Remote Sensing Applications: Society and Environment 40 (2025) 101764

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Change detection in Sentinel-2 images using deep learning ensembles

Ewa Kopec [a], Agata M. Wijata [b,c], Jakub Nalepa [a,c,*]

[a] Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland
[b] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
[c] KP Labs, Gliwice, Poland

ARTICLE INFO

Dataset link: https://rcdaudt.github.io/oscd/, https://zenodo.org/records/10674011

Keywords:
Multispectral image
Change detection
Machine learning
Fully convolutional neural network
Ensemble learning

ABSTRACT

The recent advancements in satellite imaging bring various possibilities in Earth observation in numerous domains, including the analysis of the evolution of urban areas, precision agriculture, environmental monitoring, event detection and tracking, and many more. Change detection plays a key role in a multitude of applications, as it allows for precisely monitoring the changes within an area of interest. In this article, we tackle this issue and introduce deep learning ensembles for change detection in Sentinel-2 times series of multispectral images—the proposed ensembles benefit from different deep learning model architectures. The experimental study performed over the widely-adopted benchmark datasets showed that the ensembles combine the strengths of the individual models, thus they reduce false positives and false negatives of base learners. The ensembles compensated the under-performing models, ultimately obtaining the change detection accuracy that exceeds 95% over the unseen test scenes.

1. Introduction

Earth observation (EO) through remote sensing images brings lots of opportunities in a variety of areas (Mertes et al., 2015), as images acquired in orbit intrinsically offer significant spatial scalability. To extract useful and actionable insights from satellite images, data-driven algorithms targeting different downstream tasks have been emerging at a steady pace. Change detection (CD) is one of the most important examples of tasks that play an important role in practical EO, as it allows for accurate tracking of the characteristics of a region of interest (Caye Daudt et al., 2018). By comparing images taken at different points in time, one can detect and quantify changes in land cover, land use, vegetation, infrastructure, and natural phenomena (Singh, 1989; Daudt, 2020). This process enables the identification of both subtle and significant modifications, providing valuable information for numerous applications such as environmental monitoring (Lee et al., 2020), urban planning (Mertes et al., 2015), agriculture (Segarra et al., 2020), disaster management (Hussain et al., 2013) and security (Fakhri and Gkanatsios, 2021). It also facilitates informed decision-making, supports sustainability initiatives, and contributes to a better understanding of the planet's dynamic processes (Basavaraju et al., 2022).

1.1. Real-world examples of change detection applications

The expansion of the satellite remote sensing technology has enabled the utilization of vast amounts of in-orbit data for CD, including multispectral images (MSIs) (Qiqi Zhu and Li, 2024), offering critical insights into various environmental issues, such as

* Corresponding author at: Department of Algorithmics and Software, Silesian University of Technology, Gliwice, Poland.
E-mail addresses: ewakope946@student.polsl.pl (E. Kopec), Agata.Wijata@polsl.pl (A.M. Wijata), Jakub.Nalepa@polsl.pl (J. Nalepa).

https://doi.org/10.1016/j.rsase.2025.101764
Received 12 February 2025; Received in revised form 22 September 2025; Accepted 16 October 2025
Available online 28 October 2025

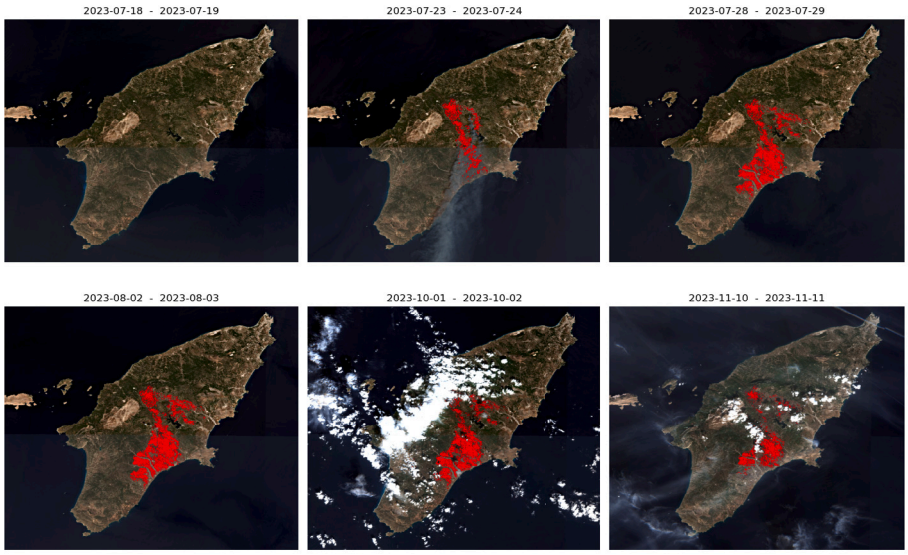

Fig. 1. A time series of Sentinel-2 images showing wildfires in Rhodes, Greece 2023. The red color indicates burning areas in the consecutive time points within this time series.

land use, vegetation cover, water quality, and urbanization (Mertes et al., 2015). Here, urbanization is a key driver of the global environmental change, but its current understanding is largely limited to local sites. Remote sensing provides a means to study urban expansion over larger areas, uncovering critical spatial and temporal growth patterns influenced by socioeconomic and political factors (Mertes et al., 2015) that are indeed vital for assessing the social, environmental, and climatic impact of urbanization. However, mapping urban expansion on a global scale is challenging due to the small and variable footprint of urban areas, their heterogeneous composition, and the diverse forms and rates of development (Mertes et al., 2015).

There are many other real-world Earth observation use cases that can directly benefit from CD analysis. In de Bem et al. (2020), the authors focused on the problem of tracking deforestation in the Brazilian Amazon using deep neural networks operating on the Landsat data. It may help us to better understand both legal and illegal deforestation activities in the region of interest, and to ultimately respond to such actions. This is of paramount environmental importance, as deforestation has a direct impact on the climate change through greenhouse gas emissions. Similarly, Kuck et al. (2021) presented an approach toward CD of selective logging in the Brazilian Amazon using synthetic aperture radar (SAR) data, and further highlighted the importance of this task, emphasizing that the forest degradation substantially contributes to the loss of gross above-ground biomass. The SAR data was also targeted in the CD task in Silva et al. (2022), where a system for identifying clear-cut deforested areas was introduced. Alshehri et al. (2023), on the other hand, adopted a transformer-based CD model, to detect deforestation in the Brazilian Amazon from Sentinel-2 imagery. Overall, deforestation — as one of the major concerns with regard to climate change and the maintenance of biodiversity — has been thoroughly investigated in the context of practical CD EO tasks (Bragagnolo et al., 2021). Apart from deforestation tracking, there are other downstream environmental tasks which might benefit from precise CD based on remote sensing data. In Zhao et al. (2024), Zhao et al. introduced an adaptive glacier extraction index to investigate the spatiotemporal patterns of the glaciers' retreat on the Tibetan Plateau, emphasizing that this phenomenon impacts the regional water cycle and the occurrence of natural hazards. Other prominent examples of CD tasks (You et al., 2020) include sea ice concentration monitoring (Stofa et al., 2025), continuous monitoring and sub-annual CD in high-latitude forests (Mulverhill et al., 2023), land use land cover applications (Chughtai et al., 2021; He et al., 2024), estimating spatiotemporal crop field variations (Dahiya et al., 2024), and many others (Bai et al., 2023; Afaq and Manocha, 2021). Fig. 1 renders an example use case of exploiting a Sentinel-2 MSI time series in the context of change detection for security and environmental monitoring (wildfires outbreaks happened in Rhodes, Greece in 2023). All of the above-mentioned downstream tasks clearly demonstrate the importance of practical CD in EO scenarios, underscoring the societal and environmental impact of CD—precise CD may lead to designing and implementing appropriate actions in response to observed changes in the area of interest in a scalable and fast way (Asokan and Anitha, 2019) which is reproducible and free from any human bias concerned with manual analysis of image time series.

1.2. Related literature

CD involves analyzing multi-temporal images of a geographic area captured at distinct time points to identify changes (Qiqi Zhu and Li, 2024). Indeed, various techniques have been developed to meet the specific EO requirements for this task (Hussain et al., 2013; Jeevan and Shanthi, 2024; Afaq and Manocha, 2021). Significant efforts have been dedicated to the advancement of different

methodologies, ranging from traditional pixel-based approaches (Singh, 1989) to object-based methods (Hussain et al., 2013; Liu et al., 2021). With the rise of machine learning (ML), it is now possible to apply data-driven algorithms that can automatically learn CD processes, offering greater accuracy and flexibility compared to conventional methods. Information about changes obtained from series of EO images can be categorized into two levels: (*i*) identifying simple binary change (i.e., *change* vs. *no change*), and (*ii*) determining detailed "from-to" changes (e.g., post-classification comparison) (Hussain et al., 2013).

One of the most common approaches to CD is pixel-based change detection (PBCD), which focuses on analyzing individual pixels in an image to identify changes (Hussain et al., 2013). This method is relatively straightforward and can be implemented using various techniques (Hussain et al., 2013). Traditional CD algorithms are typically effective for low- to medium-resolution imagery but struggle with high-resolution images. In object-based change detection (OBCD), changes are identified at the object level, rather than at the pixel level (Hussain et al., 2013; Gang Chen and Wulder, 2012). Considering the spatial and contextual information of objects, object-based CD can provide more meaningful results (Gang Chen and Wulder, 2012). The object-based approach has been shown to outperform pixel-based methods in land cover CD (Zhou et al., 2008). In addition to pixel-based and object-based approaches, CD can be performed using threshold-based methods. They involve setting a threshold value to determine whether a change has occurred or not (Singh, 1989).

CD methods exploiting ML can be divided into two high-level groups, including supervised and unsupervised techniques (Hussain et al., 2013). Supervised CD methods build upon the labeled ground-truth data (Hussain et al., 2013), and such algorithms learn from training data to classify *change* vs. *no-change* areas (Hussain et al., 2013; Somvanshi et al., 2016). Unsupervised CD methods, on the other hand, do not require labeled data and rely on the analysis of the inherent features and changes in the images themselves (Ghosh et al., 2011), e.g., through unsupervised clustering (Bovolo and Bruzzone, 2007). Overall, the choice between supervised and unsupervised CD algorithms depends on the availability of the labeled ground truth data, the desired level of detail in the change information, and the specific requirements of the downstream application. Supervised methods often provide high accuracy and specificity, but they require extensive labeled datasets, which are costly and time-consuming to create (Daudt, 2020). Unsupervised methods — as they do not need ground truth labels — are more flexible and easier to apply to new datasets. Additionally, they are useful for exploratory analysis, but they may lack precision in detecting subtle changes within time series data (Daudt, 2020).

There are many works benefiting from recent deep learning advancements (Chen et al., 2019; Pegia et al., 2022), especially convolutional neural networks (CNNs) or fully convolutional neural networks to find changes in MSIs (Wang et al., 2023; Song and Jiang, 2021; Pomente et al., 2018; Dahiya et al., 2024) (also trained in a self-supervised way (Leenstra et al., 2021)). Such techniques exploit automated representation learning—thus, they do not require designing feature extractors that would be used to elaborate representative features. This is in contrast to classic ML pipelines (Bhatt et al., 2015), in which manually-crafted feature extractors are commonly followed by feature selectors to reduce the dimensionality of the feature space—feature vectors are then used to train classic ML models (Chafiq et al., 2024). For CD, a particular emphasis is put on CNNs (Daudt et al., 2018) thanks to their exceptional ability to learn features from satellite imagery (Basavaraju et al., 2022). A very interesting approach was introduced in Camalan et al. (2022), where the authors combined CNNs with recurrent neural networks in their deep learning architecture targeting the problem of detecting and categorizing the changes in mining ponds created by the artisanal and small-scale gold mining activities. This technique showed that combining different architectural solutions in the deep learning-powered CD systems may help in benefiting from the advantages of such different models.

This work follows this research pathway, and not only designs new deep learning architectures for CD, but also builds heterogeneous deep learning ensembles (i.e., the ensembles including deep learning models of different architectures) for this task, hypothesizing that combining the strengths of fundamentally different models could lead to CD performance improvements. Of note, various deep learning models suffer from different shortcomings, e.g., related to handling complex scenarios like cloud cover or seasonal variations within the area of interest (de Albuquerque et al., 2021), or the lack of robustness against varying-quality input data, or the capabilities of the base models trained on imbalanced training sets (Khan et al., 2024). Therefore, we hypothesize that ensemble learners can improve the operational abilities of separate (base) artificial intelligence models (i.e., can lead to higher quality metrics obtained over the unseen test data, such as precision, recall, Dice scores, and so forth) by mutually addressing and compensating their weaknesses. This hypothesis is grounded directly in the very characteristics of the deep learning models—as an example, the spectral deep learning models (operating over the spectral dimension of each pixel only) do not incorporate any spatial information concerning the scene, which might be extracted and used by spectral-spatial deep learning architectures. It may, in turn, help deal with scenes obscured by clouds, as determining the cloudy areas (which are not "changes" within the image time series) may require analyzing the spatial context of particular objects. Although there have been attempts to exploit ensemble modeling for CD (Lim et al., 2018; Zhao et al., 2019; Gong et al., 2017), this area remains heavily under-researched (Peng et al., 2025), although the clear benefits of using ensemble learning in a variety of other data analysis tasks (Khan et al., 2024), including Earth observation (He et al., 2025). We address this scientific gap in this article, with the following objectives summarizing our study: (*i*) to introduce heterogeneous deep learning ensembles for CD, (*ii*) to verify whether combining the strengths of separate models could lead to improved CD performance, and to "compensating" the under-performing models, (*iii*) to ensure reproducibility of the experiments, and to (*iv*) verify the generalizability of the suggested techniques.

1.3. Contribution

We tackle the task of CD in satellite images, and our primary objective is to develop and thoroughly evaluate deep learning ensembles for CD in Sentinel-2 time series. Although building ensemble learners has shown potential in other domains (Yang et al., 2023; Kucharski et al., 2024; Ganaie et al., 2022) and downstream tasks in satellite data analysis (Iyer et al., 2021; Nalepa et al.,

2021b), it remains under-researched in CD. In this article, we hypothesize that benefiting from heterogeneous ensembles, containing (deep) ML models of various architectures can bring improvements in the overall CD capabilities of the system (reflected in the higher quality scores when compared with the base models included in the ensemble). Additionally, we hypothesize that building ensemble learners may not only enhance the capabilities of the CD framework, but can be used to compensate the worse individual models during the operation of the system (i.e., that the ensemble model, which incorporates several base models, would lead to notably better quality scores that the under-performing individual models, hence their low-quality performance would be "compensated" by the other models). This might be of importance in situations where extensive and representative test sets (ideally geographically scattered, heterogeneous and captured in various time points) are not available, and objectively understanding which individual (base) models are *best* is troublesome or even impossible.

