Mathematical and Computational Applications

MDPI

Article

Hybrid ConvLSTM U-Net Deep Neural Network for Land Use and Land Cover Classification from Multi-Temporal Sentinel-2 Images: Application to Yaoundé, Cameroon

Ange Gabriel Belinga [1,*], Stéphane Cédric Tékouabou Koumetio [2,3,*] and Mohammed El Haziti [1,4]

[1] Laboratory of Research in Computer Science and Telecommunications (LRIT), Faculty of Sciences in Rabat, Mohammed V University in Rabat, Rabat 10000, Morocco; elhazitim@gmail.com
[2] Research Laboratory in Computer Science and Educational Technologies (LITE), University of Yaoundé I, Yaoundé P.O. Box 812, Cameroon
[3] Department of Computer Science and Educational Technologies (DITE), University of Yaoundé I, Yaoundé P.O. Box 812, Cameroon
[4] High School of Technology, Mohammed V University in Rabat, Sale 11000, Morocco
* Correspondence: agbelinga@gmail.com (A.G.B.); ctekouaboukoumetio@gmail.com or cedric.tekouabou@univ-yaounde1.com (S.C.T.K.)

Abstract

Accurate mapping of land use and land cover (LULC) is crucial for various applications such as urban planning, environmental management, and sustainable development, particularly in rapidly growing urban areas. African cities such as Yaoundé, Cameroon, are particularly affected by this rapid and often uncontrolled urban growth with complex spatio-temporal dynamics. Effective modeling of LULC indicators in such areas requires robust algorithms for high-resolution images segmentation and classification, as well as reliable data with great spatio-temporal distributions. Among the most suitable data sources for these types of studies, Sentinel-2 image time series, thanks to their high spatial (10 m) and temporal (5 days) resolution, are a valuable source of data for this task. However, for an effective LULC modeling purpose in such dynamic areas, many challenges remain, including spectral confusion between certain classes, seasonal variability, and spatial heterogeneity. This study proposes a hybrid deep learning architecture combining U-Net and Convolutional Long Short-Term Memory (ConvLSTM) layers, allowing the spatial structures and temporal dynamics of the Sentinel-2 series to be exploited jointly. Applied to the Yaoundé region (Cameroon) over the period 2018–2025, the hybrid model significantly outperforms the U-Net and ConvLSTM models alone. It achieves a macro-average F1 score of 0.893, an accuracy of 0.912, and an average IoU of 0.811 on the test set. These segmentation performances reached up to 0.948, 0.953, and 0.910 for precision, F1-score, and IoU, respectively, on the built-up areas class. Moreover, despite its better performance, in terms of complexity, the figures confirm that the hybrid does not significantly penalize evaluation speed. These results demonstrate the relevance of jointly integrating space and time for robust LULC classification from multi-temporal satellite images.

Keywords: land use; land cover; U-Net; ConvLSTM; multispectral classification; image segmentation; Yaoundé; Cameroon

Check for updates

Academic Editor: Efrén Mezura-Montes

Received: 22 December 2025
Revised: 20 January 2026
Accepted: 21 January 2026
Published: 26 January 2026

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

1. Introduction

Accurate monitoring of land use and land cover (LULC) is essential for several applications, such as urban planning, environmental management, and agriculture [1–3]. For more

effective modelling of the LULC indexes, robust models for high-resolution images segmentation and classification are needed, as well as reliable data with better spatio-temporal distributions [4]. Among the various satellites providing this type of data, the Sentinel-2 satellites, part of the European Copernicus program, give free high-resolution (10 m) and multispectral images, which are ideal for mapping land cover [5]. Each Sentinel-2 image has 13 spectral bands covering the visible, near-infrared (NIR), and short-wave infrared (SWIR) ranges, providing a wealth of information for differentiating land cover classes. However, several technical challenges remain in LULC classification: spectral similarity between certain classes (e.g., bare soil vs. urban areas), spatial heterogeneity of landscapes, and seasonal or inter-annual temporal variations, as in our area of interest, the city of Yaoundé in Cameroon [6]. In light of these technical challenges, traditional methods (classical remote sensing or standard machine learning) struggle to effectively manage these complexities. Hence, hybrid deep learning techniques have been employed, which have proven to be more robust in extracting complex spatio-temporal features and improving the accuracy of LULC classification.

Recent years have seen U-Net convolutional neural networks (CNNs) become among the most popular for satellite image segmentation and LULC mapping [7,8]. The U-Net architecture, based on an encoder-decoder with skip connections, excels at summarizing patterns in the spectral and spatial dimensions of an image [9]. Numerous studies have demonstrated the success of U-Nets in identifying landscape elements (forests, crops, built-up areas, etc.) from remote sensing images. However, a standard U-Net operates image by image and does not directly exploit the temporal information available in multi-temporal image series [10,11].

On the other hand, integrating the temporal dimension can significantly improve LULC class discrimination [7]. Indeed, phenological characteristics and seasonal changes provide additional clues to distinguish, for example, agricultural crops (which turn green and then wither) from other vegetation cover, or to detect temporaly flooded areas. Therefore, remote sensing classification approaches seek to exploit three types of information: spatial, spectral, and temporal (in the case of time series of images) [5]. Deep learning techniques adapted to image time series typically include recurrent neural networks (RNNs) capable of assimilating multi-temporal image sequences. In particular, long short-term memory (LSTM) networks and their convolutional variants (ConvLSTM) allow temporal dependencies to be modeled while processing structured spatial data (images) [11,12]. Recent work in remote sensing has demonstrated the effectiveness of networks combining spatial convolutions and temporal recurrence for multi-temporal LULC classification: for example, Rubwurm and Körner proposed a convolutional recurrent layer encoder (ConvLSTM) for crop classification based on Sentinel-2 time series [11,12].

Such deep spatio-temporal approaches generally outperform more traditional machine learning methods in terms of overall accuracy, making better use of the richness of the data while overcoming the limitations of standard RNNs (the problem of vanishing gradients mitigated by the memory mechanisms of LSTMs). Accordingly, we present a method for predicting land cover from multi-temporal Sentinel-2 images based on an original deep learning model combining a U-Net and ConvLSTM layers. Our main purpose is to show how the combined use of spectral information (Sentinel-2 multispectral bands) and temporal information (images taken on different dates) can improve LULC classification in the study area.

Despite the increasing availability of high-resolution multi-temporal satellite imagery, land use and land cover (LULC) segmentation remains a challenging task, particularly in tropical and rapidly urbanizing regions. Sentinel-2 image time series are frequently affected by cloud contamination and atmospheric disturbances, leading to irregular temporal sam-

pling and missing observations as widely reported in the remote sensing literature [13,14]. In addition, temporal redundancy among closely spaced acquisitions may introduce limited new information while increasing computational complexity [9,12]. Spectral confusion between certain land cover classes, such as built-up areas and bare soil or dry vegetation, further complicates discrimination, especially in heterogeneous urban and peri-urban environments [15]. Moreover, the scarcity and uncertainty of ground-truth labels, often derived from manual interpretation or outdated ancillary data, introduce additional noise into supervised learning frameworks. These challenges motivate the development of spatio-temporal deep learning approaches capable of jointly exploiting spatial structure and temporal dynamics while remaining robust to noisy observations and limited supervision.

It is important to note that the proposed approach does not aim to introduce a novel neural network architecture at the block level. Instead, this work focuses on a task-specific instantiation of a ConvLSTM–U-Net framework for multi-temporal Sentinel-2 land use and land cover segmentation, combined with a systematic empirical evaluation against spatial-only and temporal-only baselines. The contribution is therefore primarily methodological and experimental, emphasizing robustness and applicability in data-scarce African urban environments.

Subsequently, in the following sections, we will first analyze the theoretical framework and related work in Section 2. Then, Section 3 will describe the data used and the methodology implemented before the results obtained are analyzed in Section 4. Finally, Sections 5 and 6 will be devoted to the discussion of the results and the conclusion of our work, respectively.

2. Related Work and Critical Literature Review

Land Use and Land Cover (LULC) classification from satellite images has become a crucial research topic, particularly with the availability of high-resolution multispectral datasets such as Sentinel-2 [4]. In recent years, deep learning methods have outperformed traditional classifiers (e.g., Random Forest, SVM) due to their ability to automatically learn spectral–spatial features and capture complex patterns in heterogeneous landscapes [16,17]. However, challenges such as seasonal variability, cloud cover, and spectral similarity between classes still hinder robust classification performance, especially in tropical regions like Central Africa. To provide a clearer understanding of the rest of this literature review, we have created Table 1 which summarizes the key contributions, data sources, methods, and findings of recent studies relevant to this work.

Table 1. Summary of key related studies on Conv RNN and U-Net for LULC classification.

Ref	Data Source	Method	Key Findings/Limitations
[12]	Sentinel-2 (time series)	Recurrent encoders	Effective for temporal dynamics, but limited spatial detail.
[11]	Sentinel imagery	ConvLSTM	Cloud-robust segmentation, but less efficient for spatial context.
[8]	Sentinel-2	CNN	Detects artisanal mining, limited temporal modeling.
[7]	High-resolution imagery (Ethiopia)	CNN + DL classifiers	Good accuracy for deforestation monitoring, seasonal variation issues.
[18]	Multi-source imagery	Improved U-Net	Enhances segmentation accuracy with structural modifications.
[19]	Gaofen-2	ConvNeXt-U (U-Net variant)	Lightweight U-Net for cropland extraction, efficient but not temporal.

Table 1. *Cont.*

Ref	Data Source	Method	Key Findings/Limitations
[20]	Multi-temporal datasets	Trans-ConvLSTM	Strong spatio-temporal modeling, higher complexity.
[21]	Sentinel-1 & Sentinel-2	Multimodal CNN	Improved robustness, but requires multiple sensors.
[22]	UAV video	ConvLSTM	Handles complex urban dynamics, lacks transferability to satellite imagery.
[23]	Hyperspectral images	Sparse Coding U-Net + Attention	High accuracy, computationally expensive.
[17]	Sentinel-2 (Colombia)	Deep CNN	Good performance in mountainous regions, not temporal.
[24]	RS imagery	3D FCN + attention	Strong feature extraction, but high GPU demand.
[25]	Multi-scale imagery	FCN	Handles multi-scale, but ignores temporal correlations.

2.1. Related Background

2.1.1. Data for LULC Classification

A key feature of any data is its source, and in the case of urban data, this source can be either sensors, surveys, or both [4]. However, in urban studies, satellite data is generally referred to as coming from satellites, which are simply a means of transporting sensors into space for better spatial coverage. Satellite data has become so popular that, over time, programs have multiplied for various applications. For LULC classification, the data used generally comes from satellites such as LanSAT [26], MODI [27], Sentinel, etc., [5]. Recent literature shows that data from the Sentinel 2 satellite is increasingly being used for this purpose, especially in study areas located in developing countries where other satellites often provide low-resolution data.

Moreover, the availability of images throughout the year makes it possible to capture seasonal changes in land cover: certain land use classes are more distinguishable at certain times of year (e.g., vegetation vigor in wet vs. dry seasons). Therefore, using image series rather than a single image makes it possible to improve classification by taking advantage of these temporal variations.

Several global and operational land cover products based on Sentinel-2 imagery have been released in recent years and now constitute important references for land use and land cover (LULC) mapping at high spatial resolution. Among the most widely used are Google Dynamic World, ESA WorldCover, and Esri Global Land Cover, all of which provide global coverage at 10 m resolution with multi-class land cover taxonomies.

