Assessment of Different Architectures based on 2D-UNet, 3D-UNet and UNet-ConvLSTM for Land Use Land Cover Classification Using Multi-modal and Multi-temporal Satellite Images

Fatemeh Saba[1,2], Pedro Achanccaray[2], Markus Gerke[2]

[1] Institute of Geography, University of Hamburg, Hamburg, Germany,
[2] Institute of Geodesy and Photogrammetry, Technische Universität Braunschweig, Braunschweig, Germany
Fatemeh.saba@uni-hamburg.de; (Fatemeh.saba, p.diaz, m.gerke)@tu-braunschweig.de

Keywords: Land cover land use classification, UNet segmentation, Remote sensing dataset

Abstract

Land Use/Land Cover (LULC) classification is essential for environmental monitoring and sustainable development. However, accurately classifying complex and heterogeneous land cover types remains challenging, particularly in capturing spatial–temporal dynamics. This study evaluates different UNet-based semantic segmentation models trained on multi-temporal and multi-modal (optical Sentinel-2 and radar Sentinel-1) images for LULC classification. We assessed the performance of two 2D-UNet approaches: processing each time step separately (2D-UNet) and stacking time steps as additional channels (2D-UNet-multitemporal). Also, 3D-UNet models are explored, including a standard version and a 3D-UNet with attention block (3D-UNet-attention). Finally, UNet variants with Convolutional Long Short-Term Memory (ConvLSTM) are evaluated as well: replacing all convolutional layers with ConvLSTM (UNet-ConvLSTM-all), adding ConvLSTM at the bottleneck (3D-UNet-ConvLSTM-bottleneck), and incorporating it at the end (3D-UNet-ConvLSTM-end). The results showed that 3D-UNet models outperformed 2D-UNets, with 4–5 % improvements in mean F1-score, particularly for vegetation-related classes. The 3D-UNet-ConvLSTM-bottleneck indicated a slight improvement over 3D-UNet-ConvLSTM-end and outperformed UNet-ConvLSTM-all, demonstrating the effectiveness of combining the 3D convolutions with ConvLSTM for spatiotemporal feature extraction. Notably, the 3D-UNet-attention model achieved the best results for LULC classification, particularly for detecting and monitoring dead tree areas, which is the most challenging class in our application. Overall, this study demonstrates the superior capability of 3D-UNet architectures in capturing temporal dynamics and enhancing classification accuracy for complex LULC tasks. Leveraging freely accessible satellite data, these findings provide valuable insights for forest managers, tree mortality assessment, and the identification of effective control measures.

1. Introduction

Land Use/Land Cover (LULC) classification is essential for providing information on the spatial distribution of the Earth's surface attributes. Recent extreme weather events in central Europe, such as droughts (2018-2022) (Knapp et al., 2024), and severe flooding (Mohr et al., 2023), underscore the critical relationship between LULC changes and disaster vulnerability (Naha et al., 2020). Therefore, the increasing frequency, intensity, and persistence of extreme hydrological events emphasize the need to monitor LULC changes, including the type and status of vegetation as these maps are crucial for hydrological and environmental models (Ghaseminik et al., 2021). Conventionally, field surveys face challenges in monitoring LULC changes due to their high cost, time consumption, labor intensity, and low sampling frequency. Remote sensing (RS) is a practical tool for providing consistent, cost-efficient, and time-efficient datasets (Lefulebe et al., 2022). The rapid development of RS and geospatial technologies has fostered research on monitoring and mapping LULC using publicly available satellite data, such as Landsat (Boonpook et al., 2023) and Sentinel-2 (Saba et al., 2024, Nguyen et al., 2020), as well as commercial sources like Planetscope (Acharki, 2022). Because of their free availability, most studies have applied Landsat-8, MODIS, and Sentinel-2 data to generate LULC maps, the latter is preferred due to its high spatial and temporal resolutions for mapping LULC (Andrade et al., 2021).Numerous studies have revealed that integrating optical Sentinel-2 with radar-based Sentinel-1 data can enhance the classification performance (Vizzari, 2022, Saba et al., 2022, Tavares et al., 2019, Kpienbaareh et al., 2021). Additionally, multi-temporal data enable the tracking of dynamic changes over time, which is critical for accurate LULC monitoring, and in conjunction with multi-sensor data (e.g., optical and radar) can improve LULC classification accuracy and reliability in terms of robustness against imbalanced datasets.Recent advances in supervised learning have increasingly relied on deep learning (DL) methods, particularly convolutional neural networks (CNNs) (Turkulainen et al., 2023, Minařík et al., 2021) and recurrent neural networks (RNNs) (Mou et al., 2019). UNet (Ronneberger et al., 2015) has become a widely used CNN-based architecture for semantic segmentation. In the context of spatial feature extraction, a typical 2D-UNet employs two-dimensional convolutions to capture spatial information. This approach can be extended to 3D convolutions, as in the case of 3D-UNet, which allows for the extraction of both spatial and temporal information, further enhancing its capability to handle multi-temporal RS data. On the other hand, RNN models such as Long Short-Term Memory (LSTM) networks effectively handle sequential dependencies (Zhou et al., 2019, You et al., 2020). To maximize their capabilities for spatial-temporal data, convolutional LSTM (ConvLSTM) networks have been developed by combining CNNs and RNNs to handle both spatial and temporal dependencies (Shi et al., 2015). These networks have been combined with other DL architectures, such as UNet, to simplify its complex structure (He and Luo, 2022).

Accurate LULC classification using multi-temporal and multi-sensor remote sensing data remains challenging, particularly for complex classes such as dead trees or dense forests or heterogeneous landscapes. While DL architectures like UNet, and

LSTM-based models have shown promise, their comparative performance in capturing spatial-temporal dynamics for LULC classification has not been systematically evaluated. Therefore, this study aims to fill this gap by assessing the effectiveness of different DL-based architectures, including 2D-UNet, 3D-UNet, and diverse integrations of UNet with ConvLSTM in enhancing LULC classification accuracy using freely available satellite data. The contributions of this study can be summarized as follows:

- Performance assessment of 2D-UNet for LULC classification by comparing two approaches: processing each time step separately and stacking time steps as additional channels.
- Evaluation of the impact of attention mechanisms in 3D-UNets for LULC classification: employing 3D-UNet with and without attention mechanisms to capture spatiotemporal features.
- Evaluation of ConvLSTM layers in a UNet scheme for LULC classification: replacing all convolutional layers with ConvLSTM layers, adding a ConvLSTM layer at the final stage of UNet, and incorporating ConvLSTM in the UNet's bottleneck.