Our contributions may be summarized by the following bullet points:

- We introduce **flexible CD ensembles** which build upon **heterogeneous architectures** of individual deep ML models for CD and aggregate their individual predictions using **weighted voting**, in which each base prediction is assigned a weight (i.e., the predictions with higher weights should affect the final outcome more).
- We ensure **reproducibility of our study** by making the implementation publicly available at https://github.com/smile-research/DeepEnsembles4ChangeDetection.
- We validate the capabilities of the ensembles and individual models following the validation procedure developed for the open-source and widely-adopted **Onera Satellite Change Detection (OSCD) dataset** (Caye Daudt et al., 2019).
- We utilize the **MUlti-modal MUlti-class Change Detection (MUMUCD)** dataset (Serva et al., 2025) to assess the generalization capability of models trained on the OSCD dataset. This evaluation enables us to investigate how well pretrained models handle previously unseen changes in heterogeneous environments under varying temporal and geographical conditions.

Ensuring reproducibility is our immediate answer to the *reproducibility crisis* (Kapoor and Narayanan, 2023) that we are currently — as a community — facing in the ML research. We believe that our efforts will encourage other research groups to share their implementations and validation procedures to make the comparison across the emerging CD algorithms straightforward and unbiased.

1.4. The roadmap: Structure of this article

The remaining part of this article is structured as follows. The change detection benchmark dataset used in our study, and the algorithms for CD are presented in Section 2. The experimental results are reported and discussed in Section 3. Finally, Section 4 concludes the article.

2. Materials and methods

2.1. Datasets' description

2.1.1. The ONERA satellite change detection dataset

This study utilizes the ONERA Satellite Change Detection (OSCD) dataset. It provides a standardized set of labeled Sentinel-2 data targeting urban CD (Daudt, 2020; Caye Daudt et al., 2019; Daudt et al., 2018). OSCD captures 24 regions selected based on visible urban changes. Each region is ca. 500 × 500 pixels at 10 m resolution, encompassing various levels of urbanization in different countries and continents. For each region, two images were selected, and each image consists of 13 Sentinel-2 spectral bands, resulting in a total of 26 images per region. Additionally, the dataset specifies the training-test split (14 scenes for training, 10 for testing) which is followed in this study. The example images included in OSCD are shown in Fig. 2. Of note, the authors of this dataset made sure (manually) that the number of cloud-covered pixels is not too large in the included images. Pixel-level ground truth labels were manually created by visually comparing the true-color images in each pair of images. To further improve the quality of the dataset, the GEFolki toolkit (Brigot et al., 2016) was used to achieve more precise image co-registration than that provided by the Sentinel system itself (the older image was always used as the reference image, and the newer one is aligned to it).

2.1.2. The MUlti-modal MUlti-class Change Detection (MUMUCD) dataset

To evaluate the capabilities of the models in the change detection task, we also utilize the MUlti-modal MUlti-class Change Detection (MUMUCD) dataset (Serva et al., 2025). MUMUCD provides a globally distributed collection of 70 geo-referenced bi-temporal image pairs which were acquired between 2019 and 2024 using multiple satellite sensors, including Sentinel-1 (SAR), Sentinel-2 (multispectral images), and PRISMA (hyperspectral images)—for the detailed data harmonization and pre-processing process, see Serva et al. (2025). Each scene covers an area of 1536 × 1536 pixels at a spatial resolution of 10 m, and is curated to include various terrain types, such as urban, rural, forested, and desert regions (Serva et al., 2025). The dataset indeed allows for extracting numerous non-overlapping 128 × 128 image patches which are suitable for learning tasks. To ensure seasonal diversity, scenes were selected to span the full calendar year and represent a variety of global environmental contexts. In addition to the co-registered image data, MUMUCD also provides ancillary information, including the land cover, surface elevation, and binary change labels derived from the Dynamic World maps (Serva et al., 2025). This study uses Sentinel-2 images from MUMUCD as an evaluation resource to test models trained on OSCD, enabling a comparative assessment of model generalization across datasets with differing scene diversity. The example image pairs from MUMUCD are rendered in Fig. 3.

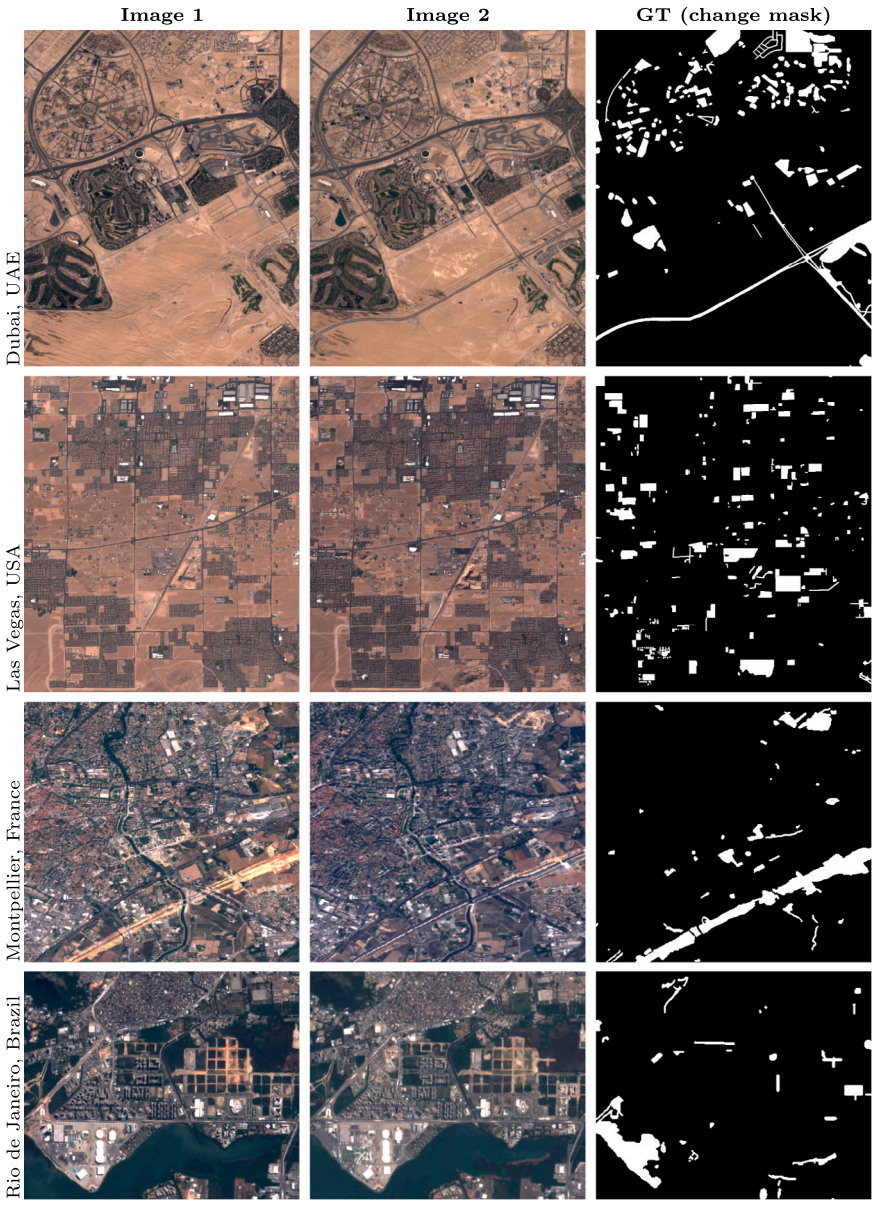

Fig. 2. The OSCD dataset—two images taken on different dates, together with the corresponding ground-truth (GT) change map (Caye Daudt et al., 2019).

2.2. Deep learning models for change detection

We use four deep learning models, and build upon the three fully-convolutional neural network architectures for CD introduced in Caye Daudt et al. (2018). The models are based on an U-Net architecture, with skip connections linking layers at the same sub-sampling scale before and after the encoding phase of an encoder–decoder architecture. This study builds upon U-Nets, as such architectures have indeed established the state of the art in a variety of fields (Isensee et al., 2021), and are commonly one of the first choices in semantic segmentation tasks, also in EO image analysis (Maurya et al., 2023; Suresh et al., 2024). This approach aims to enhance the abstract, less localized encoded information with the spatial details from the earlier network layers, resulting in accurate class predictions with precise boundaries in the output image (Caye Daudt et al., 2018). Three variants of the models are exploited:

- **Fully Convolutional Early Fusion (FC-EF)**—it includes four max-pooling and four up-sampling layers. The input for this network is a concatenation of the two images in the pair being compared (Caye Daudt et al., 2018; Daudt, 2020).

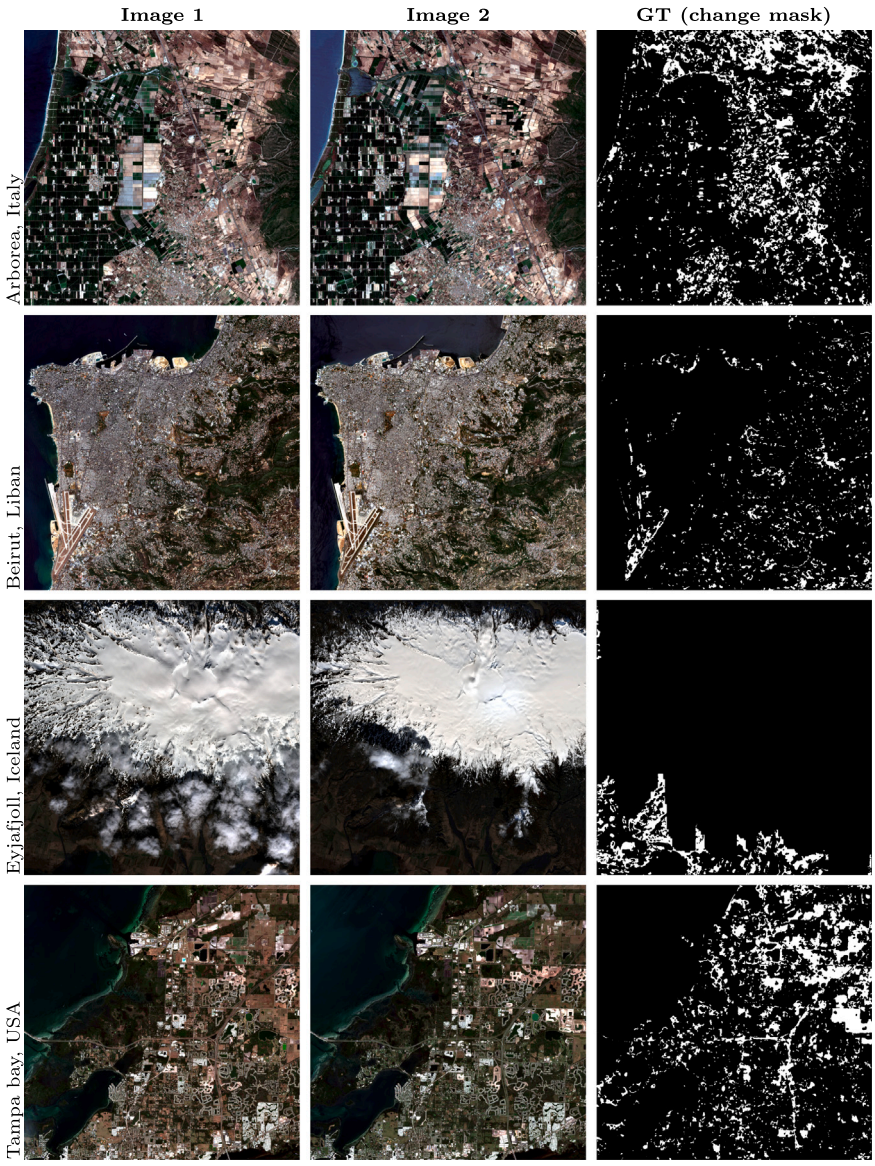

Fig. 3. The MUMUCD dataset—two images taken on different dates, together with the corresponding ground-truth (GT) change map (Serva et al., 2025).

- **Fully Convolutional Siamese-Concatenation (FC-Siam-conc)**—it concatenates the skip connections during the decoding steps, with each connection originating from one of the encoding streams. Given that the objective of change detection is to identify differences between the two images, an alternative heuristic was applied for combining the skip connections.
- **Fully Convolutional Siamese-Difference (FC-Siam-diff)**—instead of concatenating both connections from the encoding streams, the absolute value of their difference is concatenated.

It can be observed that FC-Siam-conc and FC-Siam-diff are the Siamese extensions of FC-EF, where the encoding layers are divided into two streams with identical structures and shared weights, as in a traditional Siamese network. For CD, each image in the pair is fed into one of these streams.

The fourth deep learning model, the urban CD network (UCDNet) introduced in Basavaraju et al. (2022) (UCDNet unfolds to the **u**rban **c**hange **d**etection **net**work), has been inspired by FC-Siam-diff and it adheres to its encoder–decoder structure (Fig. 4). This model has been selected, as it was showed that it offers very high quality CD. The encoder is split into two streams with identical architectures and shared weights (Fig. 5). Each input image is processed by one of these streams, consisting of convolutional and pooling layers, creating four stages per stream with three pooling layers. To identify changes between the two images, the difference in features is used as input for modified residual connections at each encoder stage. In these connections, the feature difference from

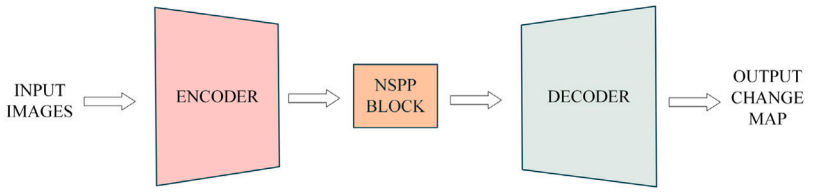

Fig. 4. The encoder–decoder architecture with the New Spatial Pyramid Pooling (NSPP) block in the urban change detection network (UCDNet).

