true deforestation cases.

The potential operational impact of this filtering strategy is substantial. For the 10-month deployment volume of 1,508,800 alerts, forwarding only 13% of alerts would reduce the analyst review queue to approximately 196,144 alerts, excluding about 1.31 million low-probability alerts from manual inspection. Assuming the human-only processing time of 4.742 s per alert reported in Table 5, this would correspond to an estimated saving of approximately 1729 analyst-hours over the deployment period.

Even under a more conservative calibration designed to retain 99.9% of true deforestation cases, analysts would need to inspect only 41% of the total alert volume. This would reduce the review queue by 59%, corresponding to an estimated saving of approximately 1173 analyst-hours. See also Figure 13 for the substantial gains available in terms of analyst workload through shifting the probability threshold of the model.

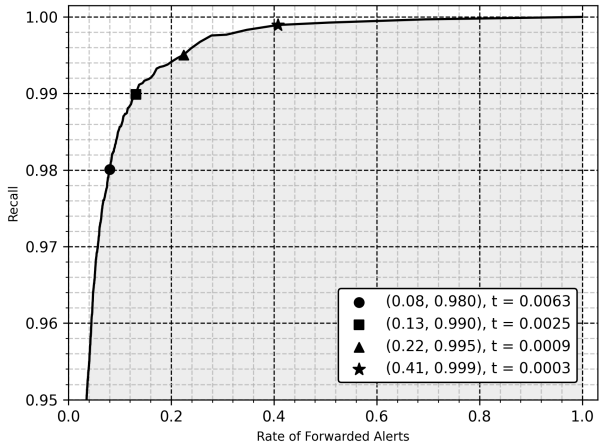

Figure 12. Workload–recall trade-off for potential human-in-the-loop alert filtering evaluated using deployment data. The curve shows model performance when used as a screening layer to determine which alerts are forwarded for expert review. The x-axis represents the positive prediction rate, interpreted operationally as the proportion of alerts retained in the analyst review queue, while the y-axis represents producer's accuracy, or the proportion of true deforestation cases preserved. Highlighted points correspond to thresholds that retain 98%, 99%, 99.5%, and 99.9% of positive labels, illustrating the trade-off between reducing analyst workload and minimizing missed deforestation events.

These results indicate that the model can support multiple operational modes. At moderate thresholds, it functions as a binary decision-support tool that helps analysts prioritize likely deforestation cases. At lower thresholds, it can operate as a conservative filtering system that substantially reduces workload while maintaining very high recall. In both cases, the model remains embedded within a human-in-the-loop workflow: analysts retain final authority over validation, while the AI system adjusts the volume and ordering of alerts according to the urgency, risk tolerance, and resource constraints of each monitoring context.



Figure 13. Workload–recall trade-off curve. The curve shows model performance, measured as deforestation recall, versus analyst workload. Note the workload, measured as monthly hours, drops quickly for small threshold levels, while recall only decreases slightly, indicating the potential for substantial gains for a human-in-the-loop algorithm.

4. Discussion

Our results demonstrate that fusing spatial and temporal features through a transparent logistic-regression layer can provide both high classification performance and operational interpretability for deforestation detection. The model leverages the spatial information contained in Sentinel-2 imagery while incorporating NDFI time-series statistics that help suppress false positives associated with seasonal vegetation dynamics.

The baseline comparisons show that the advantage of neural models depends on the input data. The improvement over gradient boosting was larger for Sentinel-2 imagery than for NDFI time series. De Bem et al. [18] likewise found that CNNs outperformed random forests in deforestation mapping. However, our ResNet learns from image patches, whereas the tree models use channel summaries, so the difference cannot be attributed to architecture alone. With all eight sources, neural fusion achieved 96.2% accuracy compared with 95.4% for gradient boosting. This modest gap shows that classical methods remain competitive for this task.

Although deep neural network components, including CNN- and RNN-based modules, are used for feature extraction, the final prediction remains interpretable because the contribution of each input source can be examined separately. This interpretability is valuable not only for post hoc explanation but also for analyst calibration in operational settings. When experts can assess whether a prediction is driven primarily by spatial clearing patterns, temporal vegetation dynamics, or proximity to water, they can more effectively identify seasonal artifacts, cloud contamination, flood-related changes, and ambiguous degradation events. Such transparency is particularly important in a human-in-the-loop workflow, where analysts must decide when to trust the model, when to override it, and when to flag cases for further review. In this sense, the proposed architecture helps bridge the gap between fully automated black-box methods and manual expert interpretation, accelerating alert review while preserving expert oversight.