2. Related works

In recent years, DL techniques have been widely applied across different RS applications, including land cover classification, demonstrating their superior performance over traditional machine learning (ML) approaches (Alem and Kumar, 2020, Trujillo-Jiménez et al., 2022, Vali et al., 2020). Their advantage lies in capturing spatial patterns across neighboring pixels, eliminating the need for manual layer-by-layer design and feature extraction, which is an aspect often overlooked by pixel-based ML approaches (Sefrin et al., 2021). CNNs have emerged as the leading approach for DL-based semantic segmentation, primarily due to their robust feature extraction capabilities. For example, Turkulainen *et al.* (Turkulainen et al., 2023) demonstrated that 2D and 3D CNN models achieved exceptional performance in detecting dead and healthy trees, with F1-scores of 0.880 and 0.928, respectively. Minařík *et al.* (Minařík et al., 2021) compared Random Forest (RF) and CNN for bark beetle symptom classification, and found that CNN achieved the highest accuracy, with F1-scores of 0.94 for infested spruces and 0.86 for healthy spruces. Ji *et al.* (Ji et al., 2018) applied a 3D-CNN for classification using Gaofen-2 multi-temporal images, showing the superior performance of 3D-CNN in comparison to the 2D-CNN and other conventional methods. Subsequently, many researchers employed UNet for semantic segmentation, an encoder-decoder architecture built on CNNs, which effectively captures semantic information across multiple scales. By utilizing feature maps at various levels, UNet mitigates issues such as resolution degradation and multiscale complexity. It is particularly suitable for semantic segmentation using RS data, as it performs better with a small sample sizes and unbalanced classes than other CNN models (Wittstruck et al., 2024). Several studies have utilized 2D-UNet for LULC classification, leveraging optical and/or SAR data. For example, Shreehari *et al.* (Shreehari et al., 2024) achieved an overall F1-score of 0.89 using 2D-UNet, while Sawant *et al.* (Sawant and Ghosh, 2024) reported the highest overall accuracy of 0.94 with a 2D-UNet-ResNet50 architecture. Similarly, Solórzano *et al.* (Solórzano et al., 2021) achieved an F1-score of 0.58 and an accuracy of 0.76 based on SAR and optical data. Moreover, 3D-UNet has been employed to capture spatiotemporal features in the same application. Wittstruck *et al.* (Wittstruck et al., 2024) compared 2D-UNet and 3D-UNet models, showing that 3D-UNet outperformed 2D-UNet when using optical and SAR data. Additionally, Xue *et al.* (Xue et al., 2023) applied 3D-UNet to multispectral UAV data, achieving an IoU of 89.9%, highlighting its effectiveness for handling complex spatiotemporal datasets. Some studies have integrated attention mechanisms to enhance the DL model performance. The attention mechanism dynamically assigns weights to the input data, emphasizing important features while reducing the focus on less significant parts. This adaptive approach improves deep learning models, improving performance across various computer vision tasks (Woo et al., 2018). For instance, Sun *et al.*(Sun et al., 2022) applied an attention mechanism in a 2D-UNet and improved 4% in accuracy compared to the standard 2D-UNet. Similarly, Zhang *et al.* (Zhang et al., 2023) reported an increase of 0.5% in accuracy compared to 2D-UNet when integrating a modified 2D-UNet with an attention mechanism based on Sentinel-1/2 data. In addition, LSTM and ConvLSTM models are specifically designed for processing sequential temporal data. In contrast to LSTM, ConvLSTM incorporates convolutional operations in place of fully connected operations in LSTMs. As a result, the data that flows through the ConvLSTM cells maintain the input dimension instead of being a 1D vector with features. Therefore, the model can simultaneously learn the spatial and temporal features throughout the entire network (see Sun *et al.* (Sun et al., 2022) for more details). ConvLSTM has been widely applied in RS tasks such as LULC classification and change detection, as it effectively utilizes temporal interdependencies in multi-temporal data, improves accuracy and reduces pre-processing efforts (Zhu et al., 2021). For instance, Zhu *et al.* (Zhu et al., 2021) integrated ConvLSTM layers into the decoder of a UNet-based model based on WorldView-2 and QuickBird images for multi-temporal LULC classification. Their approach resulted in a performance improvement of 9% for one dataset, while no improvement was observed for another dataset. Teimouri *et al.* (Teimouri et al., 2019) adopted ConvLSTM at the last stage of a fully convolutional network (FCN) for classifying different crop types from radar data and confirmed the model's performance, achieving an average accuracy of 0.86.

3. Methodology

The methodology followed in this study is based on semantic segmentation DL-models for LULC classification with a focus on dead tree areas. In the following subsections, details about the DL-models and their components are explained.

3.1 UNet

We used a variant of UNet as baseline, which consists of three main parts: encoder, bottleneck and decoder. In the encoder, the input image passes three convolutional blocks, each containing two convolutional layers, followed by batch normalization, a rectified linear unit (ReLU) activation function, and a max-pooling layer. This results in reduced spatial resolution and increased feature maps depth due to the down-sampling and filters. The encoder is linked to the decoder by the bottleneck, which is another convolutional block without down-sampling. The decoder processes the image through additional convolutional blocks (convolution + transpose convolutions), thereby reversing the effect of the encoder. Consequently, the feature maps decrease their depth and increase their spatial resolution to achieve the final LULC classification. Moreover, the encoder and decoder are connected by skip connections, which involve combining the encoder feature map with the decoder (see (Ronneberger et al., 2015) for more details).

3.2 Attention mechanism

The attention block in UNet enhances feature selection by ensuring that only the most relevant spatial information is passed from the encoder to the decoder. It achieves this by extracting Query (Q), Key (K), and Value (V) from different parts of the network. The Q is derived from the decoder's upsampled feature maps, while the K and V come from the encoder's feature maps before being sent through the skip connections. To determine the importance of different spatial regions in the encoder's feature maps, the attention block computes attention weights by comparing the Q with the K. A gating mechanism from the decoder further refines this selection, ensuring that only the most relevant features are emphasized. Once the attention weights are computed, they are applied to scale the V, suppressing less relevant regions while enhancing important ones. The refined features are then passed through the skip connection and merged with the decoder's upsampled feature maps, improving segmentation accuracy by providing more contextually significant information (see (Oktay et al., 2018) for more details).

