MS-HySAN: A Hybrid Multi-Scale Siamese Attention Network for Interpretable Change Detection in High-Resolution Satellite Imagery

Shailendra Dabral *, Anam Sabir, Unmesh Khati

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1851.v1

Keywords: change detection; hybrid CNN–LightGBM; Siamese attention network; SHAP interpretability; PlanetScope; spectral indices; class imbalance; multi-temporal analysis

 doi:10.20944/preprints202605.1851.v1

Article

MS-HySAN: A Hybrid Multi-Scale Siamese Attention Network for Interpretable Change Detection in High-Resolution Satellite Imagery

Shailendra Dabral *, Anam Sabir and Unmesh Khati

Indian Institute of Technology Indore, Indore 453552, India
* Correspondence: ms2404121005@alum.iiti.ac.in

Abstract

Remote sensing-based change detection for infrastructure monitoring demands methods that are simultaneously accurate, robust to severe class imbalance, and transparent in their decision logic. This study proposes MS-HySAN, a hybrid change-detection framework that addresses these requirements through three coordinated design decisions: (i) a truncated, attention-augmented Siamese encoder that serves as a frozen feature extractor rather than an end-to-end pixel classifier, (ii) a latent–physical fusion strategy that concatenates multi-scale CNN difference features with physically interpretable spectral-index differences, and (iii) a LightGBM classifier that performs internal sparse feature selection and exposes gradient-based SHAP attributions for post-hoc analysis. The framework is evaluated on high-resolution PlanetScope imagery (4-band and 8-band) over a national highway construction corridor in Indore, India, using 21 acquisitions from 2022–2025 with geographic *k*-fold cross-validation to enforce spatial independence. Experimental results show that the proposed hybrid model consistently outperforms conventional deep learning baselines including U-Net and Siamese U-Net across bi-temporal multi-class change-detection tasks, and competes with bi-temporal architectures (ChangeFormer, SNUNet, BIT) under the same training conditions. A SHAP interpretability analysis reveals complementary and physically meaningful contributions from the learned deep features and the handcrafted spectral indices, validating the fusion strategy. In the best-case setting, MS-HySAN (bi-temporal, indices + reflectance) achieves an overall mean F1-score of 0.95 (Kappa: 0.90), outperforming the corresponding deep baseline by +6 F1 points while maintaining stable cross-fold performance.

Keywords: change detection; hybrid CNN–LightGBM; Siamese attention network; SHAP interpretability; PlanetScope; spectral indices; class imbalance; multi-temporal analysis

1. Introduction

Remote sensing-based change detection (CD) identifies significant differences between images acquired at different times and supports urbanization monitoring, land-use change analysis, disaster assessment, and environmental impact evaluation [1,2]. High-resolution, high-cadence platforms such as PlanetScope enable scalable operational monitoring of construction corridors and infrastructure development, where ground surveys are accurate but limited in coverage and revisit frequency [3].

Classical CD methods span threshold-based difference mapping [2,4], Change Vector Analysis [5,6], Multivariate Alteration Detection [7], and supervised classifiers including support vector machines [8], random forests [9], and gradient boosting [10,11]. Deep learning architectures—encoder–decoder CNNs [12,13], Siamese networks [14], and transformer variants [15,16] have subsequently demonstrated strong performance on benchmark datasets [1], and comprehensive surveys document these advances.

Despite strong benchmark performance, deep learning models face three interrelated challenges in real-world operational settings. First, they behave as black-boxes with limited transparency, constraining adoption in infrastructure-monitoring workflows that require audit-ready decision logic [17,18].

 doi:10.20944/preprints202605.1851.v1

Second, they are data-hungry and degrade under extreme class imbalance, a pervasive condition in CD where the no-change class typically dominates by 90%+ [19–21]. Third, the quadratic self-attention cost of large transformer models limits their applicability to high-dimensional multi-spectral imagery with restricted labeled budgets [22,23].

Hybrid pipelines coupling deep feature extractors with classical classifiers have been explored in the remote sensing literature [24,25], but prior work typically applies the CNN end-to-end with the classical model used only for post-processing, or does not couple physical spectral-index differences with learned bottleneck features through a unified, jointly SHAP-attributable fusion vector. The present work addresses these gaps simultaneously.

We propose MS-HySAN (Multi-Scale Hybrid Siamese Attention Network), whose key design decisions are:

1. A *decoupled two-stage training protocol*: the Siamese backbone is pre-trained with focal loss and frozen; LightGBM is fitted independently on the resulting structured feature tensors. This eliminates saddle-point instabilities at the CNN–GBDT interface and reduces memory consumption to levels feasible in resource-constrained environments.
2. A *latent–physical hybrid fusion vector* ($D = 454$): multi-scale CNN absolute-difference features are concatenated with physically interpretable spectral-index differences, enabling LightGBM to perform internal sparse feature selection via information gain.
3. A *hierarchical mixture-of-experts inference strategy*: a general multi-class Main Model is combined with a binary Expert Model specialized for tree-cut detection via priority-based decision fusion, substantially improving recall for rare but environmentally critical events without retraining the backbone.
4. A *SHAP attribution analysis* over the joint latent–physical feature space, providing partial post-hoc transparency by directly attributing spectral-index contributions while characterizing aggregate reliance on learned spatial context.

In addition to the proposed hybrid framework, we conduct controlled baseline experiments using BIT [16], ChangeFormer [15], and SNUNet [26]. It is important to note architectural distinctions among these baselines: BIT and ChangeFormer employ self-attention over image tokens; SNUNet is a fully convolutional architecture based on nested dense skip connections (UNet++)-it is included as a strong convolutional reference. All baselines are evaluated in pairwise bi-temporal configurations for architectural and computational consistency.

The specific objectives are to: (1) quantify performance differences across binary, multiclass, and multi-temporal CD scenarios; (2) evaluate the impact of spectral indices and spectral resolution (4-band vs. 8-band) on model efficacy; (3) assess post-hoc transparency via SHAP analysis over the joint feature space; and (4) validate the hybrid model's robustness to limited training data and severe class imbalance.

2. Study Area and Data

2.1. Geographic Setting

The study area is located along Khandwa Road in Indore, Madhya Pradesh, India, encompassing an active national highway development corridor (Figure 1). The area spans from Tejaji Nagar Cross in the north to Bheru Ghat (adjacent to the IIT Indore campus) in the south, capturing intensive construction activities associated with the Indore-Khandwa transportation corridor connecting central and southern India. The section studied features complex terrain including elevated structures and tunnels, providing a diverse and dynamic landscape for evaluating CD algorithms [1].

 doi:10.20944/preprints202605.1851.v1

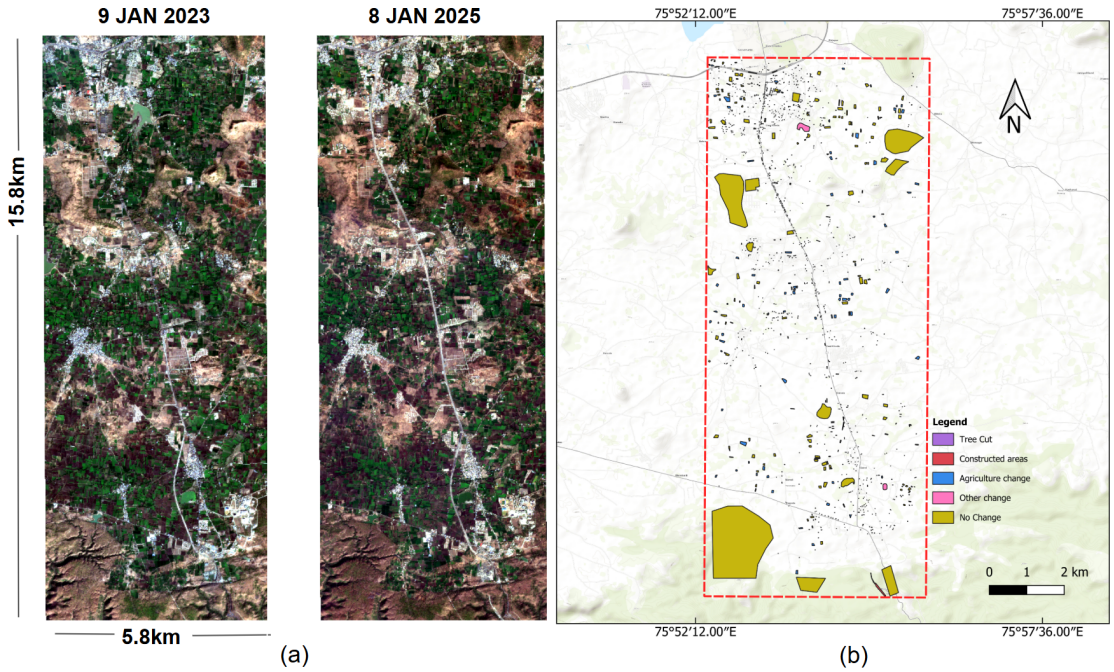

Figure 1. (**a**) Bi-temporal PlanetScope images of the study area with acquisition dates. (**b**) Training labels digitized in QGIS across five classes: Trees Cut, Built-up, Agriculture Change, Other Change, and No Change. Training polygons cover 7.20% (change) and 92.80% (no-change) of the labeled area, reflecting severe class imbalance inherent to infrastructure monitoring scenarios.

For bi-temporal experiments, a rectangular area of interest approximately 15.8 km × 5.8 km (north-south × east-west) captures the core construction zone. For multi-temporal analysis, the same corridor was monitored continuously across 21 PlanetScope acquisitions spanning 9 January 2023 to 8 January 2025 [3].

2.2. Reference Data and Class Distribution

Training and validation data were generated by manually digitizing 1,395 polygon-based sampling units in QGIS, guided by visual interpretation of high-resolution imagery and field surveys conducted between December 2022 and April 2023. Five mutually exclusive change classes were defined: *No Change* (stable land cover including undisturbed vegetation, water bodies, and built-up areas); *Trees Cut* (areas where tree removal was confirmed by field surveys as associated with highway construction); *Built-up* (new impervious surfaces including highway segments, service roads, viaducts, and bridges); *Agriculture Change* (parcels undergoing land-use modification); and *Other Changes* (landscape modifications such as changes in lake extent).