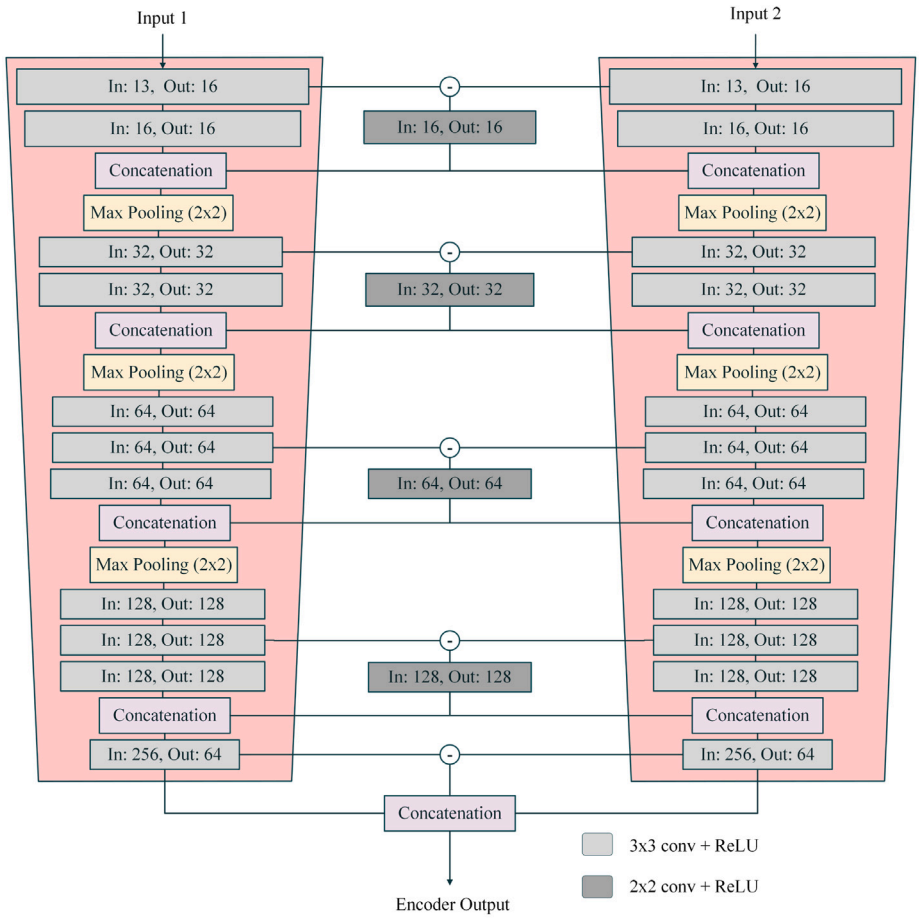

Fig. 5. The architecture of the urban change detection network (UCDNet) encoder.

the two streams is calculated after the first convolution, convolved point-wise, and then concatenated channel-wise with the output feature before passing it to the max pooling operation (Basavaraju et al., 2022).

In UCDNet, a New Spatial Pyramid Pooling (NSPP) block takes the features extracted by the encoder as input (Fig. 6a). It extracts features at various scales to enhance global context awareness, preserve the shape of the changed areas, and predict boundary pixels more accurately (Basavaraju et al., 2022). In contrast to the average pooling, batch normalization, rectified linear unit (ReLU) activation, and resizing operations in each parallel path of the Spatial Pyramid Pooling (SPP) block, the parallel paths of the NSPP consist of a pooling block and a global pooling block. The pooling blocks generate features at four different scales through four parallel paths. The utilization of global pooling aims to enhance the awareness of global information and mitigate degradation issues. By incorporating global context feature information, the performance of change detection is expected to improve, resulting in more accurate pixel classification (Basavaraju et al., 2022). To make the NSPP block applicable to OSCD, we slightly changed its architecture—three parallel paths with another kernel size in Pooling Block paths and one "clear" encoder output are implemented (Fig. 6b). Finally, the UCDNet decoder consists of three batch normalization and up-sampling layers, with the final layer of the model incorporating the Softmax activation function (Basavaraju et al., 2022) (Fig. 7).

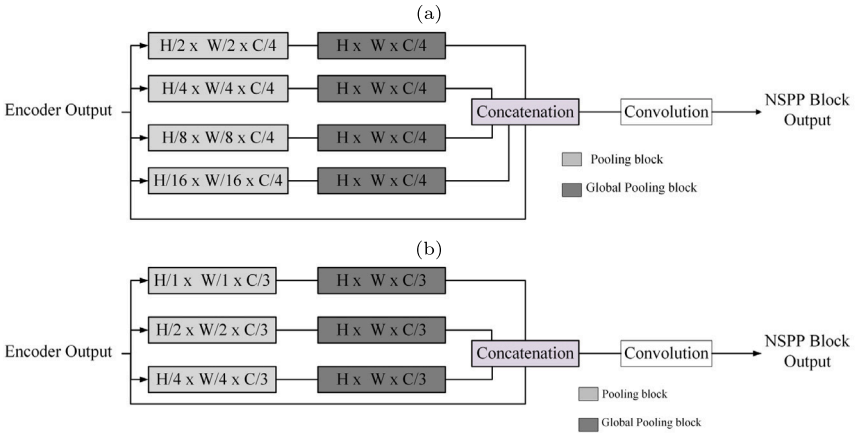

Fig. 6. Diagram of the (a) original NSPP (Basavaraju et al., 2022), and (b) NSPP implemented in this study.

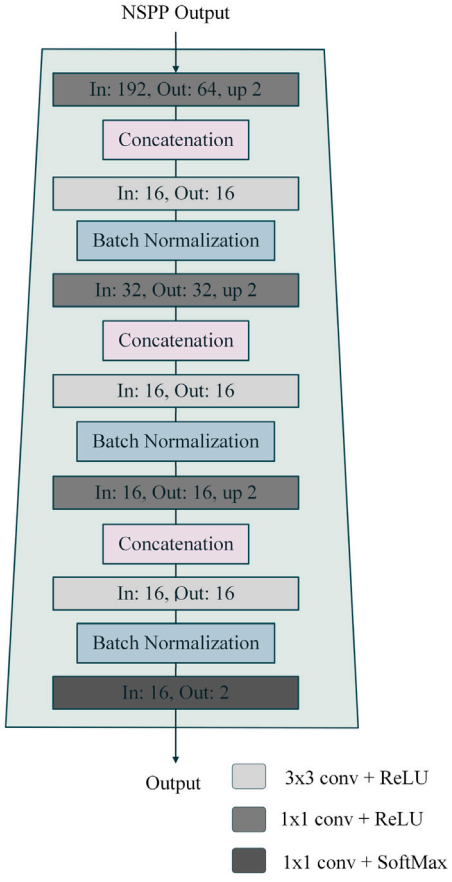

Fig. 7. The architecture of the urban change detection network (UCDNet) decoder.

2.3. Deep learning ensembles for change detection

Ensemble learning has gained attention in recent years due to its ability to improve the performance and robustness of predictive models (Rincy and Gupta, 2020). It involves combining multiple individual models, often referred to as base learners, to create a more accurate and reliable prediction (González et al., 2020). By exploiting the strengths of diverse models, ensemble learning overcomes the limitations of individual learners, leading to enhanced accuracy, stability, and generalization.

Table 1
The quality metrics used in this study.

Metric	Abbreviation	Formula
Accuracy	Acc	$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
Precision	Pr	$\text{Pr} = \frac{\text{TP}}{\text{TP+FP}}$
Recall	Re	$\text{Re} = \frac{\text{TP}}{\text{TP+FN}}$
Dice coefficient	Dice	$\text{Dice} = \frac{2\cdot\text{TP}}{2\cdot\text{TP+FP+FN}}$
Cohen's kappa	κ	$\kappa = \frac{p_o - p_e}{1-p_e}$
Intersection over union	IoU	$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}}$

One of the prominent ensemble techniques is weighted voting, which involves assigning different weights to the predictions (probabilities) elaborated by the individual ML models and then combining them to make the final prediction—this approach is followed in the reported study. It allows us for giving more importance to the deep learning models that are more accurate or reliable (as quantified over the available, possibly limited test set), thereby improving the performance of a combined learner. Also, as some of the base learners might be under-performing and could be unable to generalize, combining multiple ML algorithms might help compensate those *low-quality* CD algorithms. This is important to emphasize that this compensation could be of great importance in practical scenarios, in which the accurate quantification of the capabilities of the base models is difficult (or impossible) due to the limited validation and test sets.

In our ensembles, the weights are determined based on the final value of the loss function of each base (individual) model calculated over the training set—the better the loss function becomes (here, the negative log-likelihood loss (Yao et al., 2020), as this study is targeting the classification task), the higher weight is assigned to the corresponding deep learning model. Thus, the weights are determined once, and they are not subject to dynamically change later. Although there are other possible schemes that could be used while aggregating separate predictions of base models within an ensemble, ranging from simple hard voting techniques to more sophisticated supervised fusers (Nalepa et al., 2021b), our approach offers operational advantages: (*i*) the contribution of each model's to the final prediction of an ensemble is easy to interpret, and (*ii*) the ensemble content can be conveniently expanded without the need of retraining any additional data-driven fusers.

2.4. Evaluation metrics

To quantify the performance of the investigated CD models, we calculate the classic quality metrics extracted from the confusion matrix: accuracy (Acc), precision (Pr), recall (Re) and the Dice coefficient, being the overlap metric quantifying the agreement between the predicted and ground truth change areas (the larger overlaps show the areas for which changes were delineated more precisely by the corresponding model). In addition, Cohen's kappa (κ) was used to provide a more reliable assessment of model performance by accounting for the agreement occurring by chance, which is particularly important in imbalanced datasets such as those used in CD (also, this metric is reportedly presented in the CD works (Basavaraju et al., 2022)). Moreover, intersection over union (IoU) was added to compare the ability of the models to delineate change boundaries. All these metrics should be maximized (↑), with one indicating the perfect score.

The metrics are defined in Table 1, where TP, TN, FP and FN denote true positives (correctly classified *change* pixels), true negatives (correctly classified *no change* pixels), false positives (incorrectly classified *no change* pixels), and false negatives (incorrectly classified *change* pixels), respectively. Finally, for the Cohen's kappa, p_o represents the observed agreement, being the proportion of instances where the predicted and ground truth labels match, and p_e corresponds to the expected agreement, assuming both labels are assigned randomly (here, p_e is estimated using the empirical prior over the class labels (Artstein and Poesio, 2008)).

3. Experiments

The primary objective of the experimental study is to rigorously validate the base deep learning models, as well as the proposed ensembles in the CD task. The study is split into separate experiments, which are discussed in this section.

3.1. Experimental settings

The models were implemented in Python with the PyTorch library, the experiments were run on the Google Colab platform. Throughout the experimental study, both versions of Google Colab were used (the free version and the Colab Pro version). Colab Pro was needed to train models for a longer time (processing 50 epochs took around 8 h)—it provides more memory and faster GPUs. Software and environment configurations are shown in Table 2. To ensure full reproducibility of our study, the FC-EF, FC-Siam-conc and FC-Siam-diff models were downloaded from the original repository, whereas the UCDNet model implementation is open sourced through our GitHub repository available at: https://github.com/smile-research/DeepEnsembles4ChangeDetection. Overall, this repository contains the implementation and a step-by-step guideline showing how to re-execute the experiments.

Table 2
Software and environment configuration.

Component	Version
Environment	Colab Free/Pro version
GPU	T4 GPU, L4 GPU provided by Colab workspace
Python	3.10.12
PyTorch	2.3.0
CUDA	12.4

Table 3
The summary of the training configuration.

Summary of the training parameters	
Optimizer and learning rate	Adam, decay=1e−4
Batch size	32
Patch size	96
Number of epochs	10/30/50 (depends on an experiment)
Loss function	The negative log likelihood loss
Scheduler	ExponentialLR

3.2. Experimental methodology

Each experiment followed a structured methodology comprising data preprocessing, model configuration, training, and evaluation. Sentinel-2 imagery from the OSCD dataset was first preprocessed by normalization, augmentation, and partitioning into training and testing subsets (14 image pairs for training and 10 for testing, as suggested by the authors of the benchmark). All models were trained and verified under standardized conditions, using the same optimizer, loss functions, and evaluation metrics. Table 3 shows all pivotal training hyperparameters. To account for the significant class imbalance in the training set during the training process, we assigned a weight inversely proportional to the frequency of a given class (*change* vs. *no change*), meaning that the minority class (*change*) received a higher weight. The models were evaluated using the quality metrics described in detail in Section 2.4. The metrics obtained were calculated by each model training run, except Experiment 3 (Section 3.4.3), where the metrics are an average of five training runs per model. In total, 38 configurations were run to investigate the influence of architectural and training variations on the CD outcomes.

3.3. Reproducibility limitations

Our experiments were conducted on Google Colab using both L4 and T4 GPUs. During training, two forms of data augmentation were applied: random horizontal flipping and random 90° rotations—both rely on the Python's random module. No fixed random seed was used in the experiments, which introduces stochasticity in data augmentation, weight initialization, and training behavior. Furthermore, GPU-based operations can be non-deterministic (e.g., due to cuDNN optimizations), and hardware allocation in Colab may vary between runs. Consequently, exact reproducibility cannot be guaranteed, and repeated runs may yield slightly different results.

3.4. Results and discussion

3.4.1. Experiment 1: Baseline model with RGB bands

In this experiment, the models are validated using only the RGB bands of the Sentinel-2 imagery. The objective is to establish a baseline set of results elaborated with a minimal input configuration which is practically always available in EO images. All models are trained for 50 epochs.

The results obtained for the deep learning models operating on the RGB bands (Table 4 and Fig. 8) indicate that UCDNet demonstrated the best overall CD performance. It achieved the highest accuracy (95.74%) and precision (0.77), indicating strong capability in correctly identifying true positives (correctly determined *change* pixels). It also had the highest class accuracy for no change at 99.61%, though it struggled more with the change class (23.96%). In terms of recall, UCDNet elaborated a lower score of 0.23 when compared to recall obtained using FC-Siam-diff: 0.31. Finally, the Dice coefficient for UCDNet was 0.37, and it also had the highest Kappa score amounting to 0.35. The predicted outputs from these models rendered in Fig. 9 confirm worse results in subtle CD—only significant changes were correctly detected in most cases.

3.4.2. Experiment 2: Enhancing the models with the near infrared band band

This experiment explores the performance of the deep learning models utilizing the spectral bands beyond RGB, including the Near Infrared Band (NIR). By incorporating this band, the aim is to assess whether the increased spectral information improves the models' capabilities in CD. The models undergo training for 50 epochs.