- Google Dynamic World [28] delivers near-real-time land cover maps derived from Sentinel-2 imagery using a deep learning model trained on large-scale automatically labeled data. It distinguishes nine land cover classes, including built-up areas, crops, grass, trees, water, and bare ground. Its main strength lies in its high temporal frequency and global consistency, making it suitable for large-scale monitoring and rapid change detection.
- ESA WorldCover [29] provides a global land cover map at 10 m resolution with eleven classes, generated through a supervised classification framework combining Sentinel-1 and Sentinel-2 data. The product emphasizes thematic consistency at the

global scale and is optimized for interoperability with climate, environmental, and biodiversity studies.

- Esri Global Land Cover [30] is another global product derived from Sentinel-2 imagery, designed for integration into GIS workflows. It offers a simplified multi-class legend and visually smooth maps that are particularly suited for cartographic and visualization purposes.

While these products represent the current state of operational global LULC mapping, their objectives differ from the scope of the present study. Global products are designed to ensure worldwide consistency and broad thematic coverage, often at the cost of local specificity. In contrast, our work focuses on local urban dynamics in Yaoundé, where spectral confusion, peri-urban expansion, and heterogeneous land use patterns pose significant challenges. In addition, differences in class definitions limit direct quantitative comparison. Our study deliberately aggregates land cover into three dominant classes (water, vegetation, and built-up areas) to reduce label noise and ensure reliable supervision in a data-scarce context, whereas global products rely on more detailed but sometimes ambiguous urban subclasses.

In tropical regions, persistent cloud cover and atmospheric variability often result in incomplete or irregular Sentinel-2 time series, which has motivated the use of multi-temporal aggregation strategies and temporal modeling techniques to improve robustness to missing or noisy observations.

2.1.2. Convolutional Neural Networks (CNNs) and U-Net Variants

CNN-based approaches have been widely applied to remote sensing data due to their strong spatial feature extraction capabilities. U-Net and its variants, initially developed for biomedical image segmentation, have been successfully adapted for LULC tasks [18,19]. For instance, Ref. [8] developed a multispectral CNN for detecting artisanal mining sites in Ghana, demonstrating the power of deep models in identifying small-scale anthropogenic activities. Similarly, Refs. [24,25] introduced multi-scale and attention-based CNN architectures to better capture fine-grained spatial structures. Nonetheless, pure CNN models lack mechanisms to effectively incorporate temporal dynamics in multi-temporal imagery.

2.1.3. Recurrent and ConvLSTM-Based Models

To address temporal dependencies, recurrent architectures such as LSTM and ConvLSTM have been introduced for multi-temporal classification. Ref. [12] pioneered the use of recurrent encoders for sequential land cover classification, while Ref. [11] extended ConvLSTMs for cloud-robust segmentation of Sentinel imagery. Later, Ref. [20] proposed a Trans-ConvLSTM for spatio-temporal segmentation, confirming the advantage of combining convolutional filters with sequential modeling. In urban contexts, ConvLSTM has also been applied to UAV video segmentation [22], underlining its flexibility across data modalities. Yet, ConvLSTM models alone may struggle with capturing multi-scale spatial context compared to encoder–decoder CNN architectures like U-Net.

2.1.4. Hybrid CNN–RNN Models for Multi-Temporal Remote Sensing

Recent studies have sought to integrate CNNs with recurrent architectures to leverage both spatial and temporal information. For instance, multimodal approaches combining Sentinel-1 and Sentinel-2 have improved robustness to cloud cover and seasonal variations [21]. In addition, hybrid U-Net–based designs enhanced with attention or sparse coding mechanisms have demonstrated state-of-the-art performance in hyperspectral and multispectral segmentation [23]. Similarly, case studies in Ethiopia and Colombia show how hybrid deep learning pipelines can support environmental monitoring and land

management [7,17]. Despite these advances, relatively few studies have explored hybrid ConvLSTM–U-Net architectures specifically tailored for the spatio-temporal dynamics of Sentinel-2 data in Sub-Saharan Africa.

Beyond recurrent and hybrid CNN–RNN architectures, Ref. [31] has explored volumetric segmentation approaches based on 3D convolutional neural networks, particularly 3D U-Net variants, to jointly model spatial and temporal dimensions. These architectures treat temporal stacks of images as volumetric inputs and have demonstrated strong performance in structural and contextual analysis tasks, including medical imaging and, more recently, remote sensing applications. For instance, Mahmud et al. proposed a 3D U-Net–based framework for volumetric image segmentation, highlighting its effectiveness in capturing spatial continuity and contextual dependencies across dimensions. While such approaches provide powerful volumetric feature extraction, they generally assume regular temporal sampling and fixed-length input volumes. In contrast, the proposed ConvLSTM–U-Net architecture explicitly models temporal dependencies through recurrent memory mechanisms, making it more suitable for irregularly sampled satellite image time series and scenarios with missing or uneven temporal observations.

2.1.5. Remote Sensing Foundation Models and Vision Transformers

Recent years have witnessed the emergence of remote sensing foundation models (RSFMs), inspired by large-scale vision foundation models developed in computer vision. These models rely on self-supervised or weakly supervised pre-training over massive collections of remote sensing imagery, with the objective of learning generalizable representations that can be transferred to a wide range of downstream Earth observation tasks. A recent comprehensive survey by Lu et al. [32] provides an extensive overview of these models, highlighting architectural trends, pre-training strategies, and application domains in remote sensing. Most RSFMs adopt Vision Transformer (ViT) or hybrid CNN–Transformer architectures, often trained using masked image modeling, contrastive learning, or multimodal objectives. Representative examples include masked autoencoder-based models trained on Sentinel-2 imagery, large-scale ViT backbones pre-trained on multi-sensor datasets, and vision–language models that align satellite imagery with textual or semantic information. These approaches have demonstrated strong performance and generalization capabilities across multiple tasks, including land cover classification, change detection, and object recognition. In parallel, several curated repositories, such as the Awesome Remote Sensing Foundation Models collection [33–37], document the rapid expansion of RSFMs, covering vision-only, vision–language, generative, and temporal models. These works collectively reflect a paradigm shift from task-specific architectures toward generalist models capable of transfer learning across heterogeneous remote sensing problems.

Despite their strong potential, current RSFMs typically require large-scale unlabeled datasets, significant computational resources, and complex pre-training pipelines, which may limit their direct applicability in data-scarce or operational contexts, particularly in developing regions. Moreover, many existing foundation models focus primarily on spatial representation learning, while explicit modeling of fine-grained temporal dynamics remains an active research challenge. In contrast, our work adopts a ConvLSTM-based spatio-temporal architecture, which explicitly models temporal dependencies through recurrent memory mechanisms while preserving spatial structure via convolutional operations. This design choice is motivated by the need to effectively capture seasonal and phenological variations in Sentinel-2 time series under limited supervision, while maintaining computational efficiency and training stability. Rather than competing directly with foundation models, our approach aims to provide a strong, interpretable, and data-efficient baseline for multi-temporal land use and land cover classification in urban environments.

2.2. Research Gap, Motivations, and Challenges

2.2.1. Research Gab

Deep learning has significantly advanced the field of land use and land cover (LULC) classification from remote sensing data. Convolutional neural networks (CNNs) and U-Net-based architectures have been widely applied with encouraging results across diverse landscapes [18,19]. Nevertheless, most of these studies often focus on regions with abundant annotated data and relatively stable landscapes, while African urban areas remain underrepresented in this growing body of literature [7,17]. This underrepresentation is problematic, given the rapid urban expansion and environmental pressures in cities such as Yaoundé, Cameroon, where planning decisions depend heavily on accurate and up-to-date land cover information.

Traditional land cover classification techniques often rely on single-date imagery and heuristic or rule-based models, which lack the capacity to capture the temporal dynamics of land cover change [16,25]. Such limitations are particularly critical in tropical urban environments where agricultural cycles, seasonal vegetation growth, and human-induced land use transitions occur rapidly. Recent published works have demonstrated the value of multi-temporal approaches and sequential models in improving classification performance [12,20], but these methods have not yet been systematically applied to Central African urban contexts. Furthermore, convolutional models, although effective in extracting spatial features, struggle to represent temporal dependencies unless hybridized with recurrent architectures such as ConvLSTM [11,22].

Therefore, this paper will try to fill the research gap lied in the limited exploration of spatio-temporal deep learning models, particularly ConvLSTM-based U-Nets, for urban land cover classification in African settings using Sentinel-2 time series. This gap is further reinforced by the scarcity of ground truth data and the high heterogeneity of urban landscapes in cities like Yaoundé, which combine informal settlements, fragmented vegetation patches, and fast-changing land uses [21,24].

Accordingly, the present study does not seek to propose a new spatio-temporal architecture, but rather to evaluate the practical relevance of an established ConvLSTM–U-Net design under consistent experimental conditions and in an underexplored geographic context.

2.2.2. Motivation and Challenge

Thus, this work is driven by three motivations. First, accurate mapping of LULC is essential for urban planning and sustainable development in Yaoundé, a city experiencing rapid population and spatial growth. Second, environmental monitoring in these areas requires robust mapping tools to track deforestation, flood zones, and changes in green spaces, which are essential for climate change adaptation and resilience strategies [7,16]. Third, the availability of freely accessible, high-resolution, multi-temporal Sentinel-2 data, combined with the emergence of hybrid deep learning models, presents a unique opportunity to bridge the existing methodological gap.

Nevertheless, deploying ConvLSTM U-Net models in this context raises several challenges. Sentinel-2 images over tropical regions suffer from persistent cloud coverage and atmospheric disturbances, which affect time series consistency [11]. In addition, spectral confusion remains a critical issue, as built-up areas, bare soil, and certain vegetation types often share similar spectral signatures, leading to classification errors [8,25]. On the computational side, training deep recurrent architectures on multi-temporal datasets is computationally intensive and requires carefully optimized strategies to avoid overfitting [23]. Finally, the lack of updated and detailed ground truth data for African urban environments hampers supervised learning and validation [17,21].

Overall, these motivations and challenges highlight the need for a hybrid ConvLSTM U-Net architecture tailored to the classification of Sentinel-2 multitemporal images in Yaoundé, which not only takes into account the spatiotemporal dynamics of land cover changes but also helps to reduce existing geographical and methodological gaps in research.

3. Materials and Methods

This Section presents the empirical and technical foundations of our study aimed at developing a hybrid deep learning approach combining ConvLSTM and U-Net for land use/land cover (LULC) classification in the city of Yaoundé, Cameroon. We first describe the study area and the data collection methods used for training, validating, and testing the model. Next, we outline the methodology implemented, including data preparation, neural network design, and training and evaluation parameters. This structure highlights the link between the geographical context and the technological choices made to address the specific challenges of LULC classification in African urban environments.

3.1. Study Area and Data Collection

The choice of study area and the quality of the data collected are very important for spatio-temporal analyses. The literature shows that areas in sub-Saharan Africa have been largely overlooked in terms of modeling urban land use and occupation [1,4].

3.1.1. Study Area

We experimented with our approach in the region of Yaoundé, the capital of Cameroon (approx. 3.75–3.95° N, 11.4–11.65° E), characterized by a tropical urban environment surrounded by forest and agricultural areas (see Figure 1). Yaoundé is subdivided into 7 subdivision municipalities (Yaoundé I, Yaoundé II, Yaoundé III, Yaoundé IV, Yaoundé V, Yaoundé VI, and Yaoundé VII) (Figure 1), and its population was estimated to be 1,817,524 in 2012 [6].