The class distribution exhibits severe imbalance: no-change regions account for 92.80% of the labeled area, while all change categories together represent only 7.20%. Agriculture change (5.02%) is the largest change component, and trees cut (0.09%) constitutes the rarest but environmentally most critical class. The total labeled area is sufficient for the LightGBM-based hybrid pipeline operating at the pixel level, but imposes a well-known data-volume constraint on high-capacity end-to-end deep learning baselines; this constraint is inherent to the real-world operational monitoring scenario and is analyzed in the Results.

For multi-temporal analysis, the reference dataset was expanded to 1,776 polygons distributed across the same corridor, enabling evaluation over extended temporal sequences [21].

2.3. PlanetScope Imagery and Spectral Configuration

PlanetScope imagery was analyzed in two spectral configurations. The standard four-band configuration comprises blue (465-515 nm), green (547-583 nm), red (650-680 nm), and near-infrared (845–885 nm) bands. The extended eight-band configuration additionally includes coastal blue (431-

 doi:10.20944/preprints202605.1851.v1

452 nm), green I (513-549 nm), yellow (600-620 nm), and red-edge (697-713 nm) channels [3]. The red-edge band is particularly relevant for distinguishing chlorophyll loss and vegetation stress from illumination or seasonal variability [27,28], and is hypothesized to provide a distinctive spectral fingerprint for tree-removal events. For bi-temporal image analysis the different configuration of four bands is compared and for multitemporal analysis four-bands vs eight-bands experiments is done.

The temporal gaps between successive image pairs range from approximately 9-15 days at short revisit intervals to approximately four months for longer seasonal gaps, depending on satellite acquisition conditions. This design captures both rapid construction-driven changes (road excavation, structural development) and slower land-cover transitions (vegetation removal, agricultural conversion) [29–31].

2.4. Spectral Indices

All model variants were evaluated using a set of standardized spectral indices computed from the reflectance bands:

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}}, \tag{1}$$

$$\mathrm{EVI} = 2.5\,\frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + 6\,\rho_{\mathrm{Red}} - 7.5\,\rho_{\mathrm{Blue}} + 1}, \tag{2}$$

$$\mathrm{SAVI} = 1.5\,\frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}} + 0.5}, \tag{3}$$

$$\mathrm{NDWI} = \frac{\rho_{\mathrm{Green}} - \rho_{\mathrm{NIR}}}{\rho_{\mathrm{Green}} + \rho_{\mathrm{NIR}}}. \tag{4}$$

For eight-band imagery, two additional red-edge indices were incorporated to exploit chlorophyll-sensitive wavelengths:

$$\mathrm{NDRE} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{RedEdge}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{RedEdge}}}, \tag{5}$$

$$\mathrm{CIRE} = \frac{\rho_{\mathrm{NIR}}}{\rho_{\mathrm{RedEdge}}} - 1. \tag{6}$$

3. Methods

3.1. Experimental Framework

Three CD classification scenarios are investigated: (i) binary change detection (change vs. no change); (ii) binary category-specific detection (built-up expansion, tree-cut vs. no change); and (iii) multiclass categorical detection with five mutually exclusive classes. All models are trained and evaluated using geographic k-fold cross-validation ($k = 5$) with both vertical and horizontal spatial blocking to enforce spatial independence and minimize spatial autocorrelation [32–34].

Folds are constructed by partitioning the scene bounds Ω into K vertical and K horizontal strips. Each polygon centroid (x_p, y_p) is assigned to a fold via:

$$ID_{\mathrm{vert}} = \left\lfloor \frac{y_p - y_{\min}}{\Delta y} \right\rfloor, \qquad ID_{\mathrm{horz}} = \left\lfloor \frac{x_p - x_{\min}}{\Delta x} \right\rfloor, \tag{7}$$

yielding $2K = 10$ spatially disjoint folds. For multi-temporal analysis, temporal leakage is prevented by a stratified group k-fold scheme in which all patches from a given acquisition are assigned exclusively to either training or validation, with stratification by the dominant change class [19,21].

 doi:10.20944/preprints202605.1851.v1

3.2. Deep Learning Baselines

3.2.1. U-Net

The U-Net architecture processes concatenated bi-temporal image pairs through five downsampling levels, a bottleneck, and a mirrored decoder with skip connections [12,13]. Pixel-wise class probabilities are produced by a final 1×1 convolution on 256×256 patches.

3.2.2. Siamese U-Net

The Fully Convolutional Siamese U-Net applies a shared-weight encoder independently to each temporal input [14]. Absolute feature differences across encoder scales (64, 128, 512, and 1024-channel bottleneck) are passed to a mirrored decoder, producing pixel-wise change probabilities.

3.2.3. State-of-the-Art Baselines

Three additional baselines were evaluated in bi-temporal multiclass experiments. BIT [16] and ChangeFormer [15] are transformer-based models employing self-attention over image tokens. SNUNet [26] is a fully convolutional architecture based on nested dense skip connections and is included as a strong convolutional reference. All baselines were restricted to pairwise bi-temporal configurations: SNUNet's dense skip connections are engineered for exactly two branches, and the quadratic self-attention cost of transformer models constrains feasible patch sizes under the available computational budget [22]. Hyperparameters followed respective original publications, with coarse grid-search adjustment under an identical computational budget.

3.3. MS-HySAN Architecture

3.3.1. Feature-Extraction Backbone

The MS-HySAN backbone is a compressed Siamese encoder that differs from the full Siamese U-Net in two deliberate ways: (i) the decoder is removed entirely, so the backbone acts as a frozen feature extractor after pre-training; and (ii) the bottleneck width is reduced from 1024 to 256 channels and the two deepest encoder blocks are removed. This compression reduces feature dimensionality fed to LightGBM, improves runtime, and is motivated by the fact that excessively high-dimensional feature tensors slow downstream GBDT training without commensurate accuracy gains.

The encoder extracts hierarchical features at two scales (s_1: 64 channels; s_2: 128 channels) and a 256-channel bottleneck. At the bottleneck, a Pyramid Attention Module (PAM) captures long-range spatiotemporal correspondences (Figure 2):

$$\widehat{\mathbf{B}}_1 = \mathbf{B}_1 + \gamma\,\mathrm{Softmax}\left(\frac{\mathbf{Q}_{\mathbf{B}_1}\mathbf{K}_{\mathbf{B}_2}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{\mathbf{B}_2}, \tag{8}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are learned linear projections and γ is a learnable scaling parameter. The attention-induced bottleneck change descriptor is:

$$\Delta\mathbf{B} = \left|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\right|. \tag{9}$$

At each encoder scale, absolute feature differences

$$\Delta\mathbf{F}^{(l)} = \left|\mathbf{F}_2^{(l)} - \mathbf{F}_1^{(l)}\right|, \quad l \in \{1,2\}, \tag{10}$$

are computed and upsampled to a common spatial resolution via bilinear interpolation $U(\cdot)$.

 doi:10.20944/preprints202605.1851.v1

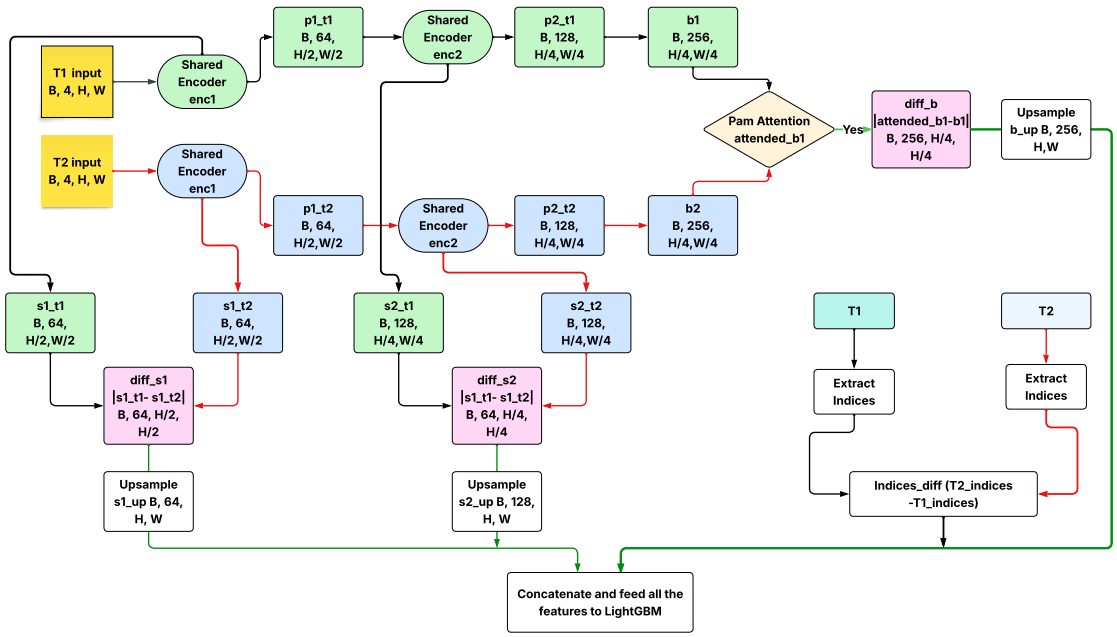

Figure 2. MS-HySAN hybrid architecture. The backbone is pre-trained with focal loss and then frozen. Attention-refined bottleneck features are concatenated with spectral-index differences and/or reflectance differences to form the hybrid feature vector supplied to LightGBM.

3.3.2. Latent–Physical Hybrid Fusion

The latent change features ($64 + 128 + 256 = 448$ channels) are augmented with a physical difference vector $\Delta\mathbf{S}$ that may include reflectance differences (e.g., $\Delta\rho_{\text{red}}$, $\Delta\rho_{\text{NIR}}$) and/or spectral-index differences (ΔNDVI, ΔNDWI, ΔEVI, ΔSAVI, ΔNDRE, ΔCIRE) [35–38]. The hybrid descriptor is:

$$\mathbf{X} = U(\Delta\mathbf{F}^{(1)}) \oplus U(\Delta\mathbf{F}^{(2)}) \oplus U(\Delta\mathbf{B}) \oplus \Delta\mathbf{S} \in \mathbb{R}^{D}, \tag{11}$$

where $\oplus$ denotes channel-wise concatenation and $D = 454$ is the pre-selection dimensionality. Four experimental configurations are evaluated: (a) indices + reflectance, (b) reflectance only, (c) spectral indices only, and (d) no additional features (deep features only).