The results obtained after adding the NIR band (Table 5 and Fig. 10) indicate that the models exhibited varied performance. FC-EF achieved the highest accuracy at 95.71%. UCDNet, while having a lower accuracy amounting to 94.15%, excelled in detecting

Table 4
The results obtained for the RGB bands using the investigated models. The best results for each metric are boldfaced.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet
Accuracy (%)	94.81	94.39	94.35	**95.74**
Class acc. (**no change/change**) (%)	98.4/28.42	97.98/27.74	97.78/**30.68**	**99.61**/23.96
Precision	0.49	0.42	0.43	**0.77**
Recall	0.28	0.27	**0.31**	0.23
Dice	0.36	0.34	0.36	**0.37**
Kappa	0.33	0.31	0.33	**0.35**

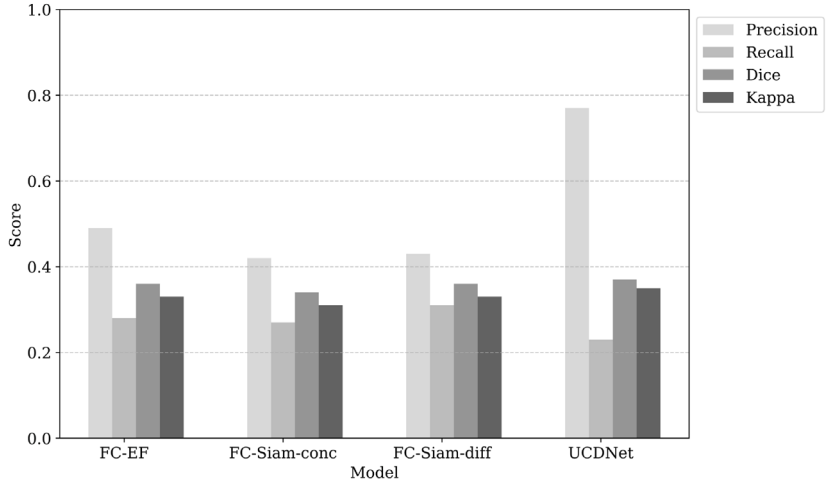

Fig. 8. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model operating on the RGB bands.

Table 5
The results obtained for the RGB and Near Infrared (NIR) bands using the investigated models. The best results for each metric are boldfaced.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet
Accuracy (%)	**95.71**	95.29	95.17	94.15
Class acc. (**no change/change**) (%)	**99.44**/26.5	98.39/37.74	98.69/29.88	95.6/**67.43**
Precision	**0.71**	0.55	0.55	0.45
Recall	0.26	0.38	0.3	**0.67**
Dice	0.39	0.45	0.39	**0.54**
Kappa	0.37	0.43	0.36	**0.51**

changes, achieving a class accuracy of 67.43% for the *change* class. In terms of precision, FC-EF outperformed the other methods with this metric amounting to 0.71. However, UCDNet outperformed in recall with a score of 0.67, despite a lower precision of 0.45. The Dice and Cohen's kappa scores were the highest for UCDNet, and they were equal to 0.54 and 0.51, respectively. With the addition of the NIR spectral band, UCDNet demonstrated strong capabilities in detecting changes and maintaining a balance between precision and recall. Comparing the visual results in Fig. 11, UCDNet model has tendency to false alarms, whereas FC-EF, FC-Siam-conc and FC-Siam-diff often predict more false negatives than false positives.

3.4.3. Experiment 3: Exploiting all Sentinel-2 bands

In this experiment, all Sentinel-2 bands are used as input to the investigated deep learning models. The goal is to maximize the utilization of the spectral information to enhance CD. The models are trained for 50 epochs. Each model was independently trained and evaluated five times, and each run was initialized separately to ensure statistical robustness. The reported results represent the average performance across these five runs.

The results obtained for all Sentinel-2 bands (Table 6 and Fig. 12) indicate that FC-EF achieved the highest overall accuracy at 95.79% and the best precision at 0.75. It also demonstrated excellent class accuracy for the *no change* class with 99.46%. In contrast, UCDNet, while slightly lower in overall accuracy at 94.95%, outperformed the other models in detecting actual changes, achieving a *change* class accuracy of 62.49%, the highest recall of 0.62, the best Dice score of 0.56, and the top Cohen's Kappa score of 0.53. Moreover, UCDNet achieved the highest IoU (0.39), confirming its strength in capturing spatial overlap between predicted and true changes. On the other hand, FC-EF and FC-Siam-diff obtained the highest AUC (0.92), reflecting better overall class separability, whereas UCDNet reached 0.80. While FC-EF performed best in terms of general classification accuracy and reducing false positives,

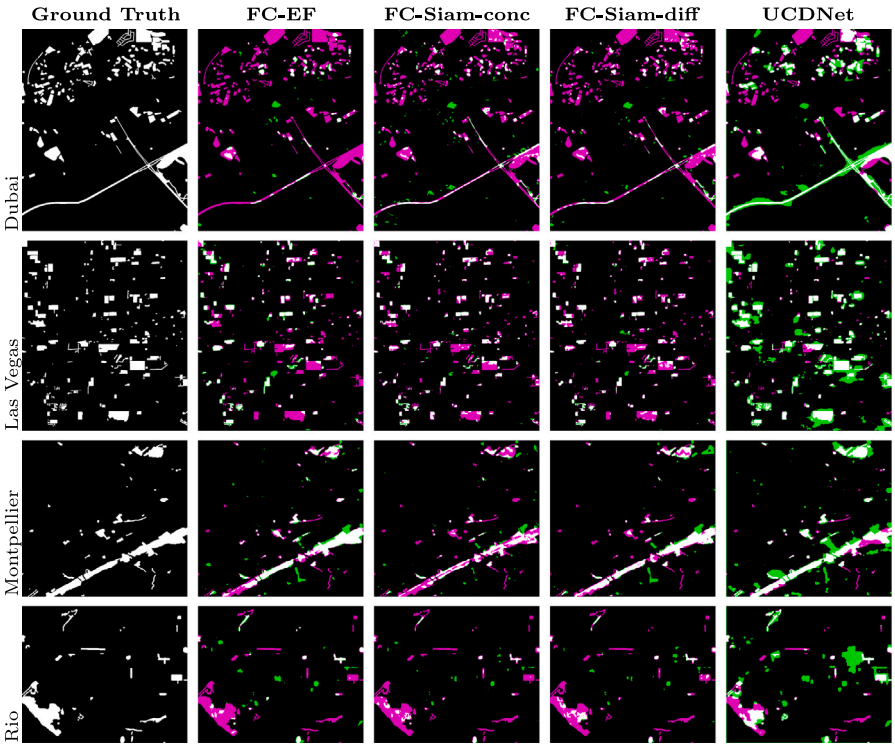

Fig. 9. Prediction results on the selected test scenes from OSCD for the models operating on the RGB Sentinel-2 bands. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels). For the high-resolution images, please see https://github.com/smile-research/DeepEnsembles4ChangeDetection.

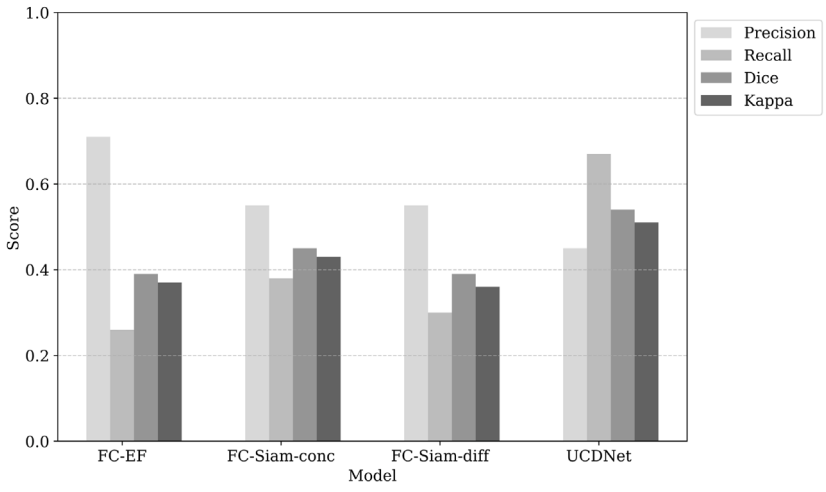

Fig. 10. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model operating on the RGB and NIR bands.

UCDNet showed superior sensitivity to true changes and better balance between precision and recall. This makes UCDNet particularly effective in identifying subtle or small-scale changes. Compared to the previous experiment (involving models trained only on RGB and NIR Sentinel-2 bands), the use of all spectral bands contributed to improved model robustness and generalization, particularly for UCDNet, which showed a notable increase in *change* class accuracy and recall. Finally, qualitative visual results are provided in Fig. 14.

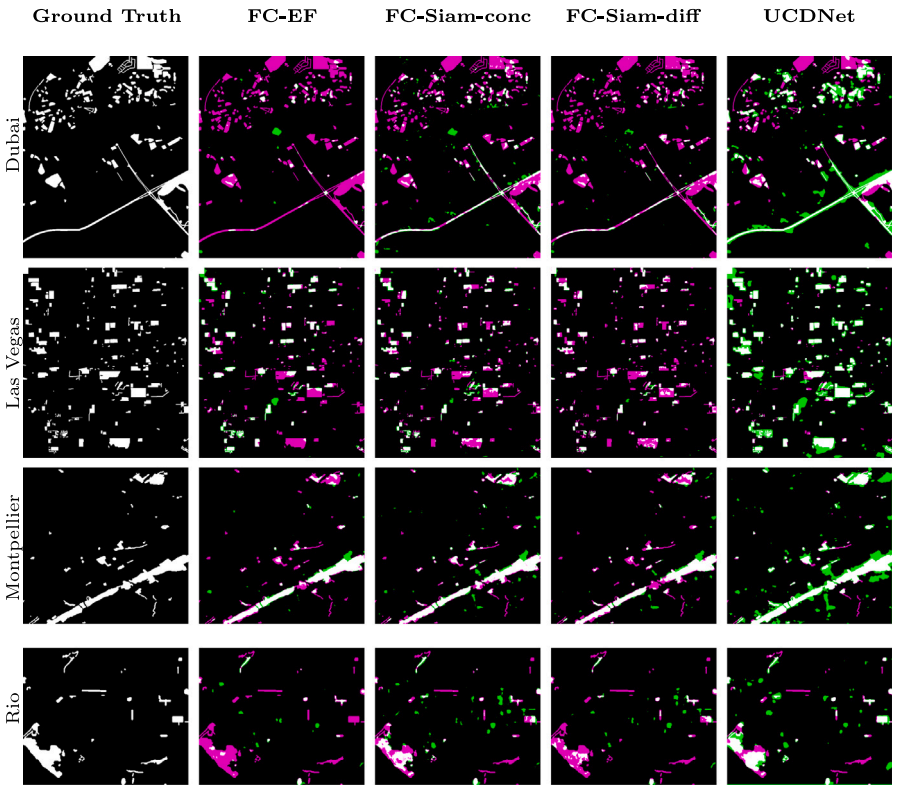

Fig. 11. Prediction results on the selected test scenes from OSCD for the models operating on the RGB and NIR Sentinel-2 bands. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels). For the high-resolution images, please see https://github.com/smile-research/DeepEnsembles4ChangeDetection.

Table 6
The results obtained for all Sentinel-2 bands using the investigated models. The best results for each metric are boldfaced. Presented results are an average and standard deviation from five runs.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet
Accuracy (%)	**95.79 (0.11)**	95.45 (0.3)	95.45 (0.02)	94.95 (0.61)
Class acc. (**change**) (%)	27.77 (5.81)	35.72 (2.25)	31.05 (1.1)	**62.49 (3.23)**
Class acc. (**no change**) (%)	**99.46 (0.23)**	98.67 (0.34)	98.92 (0.06)	96.7 (0.74)
Precision	**0.75 (0.05)**	0.6 (0.05)	0.61 (0.01)	0.51 (0.01)
Recall	0.28 (0.06)	0.36(0.02)	0.31 (0.02)	**0.62 (0.03)**
Dice	0.4 (0.06)	0.45 (0.01)	0.41 (0.01)	**0.56 (0.03)**
Kappa	0.38 (0.06)	0.42 (0.01)	0.39 (0.01)	**0.53 (0.03)**
IoU	0.26(0.02)	0.28(0.01)	0.26(0.01)	**0.39(0.02)**
AUC	**0.92(0.01)**	0.91(0.01)	**0.92(0.01)**	0.8(0.04)

Table 7
Comparison of 95% Confidence Intervals (CI) for each metric across the investigated models.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet
Accuracy (%)	[95.65, 95.93]	[95.16, 95.74]	[95.42, 95.48]	[94.10, 95.80]
Class acc. **(change)** (%)	[20.55, 35.99]	[32.93, 38.51]	[30.20, 31.90]	[60.48, 64.50]
Class acc. **(no change)** (%)	[99.18, 99.74]	[98.25, 99.09]	[98.89, 98.95]	[95.68, 97.72]
Precision	[0.69, 0.81]	[0.54, 0.66]	[0.60, 0.62]	[0.44, 0.58]
Recall	[0.21, 0.35]	[0.33, 0.39]	[0.30, 0.32]	[0.59, 0.67]
Dice	[0.33, 0.47]	[0.44, 0.46]	[0.40, 0.42]	[0.54, 0.58]
Kappa	[0.31, 0.45]	[0.41, 0.43]	[0.38, 0.40]	[0.51, 0.55]
IoU	[0.24, 0.28]	[0.27, 0.29]	[0.25, 0.27]	[0.37, 0.41]
AUC	[0.91, 0.93]	[0.90, 0.92]	[0.91, 0.93]	[0.77, 0.83]

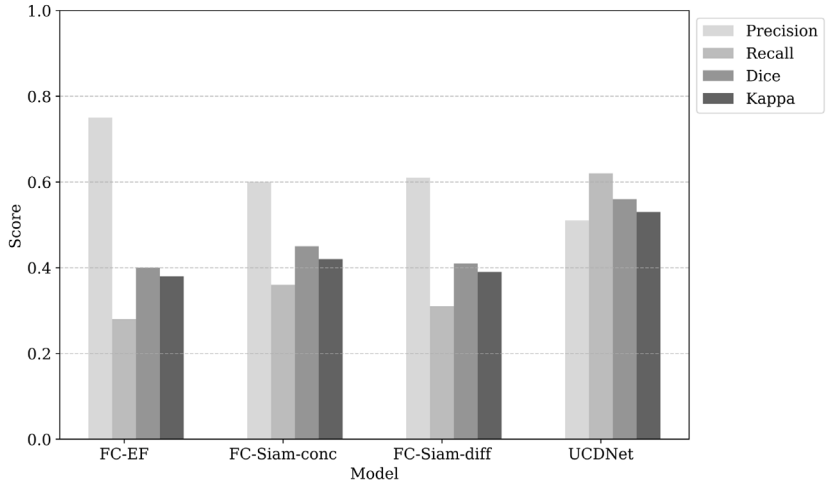

Fig. 12. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model operating on all Sentinel-2 bands.