3.3 ConvLSTM

The ConvLSTM network comprises four main components: the cell state C_t, which acts as a memory unit to retain and transmit information over time; the input gate i_t, which regulates the flow of new data into the cell state; the forget gate f_t, which determines which information should be discarded; and the output gate o_t, which processes the updated cell state and generates the final output. The equations for ConvLSTM are as follows (Shi et al., 2015):

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (2)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \quad (4)$$

$$H_t = o_t \circ tanh(C_t) \quad (5)$$

where "$*$" represents the convolution operation and "$\circ$" is the Hadamard product. Here, i, f, and o denote the input, forget, and output gates at each timestamp, respectively. H represents the hidden state, C the cell state, and X the input at each timestamp. The activation function is denoted by σ and W represents the weighted connections for each state.

3.4 Network architectures

2D-UNet : The input data is represented as a 3-dimensional tensor $[W \times H \times C]$, where W, H, and C represent the width, height, and the number of channels. In this configuration, each time step is processed separately by the model and 2D convolutions with a kernel size of $[3 \times 3]$ are applied.

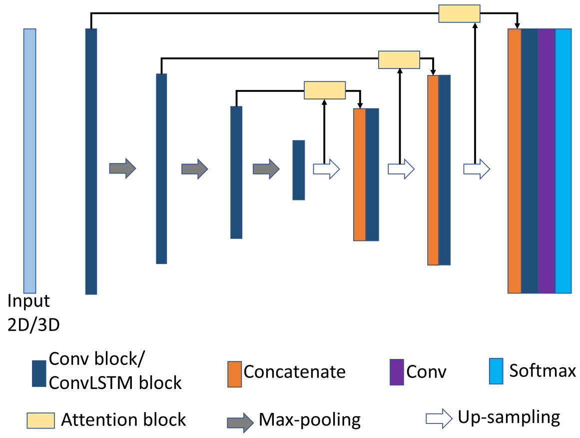

Figure 1. Representation of 2D-UNet, 2D-UNet-multitemporal, 3D-UNet, 3D-UNet-attention, and UNet-ConvLSTM-all.

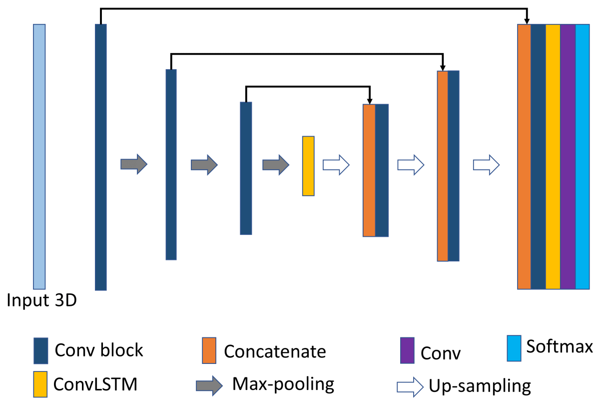

Figure 2. Representation of 3D-UNet-ConvLSTM-bottleneck and 3D-UNet-ConvLSTM-end.

2D-UNet-multitemporal : We extend 2D-UNet by stacking spectral bands from different time steps sequentially. This results in a 3-dimensional tensor of size $[W \times H \times C \cdot T]$, with a similar notation to 2D-UNet, where T represents the number of time steps. Here, we apply 2D convolutions with a kernel size of $[3 \times 3]$ to capture spatiotemporal features (Figure 1).

3D-UNet : A 3D convolution is extended by shifting the kernel in three dimensions and adding the time dimension. In 3D-UNet, each 2D convolutional layer from the baseline architecture is replaced with a 3D convolution layer (Figure 1). The input data is formatted as a 4-dimensional tensor $[W \times H \times T \times C]$, with a similar notation to the previous models. 3D convolutional kernels with a kernel size of $[3 \times 3 \times 3]$ are applied across the spatial and temporal dimensions.

3D-UNet-attention : This architecture is similar to 3D-UNet with the addition of attention blocks in the decoder. Specifically, it applies 3D convolutions, a ReLU activation, and a sigmoid gating mechanism to generate an attention map (Figure 1).

UNet-ConvLSTM-all : Here, all convolutional layers in the UNet are replaced with ConvLSTM layers (Figure 1). The input of a ConvLSTM is a set of images over time as a 4-dimensional tensor with shape $[T \times C \times H \times W]$ and kernel size $[3 \times 3]$.

3D-UNet-ConvLSTM-end : This configuration retains the conventional 3D-UNet architecture with a ConvLSTM layer with 32 kernels of size $[3 \times 3]$ added at the end. After the UNet processed the input to generate feature maps, it is passed through a ConvLSTM layer before producing the final output (Figure 2).

3D-UNet-ConvLSTM-bottleneck : In this configuration, ConvLSTM is applied at the UNet's bottleneck, where the spatial resolution is the lowest and the features are the most abstract. Specifically, after encoding the input into a low-dimensional form, the ConvLSTM layer processes this representation to capture the temporal dependencies before passing the information to the decoder (Figure 2).

4. Experiments

In the following subsections, details about the study area, labelled datasets, satellite imagery, and experimental setup are provided.

4.1 Study area

The study area is located in Central Germany in the Harz Mountains and its surroundings, a hotspot of current forest cover loss in Germany (Thonfeld et al., 2022)(see Figure 3). The Harz Mountains are characterized by an altitude ranging from approximately 200 to 640 m.a.s.l, humid conditions, a mean annual temperature of 8°C, and a mean annual precipitation sum of 760 mm (Putzenlechner et al., 2024). The entire study area covers an area of 196,886 ha, stretching from north to south for approximately 42 km and from west to east for approximately 46 km.