3.3.3. Decoupled Two-Stage Training Protocol

Training is explicitly two-stage and decoupled (Figure 3). In *Stage A*, the Siamese backbone E_θ is trained end-to-end using focal loss on labeled patch pairs. In *Stage B*, backbone weights are frozen, all training and validation patches are forward-passed to extract multi-scale difference tensors, and LightGBM is fitted on the resulting structured feature matrix augmented with $\Delta\mathbf{S}$. The two stages are never jointly optimized. This design: (i) avoids saddle-point instabilities at the CNN-GBDT interface; (ii) eliminates the need for gradient flow back into the CNN; and (iii) dramatically reduces memory consumption during GBDT fitting, enabling deployment within resource-constrained environments (Section 3.7).

 doi:10.20944/preprints202605.1851.v1

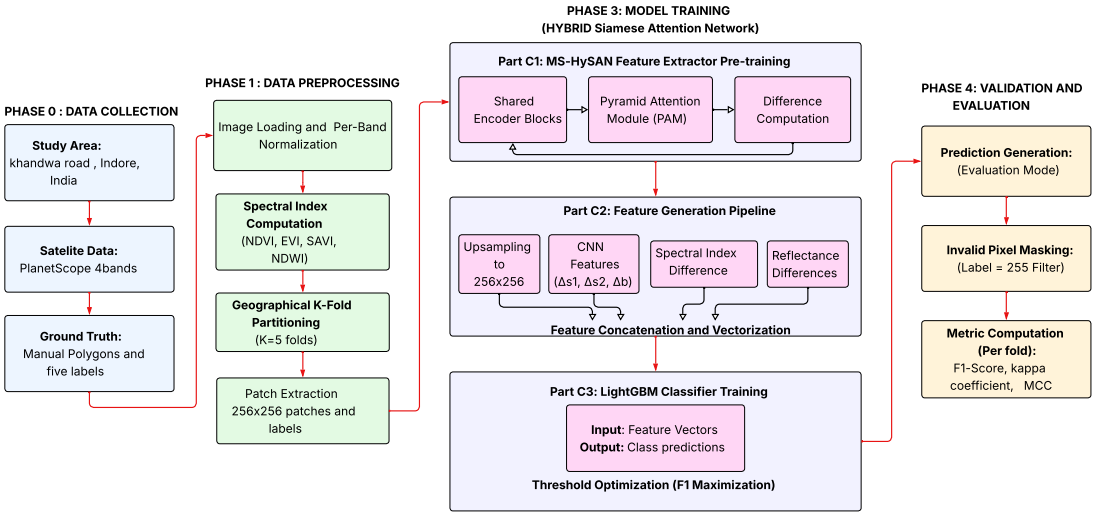

Figure 3. Complete bi-temporal MS-HySAN workflow. Stage A: Siamese backbone pre-training with focal loss. Stage B: frozen feature extraction followed by LightGBM fitting on the structured feature matrix augmented with spectral-index differences.

3.3.4. LightGBM Configuration and Feature Selection

An `LGBMClassifier` is trained per cross-validation fold with `objective` $\in$ {`binary`, `multiclass`}, 1000–1500 estimators, learning rate 0.05, and early stopping (patience = 100). Class imbalance is addressed via `scale_pos_weight`:

$$\text{scale\_pos\_weight} = \frac{N_{\text{negative}}}{N_{\text{positive}}}. \tag{12}$$

Internal feature selection via cumulative information gain yields a sparse subset $|\mathbf{X}^*| \approx$ 192-198 from the initial $D = 454$:

$$\mathbf{X}^* = \{x_j \in \mathbf{X} : \text{Gain}_j > 0\}. \tag{13}$$

For binary configurations, a validation-set threshold sweep over $[0.1, 0.9]$ selects the decision threshold maximizing F1-score.

The complete bi-temporal pipeline is summarized in Algorithm 1.

3.4. Hierarchical Two-Stage Inference and Decision Fusion

Distinguishing tree-cut events from complex background variability is particularly challenging due to seasonal vegetation changes, illumination effects, and extreme class rarity. A hierarchical mixture-of-experts strategy is adopted: (i) a multi-temporal *Main Model* for generalized multiclass detection; and (ii) a bi-temporal *Expert Model* specialized for high-precision tree-cut detection (Figure 4).

 doi:10.20944/preprints202605.1851.v1

Algorithm 1 Hybrid Multi-Scale Siamese Attention Network (MS-HySAN) with Spatial Partitioning

Require: bi-temporal 4-band imagery $\{\mathbf{I}_1, \mathbf{I}_2\} \in \mathbb{R}^{H\times W\times 4}$; reflectance/spectral maps $\{\mathbf{R}_1, \mathbf{R}_2\}$; shapefile $\mathcal{S}$; spatial bounds Ω.

Ensure: Expert ensemble $\{E_\theta, \Phi\}$; spatial performance metrics; final change map $\mathbf{M}$.

Step 1: Robust radiometric normalization

for each band $b \in \{1, \ldots, 4\}$ **do**

$P_{\min}, P_{\max} \leftarrow \text{Percentile}(\mathbf{I}_b, [2, 98])$

$\tilde{\mathbf{I}}_b \leftarrow \text{clip}\left(\dfrac{\mathbf{I}_b - P_{\min}}{P_{\max} - P_{\min} + \varepsilon}, 0, 1\right)$ {sensor-invariant scaling}

end for

Step 2: Spatial coordinate-based partitioning.

Divide scene bounds Ω into K vertical strips and K horizontal strips.

Assign each patch p to spatial folds based on centroid coordinates (x_p, y_p):

$ID_{\text{vert}} \leftarrow \left\lfloor \dfrac{y_p - y_{\min}}{\Delta y} \right\rfloor, \quad ID_{\text{horz}} \leftarrow \left\lfloor \dfrac{x_p - x_{\min}}{\Delta x} \right\rfloor$

// Total 2K spatial folds created to evaluate geographic robustness.

Stage A: Siamese backbone pre-training.

Map normalized inputs through shared encoder E_θ to extract hierarchical features:

$\mathbf{F}_t^{(1)} \in \mathbb{R}^{64}$, $\mathbf{F}_t^{(2)} \in \mathbb{R}^{128}$, $\mathbf{B}_t \in \mathbb{R}^{256}$ for $t \in \{T_1, T_2\}$.

Optimize E_θ end-to-end by minimizing focal loss $\mathcal{L}_{\text{FL}}$.

Apply PAM to bottleneck: $\widehat{\mathbf{B}}_1 \leftarrow \text{PAM}(\mathbf{B}_1, \mathbf{B}_2)$; define $\Delta\mathbf{B} \leftarrow \left|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\right|$.

Freeze E_θ after pre-training.

Stage B: Physical—Latent Hybrid Fusion and GBDT fitting.

Compute biophysical change $\Delta\mathbf{S} \leftarrow \mathbf{R}_2 - \mathbf{R}_1 \in \mathbb{R}^{d_s}$, where $d_s \in \{2, 4, 6\}$.

Form the initial hybrid vector $\mathbf{X} \in \mathbb{R}^{454}$ via upsampling $U(\cdot)$ and concatenation:

$\mathbf{X} \leftarrow \left[U(\Delta\mathbf{F}^{(1)}) \oplus U(\Delta\mathbf{F}^{(2)}) \oplus U(\Delta\mathbf{B}) \oplus \Delta\mathbf{S}\right]$.

Step 5: Spatial CV and sparse feature selection.

for each fold $k \in \{1, \ldots, 2K\}$ **do**

Train GBDT Φ_k on $\mathbf{X}$ to identify the active subset $\mathbf{X}^* = \{x_j \in \mathbf{X} : \text{Gain}_j > 0\}$.

// Observation: $|\mathbf{X}^*| \in \{192, 194, 196, 198\}$ *based on* $\Delta\mathbf{S}$ *gain optimization.*

end for

Step 6: Inference and visualization

$\hat{y} \leftarrow \arg\max \Phi(\mathbf{X}^*)$; **return** final classification map $\mathbf{M}$.

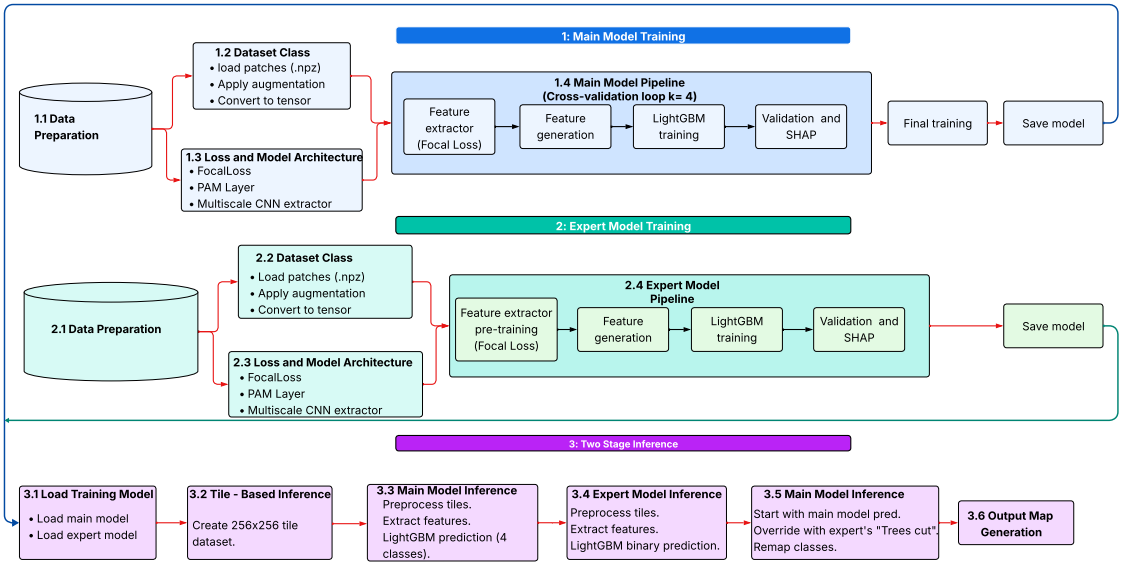

Figure 4. Multi-temporal hierarchical inference workflow. The Main Model provides general multiclass predictions; the Expert Model overrides with high-confidence tree-cut predictions via priority-based fusion.

 doi:10.20944/preprints202605.1851.v1

3.4.1. Stage 1: Multi-Temporal Main Model

Let $\mathcal{C} = \{\mathbf{I}_{d_1}, \ldots, \mathbf{I}_{d_M}\}$ denote the collection of multispectral images. Per-band normalization uses global 1st–99th percentiles over $\mathcal{C}$:

$$\tilde{x}_{i,j,b} = \text{clip}\left(\frac{x_{i,j,b} - P_1(b)}{P_{99}(b) - P_1(b) + \varepsilon},\ 0,\ 1\right). \tag{14}$$