3.1.2. Data Collection and Characterization

We retrieved a series of Sentinel-2 L2A images of this area from Google Earth Engine (GEE) [38,39]. The images were filtered between 2018 and 12 May 2025, with a cloud cover criterion of <60% to ensure sufficient visibility. Up to 144 images with virtually no cloud cover were extracted (GEE limits simultaneous exports to 100) and exported to Google Drive as shown in Table 2. Each image has a spatial resolution of 10 m (we specified scale = 10 in the GEE export, which resamples the 20 m bands to 10 m) and covers six spectral bands: Blue (B2), Green (B3), Red (B4), NIR (B8), SWIR1 (B11), and SWIR2 (B12) [11,12]. These bands correspond to the wavelengths most commonly used for discriminating between vegetation, water, and built-up areas. The pixel values are reflectances (integers in L2A data) that we normalize as floating point values (division by 10,000) for the neural network.

Table 2. Number of images per cloud cover threshold (Cloud < x%) between 2018 and 2025.

Year	<10%	<20%	<30%	<40%	<50%	<60%
2018	0	0	2	2	3	3
2019	2	5	14	23	24	27
2020	2	6	8	15	18	23
2021	2	2	8	11	19	22
2022	0	2	4	8	14	16
2023	3	4	5	12	17	19
2024	3	3	7	13	20	28
2025	0	1	1	3	5	6
Total	12	23	49	87	120	144

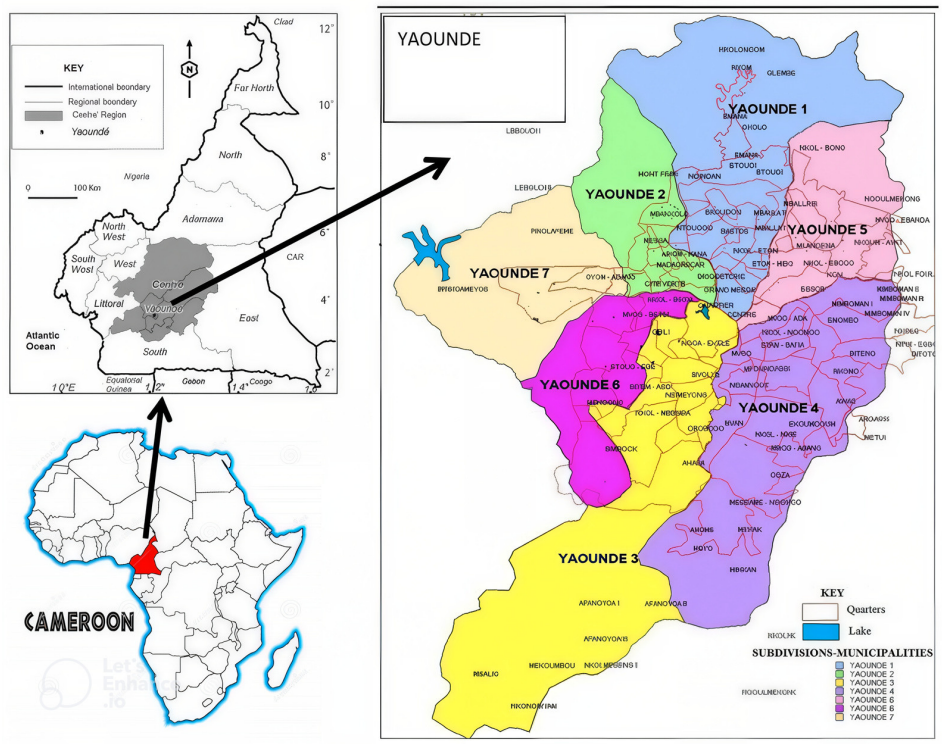


Figure 1. Map of Yaoundé city according to the "Système d'Information Géographique", Yaoundé City Council, Cameroon, (2011) [6].

At the same time, LULC reference data was prepared to train the model. For each Sentinel-2 image considered, we had a corresponding land cover map (same footprint and 10 m resolution) from a supervised learning dataset [39]. These class masks were obtained by photo interpretation and manual classification from high-resolution images and local GIS data (buildings, land use), then rasterized onto the Sentinel-2 grid. We distinguished three major classes relevant to the area: (0) Water (lakes, rivers), (1) Vegetation (forests, savannas, crops, green spaces) and (2) Built-up/urban areas (buildings, roads, impervious surfaces) [38]. These classes cover most of the land use in the city of Yaoundé and its surroundings [6]. It should be noted that the Water class is very much in the minority in the study area, while vegetation and built-up areas occupy the majority of pixels.

Although the Sentinel-2 images used in this study span the period from 2018 to 2025, this temporal range should not be interpreted as a single continuous time series modeling long-term land use evolution. Instead, the multi-year archive is used as a temporal pool to construct short and locally coherent temporal sequences. Specifically, the proposed ConvLSTM-based model processes sequences of fixed length (T = 5), composed of consecutive Sentinel-2 acquisitions, and predicts land cover labels for the last image of each sequence. Each sequence therefore captures short-term temporal context, typically corresponding to a few weeks or months depending on acquisition frequency, and is designed to model seasonal and phenological variations rather than long-term urban transformation. Applying this strategy resulted in a total of 137 independent spatio-temporal sequences, each treated as a single training sample. The use of images from multiple years is motivated by the need to mitigate persistent cloud coverage in the study area and to ensure sufficient seasonal coverage. Images from different years are not concatenated to form long trajectories; instead, they are treated as independent short sequences. This strategy improves robustness to atmospheric and seasonal variability while limiting the influence of inter-annual land use changes.

3.2. Proposed Method: Hybrid ConvLSTM U-Net Architecture

As stated in the introduction to Section 3, we present here the methodology used, including first the preparation of the data, followed by the design of the neural network, as well as the training and evaluation parameters.

3.2.1. Preparation of Spatio-Temporal Sequences of Data

From the chronological list of available Sentinel-2 images, we created sliding time series of length T = 5. In other words, each series groups together five consecutive dates (not necessarily evenly spaced, depending on cloud-free availability). For a given sequence, the first 4 images serve as history and the 5th image is the target date for which we seek to predict the LULC map [12]. The first four images in the sequence are used solely as unlabeled temporal context to provide historical information to the ConvLSTM module. Ground-truth labels for these intermediate images are neither required nor exploited during training. The sequence length is intentionally kept short to limit the likelihood of major land-cover transitions within the temporal window. In cases where minor changes occur among the historical images, the ConvLSTM mechanism implicitly learns to attenuate inconsistent temporal information through its gating functions. Explicit modeling of land-cover change across time steps is beyond the scope of the present work and is left for future investigation.

Therefore, only the final image of each temporal sequence is associated with a ground-truth land use and land cover mask (denoted as mask), while the preceding images are used solely as unlabeled temporal context. Temporal sequences of fixed length $T = 5$ are constructed using a sliding-window strategy applied to the available Sentinel-2 acquisitions. This process yields $N = M - T + 1$ spatio-temporal sequences from M images (e.g., approximately 46 sequences from 50 images). Although multiple sequences may be extracted from the same geographic region and may share identical target labels, they are not considered duplicates, as their temporal context differs due to seasonal variability, atmospheric conditions, and acquisition timing.

Each sequence is considered independently for model training. We then randomly separated these sequences into a training set (70% of the sequences), a validation set (20%) and a test set (10%), taking care to mix the dates (i.e., sequences close in time can go into different batches, ensuring validation over the entire period) [12,17].

For each sequence, class histograms Figure 2 were computed from the target LULC mask and clustered to approximate multi-class stratification. This resulted in 95 sequences for training, 28 for validation, and 14 for testing. The training, validation, and test subsets contained [122 M, 237 M, 228 M], [35 M, 71 M, 66 M], and [16 M, 35 M, 34 M] pixels per class, respectively, confirming balanced representation across datasets.

A strategy of random patches was adopted to facilitate learning while artificially increasing the volume of data. Sentinel-2 images cover a relatively large spatial area (several hundred square kilometers), and it is inefficient to provide the entire image to the network due to memory limitations and spatial redundancy [9,12]. Therefore, we defined a patch size of 256 × 256 pixels (i.e., 2.56 km on each side): each time a sample is called, a patch is randomly cut out at the same position in the five images in the sequence and in the ground truth mask. For training (random crop), the position of the patch is chosen randomly at each epoch, introducing spatial variability (data augmentation); for validation (center crop), the centered patch is systematically sampled to obtain a deterministic evaluation. Each sequential training patch is therefore a tensor of dimension (T = 5, C = 6, H = 256, W = 256) at the input, with an annotation matrix (H = 256, W = 256) at the output.

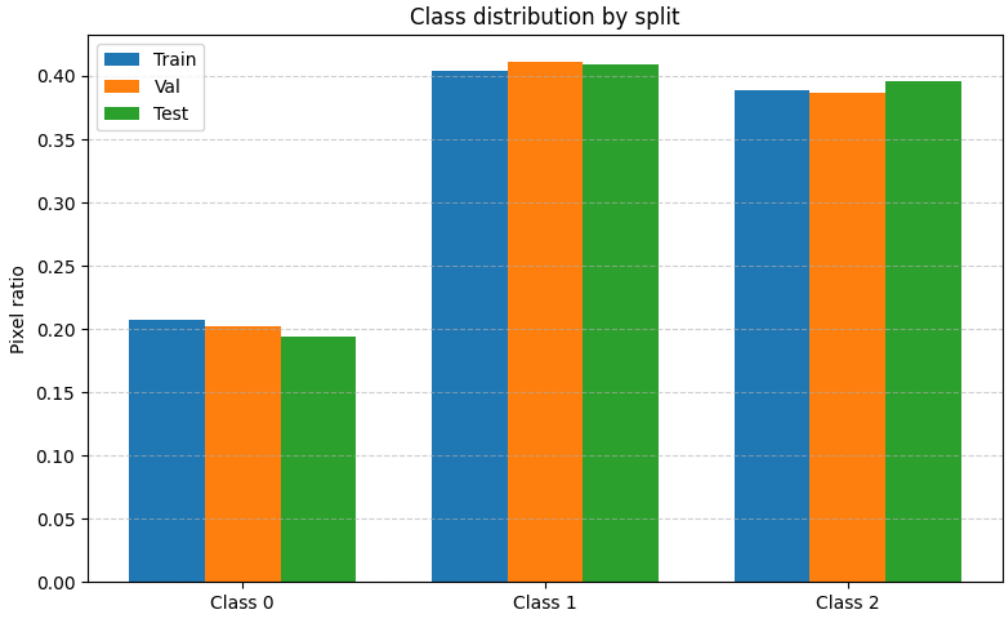


Figure 2. Distribution of data proportions in each class for the training, validation, and test datasets.

3.2.2. Hybrid ConvLSTM U-Net Model Architecture

The proposed architecture combines ConvLSTM and U-Net to jointly address temporal and spatial challenges in multi-temporal LULC segmentation. ConvLSTM is adopted for temporal modeling because it preserves spatial structure while learning temporal dependencies through convolutional recurrent gates. Unlike simple temporal aggregation strategies (e.g., temporal averaging or feature concatenation), ConvLSTM enables adaptive temporal filtering and selective memory of relevant observations. Attention-based temporal mechanisms, while powerful, typically require larger training datasets and higher computational resources, which motivated the choice of ConvLSTM in the present data-constrained and resource-limited setting. U-Net is employed for spatial feature extraction and reconstruction due to its encoder–decoder architecture with skip connections, which facilitate the preservation of fine-grained spatial information and object boundaries. This property is particularly important for LULC segmentation, where accurate delineation of urban edges and transitions between built-up areas, vegetation, and water bodies is critical. The integration of ConvLSTM with U-Net therefore enables complementary temporal context modeling and precise spatial boundary preservation.