Table 8
Comparison of the average training times for 50 epochs (in minutes), inference time (in seconds) for 10 scenes from OSCD dataset and compute units (per hour) for training for the deep learning models.

Model	Training time	Inference time	Compute units
FC-EF	469	113	1.43
FC-Siam-conc	431	99	1.64
FC-Siam-diff	424	97	1.64
UCDNet	468	108	1.43

Table 9
Confusion matrix showing prediction results obtained by the UCDNet model evaluated on the test set.

	Predicted *no change*	Predicted *change*
Actual *no change*	2585,655	81,247
Actual *change*	94,597	49,165

The performance metrics for all models were evaluated across five independent experiments, and 95% confidence intervals (CIs) were computed (Table 7). These CIs provide a range within which the true mean performance is likely to be, accounting for variability across runs. Models like FC-Siam-diff showed narrow CIs across most metrics, indicating consistent performance, whereas FC-EF and UCDNet exhibited wider intervals, particularly in metrics such as recall and change accuracy, reflecting greater variability. Notably, UCDNet achieved the highest mean change accuracy and recall, but with slightly broader CIs, suggesting a trade-off between peak performance and stability.

For each model, the training time was recorded over five independent runs, and the average training duration was calculated. In addition, the usage of computational resources was monitored in the Google Colab environment, including the average compute units consumed per hour of training, as shown in Table 8. Furthermore, inference time was measured by evaluating each model on 10 scenes from the OSCD dataset, providing insight into the real-time performance and efficiency during deployment. These metrics collectively offer a comprehensive comparison of the time efficiency, inference speed, and computational cost across all evaluated models. We wish to emphasize that, although the times obviously increase for the ensemble models, as they require training and executing more (deep) machine learning models, these times (e.g., hours for training, not days) seem still to be affordable, even for constrained hardware environments. Of note, the models are trained once, and they are later used to elaborate predictions—this process is substantially faster than training.

Table 9 presents the pixel-level confusion matrix for the UCDNet model (the best-performing one; however, the differences between the models are not statistically important when investigating κ at $p < 0.05$, according to the non-parametric Wilcoxon tests) evaluated on the test set. The model correctly classified 2,585,655 pixels as *no change* and 49,165 pixels as *change*. However, it also produced 94,597 false positives and 81,247 false negatives. These results demonstrate the strong ability of UCDNet to identify unchanged regions in high-resolution satellite imagery, while revealing the ongoing challenge of capturing all true instances of change, reflected in the lower recall and precision for the *change* class. Despite these limitations, UCDNet consistently outperformed the other evaluated models across key metrics, such as recall, Dice coefficient, and Cohen's Kappa, making it the most effective model for detecting changes at the pixel level in this experiment.

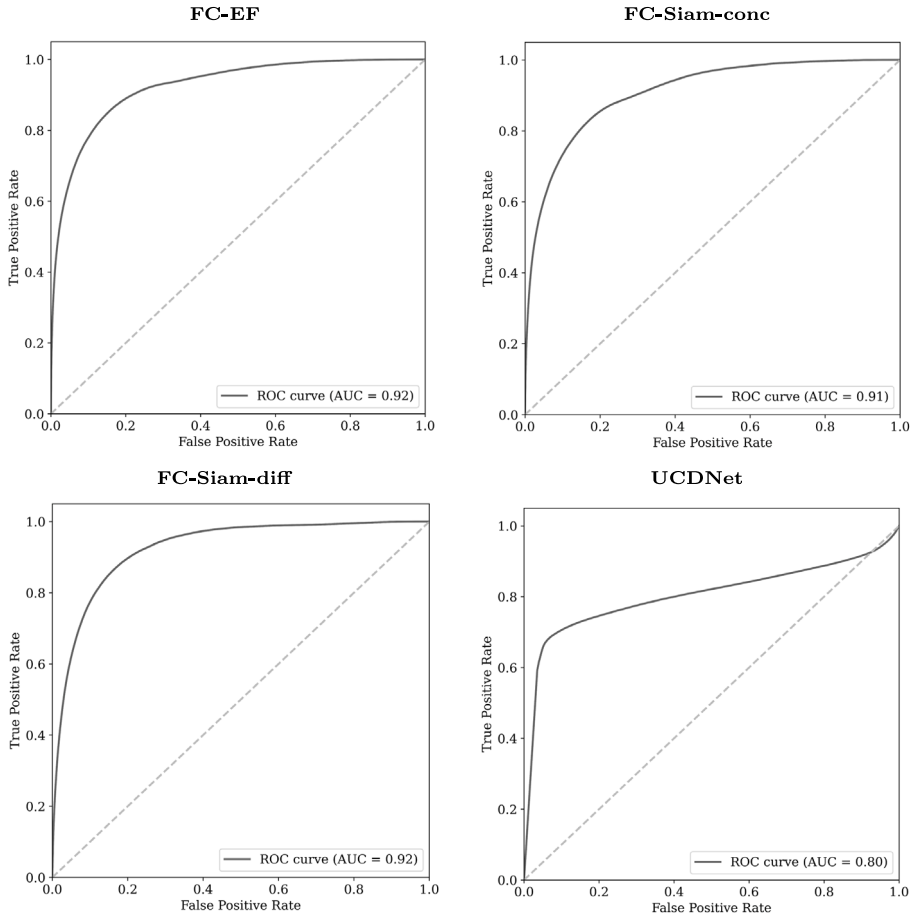

Fig. 13. ROC curves (together with the corresponding AUC values) for each deep learning model operating on all Sentinel-2 bands.

To complement the standard evaluation metrics, Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) values were included in the analysis (Fig. 13). ROC curves provide a threshold-independent assessment of the model performance by illustrating the trade-off between sensitivity and false alarm rates for different thresholds. The AUC offers a concise scalar measure of the overall separability between the *change* and *no change* classes, allowing for a more robust comparison of the investigated models. The ROC curves (Fig. 13) reveal that the Siamese-based models (FC-Siam-conc and FC-Siam-diff) and FC-EF provide high-quality class separability, with AUC values greater than 0.90, while the AUC for UCDNet amounts to 0.80. This confirms that although UCDNet achieves higher recall in detecting changes, its overall discriminative capability is weaker compared to the other investigated models.

3.4.4. Experiment 4: Comparing models trained for different numbers of epochs

In this experiment, the impact of the training duration on the resulting models' performance is investigated. This comparison seeks to determine the appropriate number of epochs for achieving the best balance between the training time (thus its computational cost) and generalizability of the deep models.

The results (Table 10 and Fig. 15) show significant differences for all deep learning models trained for different numbers of epochs. All models maintain high accuracy, generally around 95% to 96%. FC-EF shows consistent accuracy, while UCDNet achieves the highest accuracy of 96.03% after being trained for 30 epochs. Class accuracy for *no change* remains high across all models and epochs, typically above 97%. However, class accuracy for *change* varies more, with UCDNet performing better, especially at 10 epochs (69.33%). In contrast, other models show lower class accuracy for *changes*, with FC-EF reaching only 23.96% at 50 epochs. Precision values are moderate, with UCDNet achieving the highest precision (0.62) at 30 epochs. FC-EF improves precision over time, reaching 0.77 at 50 epochs. On the other hand, the recall values are more variable. The Dice coefficient scores show moderate performance, with UCDNet achieving the highest value (0.60) at 10 epochs. The Kappa score indicates moderate agreement, with UCDNet performing slightly better at 10 epochs (0.57). Overall, UCDNet demonstrates strong performance across multiple metrics, especially in class accuracy for detecting changes and in recall. FC-EF performs well in precision and accuracy but varies in recall. Finally, the FC-Siam-conc and FC-Siam-diff models show consistent but slightly lower performance.

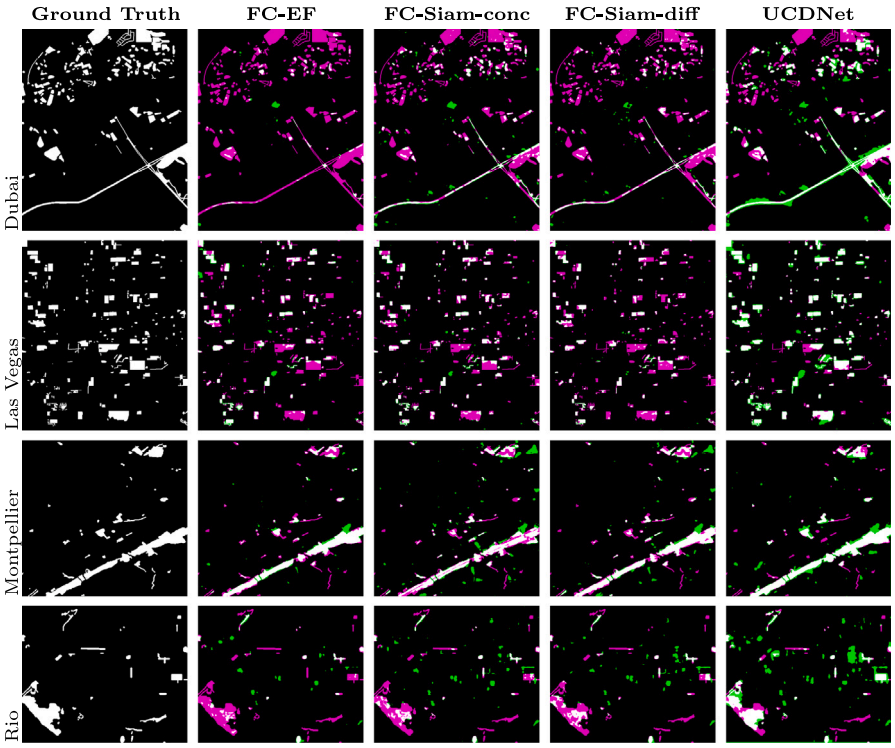

Fig. 14. Prediction results on the selected test scenes from OSCD for the models operating on all Sentinel-2 bands. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels). For the high-resolution images, please see https://github.com/smile-research/DeepEnsembles4ChangeDetection.

Table 10
The results obtained for the models trained for 10, 30, and 50 epochs. The best results for each number of processed epochs are boldfaced. The meaning of abbreviations is as follows: NC—*no change*, C—*change*. The accuracy and class accuracy are reported in %.

Epoch	Model	Metrics					
		Accuracy	Class acc. NC/C	Precision	Recall	Dice	Kappa
10	FC-EF	**95.65**	**98.71**/38.84	**0.62**	0.39	0.48	0.46
	FC-Siam-conc	95.03	98.63/28.27	0.52	0.28	0.37	0.34
	FC-Siam-diff	95.37	98.61/35.28	0.58	0.35	0.44	0.42
	UCDNet	95.16	96.55/**69.33**	0.52	**0.70**	**0.60**	**0.57**
30	FC-EF	95.75	**99.45**/26.93	**0.73**	0.27	0.40	0.38
	FC-Siam-conc	95.06	98.22/36.55	0.53	0.37	0.43	0.41
	FC-Siam-diff	95.35	98.80/31.27	0.58	0.31	0.41	0.39
	UCDNet	**96.03**	98.16/**56.47**	0.62	**0.56**	**0.59**	**0.57**
50	FC-EF	**95.75**	**99.62**/23.96	**0.77**	0.24	0.37	0.35
	FC-Siam-conc	95.27	98.58/33.87	0.56	0.34	0.42	0.40
	FC-Siam-diff	95.47	98.94/31.21	0.61	0.31	0.41	0.40
	UCDNet	95.60	97.51/**60.21**	0.57	**0.60**	**0.58**	**0.56**

Regarding the training time reported in Table 11, the increase in the number of epochs obviously results in a significant rise in the training duration for all deep learning models. For example, UCDNet's training time rises from 97 min (10 epochs) to 490 min (50 epochs). This increase must be weighed against the relatively modest gains in performance metrics. Therefore, while longer training can lead to slightly better model performance, it is essential to consider the computational effort while designing the training process. For practical applications, it may not always be profitable to train models for a longer duration if the performance gains are rather minimal.

3.4.5. Experiment 5: Deep learning ensembles for change detection

This section presents the results obtained using the proposed deep learning ensembles with the base models operating over all Sentinel-2 spectral bands. Additionally, we verify the impact of incorporating the NSPP block within the UCDNet model—Table

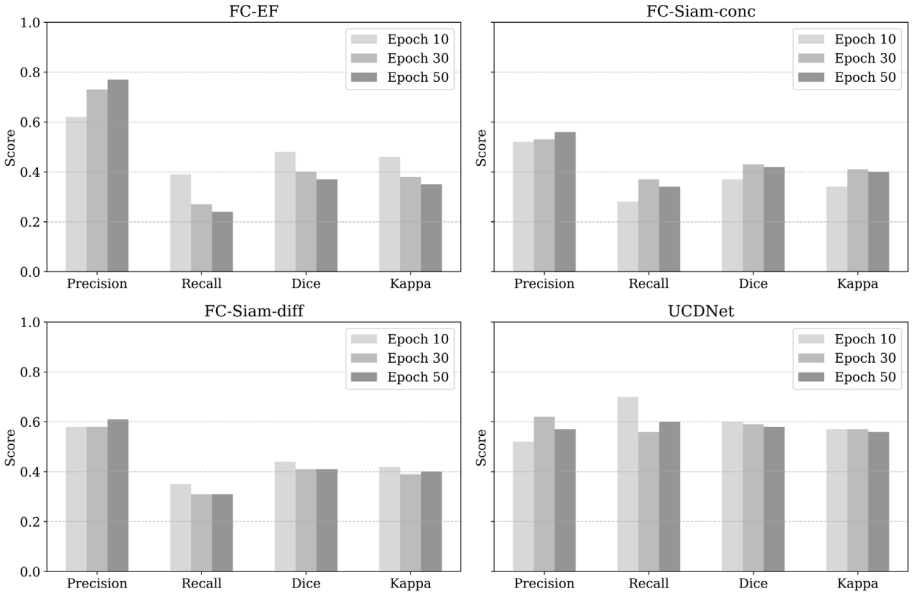

Fig. 15. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model operating on all Sentinel-2 bands and trained for different numbers of epochs.

Table 11
Comparison of the training times (in minutes) for the deep learning models trained for different numbers of epochs.

Model	Training time (min)		
	10 epochs	30 epochs	50 epochs
FC-EF	94	282	429
FC-Siam-conc	86	261	431
FC-Siam-diff	81	256	434
UCDNet	97	310	490

Table 12
The results obtained for the UCDNet model without and with the New Spatial Pyramid Pooling (NSPP) block. The best results for each metric are boldfaced.