The Siamese backbone extracts multi-scale change descriptors at $L = 3$ scales. For each bi-temporal pair $(\mathbf{I}_{t_1}, \mathbf{I}_{t_2})$:

$$\mathbf{D}^{(l)} = \left| f_\theta^{(l)}(\mathbf{I}_{t_1}) - f_\theta^{(l)}(\mathbf{I}_{t_2}) \right|, \quad l \in \{1, 2, 3\}. \tag{15}$$

At the bottleneck ($l = 3$), PAM refines the change descriptor:

$$\mathbf{D}_{\text{attn}}^{(3)} = \mathbf{D}^{(3)} + \gamma\, \text{Softmax}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V}. \tag{16}$$

The spectral difference vector $\Delta\mathbf{S} = \mathbf{S}_{t_2} - \mathbf{S}_{t_1}$ is formed from the six-index stack $\mathbf{S} = [\text{NDVI}, \text{NDWI}, \text{EVI}, \text{SAVI}, \text{NDRE}, \text{CI}$ and the initial hybrid descriptor is:

$$\mathbf{X}_{\text{init}} = U(\mathbf{D}^{(1)}) \oplus U(\mathbf{D}^{(2)}) \oplus U(\mathbf{D}_{\text{attn}}^{(3)}) \oplus \Delta\mathbf{S} \in \mathbb{R}^{454}. \tag{17}$$

Temporal leakage is prevented by group k-fold partitioning based on unique date pairs, stratified by dominant change class. A LightGBM classifier Φ is trained on the combined feature space with internal feature selection yielding $|\mathbf{X}^*| \approx 198$. The Main Model training procedure is detailed in Algorithm 2.

Stage 2: Expert Model—Centroid-Targeted Binary Learning

The Expert Model focuses on a specific bi-temporal pair $(\mathbf{I}_{T1}, \mathbf{I}_{T2})$ and uses centroid-targeted sampling on ground-truth change polygons $p \in \mathcal{P}$. For a polygon with vertices $(x_j, y_j)_{j=1}^n$, the centroid is

$$\mathbf{c} = \left(\frac{1}{n}\sum_{j=1}^{n} x_j,\ \frac{1}{n}\sum_{j=1}^{n} y_j\right), \tag{18}$$

and local patches $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{\omega\times\omega\times 8}$ are extracted around $\mathbf{c}$ with $\omega = 128$. This concentrates supervision on semantic change cores, reducing boundary mixing and improving class separability for the *Trees Cut* class.

The Expert Siamese encoder E_{exp} is trained with the binary focal loss $\mathcal{L}_{\text{exp}}$ defined above, to handle severe class imbalance [19,39,40]. Deep temporal difference features $\Delta\mathbf{F}_{\text{deep}} \in \mathbb{R}^{448}$ are extracted, and a spectral-index difference vector

$$\Delta\mathbf{s} = \mathbf{S}(\mathbf{I}_{T2}) - \mathbf{S}(\mathbf{I}_{T1}) \in \mathbb{R}^6 \tag{19}$$

is computed. These are concatenated into a hybrid vector

$$\mathbf{x}_{\text{exp}} = [\Delta\mathbf{F}_{\text{deep}} \parallel \Delta\mathbf{s}] \in \mathbb{R}^{454}. \tag{20}$$

A GBDT ensemble Φ_{exp} performs internal feature selection via information gain I_j, identifying an active subset:

$$\mathbf{x}^* = \{x_j \in \mathbf{x}_{\text{exp}} : I_j > 0\}, \quad |\mathbf{x}^*| = 198. \tag{21}$$

The Expert Model is trained on these sparse informative dimensions, and SHAP values are computed for each feature j [17]:

$$\phi_j = \sum_{S\subseteq N\setminus\{j\}} \frac{|S|!(n - |S| - 1)!}{n!}\Big[f(S\cup\{j\}) - f(S)\Big], \tag{22}$$

 doi:10.20944/preprints202605.1851.v1

Algorithm 2 Main Model: multi-temporal Hy-SAN (Hybrid Model) Training

Require: Image collection $\mathcal{C} = \{\mathbf{I}_{d_1}, \mathbf{I}_{d_2}, \ldots, \mathbf{I}_{d_M}\}$; shapefile $\mathcal{S}$ defining N temporal pairs; labels $\mathbf{Y}$; robust percentiles $P_1(b), P_{99}(b)$

Ensure: Final ensemble $\{E_\theta, \Phi\}$; SHAP importances $\boldsymbol{\phi}_{\text{avg}}$

Step 1: Pre-processing and pair generation.

From $\mathcal{C}$ and $\mathcal{S}$, construct a multi-temporal training set $\mathcal{D} = \{(\mathbf{I}_{t_1}, \mathbf{I}_{t_2}, \mathbf{Y})_n\}_{n=1}^{N}$.

for each image $\mathbf{I} \in \mathcal{C}$ **do**

$$\tilde{x}_{i,j,b} = \text{clip}\left(\frac{x_{i,j,b} - P_1(b)}{P_{99}(b) - P_1(b) + \varepsilon}, 0, 1\right) \text{ \{global normalization\}}$$

$$\mathbf{S}(\mathbf{I}) = [\text{NDVI}(\mathbf{I}), \text{NDWI}(\mathbf{I}), \ldots, \text{CIRE}(\mathbf{I})]^\top \text{ \{spectral indices\}}$$

end for

Step 2: Stratified group K-fold split.

Partition the N pairs in $\mathcal{D}$ into K sets $\{G_1, \ldots, G_K\}$ based on unique date pairs to prevent temporal data leakage, stratified by dominant change class.

Stage A: Multi-temporal backbone pre-training.

for fold $k = 1$ to K **do**

for each pair $n \in$ TrainFolds **do**

Optimize the Siamese extractor E_{θ_k} end-to-end by minimizing focal loss $\mathcal{L}_{\text{FL}}\big(E(\mathbf{I}_{t_1}, \mathbf{I}_{t_2})_n, \mathbf{Y}_n\big)$.

end for

Freeze E_{θ_k}.

Feature extraction and attention

$$\mathbf{D}^{(l)} = \left| f_{\theta_k}^{(l)}(\mathbf{I}_{t_1}) - f_{\theta_k}^{(l)}(\mathbf{I}_{t_2}) \right|, \text{ for } l \in \{1, 2, 3\}.$$

$$\mathbf{D}_{\text{attn}}^{(3)} = \text{PAM}(\mathbf{D}^{(3)}).$$

end for

Stage B: Hybrid fusion of deep and physical features.

Formulate the hybrid descriptor $\mathbf{X}_{\text{init}} \in \mathbb{R}^{454}$ via spatial alignment $U(\cdot)$:

$$\mathbf{X}_{\text{init}} = [U(\mathbf{D}^{(1)}) \oplus U(\mathbf{D}^{(2)}) \oplus U(\mathbf{D}_{\text{attn}}^{(3)}) \oplus (\mathbf{S}_{t_2} - \mathbf{S}_{t_1})].$$

Step 5: GBDT classification with internal feature selection.

Train LightGBM Φ on the combined feature space from all N pairs.

Identify the active feature subset $\mathbf{X}^* = \{x_j \in \mathbf{X}_{\text{init}} : \text{Gain}_j > 0\}$, where $|\mathbf{X}^*| = 198$.

Predict $\hat{y} = \arg\max \sum_{m=1}^{1000} g_m(\mathbf{X}^*)$.

Step 6: multi-temporal SHAP interpretability

$\phi_j = \text{ShapleyValue}(\Phi, \mathbf{X}^*)$ {deep vs. spectral contributions}

return $E_\theta, \Phi, \boldsymbol{\phi}_{\text{avg}}$

where $f(\cdot)$ is the model's prediction function.

3.4.2. Stage 2: Expert Model

The Expert Model focuses on a specific bi-temporal pair $(\mathbf{I}_{T1}, \mathbf{I}_{T2})$ and uses centroid-targeted sampling: for each ground-truth polygon with vertices $(x_j, y_j)_{j=1}^{n}$, a 128×128 patch is extracted centered on the geometric centroid

$$\mathbf{c} = \left(\frac{1}{n}\sum_{j=1}^{n} x_j,\ \frac{1}{n}\sum_{j=1}^{n} y_j\right), \tag{23}$$

concentrating supervision on semantic change cores and reducing boundary mixing. The Expert Siamese encoder is trained with a class-balanced binary focal loss:

$$\mathcal{L}_{\text{exp}} = -\big[y\,\alpha(1-p)^\gamma \log p + (1-y)(1-\alpha)\,p^\gamma \log(1-p)\big], \tag{24}$$

and a GBDT ensemble Φ_{exp} is fitted on the same $\mathbb{R}^{454}$ hybrid feature vector.

 doi:10.20944/preprints202605.1851.v1

3.4.3. Priority-Based Decision Fusion

If the Expert Model predicts a positive tree-cut label, this prediction overrides the Main Model output; otherwise, the Main Model prediction is retained. This hierarchical design improves recall and precision for rare tree-cut events while preserving the broad semantic coverage of the multiclass Main Model.

3.5. Loss Functions and Implementation Details

All deep models use Focal Loss with $\gamma = 2.0$:

$$\mathcal{L}_{\mathrm{FL}} = -\sum_{t} w_t \, (1 - p_t)^{\gamma} \log p_t, \quad w_t = \frac{N_{\mathrm{total}}}{2\, N_t}, \tag{25}$$

where N_t is the pixel count for class t.

Models are implemented in PyTorch, trained from random initialization with AdamW (initial lr $= 10^{-3}$, cosine annealing schedule) [41], batch size 8, and 15–20 epochs. Data augmentation comprises random horizontal and vertical flips and color jitter applied to RGB composites. Patch extraction uses 128×128 patches with stride 64; per-band normalization uses $[P_2, P_{98}]$ percentiles for bi-temporal experiments and $[P_1, P_{99}]$ for multi-temporal experiments:

$$\tilde{x}_{i,j,b} = \mathrm{clip}\left(\frac{x_{i,j,b} - P_{\alpha}(b)}{P_{\beta}(b) - P_{\alpha}(b) + \varepsilon},\ 0,\ 1 \right). \tag{26}$$

3.6. SHAP Interpretability and Performance Metrics

SHAP values are computed using the `TreeExplainer` algorithm [17,42] on a subsampled validation set after LightGBM training. Attributions are computed over the joint feature vector $\mathbf{X}^*$, which contains both physically defined spectral-index difference channels and opaque CNN-derived channels. This provides *partial* post-hoc transparency: spectral-index contributions carry direct biophysical meaning; the CNN subset quantifies aggregate reliance on learned spatial context. We do not claim individual interpretability for the CNN channels. Mean absolute SHAP values are aggregated over all folds.