We implemented an original ConvLSTM U-Net in PyTorch 3.11.13, combining six recursive encoding levels and a U-Net-type decoding structure [7]. Specifically, the encoder has six nested ConvLSTM layers: layer 0 operates on the input images (six spectral channels) and produces basic feature maps (64 filters) that are updated over time; then each subsequent layer reduces the spatial resolution by 2 (max pooling) and increases the depth (e.g., 128, 256, . . . up to 2048 filters at the smallest scale). Figure 3 shows the architecture of the hybrid model we implemented.

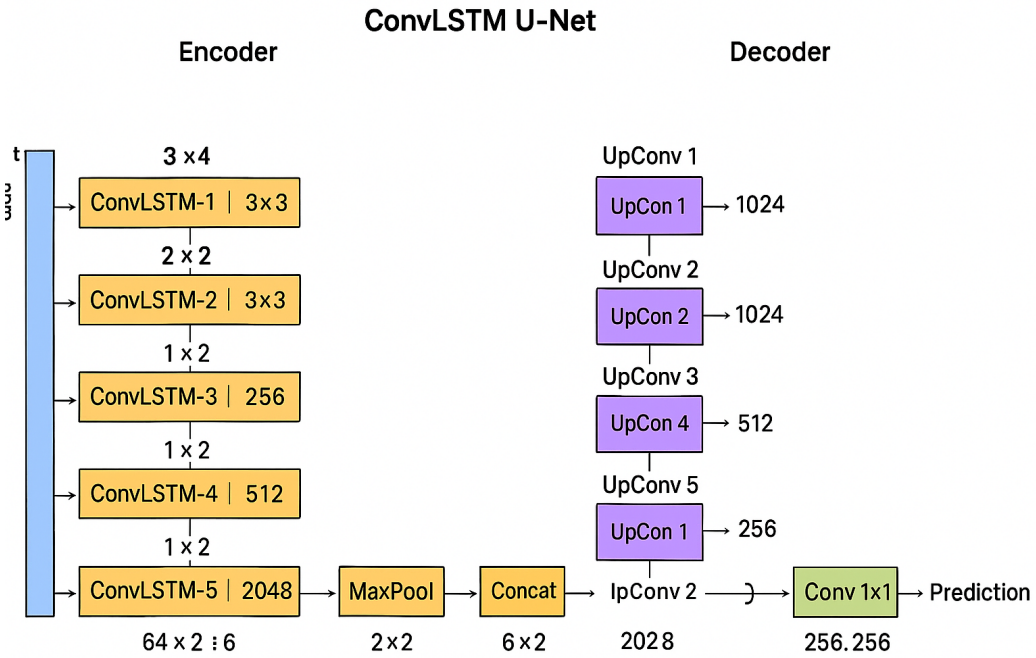


Figure 3. Hybrid ConvLSTM-UNet based Architecture.

In this hybrid architecture, the LSTM mechanism in each layer maintains a hidden state ht and a cell state ct that are propagated temporally [11]. At each time step t, the ConvLSTM layer convolves the current input (image or feature map from the previous layer) with its previous hidden state ht-1, thereby calculating input, forget, output, and candidate state gates (gates i, f, o, g). The temporal modeling component is implemented using a convolutional long short-term memory (ConvLSTM) module, following the standard formulation introduced by [10]. The ConvLSTM employs convolutional gating mechanisms to update its hidden and cell states at each time step. Given a temporal sequence of length $T = 5$, the hidden state at the final time step ($t = 5$) recursively aggregates spatial and temporal information from the preceding observations $t = 1 \ldots 4$. This final hidden representation is subsequently used as input to the U-Net decoder for LULC prediction. These tensors represent the features extracted at the final date (t = 5) but incorporating information from dates 1 to 4 [9].

The U-Net decoder then uses these features to reconstruct the segmentation. This is achieved through progressive upsampling: the hidden state at the lowest level is first upsampled by a transposed convolution (deconvolution) to double its size, then concatenated with the hidden state from the previous level [7,8]. Two 3×3 convolutions (with BatchNorm + ReLU) are applied to this concatenation to properly merge the information (this is the typical "decode block" of U-Net). This operation is repeated at each level: at each step, the upsampled feature is concatenated with the corresponding output of the encoder and processed by convolutions, until returning to the original 256×256 scale. Finally, a 1×1 convolution is applied to project the filters of the last decoder level (64 channels) into the space of the three classes to be predicted. This produces a $3 \times 256 \times 256$ output map, which is interpreted as classification logits per pixel. An argmax on the class dimension provides the predicted class for each pixel of the patch at the final date [17].

The entire model has over 87 million parameters to train (mainly in the convolutional filters of each ConvLSTM and in the convolutions of the decoder). To foster convergence, we initialized the weights in a standard way (normalized distribution) and used implicit regularization mechanisms such as batch normalization in the decoder. One challenge is that our dataset obtained from our study area (a few dozen sequences, or a few hundred

potentially distinct patches) is relatively small compared to the network's capacity. We partially overcome this problem by dividing the data into patches (each image provides many examples) and increasing the amount of data (random crops at each epoch).

4. Experiments and Results Analysis

4.1. Experimental Protocol

Our experimental protocol is shown in the diagram in Figure 4. We started with data acquisition (collection), then data preparation and splitting, followed by model training and validation, and finally testing. At each stage, the key parameters and elements are shown in the figure.

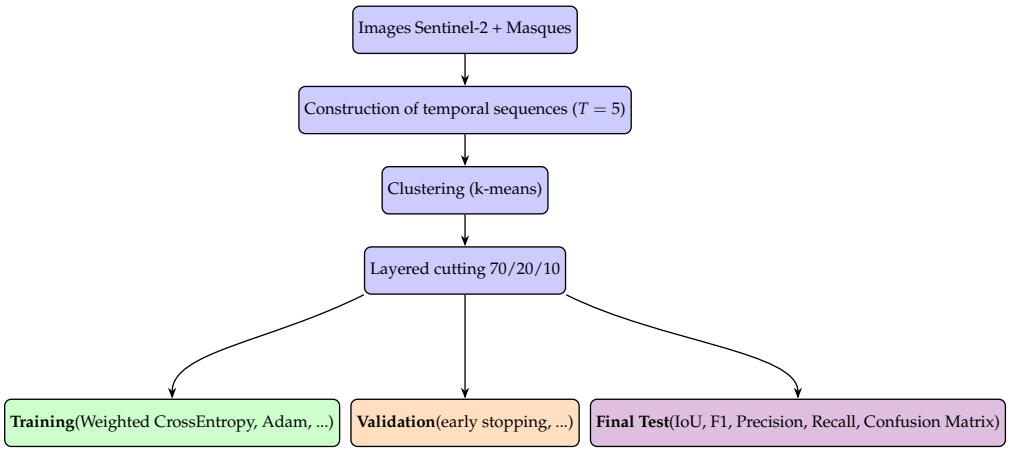


Figure 4. Experimental workflow: data preparation steps (blue), training (green), evaluation (orange), and final test (violet).

The experiments were conducted on a time series of Sentinel-2 multispectral images covering the Yaoundé region (Cameroon). The data include several acquisitions dated in GeoTIFF format as well as their associated classification masks (built-up areas, vegetation, bare soil). The images and masks are organized in separate directories for automated processing.

4.2. Model Setting and Training

We trained our hybrid ConvLSTM U-Net network on the training set on a machine with a GPU microprocessor (NVIDIA Tesla via the Kaggle environment) using Python 3.11 [40]. The number of trainable parameters for each model configuration is reported to contextualize computational complexity: approximately 69.8 M parameters for U-Net, 1.15 M for ConvLSTM, and 87.7 M for the hybrid ConvLSTM–U-Net model. For the implementation of the model, we used Rasterio [41] for reading satellite images, scikit-learn [42] for stratification and metrics, and Matplotlib 3.7.2 [43] for visualizing the results. The objective function was a pixel-by-pixel cross-entropy between the prediction and the ground truth, with class weighting to compensate for imbalance (weight of 2.5 for the Water class, 1.0 for Vegetation, and 1.0 for Urban). This choice gives greater relative importance to errors on water pixels, a minority class, to prevent them from being neglected by the model. The Adam optimizer (with an initial learning rate of 0.001) was used for gradient descent. A plateau LR reduction scheduler monitors validation loss and divides the LR by 2 in case of stagnation for 3 consecutive epochs. In addition, we used the metrics of precision, recall, F1-score, and IoU (Intersection-over-Union) on the validation set at each epoch to judge the quality of multi-class segmentation. We use the macro (unweighted) average of the metrics across the three classes as an overall indicator (average F1 per class, etc.). Finally, to optimize program termination, we implemented early stopping: if the average validation

F1 has not increased for five consecutive epochs, training is terminated prematurely to avoid overfitting. The model with the best validation F1 is saved ("best model") for use in the final evaluation (see Figure 4).

Performance Metrics Calculation

Performance is evaluated using several standard semantic segmentation metrics chosen according to the type of problem, which is a multi-class classification on an unbalanced dataset. We therefore used the following metrics: confusion matrix describe in Table 3 for error analysis, interSection over union (IoU), F1-score (per class and macro-average), precision, and recall which are given by the following formula:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The results are monitored during training on the validation set, then consolidated by a final evaluation on the test set.

Table 3. Confusion matrix. TP = number of True Positive; TN = Number of True Negative; FN = Number of False Negative; FP = Number of False Positive.

	Predicted Positive	**Predicted Negative**
Actual positive	TP	FN
Actual negative	FP	TN

Our models are based Deep neural networks which are generaly trained with simple loss functions (e.g., softmax loss). These loss functions are suitable for classical classification problems where performance is measured by the overall accuracy of the classification [44]. For the task of object category segmentation, the three classes (foreground and background) are highly unbalanced. The intersection over the union (IoU) is generally used to evaluate the performance of any object category segmentation method. The IoU metric, also known as Jaccard index, is the most commonly used performance metric for comparing the similarity between two arbitrary shapes [45]. Such a metric could be used to eliminate the performance gap between training and testing, the IoU loss has been introduced for 2D and 3D object detection [46].

4.3. Results Analysis

For the performance evaluation of our ConvLSTM-UNet hybrid model, we calculated the validation performance of the training on the validation set by calculating the values of all the standard classification metrics presented in Section 4.2. Then, we calculated the test performance values on the test dataset. It should be noted that the simulation code was configured so that execution would stop once a certain overfitting threshold was reached, which explains why we simulated the model over a small number of epochs.

4.3.1. Learning Curve Analysis

Figure 5 shows the curves representing the evolution of training-validation losses (Figure 5) and performance metrics (Figure 5), respectively, according to the different epochs.