Metric	UCDNet without NSPP block	UCDNet with NSPP block
Accuracy (%)	95.46	**95.61**
Class acc. (**no change/change**) (%)	**99.17**/26.58	97.51/**60.21**
Precision	**0.63**	0.57
Recall	0.27	**0.60**
Dice	0.37	**0.58**
Kappa	0.36	**0.56**

12 compares the performance of the UCDNet model without and with the NSPP block. The results show that utilizing it leads to improvements in several key metrics. Accuracy slightly increases from 95.46% to 95.61% with the NSPP block, and the class accuracy for detecting *changes* improves significantly from 26.58% to 60.21%, although the class accuracy for *no change* decreases from 99.17% to 97.51%. Precision decreases marginally from 0.63 to 0.57, but recall shows an improvement from 0.27 to 0.60. Furthermore, the Dice coefficient increases from 0.37 to 0.58. The Cohen's Kappa statistic also improves from 0.36 to 0.56. Overall, the inclusion of the NSPP block enhances the UCDNet model's ability to detect changes more accurately.

The results obtained using the proposed deep learning ensembles (Table 13 and Fig. 17) show a balanced enhancement in several key metrics, indicating the effectiveness of this approach. Specifically, accuracy achieved through weighted voting is 95.1%, which is competitive with the individual models. Precision for the voting method is notably high at 0.75. Recall for the voting ensemble amounts to 0.48, which, although lower than UCDNet's 0.6, suggests that the ensemble method balances between identifying true positives and minimizing false alarms. The Dice coefficient and the Cohen's Kappa statistic for the voting method are 0.44 and 0.55, respectively, showing that the ensemble maintains a good balance of precision and recall while also achieving a consistent level of agreement between the elaborated predictions and ground truth. Although the results indicate that ensembling deep learning models leads to the second best results in almost all quality metrics, it is important to emphasize that this approach allowed to outperform

Table 13
The results obtained using the proposed voting ensemble learning technique, as well as other base learners. The best results are boldfaced, and the second best are underlined.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet	Ensemble
Accuracy (%)	**95.75**	95.27	95.47	95.60	95.10
Class acc. (**change**) (%)	27.77	35.72	31.05	**62.49**	40.89
Class acc. (**no change**) (%)	**99.46**	98.67	98.92	96.7	99.11
Precision	**0.77**	0.56	0.61	0.57	0.75
Recall	0.24	0.34	0.31	**0.60**	0.48
Dice	0.37	0.42	0.41	**0.58**	0.44
Kappa	0.35	0.40	0.40	**0.56**	0.55
IoU	0.26	0.28	0.26	**0.39**	00.35
AUC	**0.92**	0.91	**0.92**	0.8	0.83

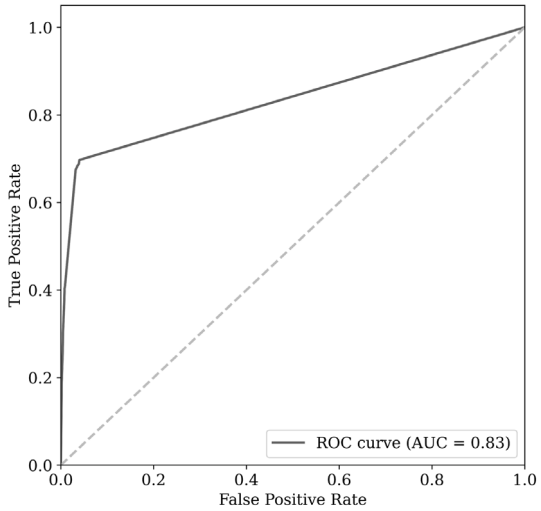

Fig. 16. The ROC curve and (and the corresponding AUC value) obtained for the proposed deep learning ensemble operating on all Sentinel-2 bands.

(thus compensate) the worst performing models—this supports our hypothesis discussed in Section 1.3. Also, this compensation may bring practical benefits of using several individual (base) models over the unseen image data. Here, determining which models are best-fitted for particular scenes/acquisition conditions may not be trivial. Thus, it would be worthwhile to benefit from the strengths of several heterogeneous models. Importantly, the ensembles offered statistically significantly better results (the κ scores), when compared to FC-EF, FC-Siam-conc, and FC-Siam-diff ($p = 0.004$, $p = 0.006$, and $p = 0.027$, according to the non-parametric Wilcoxon tests, respectively)—it shows that the ensemble can indeed compensate the under-performing models. Here, the results are statistically same for the ensemble and UCDNet (κ score, $p = 0.770$). Finally, it is worth mentioning that the results obtained using the proposed deep learning ensembles over OSCD are competitive to the results reported in other works. In Mao et al. (2024), Mao et al. presented a novel 3D swin transformer with enhanced feature aggregation framework, and confronted it over OSCD with an array of deep learning methods, including long short-term memory-inspired architectures (Papadomanolaki et al., 2021), pure transformer networks (Zhang et al., 2022), bi-branch fusion networks (Song et al., 2023), and the adjacent-level feature cross-fusion techniques with 3D convolutional neural networks (Ye et al., 2023). The authors reported the Cohen's kappa scores for OSCD ranging from 0.46 to 0.56, which is at the level of the best individual model and our ensemble, with a significantly simpler architecture when compared to the swin transformer. Similar results were recently reported in a comprehensive analysis presented by Wang and colleagues, in their article introducing HyperSIGMA (Wang et al., 2025) (it is worth noting that their large-scale foundation models, scaled to 1 billion parameters, reached the Cohen's kappa of 0.57). In Chen and Bruzzone (2022), Chen and Bruzzone introduced a self-supervised approach for CD which achieved the Cohen's kappa of 0.51 over the OSCD dataset. Therefore, one can observe that ensembling deep learning architectures into CD committees offer high quality operations which is competitive to or outperforming other, often significantly more sophisticated and larger models.

Analyzing the obtained ROC curve of the deep learning ensemble model (Fig. 16), one can observe that the shape of the curve closely resembles the ROC curve obtained for UCDNet. The corresponding AUC value of 0.83 is slightly higher than that of UCDNet and remains weaker than other models (FC-EF, FC-Siam-conc, FC-Siam-diff). However, the marginal increase in AUC indicates that ensembling contributes to a reduction of false positives while maintaining reasonable recall.

Comparing the change maps obtained using the deep learning ensemble in Fig. 18 with those produced by the individual (best) model, a noticeable reduction in both false positives and false negatives is observed. Specifically, the ensemble model significantly

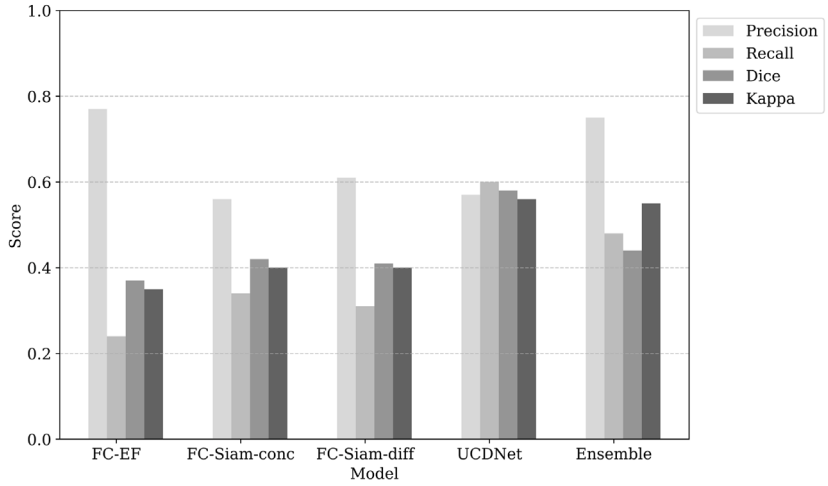

Fig. 17. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model operating on all Sentinel-2 bands, as well as obtained using the proposed deep learning ensembles.

decreases the number of false positives compared to the UCDNet results. Overall, the results indicate that the ensemble successfully combines the strengths of the base models, enhancing the reliability of CD.

A qualitative inspection of selected scenes from the OSCD dataset revealed that model performance varies depending on the spatial and environmental characteristics of the area (Fig. 19). In the Chongqing scene, which represents a dense urban environment with significant development, many urban changes were missed, resulting in a high number of false negatives. In the Saclay West scene, which is mostly agricultural with small urban changes, both false positives and false negatives were observed. While the ensemble method slightly reduced false positives in both cases, it did not detect the missed urban changes.

3.4.6. Experiment 6: Investigation of performance on MUMUCD

To assess the generalization capability of models trained on the OSCD dataset, an additional experiment was conducted using the MUMUCD dataset as the unseen test data. This evaluation aims to understand how well the models—originally optimized on OSCD—perform when applied to a different geographic and temporal context without re-training. In this setup, all available (in the MUMUCD benchmark dataset) Sentinel-2 spectral bands were utilized (a total of 12 bands). To maintain consistent input dimensions, the missing band 10 was replaced with a zero-filled array. This configuration preserves the spatial and spectral characteristics expected by the fully convolutional networks while enabling fair comparison across different datasets.

Table 14 and Fig. 20 present the evaluation results of the pre-trained models applied to the MUMUCD dataset, which was used solely for testing purposes. The test set comprises 70 image pairs, and all available Sentinel-2 bands (excluding band 10, which was zero-padded) were utilized. Although overall accuracy remains above 85% across all models—with the best result achieved by FC-EF (86.16%)—the performance in detecting actual changes is markedly low. Specifically, the accuracy for the *change* class does not exceed 3.68%, and both the recall and Dice coefficients are significantly limited, with the highest values reaching only 0.04 and 0.07, respectively. These outcomes suggest that the models struggle to generalize to the new dataset, therefore they should be fine-tuned over the new training set, e.g., following the transfer learning paradigm. Indeed, several factors may explain the reduced performance offered by the models trained over the OSCD dataset. First, the MUMUCD dataset contains different types of changes (e.g., seasonal or structural) compared to those present in the OSCD training set (which encompasses only urban changes), introducing a domain shift. The strong class imbalance, combined with subtle and complex change patterns, further complicates the detection task. Lastly, the substitution of band 10 with zero values may have disrupted the spectral integrity of the input data. Fig. 21 renders example scenes, together with the outcomes of the investigated models.

The evaluation on the MUMUCD dataset (Table 15 and Fig. 22) shows that the proposed voting ensemble achieves a competitive overall accuracy of 86.03%, which is close to the highest value obtained by the FC-EF model (86.16%). While the ensemble improves precision compared to UCDNet and FC-Siam-diff, and matches the FC-Siam-conc's 0.30 value, it does not outperform the best base learners in recall, Dice score, or Cohen's kappa. This suggests that the ensemble is more effective at reducing false positives than at detecting additional true changes, which is consistent with its design emphasis on weighted consensus rather than aggressive change identification. The ensemble maintains near-best overall accuracy while slightly improving precision, but offers limited gains in detecting additional changes. However, as already mentioned earlier, the models would certainly need further fine-tuning over the MUMUCD training set, as the results obtained over the unseen test set are significantly worse than those obtained for the training part of OSCD. Here, an interesting research avenue could include building pre-trained models (and ensembles), that could be later optimized over specific, potentially limited training sets of different characteristics. This would be in line with the current trend of building foundation models, and then fine-tuning them for specific tasks over limited (and not necessarily representative) training samples.

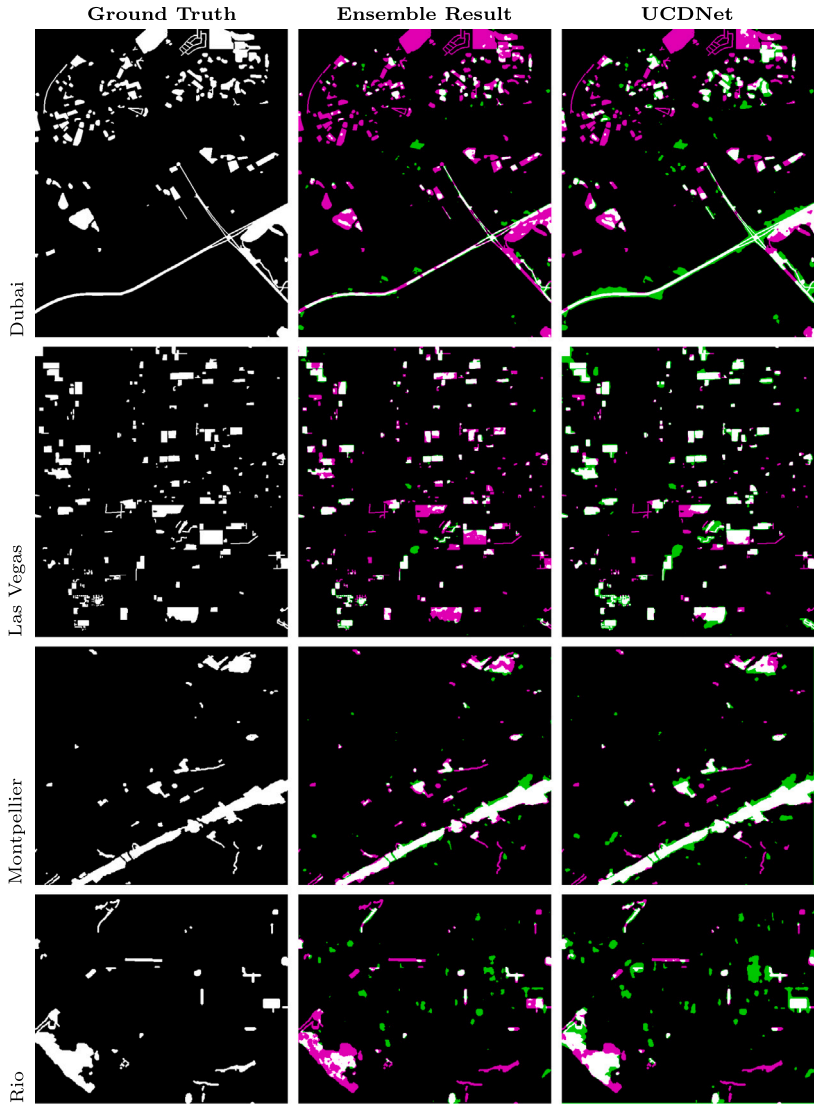

Fig. 18. The ground truth change map, the prediction obtained using the proposed deep learning ensemble and the UCDNet model for each selected test scene. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels). For the high-resolution images, please see https://github.com/smile-research/DeepEnsembles4ChangeDetection.