4.2 Satellite data

We utilized Sentinel-2 Level-2A products with 10 spectral bands and cloud cover lower than 5%. All bands were resampled to 10 m spatial resolution using nearest-neighbor resampling on the Google Earth Engine (GEE) platform. We also applied a cloud mask based on the “QA60” band's cloud and cirrus flags to improve Sentinel-2 data quality.
For radar data, we used Sentinel-1 images, a dual-polarization C-band Ground Range Detected data with a spatial resolution of 10 m. SAR images were collected in Interferometric Wide Swath mode with two polarization schemes: VV and VH. Several pre-processing steps, such as thermal noise removal, radiometric calibration, and terrain correction, have already been applied to each Sentinel-1 scene using GEE. In addition, a median filter with a window size of 3×3 pixels was applied to each SAR image to reduce the speckle effect.
The Sentinel-1 and Sentinel-2 datasets were concatenated and stacked to form a multi-source input. Both datasets were normalized to a range from 0 to 1 based on the minimum and maximum pixel values for each band. We focused on growing season images (June-August) for this study, as snow and frequent cloud cover in other seasons limit usable observations. Summer data also reduces the risk of overestimating canopy cover loss due to phenological changes in deciduous trees, which could otherwise affect the accuracy of the dead tree class (our main class of interest). Using summer observations, we ensure more accurate assessments and consider all canopy cover losses detected during this period.

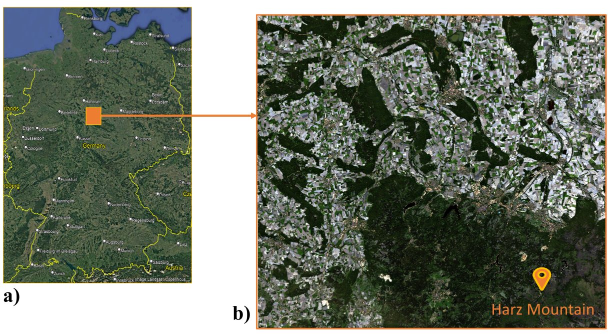

Figure 3. Location of the study area (red box) overlaid on a Google Earth image (a) and a Sentinel-2 satellite image (b).

4.3 Features

We utilized a total of 19 features for LULC classification, comprising 10 spectral bands and 7 vegetation indices (VIs) derived from optical Sentinel-2 imagery, as well as two polarizations (VV and VH) from Sentinel-1 data. The VIs used in this study are listed in Table 1.

4.4 Labels

The main LULC classes considered in our experiments were: cropland, grassland, built-up areas, and water bodies. Additionally, various tree species, including coniferous and deciduous trees, were considered, along with vegetation status, particularly dead trees. Since current LULC maps containing all the classes are not available, we integrated various data sources to create our reference labels.
In the context of LULC types, we used available LULC European Space Agency (ESA) maps with 10 m spatial resolution for 2020-2021 for this application (ESA, 2020, ESA, 2021). Regarding tree species, we utilized the dominant tree species map from 2018 generated by (Blickensdörfer et al., 2024). Then, we classified tree species into two main categories: coniferous and deciduous. Assuming that tree species remain constant over the years, we used this map as a reference for 2020 and 2021. For the dead tree samples, we visually assessed the Planetscope[1] data with a 3 m spatial resolution and manually delineated dead tree samples for the same years. In summary, we applied both ESA for land cover types, the tree species map for categorizing coniferous and deciduous trees, as well as manually collected samples of dead trees. All these maps were merged and used as references to train the DL methods in a supervised manner.

[1] https://www.planet.com/explorer

Vegetation Index	Equation
NDVI (Rouse et al., 1974)	$(NIR - Red)/(NIR + Red)$
GNDVI (Gitelson et al., 1996)	$(NIR - Green)/(NIR + Green)$
NDVI-RedEdge (Fernández-Manso et al., 2016)	$(NIR - RedEdge)/(NIR + RedEdge)$
SAVI (Huete, 1988) ($L = 0.428$)	$((NIR - Red)/(NIR + Red + L)) \cdot (1 + L)$
MNDWI (Xu, 2006)	$(Green - SWIR)/(Green + SWIR)$
NDBI (Y. Zha and Ni, 2003)	$(SWIR - NIR)/(SWIR + NIR)$
IBI (Xu, 2008)	$(NDBI - (SAVI + MNDWI)/2)/$ $(NDBI + (SAVI + MNDWI)/2)$

Table 1. Vegetation indices used in this study, where NDVI: Normalized Difference Vegetation Index, GNDVI: Green Normalized Difference Index, NDVI-RedEdge: Red Edge Normalized Difference Vegetation Index, SAVI: Soil-Adjusted Vegetation Index, MNDWI: Modified Normalized Difference Water Index, NDBI: Normalized Difference Built-up Index, IBI: Index-based Built-up Index, NIR: near-infrared, SWIR: short wave infrared,

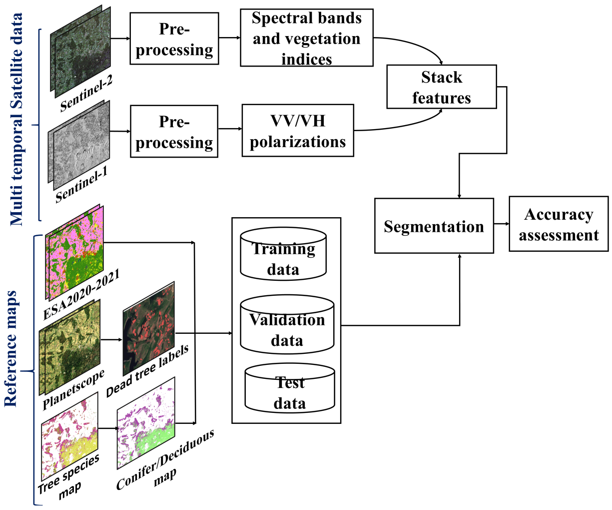

Figure 4. Workflow followed in our experiments.

4.5 Experimental setup

The created dataset was randomly split into training, validation, and test sets at a ratio of 70:10:20, respectively. During the training process, data augmentation techniques such as rotation and flip operations were randomly applied to enhance the model's ability to generalize and maintain robustness with unseen data. We also used multi-class cross-entropy as the loss function and Adam as the optimizer. It is important to note that the number of filters for all 2D-UNets and 3D-UNets is set to 64, 128, 256, and 512. However, using the same number of filters for the UNet-ConvLSTM-based architectures significantly increases the number of trainable parameters (e.g., 60M for UNet-ConvLSTM-all). To mitigate this, we reduce the number of filters by half (32, 64, 128, and 256) for all the UNet-ConvLSTM configurations. In our experiments, the DL-models were trained with the following inputs: $W = 256$, $H = 256$, $C = 19$, and $T = 3$. The training process is limited to 100 epochs; however, the process is stopped if there is no improvement in the loss function value after 10 epochs. All models were implemented using Keras/Tensorflow in Python, and the experimental environment consists of a 12th Gen Intel (R) Core (TM) i7-12700 with 32 GB RAM and a RTX 3060 12 GB GPU.
In our framework, we train different DL architectures using multi-temporal Sentinel-1/2 images acquired during 2020-2021 as the training dataset, with ESA 2020/2021 data, tree species, and dead tree labels as reference labels for LULC classification, as shown in Figure 4.
To evaluate each model's outcomes, different metrics derived from the confusion matrix, including F1-score, recall and precision, were calculated as follows:

$$\text{Precision} = \frac{X_{ij}}{X_j} \tag{6}$$

$$\text{Recall} = \frac{X_{ij}}{X_i} \tag{7}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where X_{ij} is the observation in row i and column j in the confusion matrix. X_i is the marginal total of row i and X_j is the marginal total of column j in the confusion matrix.