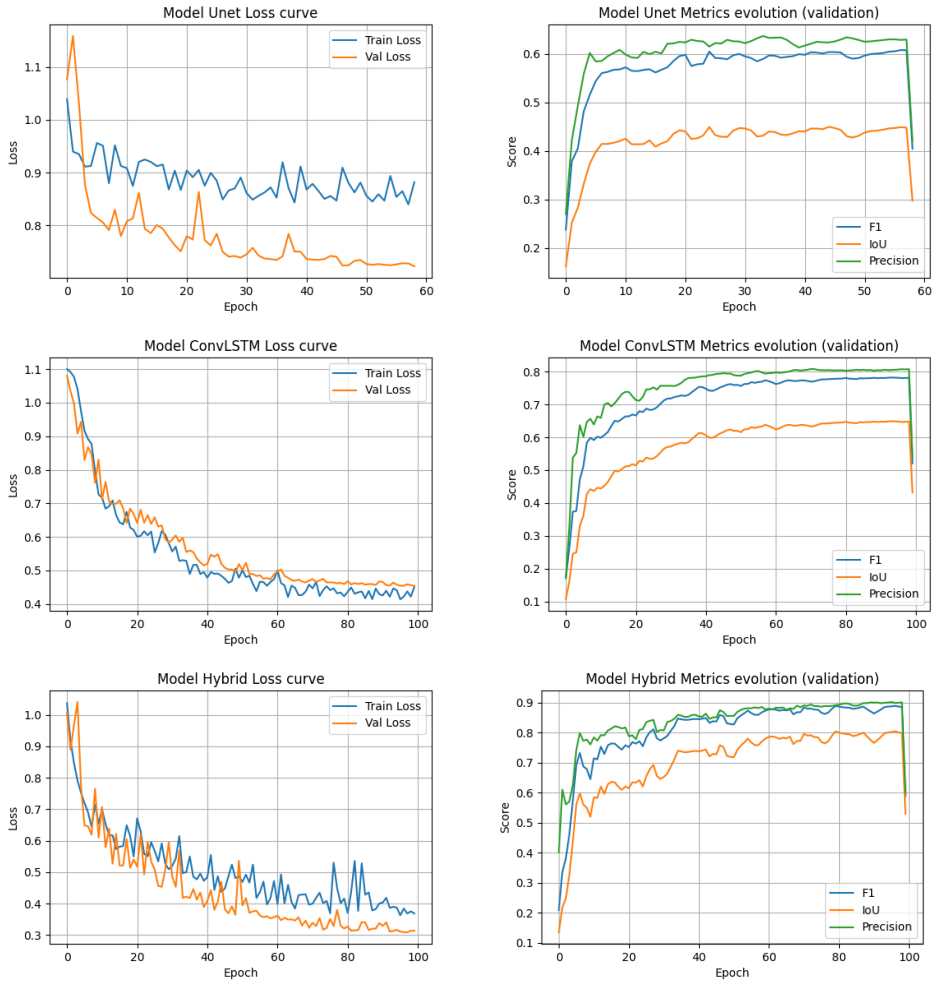


Figure 5. Illustration of the loss curve for training-validation and variation of performance metrics by epoch.

The loss curves in Figure 5 show a gradual convergence of each of the model, with a steady decrease in training and validation losses on 10 epochs. This convergence is very weak for the UNet model, which also exhibits a high rate of overfitting. It is relatively good for the ConvLSTM model, with low overfitting and loss stabilization slightly above 0.4 for both training and validation. Finally, the ConvLSTM-UNet hybrid model shows the lowest loss, stabilizing at around 0.3 for validation and around 0.35 for training.

Furthermore, the validation metric curves in Figure 5 confirm this stability which is on the left from top to bottom for UNet, ConvLSTM, and the hybrid model, respectively, showing stable performance with the curves flattening well before the last epoch. Overall, once again, the hybrid model yields the best performance curves for precision, F1-score stabilizing around 0.9, 0.88, and 0.8, respectively. It is followed by the ConvLSTM model and finally the UNet model, which performs very poorly for the same metrics. The following Section will provide more in-depth numerical analyses for each of these three architectures.

4.3.2. Performance Results Analysis

The confusion matrices shown in Figure 6 (left, right) for the model's performance during the validation and testing phases, respectively, were used to calculate the numerical values of the various metrics summarized in Table 4.

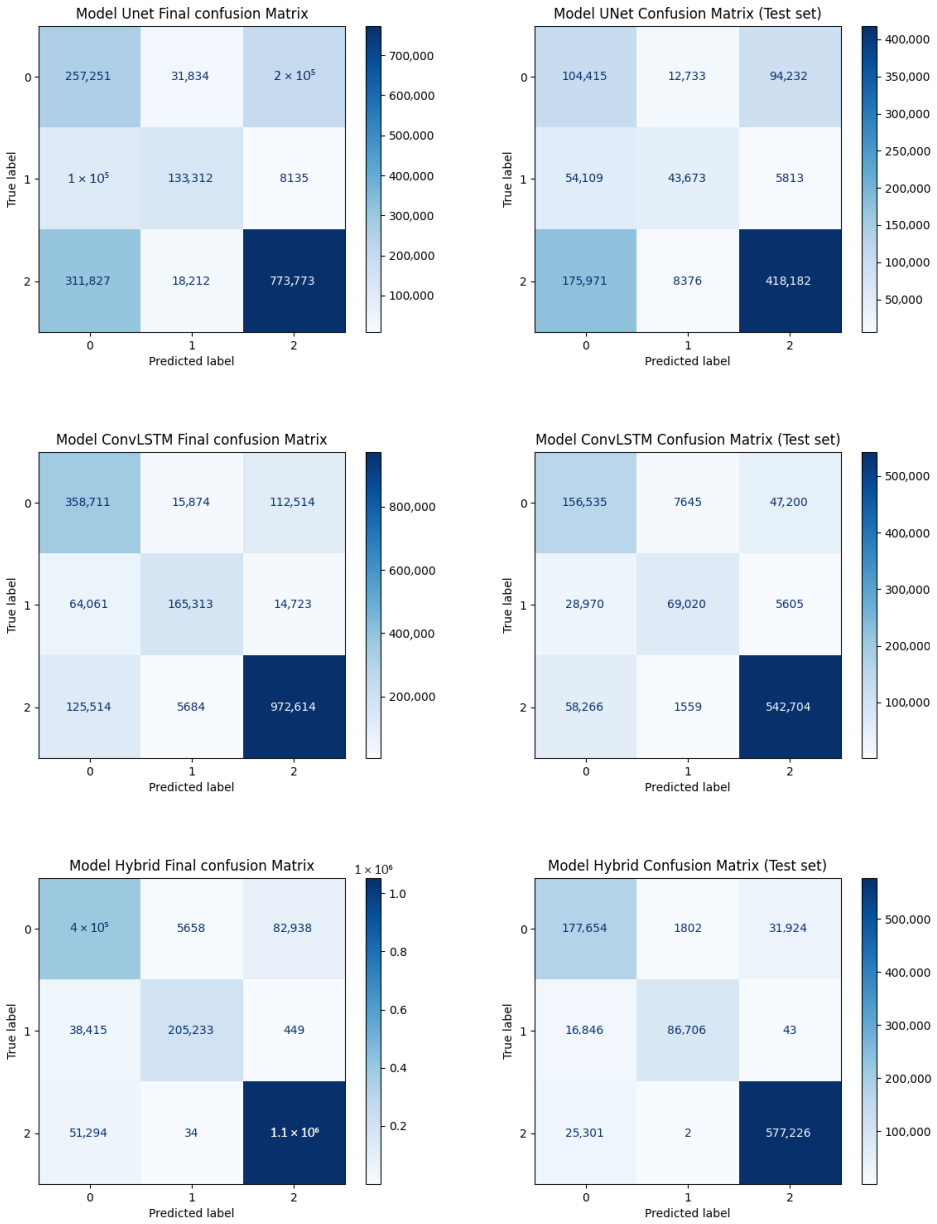


Figure 6. Confusion matrix for (**left**) validation and (**right**) test for the U-Net, ConvLSTM, Hybrid ConvLSTM-UNet as shown above each of these images.

For each of the three classes, Table 4 presents a summary of the performance of the models evaluated on the validation set (left) and test set (right) in terms of precision, recall, F1-score, and IoU for each of the three land cover classes. Our three previous models are compared: U-Net, ConvLSTM, and a hybrid model (probably ConvLSTM+U-Net). Remember that class 0 includes areas consisting of water (lakes, rivers), while class 1 includes vegetation (forests, savannas, crops, green spaces), and class 2 includes built-up areas (buildings, roads, impervious surfaces).

Table 4. Summary of model validation performance on the validation and testing sets.

	Validation Score				Test Score			
Classes	**Precision**	**Recall**	**F1-Score**	**IoU**	**Precision**	**Recall**	**F1-Score**	**IoU**
U-Net								
Class 0	0.383	0.528	0.444	0.286	0.312	0.494	0.383	0.237
Class 1	0.727	0.656	0.624	0.457	0.674	0.422	0.519	0.350
Class 2	0.790	0.701	0.743	0.590	0.807	0.694	0.746	0.595
Macro-Avg	0.633	0.592	0.603	0.444	0.598	0.537	0.549	0.394
ConLSTM								
Class 0	0.654	0.736	0.692	0.530	0.642	0.741	0.688	0.524
Class 1	0.884	0.677	0.766	0.622	0.882	0.666	0.759	0.612
Class 2	0.884	0.881	0.882	0.790	0.911	0.901	0.828	0.910
Macro-Avg	0.807	0.765	0.780	0.647	0.812	0.769	0.784	0.655
Hybrid Unet-ConLSTM								
Class 0	0.816	0.818	0.817	0.692	0.808	0.840	0.824	0.701
Class 1	0.973	0.841	0.902	0.822	0.980	0.837	0.903	0.823
Class 2	0.927	0.953	0.940	0.891	0.948	0.958	0.953	0.910
Macro-Avg	0.905	0.871	0.886	0.801	0.912	0.878	0.893	0.811

The results in Table 4 whose comparative macro average values are illustrated in the figure Figure 7, highlight contrasting performances between classes. Unsurprisingly, class 2, which is by far the most prevalent in the dataset, is globlally predicted with better accuracy by all of three models. Conversely, classes 1 and especially 0 are less well recognized, with lower recall, indicating underdetection. The confusion matrix confirms these observations: there is significant confusion between classes 0 and 2, while class 1 is scattered between correct predictions and errors toward the other two classes. We will analyse each trained model individually before concluding. We will analyse and interpret each validated and tested model individually before concluding on our experiments.

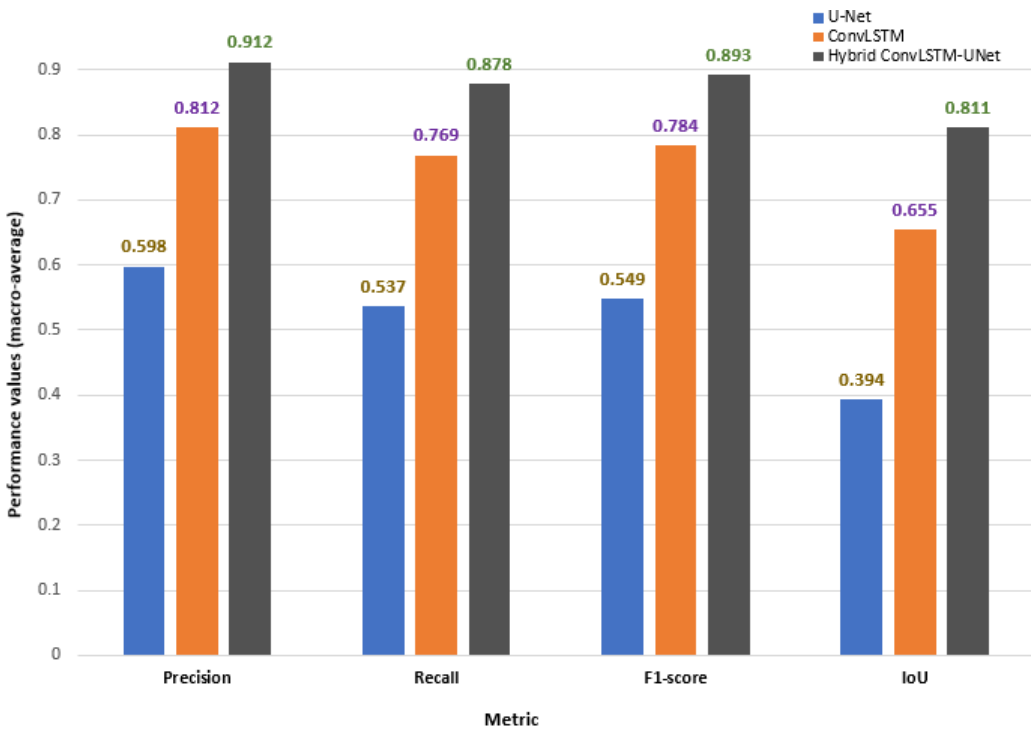


Figure 7. Macro-Average performance results of our models on the testing set.