Table 14
The results obtained for the MUMUCD benchmark for all available bands using the investigated models. The best results for each metric are boldfaced.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet
Accuracy (%)	**86.16**	85.98	85.91	85.97
Class acc. (**change**) (%)	2.68	**3.68**	3.01	1.77
Class acc. (**no change**) (%)	**99.04**	98.68	98.7	98.96
Precision	**0.3**	**0.3**	0.26	0.21
Recall	0.03	**0.04**	0.03	0.02
Dice	0.05	**0.07**	0.06	0.03
Kappa	0.03	**0.04**	0.03	0.01

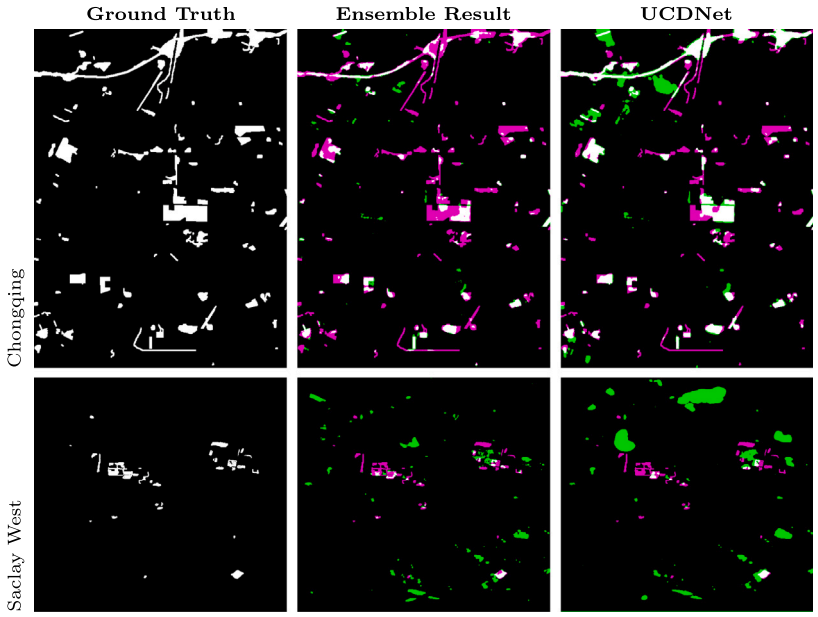

Fig. 19. The ground truth change map, the prediction obtained using the proposed deep learning ensemble and the UCDNet model for each selected test scene. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels). For the high-resolution images, please see https://github.com/smile-research/DeepEnsembles4ChangeDetection.

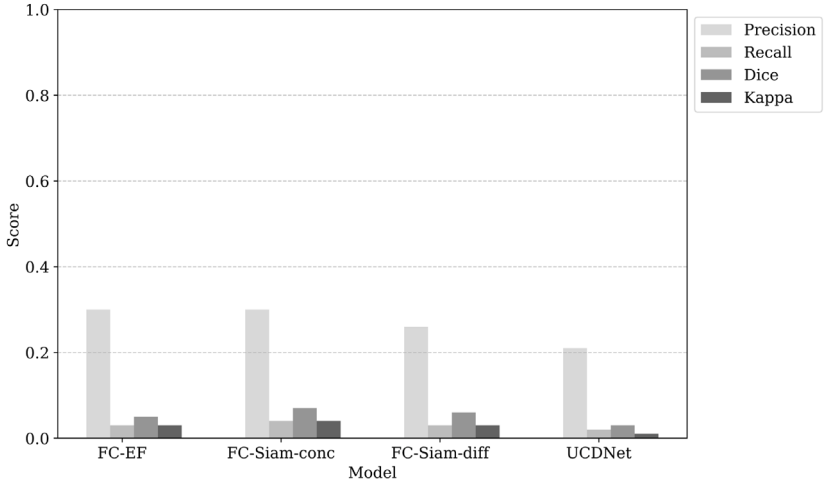

Fig. 20. Selected performance metrics (precision, recall, Dice coefficient and Cohen's kappa) obtained by each deep learning model for MUMUCD dataset.

3.4.7. Limitations and future research avenues

The most important limitation of our study is that it was performed over a limited, yet thoroughly researched benchmark datasets. Although OSCD and MUMUCD have been widely adopted by the community and become the state of the art in confronting CD algorithms (Prasad et al., 2023; Chaminda Bandara and Patel, 2025), we strongly encourage the community to further validate ensemble (deep) ML models for CD over other datasets to confirm their better generalizability when compared to individual models (see e.g., the datasets listed at https://github.com/isaaccorley/torchrs; accessed on April 10, 2025), potentially using an extended set of quality metrics, specifically targeting imbalanced classification and incorporating e.g., the Matthews Correlation Coefficient (Delgado, 2019). Here, the focus might be put on the datasets with better spatial resolution that would allow us to capture more fine-grained and subtle changes between the consecutive time points in an image time series. On top of that, it would be useful to investigate the impact of geographical locations on the CD errors obtained by the ML models—it might be useful to

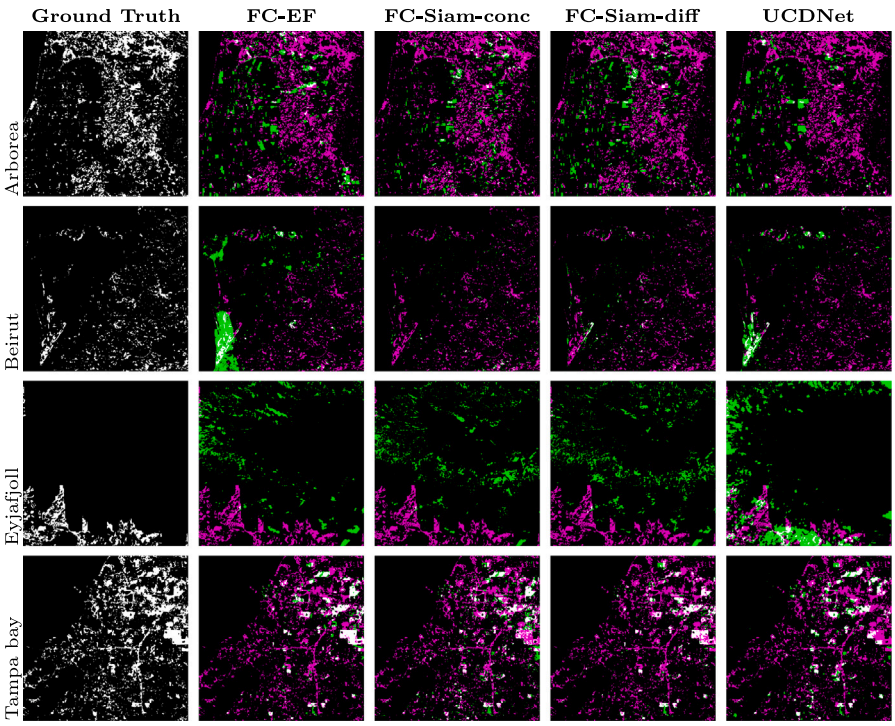

Fig. 21. Prediction results on the selected test scenes from MUMUCD for the models trained on OSCD dataset and operating on all Sentinel-2 bands. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels).

Table 15
The results obtained using the proposed voting ensemble learning technique, as well as other base learners for the MUMUCD for all available bands. The best results for each metric are boldfaced.

Metric	FC-EF	FC-Siam-conc	FC-Siam-diff	UCDNet	Ensemble
Accuracy (%)	**86.16**	85.98	85.91	85.97	86.03
Class acc. (**change**) (%)	2.68	**3.68**	3.01	1.77	2.9
Class acc. (**no change**) (%)	**99.04**	98.68	98.7	98.96	98.84
Precision	**0.3**	**0.3**	0.26	0.21	0.27
Recall	0.03	**0.04**	0.03	0.02	0.03
Dice	0.05	**0.07**	0.06	0.03	0.05
Kappa	0.03	**0.04**	0.03	0.01	0.03

build "specialized" ensembles, focusing on specific characteristics of the observed areas. Finally, it would be interesting to develop multi-class change detection techniques, in which multi-class classification of detected changes would be exploited (instead of binary classification, as exploited in this study). Such approaches could be useful in an array of downstream EO tasks, such as disaster assessment, land cover change detection, urban expansion, and many others (Kheddam and Tahraoui, 2024; Zhu et al., 2024).

Similarly, expanding ensemble CD techniques by appending other models, preferably with different underlying architectural solutions (such as e.g., vision transformers (Roy et al., 2025), foundation models (Aszkowski and Kraft, 2024; Sadel et al., 2025) and others) is an interesting research pathway which might be immediately explored based on the results reported in this article to further robustify our findings. Therefore, to enhance the capabilities of our combined approach, the further steps could include adding more well-generalizing CD models into the pool of base learners, and optimizing the content of the CD ensemble, e.g., using evolutionary algorithms (Heywood, 2024; Kucharski et al., 2024). However, it is also important to mention that including too many individual models could make the ensemble system computationally intensive, even during the inference (in practical scenarios, base models can be effectively learn offline before being deployed in the target environment).

The difficulty of practical CD is also affected by the quality of the acquired image data. As already presented in Fig. 1, satellite images might be e.g., covered by clouds—in such situations, accurate detection and tracking of changes is extremely difficult or even impossible, and may require scheduling another image acquisition. To build end-to-end CD systems, one can not only benefit from all available satellite imagery (Zhu and Woodcock, 2014), but also shall we incorporate cloud detection and semantic segmentation routines to determine the image areas which should be pruned from further downstream analysis (Kaur Buttar and Sachan, 2022).

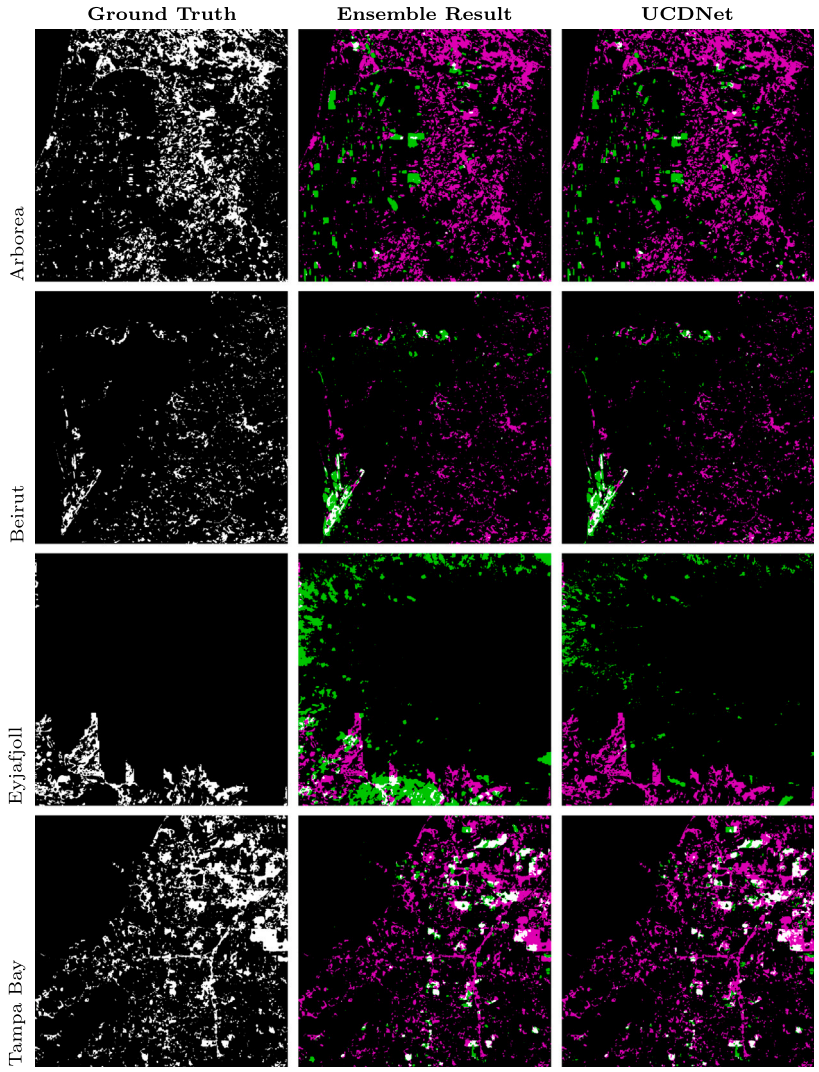

Fig. 22. Prediction results on the selected test scenes from MUMUCD for the deep learning ensemble and the UCDNet operating on all Sentinel-2 bands. The meaning of the colors is as follows: white—true positives (correctly identified *change* pixels), black—true negatives (correctly identified *no change* pixels), green—false positives (*no change* pixels incorrectly identified as those with *changes*), magenta—false negatives (*change* pixels incorrectly identified as *no change* pixels).

Such approaches may be incorporated in the on-the-ground analysis chain, but also one can deploy them in hardware-constrained edge devices after appropriate model compression and optimization (Grabowski et al., 2024). The latter approach can help in performing smart compression of satellite imagery, as well as in rescheduling the acquisition of cloudy areas for CD analysis. Also, different acquisition conditions (such as those related to the atmospheric characteristics (Nalepa et al., 2021a)) directly affect the data quality—verifying the robustness of CD models against such in-orbit conditions (as well as in-orbit noise (Reddy and Pawar, 2020) and low-quality data (Santos et al., 2021)) should be explored in more detail to show the practical abilities of the end-to-end CD analysis system.

Another challenge which is related to any ensemble model is that it requires not only training multiple base learners, but also elaborating their predictions for the incoming data that will be later fused in the voting process, or aggregated differently using other fusers. Since the models can be seamlessly added to an ensemble in an incremental way, expanding their content may be easily deployed in practical scenarios. On the other hand, if such systems are to be implemented on board edge devices, such as imaging satellites, deep learning ensembles can become challenging to run in this constrained execution environment, especially with the on-board memory and energy limitations. However, there are studies showing the advantages of bringing deep models on board satellites (Wijata et al., 2023), also for rapid CD tasks (Serief et al., 2023) (thus, such techniques might be applicable in real-time and near real-time applications). Finally, as already mentioned, there exist approaches for compressing and optimizing large-capacity learners for their further deployment in heavily constrained environments (Dantas et al., 2024). These methods may be easily

incorporated into the deployment chain, and could be utilized for compressing both individual deep learning algorithms, as well as ensemble classifiers (Holliday et al., 2017). They include, among others, knowledge distillation (Li et al., 2021; Okamoto et al., 2022) into a carefully designed architecture which would fit the specific characteristics of the flight hardware that has been already utilized in orbit. Also, inference of separate models can be parallelized, if there are available processing units, as such predictions are independent of each other. Additionally, since deep learning models are still perceived as "black boxes", introducing explainable artificial intelligence (XAI) techniques to interpret them (but also to improve them, through understanding their internals (Zaigrajew et al., 2024)) have been attracting research attention (Hall et al., 2022; Gevaert, 2022). This is an important research avenue which will help the community build more "trust" in data-driven, deep learning-powered techniques, especially in critical downstream applications (Wijata et al., 2023).