Architecture	# Filters	# Par.
2D-UNet	64,128, 256,512	7.71 M
2D-UNet-multitemporal		7.73 M
3D-UNet		21.74 M
3D-UNet-attention		21.92 M
UNet-ConvLSTM-all	32,64, 128,256	15.02 M
3D-UNet-ConvLSTM-end		5.52 M
3D-UNet-ConvLSTM-bottleneck		11.05 M

Table 2. Architectures employed in this study with their corresponding number of trainable parameters (# Par.) and number of filter per convolutional layer (# Filters)

5. Results

All architectures with their corresponding number of trainable parameters and number of filters per convolutional layer are listed in Table 2. Both 2D-UNet and 2D-UNet-multitemporal architectures were relatively lightweight, requiring fewer trainable parameters and consequently less training time, whereas the 3D-UNet architectures required more trainable parameters and longer training times. Among the UNet-ConvLSTM architectures, UNet-ConvLSTM-all demonstrated a higher computational cost, followed by 3D-UNet-ConvLSTM-bottleneck and 3D-UNet-ConvLSTM-end. It is important to note that the UNet-ConvLSTM-all model, with the same number of filters per convolutional layer as 2D and 3D-UNet models, has approximately 60 M trainable parameters. To simplify the model and reduce computational complexity, we halved the number of filters when using ConvLSTM.
Table 3 presents the results obtained on the test set by all selected architectures in terms of F1-score per class and mean F1-score (m-F1). In general terms, m-F1 ranges from 0.82 to 0.86 for LULC classification, with slight differences between models. However, taking a closer look at F1-scores per class, we can see that cropland consistently achieved the highest values, with F1-scores ranging between 0.93 and 0.95, followed by deciduous trees (0.89–0.92) and water bodies (0.79–0.82), being the less difficult classes to be distinguished by all architectures. Moreover, built-up areas and coniferous trees were classified with F1-scores exceeding 0.75 and 0.70, respectively. In contrast, dead trees and grasslands exhibited the lowest values, with F1-scores ranging from 0.47 to 0.57 and 0.66 to 0.72, respectively. Additionally, cropland, water bodies, and deciduous trees showed similar accuracies across all architectures, with minimal variation in their performance compared to other classes.

5.1 Evaluation of 2D-UNet and 3D-UNet architectures

We conducted experiments to evaluate the impact of multi-temporal input data on the classification performance based on 2D-UNet and 3D-UNet, as described in Section 3.4.
As seen from Table 3, for the 2D-UNet models, 2D-UNet-multitemporal showed a slight improvement of 1% in m-F1 compared to 2D-UNet. At the class level, built-up and coniferous trees exhibited the most significant increase (4%), followed by water (3%), cropland (2%), and deciduous trees and grassland (1% each). No change was observed for dead trees.
For the 3D-UNet models, the 3D-UNet-attention achieved a similar m-F1 compared to the 3D-UNet. However, class-wise improvements were observed for dead trees and built-up areas (4% each) as well as for water and deciduous trees (1% each). Conversely, grassland and coniferous trees experienced a slight decrease in accuracy.
When comparing the 2D and 3D-UNet models, both 3D-UNet configurations outperformed the 2D models as expected. Spe-

Architectures	F1-score per class							m-F1
	Grassland	Cropland	Built-up	Water	Conifer tree	Deciduous tree	Dead tree	
2D-UNet	0.66	0.93	0.78	0.79	0.70	0.89	0.51	0.82
2D-UNet-multitemporal	0.67	0.95	0.81	0.81	0.73	0.90	0.51	0.83
3D-UNet	0.72	0.95	0.79	0.81	0.79	0.91	0.55	0.86
3D-UNet-attention	0.69	0.95	0.82	0.82	0.77	0.92	0.57	0.86
UNet-ConvLSTM-all	0.67	0.93	0.75	0.80	0.76	0.90	0.47	0.84
3D-UNet-ConvLSTM-end	0.70	0.95	0.79	0.79	0.77	0.91	0.56	0.85
3D-UNet-ConvLSTM-bottleneck	0.72	0.95	0.80	0.81	0.79	0.91	0.56	0.86

Table 3. Results obtained on the test set by each architecture in terms of F1-score per class and mean F1-score (m-F1)

cifically, both 3D-UNet and 3D-UNet-attention achieved improvements of 4% and 5% in m-F1 score compared to 2D-UNet-multitemporal and 2D-UNet, respectively.

In a detailed class-wise comparison between 3D-UNet and 2D-UNet-multitemporal, the largest improvements were observed for grassland, coniferous trees, and dead trees, with increases of 7%, 8%, and 8%, respectively. Deciduous trees showed a minor improvement of 1%, whereas built-up areas experienced a slight decline. Water and cropland classes remained unchanged. Similarly, when comparing 3D-UNet-attention to 2D-UNet-multitemporal, notable improvements were observed for grassland (3%), coniferous trees (5%), and dead trees (12%).

Finally, when comparing 3D-UNet to 2D-UNet, improvements were observed across all classes: grassland (9%), cropland (2%), built-up areas (1%), water (2%), coniferous trees (12%), deciduous trees (2%), and dead trees (8%). Similarly, 3D-UNet-attention demonstrated significant gains over 2D-UNet, particularly for dead trees (12%) and coniferous trees (10%), while other classes showed moderate improvements ranging from 2% to 5%.

5.2 Evaluation of UNet-ConvLSTM architectures

Experiments based on UNet-ConvLSTM are also conducted to explore how the integration of ConvLSTM with the UNet architecture, whether at the end, as a replacement for all layers, or within the bottleneck, affects the results, as detailed in Section 3.4.