Among the three models, the ConvLSTM+U-Net hybrid model delivers near-optimal performance, achieving very high and balanced scores, with 0.811 and 0.893 for the macro average of the IoU and F1-score, respectively. For built-up areas (class 2), the IoU rises to 0.910 and the F1-score to 0.953, reflecting an excellent ability to segment urban structures. Vegetation (class 1) is also well captured (IoU = 0.902), probably thanks to its seasonal variability, which is well modeled by the temporal component (ConvLSTM) of the hybrid model. Finally, the Water class (class 0) benefits from a clear gain (IoU = 0.703, f1 = 0.824), confirming that the hybrid architecture also manages to better isolate these entities, which are often confused.

The ConvLSTM model shows a net gain for vegetation and built-up areas compared to the U-Net model, but still remains weaker than the hybrid model. The ConvLSTM model brings notable improvements with a macro-average IoU of 0.655 and a macro-average f1-score of 0.784. It enables better detection of vegetation (class 1) (IoU = 0.622) and, above all, built-up areas (class 2) (IoU = 0.790), thanks to the consideration of temporal dynamics, which helps to better distinguish between stable and changing patterns. However, the Water class remains difficult (IoU = 0.524), suggesting a need for specific processing for liquid surfaces (e.g., NDWI or dedicated water mask).

Ultimately, the U-Net model delivered limited performance, especially on 0 and 2 (water and buildings, relatively speaking). On the test set, U-Net achieved an average IoU of 0.394 and an average f1-score of 0.549, indicating very modest performance. The Water class (0), which is in the minority, is particularly poorly predicted (IoU = 0.237, f1 = 0.383), probably due to confusion with vegetation or built-up areas with similar spectral signatures (reflections, humidity, etc.). Built-up areas (class 2) are better detected (IoU = 0.590), but remain suboptimal, revealing the limitations of U-Net in modeling the spatial and temporal complexity of urban surfaces.

The results of our hybrid model, when compared together individual one, demonstrate that purely spatial models such as U-Net are insufficient to capture LULC dynamics. The ConvLSTM model significantly improves the temporal modeling of transitions, particularly for vegetation and urban areas. The ConvLSTM+U-Net hybrid model combines the best of both worlds, with high, robust, and balanced performance across all classes, including water (class 0), which is often neglected in conventional approaches. The results of our work highlight the relevance of hybrid spatio-temporal architectures for accurate and robust land use/land cover (LULC) prediction from multi-temporal satellite images.

4.4. Qualitative Analysis of Predictions

Figures 8 and A1 show randomly selected image samples from the validation and test sets, respectively, superimposed on the images predicted by our model after training. The visualizations also compare predicted maps and actual field data. The columns from left to right represent: RGB images (satellite view), ground truth, and corresponding model predictions. The rows represent the U-Net, ConvLSTM, and ConvLSTM+U-Net models. The colors represent the classes: black for water (class 0), green for vegetation (class 1), and blue for built-up areas (class 2).

According to these results, the U-Net model, based solely on spatial convolutions, shows very noisy predictions. Vegetation areas are fragmented, built-up areas are under-segmented, and water surfaces are largely ignored. The lack of temporal context limits its ability to distinguish stable structures (buildings, waterways) from seasonal changes (vegetation). The ConvLSTM model, on the other hand, improves the consistency of vegetated areas by incorporating the temporal dimension. However, it struggles to correctly identify water areas, which are often confused with built-up areas. This limitation can be explained by the absence of an effective spatial encoding structure. Finally, the ConvLSTM-U-Net

model effectively combines the hierarchical spatial extraction of U-Net and the temporal modeling of ConvLSTM. As a result, it produces a cleaned segmentation that is much more faithful to actual shapes and well separated between classes. Water bodies are also correctly detected, vegetation masses are well delimited, and urban areas are accurately rendered.

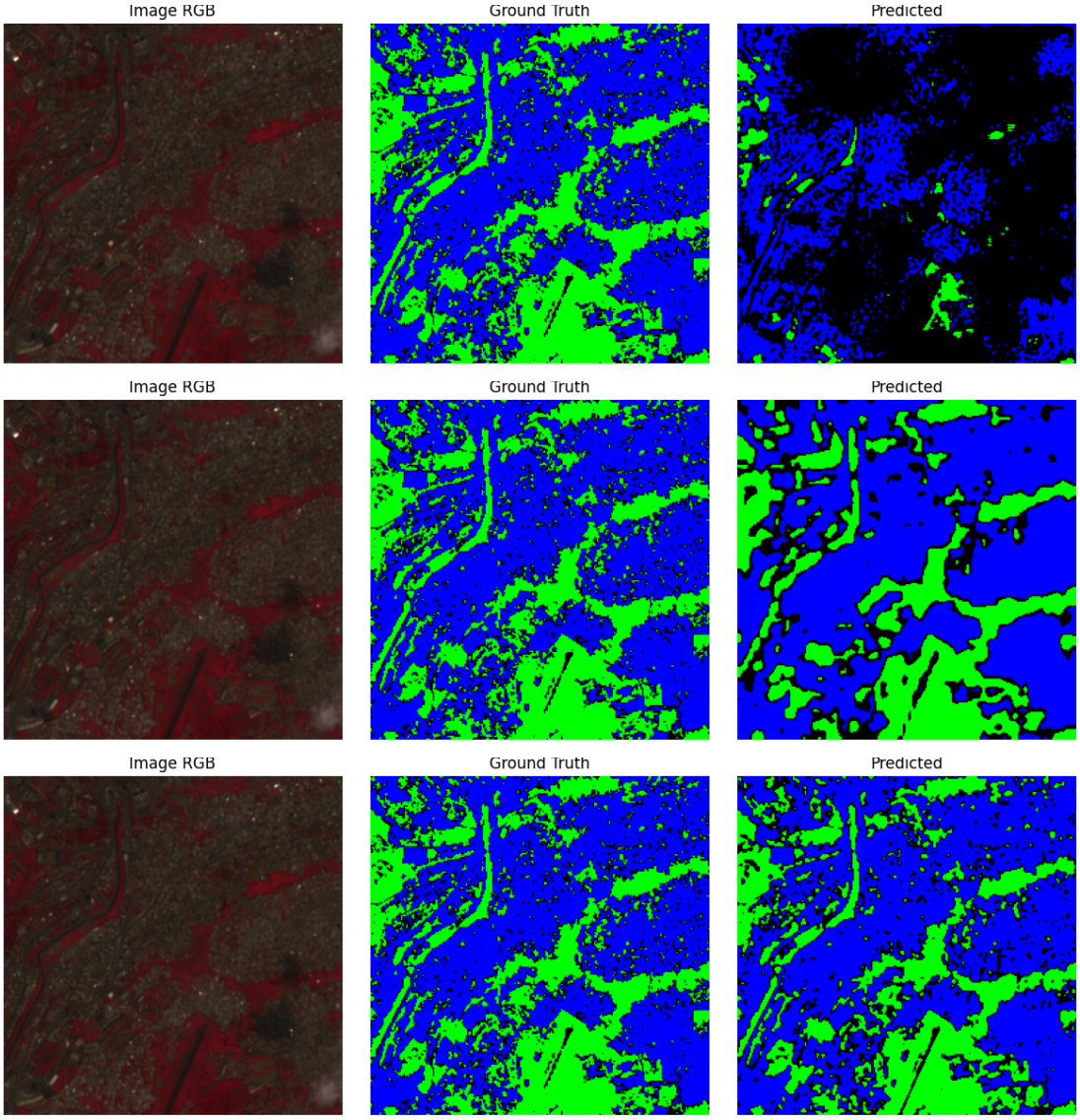


Figure 8. Visualization of a Sentinel-2 image extract overlaid on the corresponding image after validation (from validation set). The image was randomly selected. In the ground truth and predicted maps, color black corresponds to water, green to vegetation, and blue to built-up area. The RGB image is a Sentinel-2 B4-B3-B2 composition used for visual reference. The top line (1st line) shows the prediction of the U-Net model, the middle line (2nd line) shows the prediction of the ConvLSTM model, and finally the bottom line (3rd line) shows the prediction of the hybrid Unet-ConvLSTM model.

This qualitative analysis of the visual results corroborates the quantitative performance previously reported (see Table 4) and highlights the importance of a spatio-temporal architecture for robust and consistent prediction of land cover from multi-temporal satellite images.

While these qualitative results highlight improved spatial coherence and reduced noise in the proposed hybrid model, a quantitative assessment of temporal consistency across successive predictions remains outside the scope of the present study and is identified as a perspective for future evaluation.

4.5. Study of Model Complexity

The images in Figure 9 provide indicators on the complexity parameters of the different models studied. The complexity of the models is studied here in terms of resource consumption and computation time.

```
========== Model UNet Training Summary ==========  ========== Model ConvLSTM Training Summary ==========
Total Duration        : 3:33:03                     Total Duration        : 6:05:17
RAM Used (GB)         : 2.34 (initial: 1.81)        RAM Used (GB)         : 1.74 (initial: 1.48)
GPU Used (GB)         : 1.06 (initial: 0.26)        GPU Used (GB)         : 0.08 (initial: 0.00)
Last F1 Score         : 0.6035                      Last F1 Score         : 0.7809
Last mIoU             : 0.4431                      Last mIoU             : 0.6475
Last Validation Loss  : 0.7220                      Last Validation Loss  : 0.4546
Number of Epochs      : 59                          Number of Epochs      : 100
Validation Inference Time : 54.41 sec               Validation Inference Time : 53.24 sec
================================================  ================================================

          ========== Model Hybrid Training Summary ==========
          Total Duration        : 6:02:44
          RAM Used (GB)         : 1.83 (initial: 1.46)
          GPU Used (GB)         : 1.37 (initial: 0.33)
          Last F1 Score         : 0.8864
          Last mIoU             : 0.7997
          Last Validation Loss  : 0.3143
          Number of Epochs      : 100
          Validation Inference Time : 53.94 sec
          ================================================
```

Figure 9. Model training summery.

The three figures show that the total training time increases from 3 h 33 min (UNet) to 6 h 05 min for the Conv LSTM model and 6 h 02 min for the hybrid model, representing an almost twofold increase in time cost. The RAM used is actually slightly lower for the hybrid model (1.83 GB) than for U-Net (2.34 GB) and remains very slightly above that of ConvLSTM (1.83 GB). However, GPU usage increases slightly for the hybrid model (1.37 GB compared to 1.06 GB and 0.08 GB for UNet and ConvLSTM, respectively), which is consistent with a more complex model or a richer pipeline on the GPU side. The inference time on validation and testing remains almost identical (around 54 s), which means that, despite its better performance, the hybrid does not significantly penalize evaluation speed.