Performance is evaluated with three complementary metrics suited to class-imbalanced conditions:

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{27}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{28}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \cdot \mathrm{TN}) - (\mathrm{FP} \cdot \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}. \tag{29}$$

MCC is used as the primary ranking metric given its robustness to class imbalance.

3.7. Computational Environment

All experiments were conducted in the Kaggle notebook environment using an NVIDIA Tesla P100 GPU with 16 GB VRAM and 29 GB system RAM. All runs were designed to complete within the 12-hour session limit using up to 20 GB persistent disk storage.

4. Results

4.1. Binary Change Detection

Tables 1–3 report bi-temporal binary change detection results for three scenarios of increasing difficulty: overall change discrimination, category-specific built-up expansion, and the most challenging tree-cut class. The hybrid model consistently and substantially outperforms both deep learning

 doi:10.20944/preprints202605.1851.v1

baselines and standalone LightGBM variants across all three scenarios, confirming that spatial context from the CNN backbone and physical discriminability from spectral-index differences are jointly necessary for reliable CD.

Table 1. Bi-temporal binary change detection: Change vs. No-Change. The best result per metric(average) is in **bold**.

Model	MCC	Kappa	F1-score
U-Net	0.88	0.85	0.86
Siamese U-Net	0.87	0.87	0.90
LightGBM (Reflectance)	0.39	0.38	0.40
LightGBM (4 Indices)	0.73	0.72	0.74
LightGBM (Indices + Ref.)	0.67	0.67	0.68
Hybrid (Reflectance)	0.92	0.92	0.93
Hybrid (4 Indices)	0.90	0.90	0.92
Hybrid (Indices + Ref.)	**0.94**	**0.93**	**0.95**

Table 2. Bi-temporal binary change detection: Built-up Area Change. The best result per metric (average) is in **bold**.

Model	MCC	Kappa	F1-score
U-Net	0.75	0.73	0.74
Siamese U-Net	0.70	0.66	0.67
LightGBM (Reflectance)	0.44	0.40	0.41
LightGBM (4 Indices)	0.44	0.39	0.40
LightGBM (Indices + Ref.)	0.58	0.53	0.54
Hybrid (Reflectance)	0.87	0.86	0.87
Hybrid (4 Indices)	0.89	0.89	0.89
Hybrid (Indices + Ref.)	**0.92**	**0.93**	**0.93**

Table 3. Bi-temporal binary change detection: Trees-Cut Class. The best result per metric (average) is in **bold**.

Model	MCC	Kappa	F1-score
U-Net	0.37	0.29	0.30
Siamese U-Net	0.39	0.31	0.30
LightGBM (Reflectance)	0.34	0.29	0.29
LightGBM (4 Indices)	0.28	0.23	0.25
LightGBM (Indices + Ref.)	0.24	0.17	0.17
Hybrid (Indices + Ref.)	0.74	0.72	0.73
Hybrid (4 Indices)	**0.77**	**0.77**	**0.76**

4.1.1. Change vs. No-Change

The hybrid (Indices + Ref.) achieves the highest performance (MCC = 0.94, F1 = 0.95), improving over U-Net (MCC = 0.88) by +6 MCC points and over standalone LightGBM with reflectance (MCC = 0.39) by a margin of +55 points (Table 1). The striking gap between standalone LightGBM (MCC = 0.39–0.73) and the hybrid variants (MCC = 0.90–0.94) confirms that spatial context from the CNN backbone is essential for separating spectrally ambiguous no-change and change regions: spectral differences alone are insufficient to resolve spatial variability at pixel level without the learned contextual features.

 doi:10.20944/preprints202605.1851.v1

4.1.2. Built-Up Area Change

For built-up expansion (Table 2), Hybrid (Indices + Ref.) leads with MCC = 0.92, improving over U-Net (MCC = 0.75) by +17 points and over all standalone LightGBM variants (MCC = 0.44-0.58) by large margins. Notably, Siamese U-Net (MCC = 0.70) underperforms U-Net (MCC = 0.75) for this class, likely because asymmetric reflectance changes from newly paved impervious surfaces are better handled by the full-band concatenation input strategy of U-Net than by the shared-weight absolute difference encoding of the Siamese architecture.

4.1.3. Tree-Cut Detection

The tree-cut binary task is the most challenging scenario, reflecting both extreme class rarity (0.09% of labeled area) and spectral ambiguity in four-band imagery (Table 3). End-to-end deep models, U-Net (MCC = 0.37) and Siamese U-Net (MCC = 0.39)vperform only marginally above standalone LightGBM variants (MCC = 0.24-0.34), indicating that neither approach alone provides sufficient discriminative signal for this rare class. The Hybrid (4 Indices) achieves MCC = 0.77, a +38-point improvement over the best deep baseline. This result demonstrates that combining CNN spatial features with vegetation-sensitive index differences (primarily NDVI and NDWI) recovers the discriminative signal absent from either component individually. The slight advantage of Hybrid (4 Indices) over Hybrid (Indices + Ref.) suggests that raw reflectance differences introduce some redundant noise relative to the spectral index signal for this spectrally subtle class.

4.2. Bi-Temporal Multiclass Analysis

Table 4 summarizes five-class bi-temporal classification results. The hybrid model combining spectral indices and reflectance achieves the highest overall agreement (F1 = 0.95, Kappa = 0.90, MCC = 0.90), outperforming all baseline architectures. Hybrid configurations using only reflectance or only indices show reduced but still competitive performance, confirming that neither handcrafted spectral features nor deep features alone are sufficient to capture the diversity of land-cover transitions present in infrastructure-driven change scenarios.

Table 4. Bi-temporal multiclass change detection (5 classes). Results are mean ± std over $2K = 10$ spatial folds. The best result per metric among all models is in **bold**. Baselines are grouped by category.

Model	F1-score	Kappa	MCC
Proposed hybrid variants			
Hybrid (Indices + Ref.)	**0.95 ± 0.05**	**0.90 ± 0.08**	**0.90 ± 0.09**
Hybrid (Reflectance only)	0.95 ± 0.05	0.89 ± 0.09	0.89 ± 0.09
Hybrid (4 Spectral Indices)	0.95 ± 0.05	0.83 ± 0.10	0.83 ± 0.10
Hybrid (No Add-on Features)	0.92 ± 0.05	0.80 ± 0.11	0.80 ± 0.11
End-to-end deep learning			
U-Net	0.89 ± 0.06	0.73 ± 0.11	0.74 ± 0.10
Siamese U-Net	0.89 ± 0.07	0.71 ± 0.15	0.73 ± 0.13
ChangeFormer (transformer)	0.88 ± 0.05	0.67 ± 0.13	0.69 ± 0.11
SNUNet (dense-skip CNN)	0.84 ± 0.14	0.64 ± 0.20	0.67 ± 0.19
BIT (transformer)	0.65 ± 0.15	0.19 ± 0.13	0.20 ± 0.12
Classical reference			
Random Forest	0.90 ± 0.11	0.77 ± 0.19	0.75 ± 0.22

Classical encoder–decoder architectures (U-Net: MCC = 0.74; Siamese U-Net: MCC = 0.73) exhibit substantially lower Kappa and MCC values, reflecting confusion between minority classes under extreme class imbalance. Random Forest (MCC = 0.75) is included as a classical reference and performs comparably to the deep baselines, further underscoring the robustness advantage of combining deep feature extraction with GBDT classification.

 doi:10.20944/preprints202605.1851.v1

The lower performance of ChangeFormer and SNUNet relative to their published benchmark results reflects a combination of insufficient training data for high-capacity models and suboptimal hyperparameter transfer between the standard LEVIR-CD protocol and the geographically blocked setup-not architectural failure alone. BIT's severe degradation (MCC = 0.20) is attributable to its reliance on ImageNet pre-trained weights incompatible with 4/8-band PlanetScope inputs and its high quadratic self-attention cost under the available patch budget. Representative change maps are shown in Figure 5, where the hybrid framework produces spatially coherent, low-noise predictions compared to the salt-and-pepper patterns exhibited by deep baselines.

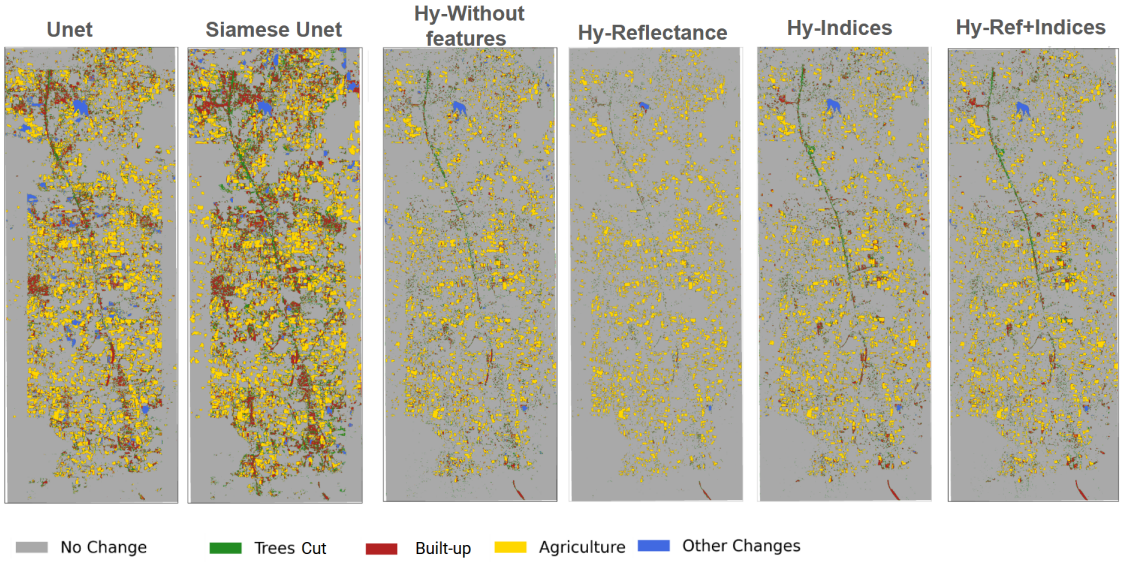

Figure 5. Bi-temporal change maps from PlanetScope imagery. Yellow: agriculture change; green: trees cut and built-up; blue: other changes. The hybrid model produces spatially coherent, boundary-localized predictions with reduced false positives compared to deep baselines.