As seen from Table 3, the results showed that the 3D-UNet-ConvLSTM-bottleneck performed similarly to the 3D-UNet-ConvLSTM-end, with only a 1% increase in the m-F1 score, while outperforming the UNet-ConvLSTM-all by 2% in m-F1. The classes in the UNet-ConvLSTM-bottleneck either improved slightly (1-2% in the F1-score for grassland, built-up, coniferous tree, and water) or remained stable (cropland, deciduous and dead tree) relative to the UNet-ConvLSTM-end. Specifically, the 3D-UNet-ConvLSTM-bottleneck showed improvements in grassland (0.72 vs. 0.70 and 0.67), built-up areas (0.80 vs. 0.79 and 0.75), and coniferous trees (0.79 vs. 0.77 and 0.76), while maintaining similar performances for cropland (0.95 vs. 0.95 and 0.93), water (0.81 vs. 0.79 and 0.80), and deciduous trees (0.91 vs. 0.91 and 0.90). Dead trees demonstrated a significant improvement, with an accuracy of 0.56 compared to 0.47.

6. Discussion

In this study, we evaluated different architectures for LULC classification based on UNet, which exploited spatial and/or temporal features. The results demonstrated that all models effectively classified LULC classes, achieving m-F1 ranging from 0.82 to 0.86. Croplands achieved the highest F1-scores (0.93–0.95), followed by deciduous trees (0.89–0.92), which can be attributed to the abundance of training data for these classes. Water bodies were also well classified, with F1-scores of 0.79–0.82, likely due to their spatial homogeneity and distinct spectral signatures in the near-infrared spectrum and radar polarizations. In contrast, dead trees (0.47–0.56) and grasslands (0.66–0.72) exhibited the lowest accuracy across all the architectures. The poor performance of dead trees may stem from limited training samples (1% of the total training samples) and misclassification with coniferous trees. It is important to note that the transition from a healthy to dead state occurs gradually, meaning that areas in intermediate stages of defoliation were not sufficiently represented in the training data. This contributed to the misclassification and reduced the accuracy of the dead tree class. Additionally, we used Sentinel-2 data with a 10-meter resolution for training, whereas dead tree annotations were derived from higher-resolution data. The coarser resolution of Sentinel-2 may result in mixed spectral signals from multiple land cover types within a single pixel rather than exclusively capturing dead trees. This issue is particularly pronounced in areas with standing defoliated trees, further decreasing the accuracy of this class.

For grasslands, the lower accuracy is likely due to confusion with cropland and coniferous trees, potentially caused by similarities in spectral characteristics, such as unused farmlands resembling grasslands, and deforested or recovering areas being misclassified as coniferous trees.

The results revealed that 2D-UNet achieved the lowest performance among all architectures, as it cannot capture temporal relationships between multi-temporal images. The 2D-UNet-multitemporal model showed only a 1% improvement in m-F1 over 2D-UNet, indicating that using temporal information—whether as separate inputs (2D-UNet) or as additional channels (2D-UNet-multitemporal)—has minimal impact on performance in our application. This suggests that stacking images is not an effective approach for exploiting and extracting temporal features, as it treats all channels as a static set of features without accounting for temporal dependencies.

A significant performance improvement was observed when transitioning from the 2D to 3D-UNet models. Both 3D-UNet and 3D-UNet-attention achieved notable improvement in m-F1, outperforming 2D-UNet by 5% and 2D-UNet-multitemporal by 4%. As expected, the use of multi-temporal data in the 3D-UNet models led to substantial improvements in the vegetation-related classes, with increases of 3–9% for grassland, 5–12% for coniferous trees, and 8-12% for dead trees. In contrast, the accuracy of the other classes, such as water, remained relatively stable or showed only minor changes. These findings underscore the superiority of 3D-UNet models in explicitly modeling temporal dynamics, particularly for vegetation classes, which exhibit strong temporal dependencies and evolving patterns over time.

When comparing the UNet-ConvLSTM models, the 3D-UNet-ConvLSTM-bottleneck performed slightly better than the

3D-UNet-ConvLSTM-end, and both outperformed the UNet-ConvLSTM-all in terms of accuracy, while being less computationally expensive. The 3D-UNet-ConvLSTM-end and 3D-UNet-ConvLSTM-bottleneck leveraged the efficiency of the standard 3D-UNet for spatial-temporal feature extraction while enhancing temporal modeling by using ConvLSTM at the final stage and bottleneck, respectively.
In addition, the superior performance of 3D-UNet-ConvLSTM at the bottleneck or end suggests that extracting temporal information using 3D convolutions is more effective than relying solely on ConvLSTM in our application. This hybrid approach balances spatial and temporal feature extraction, leading to better generalization and accuracy, particularly for complex LULC classes.
This study is subject to several limitations. Certain complex land cover classes remain challenging to classify accurately, particularly in heterogeneous landscapes. Although high spatial resolution imagery supports detailed mapping, it can also increase intra-class variability, thereby complicating classification. Additionally, the dependence on ESA-derived reference data introduces potential uncertainty due to possible inaccuracies in the ground truth labels. Furthermore, the relatively small number of dead tree samples limits the model's ability to robustly learn and generalize this class.

7. Conclusions

This study evaluated different UNet architectures for LULC classification based on a combination of multi-temporal optical Sentinel-2 and radar Sentinel-1 data. We compared 2D-UNet, 2D-Net-multitemporal (as additional channels), 3D-UNet, and 3D-UNet-attention (adding an attention block to 3D-UNet) for LULC classification. It also explored three UNet-ConvLSTM configurations (replacing all layers, adding at the bottleneck, or incorporating at the end). The results indicated that 2D-UNet-multitemporal showed a slight improvement compared to 2D-UNet. In contrast, 3D-UNet models significantly outperformed 2D-UNet models, with improvements of 4-5% in m-F1, highlighting the superiority of 3D-UNet in capturing spatial and temporal features. When comparing the UNet-ConvLSTM models, 3D-UNet-ConvLSTM-bottleneck, where spatial and temporal feature extraction were handled using 3D-UNet and temporal dependencies were enhanced using ConvLSTM at the bottleneck, outperformed the other architectures. Given that the dead tree class is the most critical in this application, 3D-UNet-attention achieved the best results. This architecture leveraged multi-spatial and temporal features through 3D-UNet, combined with an attention mechanism, to effectively capture the complex dynamics of dead tree detection.
This study used satellite images from the growing season, making the models specifically tailored to those conditions. To enhance their robustness, the models will be further evaluated using satellite data from other sensors or seasons, accounting for varying conditions and differences in image availability. This can improve generalization while minimizing potential confusion between dead trees and deciduous trees.