The figures show and confirm that the ConvLSTM+U-Net hybrid model provides a substantial gain in precision (F1 and mIoU) for a reasonable additional training cost and without a major impact on inference speed.

5. Discussion: Relevance for Urban Planning, Limitations, and Future Directions

The experimental results obtained in this study demonstrate the relevance of the proposed hybrid approach (ConvLSTM-UNet) for land cover classification based on multitemporal Sentinel-2 satellite images. The integration of ConvLSTM layers allows for better exploitation of the temporal dimension of image series, which significantly improves performance on classes subject to seasonal (e.g., vegetation) or spectral (built-up areas confused with certain soil types) variations.

The comparative analysis shows that the U-Net model, although spatially effective, suffers from a lack of temporal sensitivity. This results in increased confusion between vegetation and built-up areas, particularly in expanding peri-urban areas. The ConvLSTM model alone, although incorporating the temporal dimension, is limited in its ability to finely reconstruct spatial structures, resulting in blurred contours and localized loss of accuracy. In contrast, the ConvLSTM-UNet hybrid model achieved the best overall performance with an average F1-score of 0.889, an accuracy of 0.879, and an average IoU of 0.797, significantly outperforming the other two configurations. The "built" class (class 2), which is often critical in the context of urban sprawl analysis, benefits particularly from this architecture,

with an IoU reaching 0.868. These results reflect the complementarity of the ConvLSTM modules for temporal reasoning and the U-shaped structure for spatial localization.

A qualitative visualization of the predictions shows that the hybrid manages to preserve complex urban forms while more effectively detecting vegetated areas. The "water" class remains more difficult to predict, often confused with shadows or certain dark surfaces (IoU 0.692), which is an area for improvement.

Thus, the recent advances in remote sensing foundation models open promising perspectives for future extensions of this work. In particular, RSFMs could be leveraged as pre-trained feature extractors, enabling transfer learning or knowledge distillation toward lightweight spatio-temporal architectures such as ConvLSTM-based networks. This hybrid strategy could combine the representation power of large pre-trained models with the explicit temporal modeling and operational efficiency of recurrent convolutional networks. Future research may explore the integration of foundation models for initializing spatial encoders, self-supervised temporal pre-training, or weakly supervised urban fabric characterization, especially in contexts where detailed ground truth is unavailable. Such directions would allow bridging the gap between large-scale generalist models and locally optimized, task-specific spatio-temporal approaches.

While this study focuses on three dominant land cover classes (water, vegetation, and built-up areas), we acknowledge that urban environments are inherently heterogeneous and composed of diverse urban fabric typologies, such as dense urban cores, low-density residential areas, informal settlements, and industrial zones. The restriction to three classes was motivated by the availability of reliable, temporally consistent ground truth and by the need to limit label noise in a supervised learning setting, particularly in rapidly changing peri-urban contexts. Recent advances in remote sensing and deep learning suggest that finer-grained urban classification can be addressed even in the absence of detailed reference data. These representations can subsequently be explored through clustering or representation learning techniques to identify urban sub-structures. Another promising direction involves knowledge distillation from global land cover products, which can provide weak semantic supervision to guide local refinement at higher spatial or thematic resolution.

Limitations and Perspectives

Although the proposed framework demonstrates strong performance in terms of classical segmentation metrics such as accuracy, F1-score, and Intersection-over-Union, these indicators alone do not fully capture all aspects of multi-temporal land use and land cover segmentation. In particular, the present study focuses on short-term temporal coherence using fixed-length sequences (T = 5) and does not explicitly evaluate long-term temporal generalization across years. While seasonal variability is partially addressed through multi-year sampling of Sentinel-2 imagery, robustness to cross-season and cross-year generalization has not been quantitatively assessed. Furthermore, spatial consistency across time is analyzed qualitatively through visual inspection of successive predictions (Figures 8 and 9), but no dedicated temporal smoothness or stability metrics are currently employed. The evaluation of temporal generalization, seasonal robustness, and spatio-temporal consistency using dedicated indicators constitutes an important direction for future work.

Another important limitation of the proposed approach concerns model size and scalability. The hybrid ConvLSTM–U-Net architecture, while effective in capturing spatio-temporal dynamics, involves a relatively large number of parameters and recurrent convolutional operations, which increase GPU memory consumption and training time. Although inference time remains acceptable at the patch level, deploying such a model over

large geographic extents or at national scale may pose practical challenges, particularly in resource-constrained operational environments. Addressing these scalability issues requires strategies such as model compression, parameter pruning, or knowledge distillation to reduce computational complexity while preserving predictive performance. In this regard, the present work constitutes a high-capacity reference model, and the development of lightweight distilled variants suitable for large-scale or near-real-time applications is identified as a primary direction for future research.

Another limitation of the present study concerns the semantic granularity of the land use and land cover classification. Although Sentinel-2 imagery supports finer LULC differentiation, the classification task was intentionally restricted to three broad classes (built-up areas, vegetation, and water bodies). This choice was motivated by the limited availability of reliable and temporally consistent ground-truth annotations in the study area, as well as the need to reduce label noise in a supervised multi-temporal learning setting. Furthermore, all experiments were conducted under constrained computational resources using Kaggle-based GPU environments, where memory and training time limitations impose practical constraints on model complexity and class granularity, particularly for recurrent spatio-temporal architectures. It is important to note that the proposed framework is class-agnostic and can be naturally extended to finer LULC taxonomies. Future work will therefore investigate richer urban subclasses using weak supervision from global land cover products (e.g., Dynamic World, WorldCover), as well as self-supervised and distillation-based strategies to enable scalable multi-class learning.

A further limitation of the present study is the absence of a comprehensive ablation analysis to quantify the sensitivity of the proposed framework to key design choices, including the sequence length, the depth of the ConvLSTM module, the decoder configuration, patch size, and parameter-sharing strategies across time steps. These design parameters were selected based on prior work and empirical considerations to balance temporal context modeling, spatial resolution, and computational feasibility. In particular, the choice of a fixed sequence length T = 5 reflects a trade-off between capturing short-term temporal dynamics and limiting noise, label inconsistency, and computational cost. A systematic ablation study exploring these design dimensions would provide valuable insight into model robustness and optimal configuration, and is identified as an important direction for future work.

Although our study shows very promising prospects for hybrid models, some limitations are worth mentioning: First, there is class imbalance because minority categories (class 0 cases) suffer from low recall, reducing their usefulness in detailed monitoring scenarios (urban green spaces, linear infrastructure). Second, spectral confusion, which is the source of errors between classes 0 and 2, suggests that certain spectral signatures remain ambiguous (e.g., bare soil vs. light-colored buildings). Third and finally, the loss of spatial finesse, marked by the standardization of predictions, reduces the ability to detect small urban entities, which are crucial for detailed planning.

To strengthen the operational applicability of this type of model in urban planning, it would be relevant to explore the following in our future work: (i) increasing minority classes (through oversampling or adapted loss, e.g., focal loss), (ii) architectures that integrate spatio-temporal attention to better capture local details, and (iii) refinement post-processing (e.g., random conditional fields, morphological filtering) to improve object delineation.

6. Conclusions

The objective of this study was to build a hybrid deep learning model capable of capturing both spatial and temporal trends in land use and land cover in urban areas. The study highlights the effectiveness of a hybrid ConvLSTM+UNet model for predict-

ing land cover from multi-temporal Sentinel-2 images. The joint integration of spatial and temporal dimensions significantly improves classification accuracy, particularly in dynamic urban contexts. The performance obtained on the spatio-temporal data from the Yaoundé study area between 2018 and 2025 confirms that this approach is robust in the face of spectral heterogeneities and seasonal dynamics. The results obtained for the Yaoundé region, with a macro-average F1-score of 0.893 and an IoU of 0.811 across all classes, confirm the robustness and relevance of the model for distinguishing urban areas, vegetation, and water bodies. Moreover, despite its better performance, the figures confirm that the hybrid does not significantly penalize evaluation speed. These results pave the way for concrete applications in urbanization monitoring, urban planning, and environmental management. Prospects for improvement include extending to more LULC classes, integrating spectral indices (NDVI, NDBI) as additional channels, and exploring advanced techniques such as spatio-temporal attention or knowledge distillation for lighter and more efficient deployments.

Author Contributions: Conceptualization, A.G.B. and S.C.T.K.; methodology, A.G.B. and S.C.T.K.; software, A.G.B. and S.C.T.K.; validation, A.G.B. and S.C.T.K.; formal analysis, A.G.B. and S.C.T.K.; investigation, A.G.B. and S.C.T.K.; resources, A.G.B., M.E.H. and S.C.T.K.; data curation, A.G.B. and S.C.T.K.; writing—original draft preparation, A.G.B. and S.C.T.K.; writing—review and editing, A.G.B., M.E.H. and S.C.T.K.; visualization, A.G.B. and S.C.T.K.; supervision, M.E.H. and S.C.T.K.; project administration, A.G.B., M.E.H. and S.C.T.K.; funding acquisition, A.G.B. and S.C.T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: To all those who contributed directly or indirectly to the success of this study.

Conflicts of Interest: There are no conflicts of interest regarding the publication of this paper.

Abbreviations

The following abbreviations are used in this manuscript:

LULC	Land Use Land Cover
LSTM	Long Short Term Memory
ConvLSTM	Convolutional Long Short Term Memory
CNN	Convolutional Neural Network