4.2.1. SHAP Feature Attribution (Bi-Temporal)

Figure 6 shows SHAP feature attributions for the four bi-temporal hybrid configurations. For the indices + reflectance configuration, near-infrared and red reflectance differences are the most influential physically interpretable contributors, followed by NDWI, NDVI, EVI, and SAVI differences. This pattern reflects the importance of vegetation loss and impervious surface expansion, the primary spectral signatures of highway construction in separating change from stable land cover. When only spectral indices are used, NDWI dominates, confirming that moisture-related changes provide the strongest separability in the absence of raw reflectance. When only reflectance is used, NIR and red differences dominate. Across all configurations, the remaining substantial contributions originate from CNN-derived channels encoding higher-order spatial and contextual information not captured by handcrafted indices.

 doi:10.20944/preprints202605.1851.v1

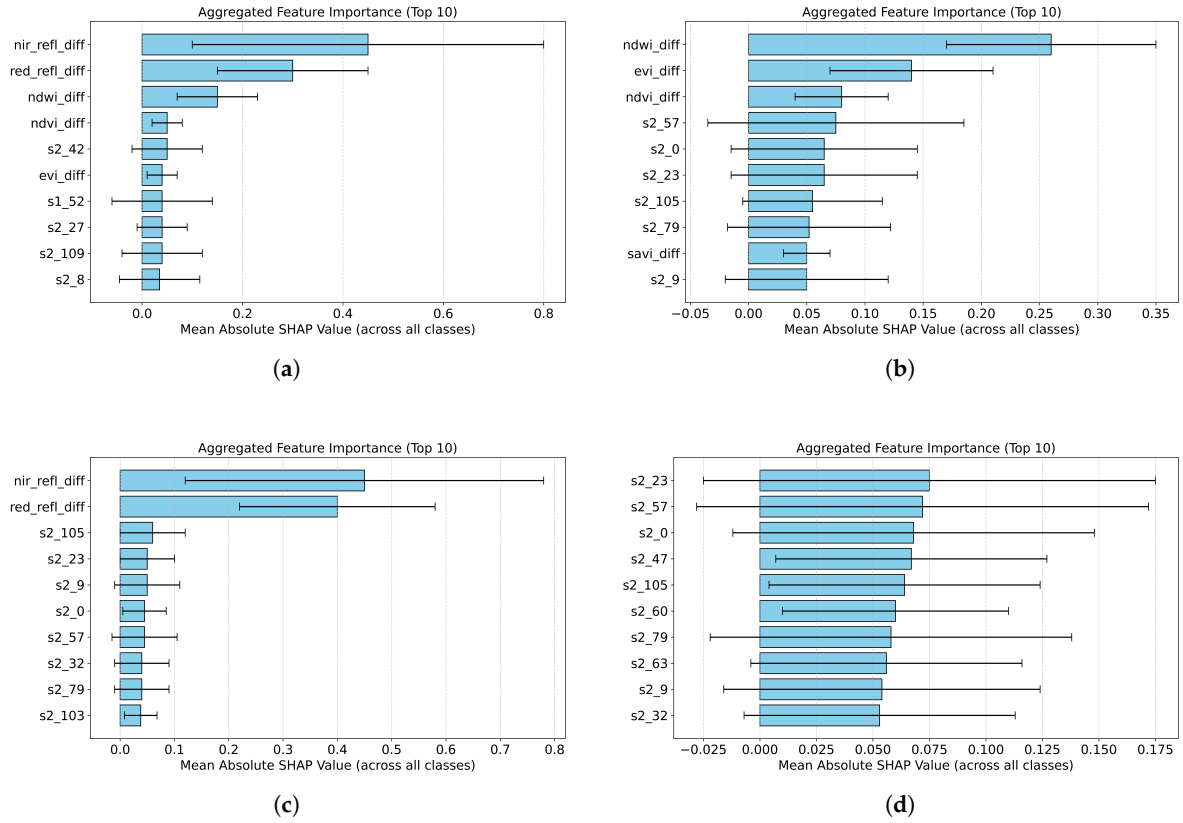

Figure 6. SHAP top-10 feature importance for bi-temporal hybrid configurations: (**a**) indices + reflectance; (**b**) indices only; (**c**) reflectance only; (**d**) deep features only. Spectral-index channels carry direct physical interpretation; CNN channels represent aggregate reliance on learned spatial context.

4.3. Multi-Temporal Analysis: 4-Band PlanetScope

Tables 5 and 6 report results for 21 sequential four-band acquisitions. In the main training regime, the hybrid (F1 = 0.96, MCC = 0.91) performs comparably to the Siamese U-Net (F1 = 0.97, MCC = 0.95); the modest gap is consistent with the main-training pattern observed in bi-temporal experiments. The critical distinction emerges in the expert-training scenario, where labeled samples for rare tree-cut events are severely limited: the hybrid framework maintains stable performance (F1 = 0.86, MCC = 0.87), while the Siamese U-Net experiences catastrophic collapse (F1 = 0.20, MCC = 0.20). Generated change maps are shown in Figure 7.

Table 5. Multi-temporal change detection: 4-band PlanetScope (main training). Results are mean ± std over K temporal folds.

Model	F1-score	Kappa	MCC
Siamese U-Net	0.97 ± 0.01	0.95 ± 0.02	0.95 ± 0.02
Hybrid (Spectral Indices)	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$
Hybrid (without indices)	0.95 ± 0.02	0.90 ± 0.03	0.91 ± 0.03
U-Net	0.93 ± 0.02	0.91 ± 0.04	0.95 ± 0.04

Table 6. Multi-temporal change detection: 4-band PlanetScope (expert training). The best hybrid result is in **bold**.

Model	MCC	Kappa	F1-score
Hybrid (Spectral Indices)	**0.87**	**0.86**	**0.86**
U-Net	0.83	0.82	0.82
Hybrid (without indices)	0.77	0.77	0.77
Siamese U-Net	0.20	0.19	0.20

 doi:10.20944/preprints202605.1851.v1

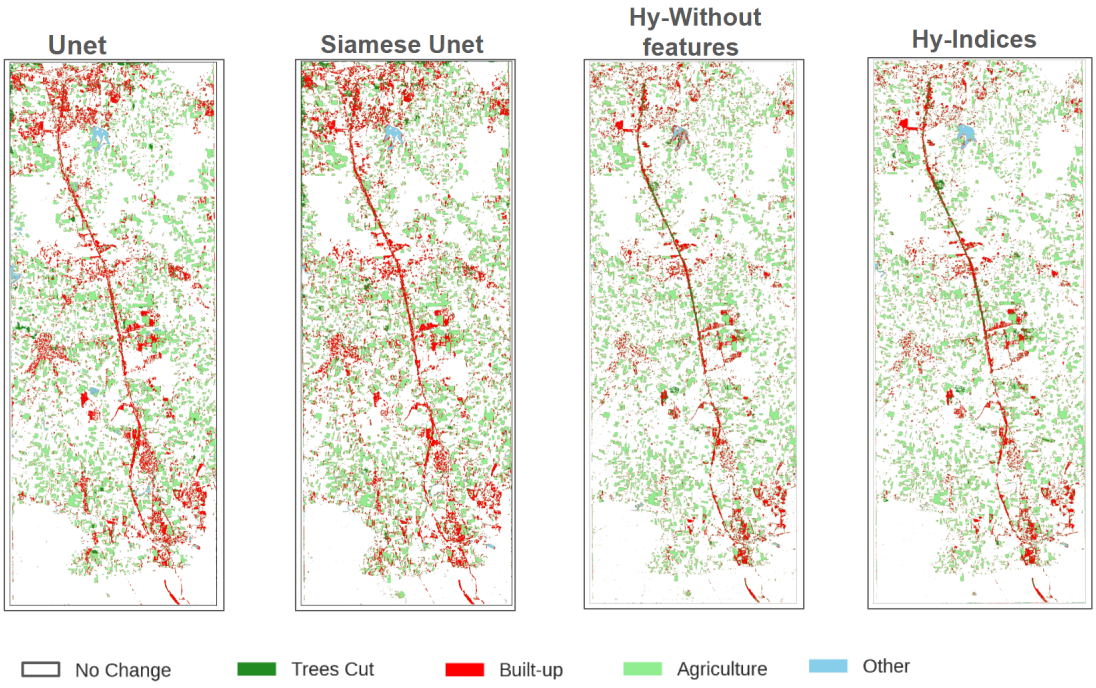

Figure 7. Change maps for 4-band bi-temporal images generated by the multi-temporal hybrid model trained on 21 sequential acquisitions.

4.4. Multi-Temporal Analysis: 8-Band PlanetScope

Tables 7 and 8 present results for eight-band imagery. A pronounced result is the dramatic recovery of the Siamese U-Net from its 4-band expert-training collapse: Kappa increases from 0.19 to 0.90 in the 8-band expert regime. In the main training regime, the Siamese U-Net (F1 = 0.97) modestly outperforms the hybrid (F1 = 0.94). In expert training, the hybrid (F1 = 0.89) essentially matches Siamese U-Net (F1 = 0.90), with both architectures benefiting from the richer spectral fingerprint. Transformer-based architectures were not evaluated in 8-band or full multi-temporal configurations, as their ImageNet-initialized encoders are incompatible with non-RGB input, and training vision transformers from scratch on high-spectral- dimensionality data with limited labels leads to unstable optimization and poor generalization [23].

Table 7. Multi-temporal change detection: 8-band PlanetScope (main training).

Model	F1-score	Kappa	MCC
Siamese U-Net	**0.97 ± 0.01**	**0.94 ± 0.02**	**0.94 ± 0.02**
Hybrid (Spectral Indices)	0.94 ± 0.03	0.87 ± 0.04	0.88 ± 0.04
Hybrid (without indices)	0.93 ± 0.03	0.85 ± 0.05	0.85 ± 0.05
U-Net	0.90 ± 0.09	0.82 ± 0.04	0.82 ± 0.04

Table 8. Multi-temporal change detection: 8-band PlanetScope (expert training).