References

Acharki, S., 2022. PlanetScope contributions compared to Sentinel-2, and Landsat-8 for LULC mapping. *Remote Sensing Applications: Society and Environment*, 27, 100774.

Alem, A., Kumar, S., 2020. Deep learning methods for land cover and land use classification in remote sensing: A review. *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 903–908.

Andrade, J., Cunha, J., Silva, J., Rufino, I., Galvão, C., 2021. Evaluating single and multi-date Landsat classifications of land-cover in a seasonally dry tropical forest. *Remote Sensing Applications: Society and Environment*, 22, 100515. https://www.sciencedirect.com/science/article/pii/S2352938521000513.

Blickensdörfer, L., Oehmichen, K., Pflugmacher, D., Kleinschmit, B., Hostert, P., 2024. National tree species mapping using Sentinel-1/2 time series and German National Forest Inventory data. *Remote Sensing of Environment*, 304, 114069. https://www.sciencedirect.com/science/article/pii/S0034425724000804.

Boonpook, W., Tan, Y., Nardkulpat, A., Torsri, K., Torteeka, P., Kamsing, P., Sawangwit, U., Pena, J., Jainaen, M., 2023. Deep Learning Semantic Segmentation for Land Use and Land Cover Types Using Landsat 8 Imagery. *ISPRS International Journal of Geo-Information*, 12(1). https://www.mdpi.com/2220-9964/12/1/14.

ESA, 2020. ESA WorldCover 2020. `https://worldcover2020.esa.int/`.

ESA, 2021. ESA WorldCover 2021. `https://worldcover2021.esa.int/`.

Fernández-Manso, A., Fernández-Manso, O., Quintano, C., 2016. SENTINEL-2A red-edge spectral indices suitability for discriminating burn severity. *International Journal of Applied Earth Observation and Geoinformation*, 50, 170-175. https://www.sciencedirect.com/science/article/pii/S0303243416300368.

Ghaseminik, F., Aghamohammadi, H., Azadbakht, M., 2021. Land cover mapping of urban environments using multispectral LiDAR data under data imbalance. *Remote Sensing Applications: Society and Environment*, 21, 100449. https://www.sciencedirect.com/science/article/pii/S2352938520306248.

Gitelson, A. A., Kaufman, Y. J., Merzlyak, M. N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58(3), 289–298. https://www.sciencedirect.com/science/article/pii/S0034425796000727.

He, L., Luo, W., 2022. 3D-ConvLSTMNet: A deep spatio-temporal model for traffic flow prediction. *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, 147–152.

Huete, A., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25(3), 295-309. https://www.sciencedirect.com/science/article/pii/003442578890106X.

Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y., 2018. 3D Convolutional Neural Networks for Crop Classification with Multi-Temporal Remote Sensing Images. *Remote Sensing*, 10(1). https://www.mdpi.com/2072-4292/10/1/75.

Knapp, N., Wellbrock, N., Bielefeldt, J., Dühnelt, P., Hentschel, R., Bolte, A., 2024. From single trees to country-wide maps: Modeling mortality rates in Germany based on the Crown Condition Survey. *Forest Ecology and Management*, 568, 122081. https://www.sciencedirect.com/science/article/pii/S0378112724003931.

Kpienbaareh, D., Sun, X., Wang, J., Luginaah, I., Bezner Kerr, R., Lupafya, E., Dakishoni, L., 2021. Crop Type and Land Cover Mapping in Northern Malawi Using the Integration of Sentinel-1, Sentinel-2, and PlanetScope Satellite Data. *Remote Sensing*, 13(4). https://www.mdpi.com/2072-4292/13/4/700.

Lefulebe, B. E., Van der Walt, A., Xulu, S., 2022. Fine-Scale Classification of Urban Land Use and Land Cover with PlanetScope Imagery and Machine Learning Strategies in the City of Cape Town, South Africa. *Sustainability*, 14(15). https://www.mdpi.com/2071-1050/14/15/9139.

Minařík, R., Langhammer, J., Lendzioch, T., 2021. Detection of Bark Beetle Disturbance at Tree Level Using UAS Multispectral Imagery and Deep Learning. *Remote Sensing*, 13(23). https://www.mdpi.com/2072-4292/13/23/4768.

Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J. E., Ehmele, F., Feldmann, H., Franca, M. J., Gattke, C., Hundhausen, M., Knippertz, P., Küpfer, K., Mühr, B., Pinto, J. G., Quinting, J., Schäfer, A. M., Scheibel, M., Seidel, F., Wisotzky, C., 2023. A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe – Part 1: Event description and analysis. *Natural Hazards and Earth System Sciences*, 23(2), 525–551. https://nhess.copernicus.org/articles/23/525/2023/.

Mou, L., Bruzzone, L., Zhu, X. X., 2019. Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2), 924-935.

Naha, S., Rico-Ramirez, M. A., Rosolem, R., 2020. Quantifying the impact of land cover changes on hydrological extremes in India. *Hydrology and Earth System Sciences Discussions*, 2020, 1–38. https://hess.copernicus.org/preprints/hess-2020-220/.

Nguyen, H. T. T., Doan, T. M., Tomppo, E., McRoberts, R. E., 2020. Land Use/Land Cover Mapping Using Multitemporal Sentinel-2 Imagery and Four Classification Methods—A Case Study from Dak Nong, Vietnam. *Remote Sensing*, 12(9). https://www.mdpi.com/2072-4292/12/9/1367.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning where to look for the pancreas.

Putzenlechner, B., Bevern, F., Koal, P., Grieger, S., Kappas, M., Koukal, T., Löw, M., Filipponi, F., 2024. Accuracy assessment of LAI, PAI and FCOVER from Sentinel-2 and GEDI for monitoring forests and their disturbance in Central Germany. *European Journal of Remote Sensing*, 57(1), 2422323. https://doi.org/10.1080/22797254.2024.2422323.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, Cham, 234–241.