Operational deployment confirmed the practical value of the system. By embedding the model directly into the existing SAD pipeline and maintaining the workflow within Google Earth Engine infrastructure, we streamlined the monthly deforestation validation process without requiring substantial additional computational resources. The model runs efficiently on modest hardware and processes spatial tiles concurrently, making the approach suitable for monitoring programs with limited infrastructure. Quality-control checks also indicated that analysts did not simply accept model outputs uncritically: false positives produced by the model were routinely discarded by human reviewers, false negatives were also audited, and we found no evidence that AI assistance introduced systematic confirmation bias. The observed 45% reduction in review time per alert suggests that similar hybrid approaches could be valuable for other remote-sensing tasks that require large-scale screening combined with expert validation.

4.1. Future Work

Building on this deployment, we identify two main directions for future development: active learning and improved classification of partial canopy disturbance.

First, we plan to develop an auditing tool to identify potential mislabels after the hybrid validation pipeline has been applied. Given a fixed auditing budget, the system would estimate predictive uncertainty using the entropy of the model's output distribution. Among cases in which the model prediction disagrees with the final human label, the tool would prioritize high-confidence disagreements, that is, cases with low predictive uncertainty, for secondary review. These cases are especially informative because they may reveal either model failures or inconsistencies in human labeling. Second, partial canopy disturbances caused by selective logging, forest fire, severe seasonal effects, or other

forms of degradation remain an important source of human–AI disagreement. Training specialized submodels to distinguish among these disturbance types is therefore a key next step. Such models could improve classification performance while also reducing training instability associated with weak or heterogeneous feature–label relationships, as discussed in Section 3.1.2.

Torres et al. [23] showed how model-generated maps can provide a starting point for expert review. Our study evaluates this form of assistance in an established production workflow at scale. Masolele et al. [24] also used expert input to improve the model, selecting images for annotation based on model uncertainty and retraining with the additional labels. A natural next step is to adapt this annotation-and-retraining cycle to SAD's ongoing operations. This would establish an active-learning cycle in which the model helps select examples for expert annotation and is subsequently retrained with those labels.

Tharwat and Schenck [36] also discuss combining active learning with explanations of model predictions. In SAD, selected alerts could be accompanied by the contributions of the spatial, temporal and river-distance models, allowing analysts to inspect how each source influences the prediction during annotation.

4.2. Weaknesses and Limitations

Despite its strong performance, the current system has several limitations that should guide future improvements. First, because the dataset relies on manual labels produced by SAD analysts, model performance necessarily reflects expert decisions and may encode analyst-specific biases. However, this limitation is mitigated by the fact that the SAD pipeline is operated by experienced analysts, is applied regularly, and is subject to quality-control procedures. We therefore do not expect label bias to be a major practical limitation, although continued auditing remains important. A related concern is the potential feedback loop between model predictions and future human labels. Because analysts view model outputs during AI-assisted validation, future training data may partly reflect model-influenced decisions rather than fully independent expert judgments. This is a common challenge in human-in-the-loop systems and can be mitigated through blind control samples, periodic independent audits, and uncertainty-based review protocols. These safeguards are important for preserving label independence and reducing the risk of confirmation bias as the system evolves. Second, the training data are specific to the Brazilian Legal Amazon. Applying the model to other tropical forest regions, such as the Congo Basin or Southeast Asia, may require fine-tuning because of regional differences in vegetation structure, land-use practices, disturbance regimes, and seasonal phenology. The random-split ablation prioritizes model selection for continued monitoring within the BLA, where recurring local patterns remain relevant, while covering all three seasonal cycles. It does not assess transfer to unseen locations. Performance on future alerts is evaluated separately during deployment using alerts reviewed without access to model predictions. This evaluation covers ten months and does not establish performance across repeated seasonal cycles or long-term changes.

Third, persistent cloud cover remains a challenge. In some months and regions, cloud contamination reduces the number of valid observations available for mosaic construction. Stricter cloud-filtering procedures could improve image quality and model confidence, but they may also delay alert generation, thereby reducing the near-real-time value of the monitoring system. Future work should therefore explore strategies that improve robustness to cloud contamination without compromising operational latency.