Model	MCC	Kappa	F1-score
Siamese U-Net	**0.91**	**0.90**	**0.90**
Hybrid (Spectral Indices)	0.88	0.89	0.89
Hybrid (without indices)	0.84	0.85	0.84
U-Net	0.81	0.80	0.80

 doi:10.20944/preprints202605.1851.v1

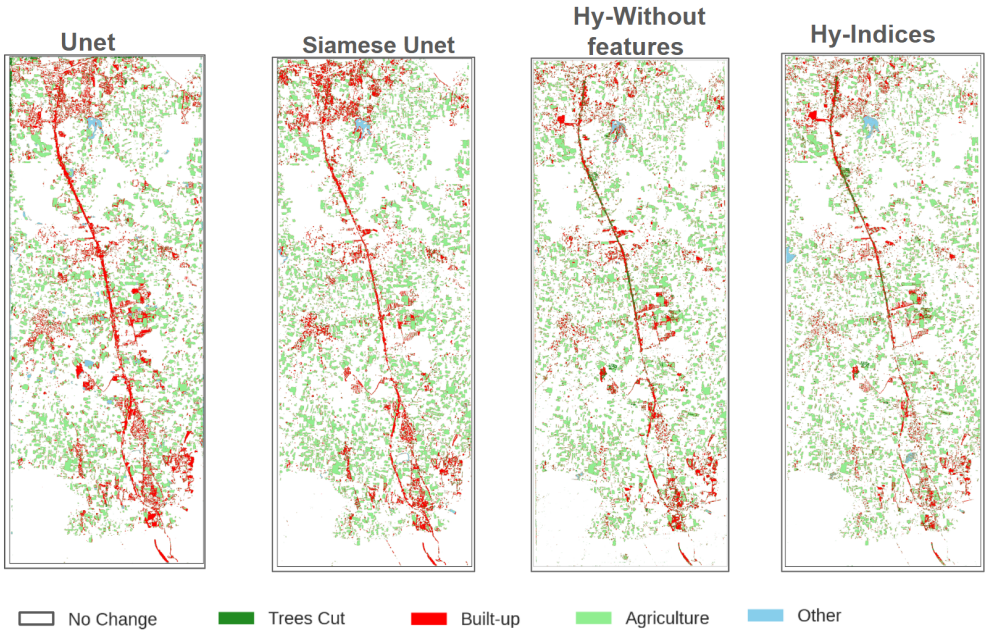

Figure 8. Change maps for 8-band bi-temporal images generated by the multi-temporal hybrid model.

4.4.1. SHAP Feature Attribution (Multi-Temporal)

Figure 9 presents SHAP attributions for multi-temporal hybrid configurations. For the 8-band configuration, NDWI difference is the strongest contributor, followed by NDRE, NDVI, CIRE, EVI, and SAVI differences. The prominence of red-edge indices (NDRE, CIRE) underscores the critical role of chlorophyll-sensitive bands in distinguishing persistent vegetation removal from seasonal variability across long temporal sequences [27,28]. For the 4-band configuration, NDWI again ranks first, with NDVI, EVI, and SAVI following; the absence of red-edge information limits spectral discrimination capacity, but moisture and greenness indices still provide robust temporal cues [35,36]. Across both configurations, the dominance of index-difference features confirms that relative spectral change is more informative than absolute reflectance for multi-temporal CD [29].

 doi:10.20944/preprints202605.1851.v1

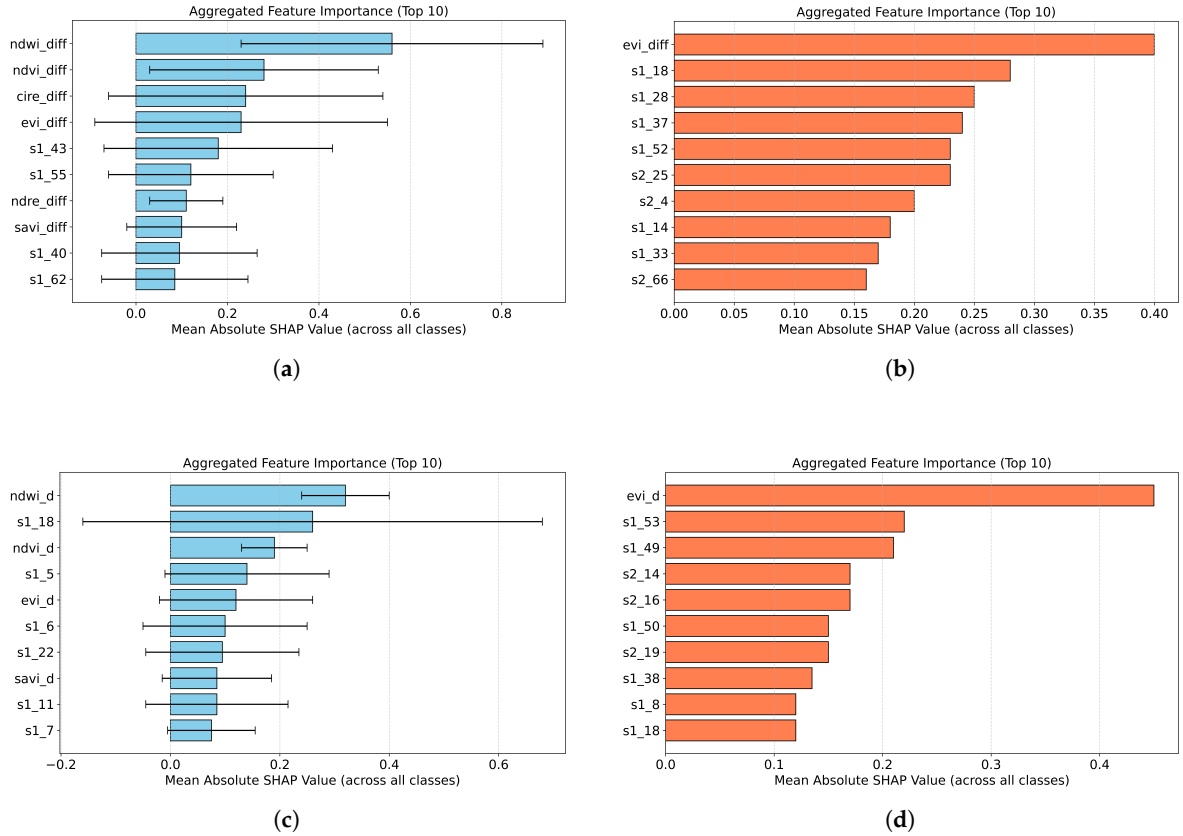

Figure 9. SHAP top-10 feature importance for multi-temporal hybrid models: (**a**) 8-band main model; (**b**) 8-band expert model; (**c**) 4-band main model; (**d**) 4-band expert model.

5. Discussion

5.1. Hybrid vs. End-to-End Deep Learning: Robustness Analysis

The results reveal a consistent pattern across all scenarios: the hybrid CNN–LightGBM framework provides superior robustness to class imbalance and limited supervision, while remaining competitive in data-abundant main-training regimes. This behavior is most dramatically illustrated in the 4-band expert-training scenario (Siamese U-Net: F1 = 0.20 → Hybrid: F1 = 0.86), but the advantage holds across binary tasks most notably for the rare tree-cut class (+38 MCC points over deep baselines).

The structural reason lies in the decoupled design. Gradient-boosted classifiers are insensitive to feature scale, do not require gradient flow through the CNN, and can directly address imbalance at the tree-split level via `scale_pos_weight` without susceptibility to the saddle-point instabilities that affect joint CNN–classifier optimization [11,43]. Furthermore, LightGBM's internal feature selection discards redundant or noisy dimensions, providing an implicit regularization that benefits sparse-label scenarios.

An important nuance from the binary built-up results is that Siamese U-Net underperforms U-Net for built-up expansion. This reflects a fundamental limitation of absolute-difference encoding: it discards directional change information, which is consequential when new impervious surfaces exhibit asymmetric spectral changes (bright across multiple bands) relative to the pre-change state. U-Net's full-band concatenation retains this directionality. The hybrid framework avoids this limitation by pairing the CNN backbone with spectral-index differences that can encode signed change direction.

5.2. Spectral Dimensionality Effects: 4-Band vs. 8-Band

A central finding is that spectral richness and architectural robustness are partially substitutable. In four-band imagery, absolute difference operations produce ambiguous change signals because tree felling, cloud shadows, seasonal browning, and illumination changes produce similar magnitude

 doi:10.20944/preprints202605.1851.v1

responses in RGB+NIR. The Siamese U-Net cannot disambiguate these signals under data scarcity, leading to the 4-band expert collapse. The hybrid model partially resolves this ambiguity through vegetation-sensitive index differences (NDVI, NDWI) even with 4-band input.

Eight-band imagery resolves the ambiguity more fundamentally by providing a distinctive red-edge spectral fingerprint for vegetation removal: tree felling simultaneously eliminates chlorophyll content (captured by red-edge bands) and vegetation structure (NIR), while increasing red and blue reflectance due to exposed soil. Seasonal changes primarily affect NIR and chlorophyll bands but preserve structural characteristics; shadow effects reduce reflectance uniformly across all bands. This high-dimensional fingerprint enables even a CNN-only decoder to learn correct decision boundaries from few samples, explaining the Siamese U-Net's recovery in the 8-band expert regime (Kappa: 0.19 $\rightarrow$ 0.90).

In the 8-band main-training regime, however, the Siamese U-Net (F1 = 0.97) slightly outperforms the hybrid (F1 = 0.94). We attribute this to a dimensionality-capacity mismatch: the absolute-difference operation discards signed change information that is more consequential for 8-band imagery, whereas the Siamese U-Net decoder retains signed multi-resolution skip features throughout. These findings establish a general principle: the hybrid design provides a data-scarcity hedge that is most valuable when the spectral signal is ambiguous (4-band), while end-to-end architectures can match or exceed hybrid performance when spectral information is rich and sufficient training data are available [31].

5.3. SHAP Attribution and Partial Interpretability

SHAP attribution over the joint latent-physical feature vector provides partial but meaningful post-hoc transparency. The spectral-index subset consistently ranks among the highest-contributing features and carries direct biophysical meaning, while the CNN subset quantifies aggregate reliance on learned spatial context. This complementary structure directly interpretable spectral dimensions alongside opaque but spatially informative deep dimensions—validates the hybrid design philosophy and distinguishes it from end-to-end CNN pipelines where no individually interpretable features exist in the decision space [18,44].