Rouse, J., Haas, R., Schell, J., Deering, D., Harlan, J., 1974. Monitoring the Vernal Advancement of Retrogradation of Natural Vegetation: Final Report. *National Aeronautics & Space Administration/Goddard Space Flight Centre, Greenbelt, MD, USA*.

Saba, F., Latifi, H., Valadan Zoej, M. J., Esmaili, R., 2022. Mapping dead understorey Buxus hyrcana Pojark using Sentinel-2 and Sentinel-1 data. *Forestry: An International Journal of Forest Research*, 96(2), 228-248. https://doi.org/10.1093/forestry/cpac049.

Saba, F., Latifi, H., Zoej, M. J. V., Heipke, C., 2024. Analysis of the spatio-temporal dynamics of Buxus hyrcana Pojark defoliation using spaceborne satellite data. *Forestry: An International Journal of Forest Research*, 97(5), 713-727. https://doi.org/10.1093/forestry/cpae005.

Sawant, S., Ghosh, J. K., 2024. Land use land cover classification using Sentinel imagery based on deep learning models. *Journal of Earth System Science*, 133(2), 101. https://doi.org/10.1007/s12040-024-02298-z.

Sefrin, O., Riese, F. M., Keller, S., 2021. Deep Learning for Land Cover Change Detection. *Remote Sensing*, 13(1). https://www.mdpi.com/2072-4292/13/1/78.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., chun Woo, W., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting.

Shreehari, A., Chandan, P. R., Vinod, P. V., Hebbar, R., 2024. Land use land cover classification using deep learning techniques: A comparative study. *2024 11th International Conference on Signal Processing and Integrated Networks (SPIN)*, 7–11.

Solórzano, J. V., Mas, J. F., Gao, Y., Gallardo-Cruz, J. A., 2021. Land Use Land Cover Classification with U-Net: Advantages of Combining Sentinel-1 and Sentinel-2 Imagery. *Remote Sensing*, 13(18). https://www.mdpi.com/2072-4292/13/18/3600.

Sun, Y., Bi, F., Gao, Y., Chen, L., Feng, S., 2022. A Multi-Attention UNet for Semantic Segmentation in Remote Sensing Images. *Symmetry*, 14(5). https://www.mdpi.com/2073-8994/14/5/906.

Tavares, P. A., Beltrão, N. E. S., Guimarães, U. S., Teodoro, A. C., 2019. Integration of Sentinel-1 and Sentinel-2 for Classification and LULC Mapping in the Urban Area of Belém, Eastern Brazilian Amazon. *Sensors*, 19(5). https://www.mdpi.com/1424-8220/19/5/1140.

Teimouri, N., Dyrmann, M., Jørgensen, R. N., 2019. A Novel Spatio-Temporal FCN-LSTM Network for Recognizing Various Crop Types Using Multi-Temporal Radar Images. *Remote Sensing*, 11(8). https://www.mdpi.com/2072-4292/11/8/990.

Thonfeld, F., Gessner, U., Holzwarth, S., Kriese, J., da Ponte, E., Huth, J., Kuenzer, C., 2022. A First Assessment of Canopy Cover Loss in Germany’s Forests after the 2018–2020 Drought Years. *Remote Sensing*, 14(3). https://www.mdpi.com/2072-4292/14/3/562.

Trujillo-Jiménez, M. A., Liberoff, A. L., Pessacg, N., Pacheco, C., Díaz, L., Flaherty, S., 2022. SatRed: New classification land use/land cover model based on multi-spectral satellite images and neural networks applied to a semiarid valley of Patagonia. *Remote Sensing Applications: Society and Environment*, 26, 100703. https://www.sciencedirect.com/science/article/pii/S2352938522000118.

Turkulainen, E., Honkavaara, E., Näsi, R., Oliveira, R. A., Hakala, T., Junttila, S., Karila, K., Koivumäki, N., Pelto-Arvo, M., Tuviala, J., Östersund, M., Pölönen, I., Lyytikäinen-Saarenmaa, P., 2023. Comparison of Deep Neural Networks in the Classification of Bark Beetle-Induced Spruce Damage Using UAS Images. *Remote Sensing*, 15(20). https://www.mdpi.com/2072-4292/15/20/4928.

Vali, A., Comai, S., Matteucci, M., 2020. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sensing*, 12(15). https://www.mdpi.com/2072-4292/12/15/2495.

Vizzari, M., 2022. PlanetScope, Sentinel-2, and Sentinel-1 Data Integration for Object-Based Land Cover Classification in Google Earth Engine. *Remote Sensing*, 14(11). https://www.mdpi.com/2072-4292/14/11/2628.

Wittstruck, L., Jarmer, T., Waske, B., 2024. Multi-Stage Feature Fusion of Multispectral and SAR Satellite Images for Seasonal Crop-Type Mapping at Regional Scale Using an Adapted 3D U-Net Model. *Remote Sensing*, 16(17). https://www.mdpi.com/2072-4292/16/17/3115.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. CBAM: Convolutional Block Attention Module.

Xu, H., 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. https://doi.org/10.1080/01431160600589179.

Xu, H., 2008. A new index for delineating built-up land features in satellite imagery. *International Journal of Remote Sensing*, 29(14), 4269–4276. https://doi.org/10.1080/01431160802039957.

Xue, J., Sun, S., Zhao, H., Chen, W., 2023. Vegetation classification of UAV multispectral remote sensing images based on deep learning. *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 768–771.

Y. Zha, J. G., Ni, S., 2003. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583–594. https://doi.org/10.1080/01431160304987.

You, H., Tian, S., Yu, L., Lv, Y., 2020. Pixel-Level Remote Sensing Image Recognition Based on Bidirectional Word Vectors. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2), 1281-1293.

Zhang, G., Roslan, S. N. A. b., Wang, C., Quan, L., 2023. Research on land cover classification of multi-source remote sensing data based on improved U-Net network. *Scientific Reports*, 13(1), 16275. https://doi.org/10.1038/s41598-023-43317-1.

Zhou, Y., Luo, J., Feng, L., Yang, Y., Chen, Y., Wu, W., 2019. Long-short-term-memory-based crop classification using high-resolution optical images and multi-temporal SAR data. *GIScience & Remote Sensing*, 56(8), 1170–1191. https://doi.org/10.1080/15481603.2019.1628412.

Zhu, Y., Geiß, C., So, E., Jin, Y., 2021. Multitemporal Relearning With Convolutional LSTM Models for Land Use Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 3251-3265.