The operational efficiency comparison used the same months in both periods and involved experienced analysts. Nevertheless, because the comparison was historical, we cannot isolate the contribution of AI assistance from other differences between the periods.

Finally, the model was designed for human-in-the-loop assistance rather than fully automated deforestation detection. At the default threshold of 0.5, which approximately balances false positives and false negatives in absolute terms, the model achieves a true positive rate of 69.7%. Although lowering the threshold would increase recall, the very low base rate of deforestation means that a fully automated system would still generate a large number of false positives. For this reason, the model is best understood as a decision-support tool that improves the efficiency of expert validation, rather than as a replacement for human analysts.

5. Conclusions

We developed and deployed an interpretable deep learning model for human-in-the-loop deforestation monitoring in the Brazilian Legal Amazon. Trained on 2.8 million manually labeled Sentinel-2 image pairs, the architecture fuses spatial features extracted by ResNet modules with temporal vegetation dynamics captured by LSTM models, using a transparent logistic-regression layer for final prediction. This design enables the model to distinguish true deforestation from seasonal vegetation change, atmospheric artifacts, and other sources of false-positive alerts while preserving source-level interpretability.

During operational deployment, the model achieved an F1-score of 0.75 and balanced accuracy of 84.8% when evaluated against final analyst labels, with a true positive rate of 69.7% and a false positive rate of 0.1% at the default threshold. More importantly, deployment within Imazon's SAD pipeline showed that the hybrid workflow can substantially amplify expert capacity: AI assistance reduced the average manual review time per alert by 45% while maintaining analyst oversight over final validation decisions.

Author Contributions: Conceptualization, P.O. and C.M.S.J.; methodology, R.S., F.G., P.O., L.S.V.B.A., R.F.F. and M.A.; software, R.S. and F.G.; validation, R.S., F.G., P.O., L.S.V.B.A., R.F.F. and M.A.; investigation, R.S., F.G., P.O., L.S.V.B.A., R.F.F. and M.A.; data curation, R.S., F.G., L.S.V.B.A., R.F.F. and M.A.; writing—original draft preparation, R.S., F.G., L.S.V.B.A., R.F.F. and M.A.; writing—review and editing, P.O., L.S.V.B.A., C.M.S.J., R.F.F. and M.A.; visualization, R.S., F.G., L.S.V.B.A., R.F.F. and M.A.; supervision, C.M.S.J.; project administration, P.O. and C.M.S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Norwegian Agency for Development Cooperation (Norad), represented by the Department for Civil Society and the Private Sector Grant BRA-2049 BRA-21/0004; IMPA's Center for Projects and Innovation (Centro Pi); a Google Research Scholar Program award; and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Data Availability Statement: The source code used to train and evaluate the AI model is available at: https://github.com/rlschuller/dsad-rs (accessed on 1 November 2026). The repository includes scripts for data preprocessing, model training, threshold calibration and performance evaluation. Additionally, a dataset with the labeled polygons is available at https://doi.org/10.7910/DVN/LA2KC6 (accessed on 1 November 2026).

Acknowledgments: We thank the providers of the satellite and geospatial datasets used in this study, including the European Space Agency/Copernicus Programme for Sentinel-1 and Sentinel-2 imagery, the U.S. Geological Survey and NASA for Landsat imagery, and the Google Earth Engine platform for enabling large-scale geospatial processing.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lovejoy, T.E.; Nobre, C. Amazon tipping point. *Sci. Adv.* **2018**, *4*, eaat2340. https://doi.org/10.1126/sciadv.aat2340.
2. Salati, E.; Dall'Olio, A.; Matsui, E.; Gat, J.R. Recycling of water in the Amazon Basin: An isotopic study. *Water Resour. Res.* **1979**, *15*, 1250–1258. https://doi.org/10.1029/WR015i005p01250.