Appendix A

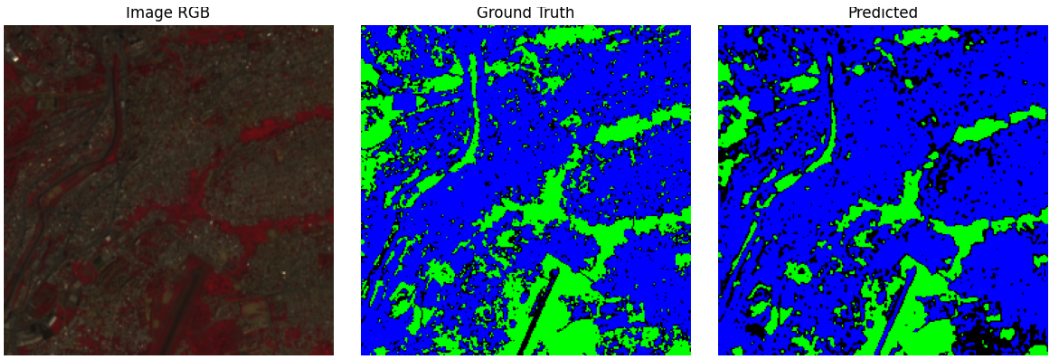


Figure A1. *Cont.*

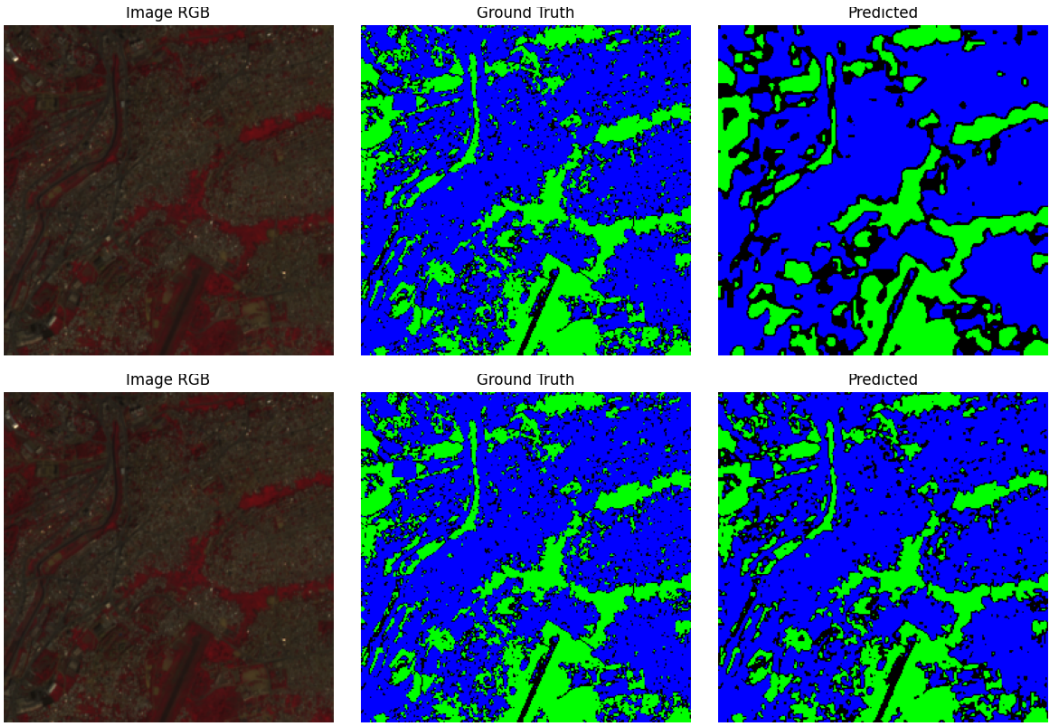


Figure A1. Visualization of a Sentinel-2 image extract overlaid on the corresponding image after validation (from validation set). The image was randomly selected. In the ground truth and predicted maps, the color black corresponds to water, green to vegetation, and blue to built-up area. The RGB image is a Sentinel-2 B4-B3-B2 composition used for visual reference.

References

1. Belinga, A.G.; El Haziti, M. Overviewing the emerging methods for predicting urban sprawl features. *E3S Web of Conf.* **2023**, *418*, 03008. [CrossRef]
2. Belinga, A.G.; Koumetio, S.C.T.; El Haziti, M. Exploring the potentialities and challenges of deep learning for simulation and prediction of urban sprawl features. *Data Policy* **2025**, *7*, e2. [CrossRef]
3. Tekouabou, S.C.K.; Diop, E.B.; Azmi, R.; Jaligot, R.; Chenal, J. Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges. *J. King Saud-Univ.-Comput. Inf. Sci.* **2022**, *34*, 5943–5967. [CrossRef]
4. Tékouabou, S.C.; Chenal, J.; Azmi, R.; Toulni, H.; Diop, E.B.; Nikiforova, A. Identifying and classifying urban data sources for machine learning-based sustainable urban planning and decision support systems development. *Data* **2022**, *7*, 170. [CrossRef]
5. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 data for land cover/use mapping: A review. *Remote Sens.* **2020**, *12*, 2291. [CrossRef]
6. Kamda Silapeux, A.; Ponka, R.; Frazzoli, C.; Fokou, E. Waste of fresh fruits in Yaoundé, Cameroon: Challenges for retailers and impacts on consumer health. *Agriculture* **2021**, *11*, 89. [CrossRef]
7. Masolele, R.N.; De Sy, V.; Marcos, D.; Verbesselt, J.; Gieseke, F.; Mulatu, K.A.; Moges, Y.; Sebrala, H.; Martius, C.; Herold, M. Using high-resolution imagery and deep learning to classify land-use following deforestation: A case study in Ethiopia. *GISci. Remote Sens.* **2022**, *59*, 1446–1472. [CrossRef]
8. Gallwey, J.; Robiati, C.; Coggan, J.; Vogt, D.; Eyre, M. A Sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in Ghana: Applying deep learning to shallow mining. *Remote Sens. Environ.* **2020**, *248*, 111970. [CrossRef]
9. Pelletier, C.; Webb, G.I.; Petitjean, F. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sens.* **2019**, *11*, 523. [CrossRef]
10. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1.
11. Rußwurm, M.; Körner, M. Convolutional LSTMs for cloud-robust segmentation of remote sensing imagery. *arXiv* **2018**, arXiv:1811.02471. [CrossRef]
12. Rußwurm, M.; Körner, M. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 129. [CrossRef]

13. Li, Z.; Shen, H.; Cheng, Q.; Liu, Y.; You, S.; He, Z. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 197–212. [CrossRef]
14. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
15. Zhang, H.; Li, Y.; Zhang, Y.; Shen, Q. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote Sens. Lett.* **2017**, *8*, 438–447. [CrossRef]
16. Sefrin, O.; Riese, F.M.; Keller, S. Deep learning for land cover change detection. *Remote Sens.* **2020**, *13*, 78. [CrossRef]
17. Arrechea-Castillo, D.A.; Solano-Correa, Y.T.; Muñoz-Ordóñez, J.F.; Pencue-Fierro, E.L.; Figueroa-Casas, A. Multiclass land use and land cover classification of Andean Sub-Basins in Colombia with Sentinel-2 and Deep Learning. *Remote Sens.* **2023**, *15*, 2521. [CrossRef]
18. Zhang, G.; Roslan, S.N.A.B.; Wang, C.; Quan, L. Research on land cover classification of multi-source remote sensing data based on improved U-net network. *Sci. Rep.* **2023**, *13*, 16275. [CrossRef] [PubMed]
19. Liu, S.; Cao, S.; Lu, X.; Peng, J.; Ping, L.; Fan, X.; Teng, F.; Liu, X. Lightweight Deep Learning Model, ConvNeXt-U: An Improved U-Net Network for Extracting Cropland in Complex Landscapes from Gaofen-2 Images. *Sensors* **2025**, *25*, 261. [CrossRef]
20. Xu, X. Multi-temporal Land Cover Segmentation via Trans-ConvLSTM. In *Proceedings of the International Conference on Big Data Analytics for Cyber-Physical System in Smart City*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 422–430. [CrossRef]
21. Wenger, R.; Puissant, A.; Weber, J.; Idoumghar, L.; Forestier, G. Multimodal and multitemporal land use/land cover semantic segmentation on sentinel-1 and sentinel-2 imagery: An application on a multisenge dataset. *Remote Sens.* **2022**, *15*, 151. [CrossRef]
22. Majidizadeh, A.; Hasani, H.; Jafari, M. Semantic segmentation of oblique UAV video based on ConvLSTM in complex urban area. *Earth Sci. Inform.* **2024**, *17*, 3413–3435. [CrossRef]
23. Yele, V.P.; Badhe, N.B.; Alegavi, S.; Sedamkar, R. Multi Attention Convolutional Sparse Coding U-Net for Enhanced Land-Use and Land-Cover Segmentation Using Hyperspectral Images. *Sens. Imaging* **2025**, *26*, 69. [CrossRef]
24. Buttar, P.K.; Sachan, M.K. Land Cover Segmentation Using 3-D FCN-Based Architecture With Coordinate Attention. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502905. [CrossRef]
25. Li, R.; Zheng, S.; Duan, C.; Wang, L.; Zhang, C. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-Spat. Inf. Sci.* **2022**, *25*, 278–294. [CrossRef]
26. Alam, A.; Bhat, M.S.; Maheen, M. Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley. *GeoJournal* **2020**, *85*, 1529–1543. [CrossRef]
27. Usman, M.; Liedl, R.; Shahid, M.; Abbas, A. Land use/land cover classification and its change detection using multi-temporal MODIS NDVI data. *J. Geogr. Sci.* **2015**, *25*, 1479–1506. [CrossRef]
28. Brown, C.F.; Brumby, S.P.; Guzder-Williams, B.; Birch, T.; Hyde, S.B.; Mazzariello, J.; Czerwinski, W.; Pasquarella, V.J.; Haertel, R.; Ilyushchenko, S.; et al. Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci. Data* **2022**, *9*, 251. [CrossRef]
29. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; et al. *ESA WorldCover 10 m 2021 v200*; Zenodo: Geneva, Switzerland, 2022. [CrossRef]
30. Karra, K.; Kontgis, C.; Statman-Weil, Z.; Mazzariello, J.C.; Mathis, M.; Brumby, S.P. Global land use/land cover with Sentinel 2 and deep learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 12–16 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4704–4707.
31. Mahmud, B.U.; Hong, G.Y.; Mamun, A.A.; Ping, E.P.; Wu, Q. Deep learning-based segmentation of 3D volumetric image and microstructural analysis. *Sensors* **2023**, *23*, 2640. [CrossRef]
32. Lu, S.; Guo, J.; Zimmer-Dauphinee, J.R.; Nieusma, J.M.; Wang, X.; vanValkenburgh, P.; Wernke, S.A.; Huo, Y. Vision foundation models in remote sensing: A survey. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 190–215. [CrossRef]
33. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 16–22 June 2024; pp. 27672–27683.
34. Li, Y.; Tan, J.; Dang, B.; Ye, M.; Bartalev, S.A.; Shinkarenko, S.; Wang, L.; Zhang, Y.; Ru, L.; Guo, X.; et al. Unleashing the potential of remote sensing foundation models via bridging data and computility islands. *Innovation* **2025**, *6*, 100841. [CrossRef]
35. Wu, K.; Zhang, Y.; Ru, L.; Dang, B.; Lao, J.; Yu, L.; Luo, J.; Zhu, Z.; Sun, Y.; Zhang, J.; et al. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. *Nat. Mach. Intell.* **2025**, *7*, 1235–1249. [CrossRef]
36. Zhu, Q.; Lao, J.; Ji, D.; Luo, J.; Wu, K.; Zhang, Y.; Ru, L.; Wang, J.; Chen, J.; Yang, M.; et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 14733–14744.
37. Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv* **2024**, arXiv:2406.10100.

38. Chen, Z.; Zhao, S. Automatic monitoring of surface water dynamics using Sentinel-1 and Sentinel-2 data with Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 103010. [CrossRef]
39. Nasiri, V.; Deljouei, A.; Moradi, F.; Sadeghi, S.M.M.; Borz, S.A. Land use and land cover mapping using Sentinel-2, Landsat-8 Satellite Images, and Google Earth Engine: A comparison of two composition methods. *Remote Sens.* **2022**, *14*, 1977. [CrossRef]
40. Azad, R.; Asadi-Aghbolaghi, M.; Fathy, M.; Escalera, S. Bi-directional ConvLSTM U-Net with densley connected convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
41. Gillies, S. *Rasterio Documentation*; MapBox: San Francisco, CA, USA, 2019; Volume 23.
42. Kramer, O. Scikit-learn. In *Machine Learning for Evolution Strategies*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 45–53. [CrossRef]
43. Bisong, E. Matplotlib and seaborn. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 151–165. [CrossRef]
44. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *Advances in Visual Computing, Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 12–14 December 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 234–244. [CrossRef]
45. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized interSection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 658–666.
46. Zhou, D.; Fang, J.; Song, X.; Guan, C.; Yin, J.; Dai, Y.; Yang, R. IoU loss for 2D/3D object detection. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Québec City, QC, Canada, 15–18 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 85–94. [CrossRef]

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