The multi-temporal SHAP results additionally reveal that relative spectral change (index differences) is more informative than absolute reflectance for long temporal sequences, a finding consistent with known advantages of ratio-based features for normalization-invariant CD [29,35].

5.4. Limitations and Future Work

Several limitations warrant acknowledgment. First, the study area covers a single geographic corridor in central India; out-of-region generalization to markedly different terrain, climate, or sensor configurations remains an open empirical question. Second, the absolute-difference operation used to construct the hybrid feature vector discards directional change information, a structural limitation that may explain the modest advantage of the Siamese U-Net in 8-band main-training regimes. Third, CNN-derived channels in the hybrid vector are individually opaque; SHAP provides *partial*, not complete, transparency. Fourth, transformer-based baselines (BIT, ChangeFormer) were not evaluated in 8-band or multi-temporal configurations due to RGB-initialized encoder incompatibility and self-attention memory constraints. Finally, evaluation was restricted to the proprietary PlanetScope dataset rather than standard public benchmarks (e.g., LEVIR-CD, WHU-CD). This is because standard benchmarks are limited to RGB imagery and cannot support the multispectral (4-band and 8-band) spectral-index fusion that is central to the proposed methodology.

Future work will investigate: (i) signed change features to recover directional information discarded by the absolute-difference operation; (ii) systematic comparison across classical ML, hybrid deep+ML, and end-to-end deep architectures for different operational CD scenarios (urban expansion, disaster response); (iii) improved interpretability of CNN-derived channels via gradient-weighted class activation mapping and activation maximization; (iv) extensions to larger geographic extents and cross-sensor evaluation; and (v) quantitative analysis of training set size vs. model capacity interactions.

 doi:10.20944/preprints202605.1851.v1

6. Conclusions

This study presented MS-HySAN, a hybrid CNN–LightGBM framework for bi-temporal and multi-temporal change detection in high-resolution PlanetScope imagery, evaluated on a demanding infrastructure monitoring scenario in Indore, India. Three coordinated design decisions define the framework: a decoupled two-stage training protocol, a latent-physical hybrid fusion vector combining multi-scale CNN difference features with spectral-index differences, and a hierarchical Expert Model for rare-class recovery.

Across binary, multiclass, and multi-temporal scenarios, MS-HySAN consistently outperforms end-to-end deep learning architectures under conditions of severe class imbalance and limited labeled data. The most pronounced advantage occurs in binary tree-cut detection (+38 MCC points over deep baselines) and in expert-training regimes, where the hybrid model (F1 = 0.86) fully recovers from the catastrophic collapse (F1 = 0.20) of the Siamese U-Net baseline. In data-abundant main-training regimes, the hybrid remains competitive with Siamese U-Net across both spectral configurations.

SHAP attribution confirms that spectral-index differences and CNN features play complementary, physically consistent roles. This partial transparency-directly attributable spectral dimensions alongside aggregate spatial context-is a property absent in end-to-end deep pipelines and is essential for operational infrastructure monitoring and policy-relevant land-use analysis. The analysis further establishes that spectral richness (red-edge bands) and architectural robustness are partially substitutable: 8-band imagery largely resolves the ambiguity that causes deep-only baselines to collapse under data scarcity, while the hybrid design remains a robust choice when spectral information is limited.

These results establish a transparent, data-efficient framework applicable to operational monitoring tasks characterized by severe class imbalance, limited labeled budgets, and the need for audit-ready decision logic.

Funding: This research received no external funding.

Data Availability Statement: The PlanetScope imagery used in this study is commercially available from Planet Labs, Inc. Training labels generated for this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge Kaggle for providing access to the NVIDIA Tesla P100 GPU used in all computational experiments.

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

Abbreviations

The following abbreviations are used in this manuscript:

CD	Change Detection
CIRE	Chlorophyll Index Red Edge
CNN	Convolutional Neural Network
CVA	Change Vector Analysis
EVI	Enhanced Vegetation Index
GBDT	Gradient Boosted Decision Tree
LightGBM	Light Gradient Boosting Machine
MCC	Matthews Correlation Coefficient
MS-HySAN	Multi-Scale Hybrid Siamese Attention Network
NDRE	Normalized Difference Red-Edge Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near-Infrared
PAM	Pyramid Attention Module
SAVI	Soil-Adjusted Vegetation Index

 doi:10.20944/preprints202605.1851.v1

SHAP	SHapley Additive exPlanations
SOTA	State of the Art

References

1. Cheng, G.; Huang, Y.; Li, X.; Lyu, S.; Xu, Z.; Zhao, H.; Zhao, Q.; Xiang, S. Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing* **2024**, *16*, 2355. https://doi.org/https://www.mdpi.com/2072-4292/16/13/2355#.
2. Singh, A. Review article digital change detection techniques using remotely-sensed data. *International journal of remote sensing* **1989**, *10*, 989–1003. https://doi.org/https://doi.org/10.1080/01431168908903939.
3. Houborg, R.; McCabe, M.F. A Cubesat enabled Spatio-Temporal Enhancement Method (CESTEM) utilizing Planet, Landsat and MODIS data. *Remote Sensing of Environment* **2018**, *209*, 211–226. https://doi.org/10.1016/j.rse.2018.02.067.
4. Lu, D.; Mausel, P.; Brondizio, E.; Moran, E. Change detection techniques. *International journal of remote sensing* **2004**, *25*, 2365–2401. https://doi.org/https://doi.org/10.1080/0143116031000139863.
5. Bruzzone, L.; Prieto, D.F. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote sensing* **2002**, *38*, 1171–1182. https://doi.org/https://doi.org/10.1109/36.843009.
6. Hussain, M.; Chen, D.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing* **2013**, *80*, 91–106. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2013.03.006.
7. Nielsen, A.A.; Conradsen, K.; Simpson, J.J. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment* **1998**, *64*, 1–19. https://doi.org/https://doi.org/10.1016/S0034-4257(97)00162-4.
8. Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.
9. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
10. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794. https://doi.org/https://doi.org/10.1145/2939672.2939785.
11. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017**, *30*.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference. Springer, 2015, pp. 234–241.
13. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Proceedings of the International workshop on deep learning in medical image analysis. Springer, 2018, pp. 3–11. https://doi.org/https://doi.org/10.1007/978-3-030-00889-5_1.
14. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE international conference on image processing (ICIP). IEEE, 2018, pp. 4063–4067. https://doi.org/https://doi.org/10.1109/ICIP.2018.8451652.
15. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2022, pp. 207–210. https://doi.org/https://doi.org/10.1109/IGARSS46834.2022.9883686.
16. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–14. https://doi.org/https://doi.org10.1109/TGRS.2021.3095166.
17. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
18. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18. https://doi.org/https://doi.org/10.3390/e23010018.
19. Bria, A.; Marrocco, C.; Tortorella, F. Addressing class imbalance in deep learning for small lesion detection on medical images. *Computers in biology and medicine* **2020**, *120*, 103735. https://doi.org/https://doi.org/10.1016/j.compbiomed.2020.103735.
20. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *Journal of big data* **2019**, *6*, 27.

 doi:10.20944/preprints202605.1851.v1

21. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **2002**, *16*, 321–357. https://doi.org/https://doi.org/10.1613/jair.953.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*. https://doi.org/https://doi.org/10.1016/j.jag.2024.103767.
23. Matsoukas, C.; Haslum, J.F.; Söderberg, M.; Smith, K. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038* **2021**. https://doi.org/https://doi.org/10.48550/arXiv.2108.09038.
24. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing* **2019**, *152*, 166–177. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.04.015.
25. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment* **2020**, *241*, 111716. https://doi.org/https://doi.org/10.1016/j.rse.2020.111716.
26. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters* **2021**, *19*, 1–5. https://doi.org/https://doi.org/10.1109/LGRS.2021.3056416.
27. FILELLA, I.; PENUELAS, J. The red edge position and shape as indicators of plant chlorophyll content, biomass and hydric status. *International Journal of Remote Sensing* **1994**, *15*, 1459–1470. https://doi.org/10.1080/01431169408954177.
28. Gitelson, A.A.; Merzlyak, M.N. Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. *Journal of plant physiology* **1996**, *148*, 494–500. https://doi.org/https://doi.org/10.1016/S0176-1617(96)80284-7.
29. Mou, L.; Bruzzone, L.; Zhu, X.X. Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 924–935. https://doi.org/10.1109/TGRS.2018.2863224.
30. Lyu, H.; Lu, H.; Mou, L. Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection. *Remote Sensing* **2016**, *8*. https://doi.org/10.3390/rs8060506.
31. Hirschmugl, M.; Deutscher, J.; Sobe, C.; Bouvet, A.; Mermoz, S.; Schardt, M. Use of SAR and optical time series for tropical forest disturbance mapping. *Remote Sensing* **2020**, *12*, 727. https://doi.org/https://doi.org/10.3390/rs12040727.
32. Pohjankukka, J.; Pahikkala, T.; Nevalainen, P.; Heikkonen, J. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science* **2017**, *31*, 2001–2019. https://doi.org/https://doi.org/10.1080/13658816.2017.1346255.
33. Meyer, H.; Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* **2021**, *12*, 1620–1633. https://doi.org/https://doi.org/10.1111/2041-210X.13650.
34. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications* **2020**, *11*, 4540. https://doi.org/https://doi.org/10.1038/s41467-020-18321-y.
35. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* **1979**, *8*, 127–150. https://doi.org/https://doi.org/10.1016/0034-4257(79)90013-0.
36. Gao, B.C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment* **1996**, *58*, 257–266. https://doi.org/https://doi.org/10.1016/S0034-4257(96)00067-3.
37. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote sensing of environment* **2002**, *83*, 195–213. https://doi.org/https://doi.org/10.1016/S0034-4257(02)00096-2.
38. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote sensing of environment* **1988**, *25*, 295–309. https://doi.org/https://doi.org/10.1016/0034-4257(88)90106-X.
39. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988. https://doi.org/https://doi.org/10.48550/arXiv.1708.02002.

 doi:10.20944/preprints202605.1851.v1

40. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **2009**, *21*, 1263–1284. https://doi.org/https://doi.org/10.1109/TKDE.2008.239.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**. https://doi.org/https://doi.org/10.48550/arXiv.1412.6980.
42. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144. https://doi.org/https://doi.org/10.1145/2939672.2939778.
43. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, pp. 1189–1232.
44. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.