3. Brazilian National Congress. Law No. 12651 of 25 May 2012 (Brazilian Forest Code). Available online: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12651.htm (accessed on 13 August 2025).
4. Aragão, L.E.O.C.; Anderson, L.O.; Fonseca, M.G.; Rosan, T.M.; Vedovato, L.B.; Wagner, F.H.; Silva, C.V.J.; Silva Junior, C.H.L.; Arai, E.; Aguiar, A.P.; et al. 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. *Nat. Commun.* **2018**, *9*, 536. https://doi.org/10.1038/s41467-017-02771-y.
5. Souza, C.M.; Oliveira, L.A.; De Souza Filho, J.S.; Ferreira, B.G.; Fonseca, A.V.; Siqueira, J.V. Landsat sub-pixel land cover dynamics in the Brazilian Amazon. *Front. For. Glob. Change* **2023**, *6*, 1294552. https://doi.org/10.3389/ffgc.2023.1294552.
6. Nobre, C.A.; Sampaio, G.; Borma, L.S.; Castilla-Rubio, J.C.; Silva, J.S.; Cardoso, M. Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 10759–10768. https://doi.org/10.1073/pnas.1605516113.
7. Souza, C.M.; Roberts, D.A.; Cochrane, M.A. Combining spectral and spatial information to map canopy damage from selective logging and forest fires. *Remote Sens. Environ.* **2005**, *98*, 329–343. https://doi.org/10.1016/j.rse.2005.07.013.
8. Valeriano, D.M.; Mello, E.M.K.; Moreira, J.C.; Shimabukuro, Y.E.; Duarte, V.; Souza, I.M.; Santos, J.R.; Barbosa, C.C.F.; Souza, R.C.M. Monitoring tropical forest from space: The PRODES digital project. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *35*, 272–274.
9. Diniz, C.G.; Souza, A.A.d.A.; Santos, D.C.; Dias, M.C.; da Luz, N.C.; de Moraes, D.R.V.; Maia, J.S.; Gomes, A.R.; Narvaes, I.d.S.; Valeriano, D.M.; et al. DETER-B: The new Amazon near real-time deforestation detection system. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3619–3628. https://doi.org/10.1109/JSTARS.2015.2437075.
10. Hansen, M.C.; Potapov, P.V.; Pickens, A.H.; Tyukavina, A.; Hernandez-Serna, A.; Zalles, V.; Turubanova, S.; Kommareddy, I.; Stehman, S.V.; Song, X.P.; et al. Global land use extent and dispersion within natural land cover using Landsat data. *Environ. Res. Lett.* **2022**, *17*, 034050. https://doi.org/10.1088/1748-9326/ac46ec.
11. Amazon Conservation. Monitoring of the Andes Amazon Program (MAAP). Available online: https://www.maapprogram.org/ (accessed on 26 October 2025).
12. Shimabukuro, Y.E.; Batista, G.T.; Mello, E.M.K.; Moreira, J.C.; Duarte, V. Using shade fraction image segmentation to evaluate deforestation in Landsat Thematic Mapper images of the Amazon Region. *Int. J. Remote Sens.* **1998**, *19*, 535–541. https://doi.org/10.1080/014311698216152.
13. Asner, G.P. Cloud cover in Landsat observations of the Brazilian Amazon. *Int. J. Remote Sens.* **2001**, *22*, 3855–3862. https://doi.org/10.1080/01431160010006926.
14. DeVries, B.; Pratihast, A.K.; Verbesselt, J.; Kooistra, L.; Herold, M. Characterizing forest change using community-based monitoring data and Landsat time series. *PLoS ONE* **2016**, *11*, e0147121. https://doi.org/10.1371/journal.pone.0147121.
15. Healey, S.P.; Cohen, W.B.; Yang, Z.; Brewer, C.K.; Brooks, E.B.; Gorelick, N.; Hernandez, A.J.; Huang, C.; Hughes, M.J.; Kennedy, R.E.; et al. Mapping forest change using stacked generalization: An ensemble approach. *Remote Sens. Environ.* **2018**, *204*, 717–728. https://doi.org/10.1016/j.rse.2017.09.029.
16. Cué La Rosa, L.E.; van Duijvenbode, J.; Calle, Z.; Marcos, D.; Dallinga, J.; Reiche, J. Predicting near-future deforestation across the tropics using deep learning: Insights from the Forest Foresight project. *Environ. Res. Commun.* **2025**, *7*, 125002. https://doi.org/10.1088/2515-7620/ae1f69.
17. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 507–520.
18. de Bem, P.P.; de Carvalho Junior, O.A.; Guimarães, R.F.; Gomes, R.A.T. Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks. *Remote Sens.* **2020**, *12*, 901. https://doi.org/10.3390/rs12060901.
19. Irvin, J.; Sheng, H.; Ramachandran, N.; Johnson-Yu, S.; Zhou, S.; Story, K.; Rustowicz, R.; Elsworth, C.; Austin, K.; Ng, A.Y. ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery. *arXiv* **2020**, arXiv:2011.05479. https://doi.org/10.48550/arXiv.2011.05479.
20. Kaselimi, M.; Voulodimos, A.; Daskalopoulos, I.; Doulamis, N.; Doulamis, A. A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 3299–3307. https://doi.org/10.1109/TNNLS.2022.3144791.
21. Soto Vega, P.J.; Torres, D.L.; Andrade-Miranda, G.X.; da Costa, G.A.O.P.; Feitosa, R.Q. Assessing the generalization capacity of convolutional neural networks and vision transformers for deforestation detection in tropical biomes. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, *XLVIII-3-2024*, 519–525. https://doi.org/10.5194/isprs-archives-XLVIII-3-2024-519-2024.
22. Slagter, B.; Reiche, J.; Marcos, D.; Mullissa, A.; Lossou, E.; Peña-Claros, M.; Herold, M. Monitoring direct drivers of small-scale tropical forest disturbance in near real-time with Sentinel-1 and -2 data. *Remote Sens. Environ.* **2023**, *295*, 113655. https://doi.org/10.1016/j.rse.2023.113655.

23. Torres, D.L.; Turnes, J.N.; Soto Vega, P.J.; Feitosa, R.Q.; Silva, D.E.; Marcato Junior, J.; Almeida, C. Deforestation detection with fully convolutional networks in the Amazon forest from Landsat-8 and Sentinel-2 images. *Remote Sens.* **2021**, *13*, 5084. https://doi.org/10.3390/rs13245084.
24. Masolele, R.N.; Marcos, D.; De Sy, V.; Abu, I.O.; Verbesselt, J.; Reiche, J.; Herold, M. Mapping the diversity of land uses following deforestation across Africa. *Sci. Rep.* **2024**, *14*, 1681. https://doi.org/10.1038/s41598-024-52138-9.
25. Chen, S.; Woodcock, C.E.; Bullock, E.L.; Arévalo, P.; Torchinava, P.; Peng, S.; Olofsson, P. Monitoring temperate forest degradation on Google Earth Engine using Landsat time series analysis. *Remote Sens. Environ.* **2021**, *265*, 112648. https://doi.org/10.1016/j.rse.2021.112648.
26. Souza, C., Jr. Monitoring of forest degradation: A review of methods in the Amazon Basin. In *Global Forest Monitoring from Earth Observation*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 171–194. https://doi.org/10.1201/b13040-10.
27. Copernicus Programme. Sentinel-1 and 2 Imagery. Available online: https://sentiwiki.copernicus.eu/ (accessed on 13 August 2025).
28. Pekel, J.F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418–422. https://doi.org/10.1038/nature20584.
29. U.S. Geological Survey. Landsat-7 and 8 Imagery. Available online: https://www.usgs.gov/landsat-missions/landsat-collection-2 (accessed on 13 August 2025).
30. Yamazaki, D.; Ikeshima, D.; Tawatari, R.; Yamaguchi, T.; O'Loughlin, F.; Neal, J.C.; Sampson, C.C.; Kanae, S.; Bates, P.D. A high-accuracy map of global terrain elevations. *Geophys. Res. Lett.* **2017**, *44*, 5844–5853. https://doi.org/10.1002/2017GL072874.
31. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385. https://doi.org/10.48550/arXiv.1512.03385.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2017**, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980.
34. Stehman, S.V.; Foody, G.M. Key issues in rigorous accuracy assessment of land cover products. *Remote Sens. Environ.* **2019**, *231*, 111199. https://doi.org/10.1016/j.rse.2019.05.018.
35. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.G. Learning from noisy labels with deep neural networks: A survey. *arXiv* **2022**, arXiv:2007.08199.
36. Tharwat, A.; Schenck, W. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics* **2023**, *11*, 820. https://doi.org/10.3390/math11040820.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.