4. Conclusions

Change detection is one of the most important tasks in practical Earth observation, as it allows us to precisely track the changes within a particular area of interest. This article tackled this problem and introduced ensemble learning techniques for change detection in Sentinel-2 time series image data which benefit from various deep learning architectures whose separate predictions are collectively used, in a weighted voting scheme, to elaborate the final prediction and to compensate for low-quality individual models in such combined systems, effectively leading to higher-quality CD. The experimental study, performed over the ONERA Satellite Change Detection dataset, as well as the new MUMUCD benchmark, using an array of quality metrics, showed that the proposed ensemble method improves the performance by the base models, and helps in compensating the performance of low-quality (underperforming) models included in an ensemble. By leveraging the strengths of multiple base learners, the approach reduced both false positives and false negatives, resulting in more robust change maps. This is particularly important in real-world applications, where precise delineation of changes may play a key role in decision-making processes, e.g., in forestry, disaster quantification and recovery, precision agriculture, urbanization tracking and analysis, and in many other downstream use cases and Earth observation domains (Afaq and Manocha, 2021; Ning et al., 2024).

There are interesting avenues for the future work that emerge from our study. First, the base learners could be further optimized in order to enhance the capabilities of the combined classifiers. Also, other deep models can be easily appended to the ensembles—other architectures might capture other change characteristics, ultimately leading to better CD. Our current efforts are focused on building resource-frugal models that could be seamlessly deployed on board satellites equipped with artificial intelligence capabilities (Ziaja et al., 2021). It would reduce the need of downlinking large amounts of EO image data for the on-the-ground analysis, and thus to make the analysis process faster, scalable and cheaper or even feasible, if the downlink of highly-dimensional imagery is prohibitively costly. Finally, exploiting high-resolution images might offer more precise change detection, ideally coupled with the additional explainable artificial intelligence techniques to interpret potentially "black-box" deep learning models. Utilizing such methods, together with the additional uncertainty quantification, could help in building more trust of the community in data-driven algorithms deployed for practical Earth observation in emerging use cases. It, in turn, may affect a wide range of real-world applications, spanning across shaping the environmental policies, land use planning, quantifying and responding to natural disasters, and many more.

CRediT authorship contribution statement

Ewa Kopec: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Data curation. **Agata M. Wijata:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Jakub Nalepa:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Silesian University of Technology, Poland grant for maintaining and developing research potential. AMW and JN were supported by the European Funds for Silesia 2021–2027 Program co-financed by the Just Transition Fund—project entitled "Development of the Silesian biomedical engineering potential in the face of the challenges of the digital and green economy (BioMeDiG)" (project number: FESL.10.25-IZ.01-07G5/23). The authors thank the anonymous referees for their insightful comments that improved the quality of the manuscript.

Data availability

The data is publicly available through https://rcdaudt.github.io/oscd/ and https://zenodo.org/records/10674011.

References

Afaq, Y., Manocha, A., 2021. Analysis on change detection techniques for remote sensing applications: A review. Ecol. Informatics 63, 101310. http://dx.doi.org/10.1016/j.ecoinf.2021.101310, URL https://www.sciencedirect.com/science/article/pii/S1574954121001011.

Alshehri, M., Ouadou, A., Scott, G., 2023. Deforestation detection in the Brazilian Amazon using transformer-based networks. In: 2023 IEEE Conference on Artificial Intelligence. CAI, pp. 292–293. http://dx.doi.org/10.1109/CAI54212.2023.00130.

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Comput. Linguist. 34 (4), 555–596. http://dx.doi.org/10.1162/coli.07-034-R2.

Asokan, A., Anitha, J., 2019. Change detection techniques for remote sensing applications: a survey. Earth Sci. Informatics 12 (2), 143–160. http://dx.doi.org/10.1007/s12145-019-00380-5.

Aszkowski, P., Kraft, M., 2024. Streamlining crop segmentation with multispectral imaging and foundation models: Minimizing manual annotation. In: 2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing. ICCP, pp. 1–8. http://dx.doi.org/10.1109/ICCP63557.2024.10793002.

Bai, T., Wang, L., Yin, D., Sun, K., Chen, Y., Li, W., Li, D., 2023. Deep learning for change detection in remote sensing: a review. Geo-Spatial Inf. Sci. 26 (3), 262–288. http://dx.doi.org/10.1080/10095020.2022.2085633.

Basavaraju, K.S., Sravya, N., Lal, S., Nalini, J., Reddy, C.S., Dell'Acqua, F., 2022. UCDNet: A deep learning model for urban change detection from bi-temporal multispectral Sentinel-2 satellite images. IEEE Trans. Geosci. Remote Sens. 60, 1–10. http://dx.doi.org/10.1109/TGRS.2022.3161337.

Bhatt, A., Ghosh, S.K., Kumar, A., 2015. Automated change detection in satellite images using machine learning algorithms for Delhi, India. In: 2015 IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 1678–1681. http://dx.doi.org/10.1109/IGARSS.2015.7326109.

Bovolo, F., Bruzzone, L., 2007. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Trans. Geosci. Remote Sens. 45 (1), 218–236. http://dx.doi.org/10.1109/TGRS.2006.885408.

Bragagnolo, L., da Silva, R., Grzybowski, J., 2021. Amazon forest cover change mapping based on semantic segmentation by U-Nets. Ecol. Informatics 62, 101279. http://dx.doi.org/10.1016/j.ecoinf.2021.101279, URL https://www.sciencedirect.com/science/article/pii/S1574954121000704.

Brigot, G., Colin-Koeniguer, E., Plyer, A., Janez, F., 2016. Adaptation and evaluation of an optical flow method applied to coregistration of forest remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 9 (7), 2923–2939. http://dx.doi.org/10.1109/JSTARS.2016.2578362.

Camalan, S., Cui, K., Pauca, V.P., Alqahtani, S., Silman, M., Chan, R., Plemmons, R.J., Dethier, E.N., Fernandez, L.E., Lutz, D.A., 2022. Change detection of amazonian alluvial gold mining using deep learning and Sentinel-2 imagery. Remote. Sens. 14 (7), http://dx.doi.org/10.3390/rs14071746, URL https://www.mdpi.com/2072-4292/14/7/1746.

Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: 2018 25th IEEE International Conference on Image Processing. ICIP, pp. 4063–4067. http://dx.doi.org/10.1109/ICIP.2018.8451652.

Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2019. OSCD - onera satellite change detection. URL https://doi.org/10.21227/asqe-7s69.

Chafiq, T., Hmamou, M., Ouhammou, I., Azmi, R., Kumar, M., 2024. Modelling change detection for unveiling urban transitions: using machine learning algorithms and Sentinel-2 data in Larache City, Morocco. Model. Earth Syst. Environ. 10 (2), 1711–1725. http://dx.doi.org/10.1007/s40808-023-01860-w.

Chaminda Bandara, W.G., Patel, V.M., 2025. Deep metric learning for unsupervised remote sensing change detection. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 5125–5135. http://dx.doi.org/10.1109/WACV61041.2025.00501.

Chen, Y., Bruzzone, L., 2022. A self-supervised approach to pixel-level change detection in bi-temporal RS images. IEEE Trans. Geosci. Remote Sens. 60, 1–11. http://dx.doi.org/10.1109/TGRS.2022.3203897.

Chen, Y., Ouyang, X., Agam, G., 2019. ChangeNet: Learning to detect changes in satellite images. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery. GeoAI '19, Association for Computing Machinery, New York, NY, USA, pp. 24–31. http://dx.doi.org/10.1145/3356471.3365232.

Chughtai, A.H., Abbasi, H., Karas, I.R., 2021. A review on change detection method and accuracy assessment for land use land cover. Remote. Sens. Appl.: Soc. Environ. 22, 100482. http://dx.doi.org/10.1016/j.rsase.2021.100482, URL https://www.sciencedirect.com/science/article/pii/S2352938521000185.

Dahiya, N., Singh, G., Gupta, D.K., Kalogeropoulos, K., Detsikas, S.E., Petropoulos, G.P., Singh, S., Sood, V., 2024. A novel deep learning change detection approach for estimating spatiotemporal crop field variations from Sentinel-2 imagery. Remote. Sens. Appl.: Soc. Environ. 35, 101259. http://dx.doi.org/10.1016/j.rsase.2024.101259, URL https://www.sciencedirect.com/science/article/pii/S235293852400123X.

Dantas, P.V., Sabino da Silva, W., Cordeiro, L.C., Carvalho, C.B., 2024. A comprehensive review of model compression techniques in machine learning. Appl. Intell. 54 (22), 11804–11844. http://dx.doi.org/10.1007/s10489-024-05747-w.

Daudt, R.C., 2020. Convolutional Neural Networks for Change Analysis in Earth Observation Images with Noisy Labels and Domain Shifts (Ph.D. thesis). Polytechnique de Paris.

Daudt, R.C., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Urban change detection for multispectral earth observation using convolutional neural networks. In: IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 2115–2118. http://dx.doi.org/10.1109/IGARSS.2018.8518015.

de Albuquerque, A.O., de Carvalho, O.L.F., Silva, C.R.e., Luiz, A.S., de Bem, P.P., Gomes, R.A.T., Guimarães, R.F., Júnior, O.A.d., 2021. Dealing with clouds and seasonal changes for center pivot irrigation systems detection using instance segmentation in Sentinel-2 time series. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 8447–8457. http://dx.doi.org/10.1109/JSTARS.2021.3104726.

de Bem, P.P., de Carvalho Junior, O.A., Fontes Guimarães, R., Trancoso Gomes, R.A., 2020. Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks. Remote. Sens. 12 (6), http://dx.doi.org/10.3390/rs12060901, URL https://www.mdpi.com/2072-4292/12/6/901.

Delgado, X.-A., 2019. Why Cohen's Kappa should be avoided as performance measure in classification. PLOS ONE 14 (9), 1–26. http://dx.doi.org/10.1371/journal.pone.0222916.

Fakhri, F., Gkanatsios, I., 2021. Integration of Sentinel-1 and Sentinel-2 data for change detection: A case study in a war conflict area of Mosul city. Remote. Sens. Appl.: Soc. Environ. 22, 100505. http://dx.doi.org/10.1016/j.rsase.2021.100505, URL https://www.sciencedirect.com/science/article/pii/S2352938521000410.

Ganaie, M., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Eng. Appl. Artif. Intell. 115, 105151. http://dx.doi.org/10.1016/j.engappai.2022.105151, URL https://www.sciencedirect.com/science/article/pii/S095219762200269X.

Gang Chen, L.M.T.C., Wulder, M.A., 2012. Object-based change detection. Int. J. Remote Sens. 33 (14), 4434–4457. http://dx.doi.org/10.1080/01431161.2011.648285.

Gevaert, C.M., 2022. Explainable AI for Earth observation: A review including societal and regulatory perspectives. Int. J. Appl. Earth Obs. Geoinf. 112, 102869. http://dx.doi.org/10.1016/j.jag.2022.102869, URL https://www.sciencedirect.com/science/article/pii/S1569843222000711.

Ghosh, A., Mishra, N.S., Ghosh, S., 2011. Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. Inform. Sci. 181 (4), 699–715. http://dx.doi.org/10.1016/j.ins.2010.10.016, URL https://www.sciencedirect.com/science/article/pii/S0020025510005153.

Gong, M., Zhan, T., Zhang, P., Miao, Q., 2017. Superpixel-based difference representation learning for change detection in multispectral remote sensing images. IEEE Trans. Geosci. Remote Sens. 55 (5), 2658–2673. http://dx.doi.org/10.1109/TGRS.2017.2650198.

González, S., García, S., Del Ser, J., Rokach, L., Herrera, F., 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Inf. Fusion 64, 205–237. http://dx.doi.org/10.1016/j.inffus.2020.07.007, URL https://www.sciencedirect.com/science/article/pii/S1566253520303195.

Grabowski, B., Ziaja, M., Kawulok, M., Bosowski, P., Longépé, N., Le Saux, B., Nalepa, J., 2024. Squeezing adaptive deep learning methods with knowledge distillation for on-board cloud detection. Eng. Appl. Artif. Intell. 132, 107835. http://dx.doi.org/10.1016/j.engappai.2023.107835, URL https://www.sciencedirect.com/science/article/pii/S0952197623020195.

Hall, O., Ohlsson, M., Rögnvaldsson, T., 2022. A review of explainable AI in the satellite data, deep machine learning, and human poverty domain. Patterns 3 (10), 100600. http://dx.doi.org/10.1016/j.patter.2022.100600, URL https://www.sciencedirect.com/science/article/pii/S2666389922002252.

He, Y., Chen, H., Zhu, Q., Zhang, Q., Zhang, L., Liu, T., Li, W., Chen, H., 2025. A heterogeneous ensemble learning method combining spectral, terrain, and texture features for landslide mapping. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 18, 3746–3765. http://dx.doi.org/10.1109/JSTARS.2025.3525633.

He, H., Yan, J., Liang, D., Sun, Z., Li, J., Wang, L., 2024. Time-series land cover change detection using deep learning-based temporal semantic segmentation. Remote Sens. Environ. 305, 114101. http://dx.doi.org/10.1016/j.rse.2024.114101, URL https://www.sciencedirect.com/science/article/pii/S0034425724001123.

Heywood, M.I., 2024. Evolutionary ensemble learning. In: Banzhaf, W., Machado, P., Zhang, M. (Eds.), Handbook of Evolutionary Machine Learning. Springer Nature Singapore, Singapore, pp. 205–243. http://dx.doi.org/10.1007/978-981-99-3814-8_8.

Holliday, A., Barekatain, M., Laurmaa, J., Kandaswamy, C., Prendinger, H., 2017. Speedup of deep learning ensembles for semantic segmentation using a model compression technique. Comput. Vis. Image Underst. 164, 16–26, Deep Learning for Computer Vision. http://dx.doi.org/10.1016/j.cviu.2017.05.004. URL https://www.sciencedirect.com/science/article/pii/S1077314217300826.

Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: From pixel-based to object-based approaches. ISPRS J. Photogramm. Remote Sens. 80, 91–106. http://dx.doi.org/10.1016/j.isprsjprs.2013.03.006, URL https://www.sciencedirect.com/science/article/pii/S0924271613000804.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211. http://dx.doi.org/10.1038/s41592-020-01008-z.

Iyer, P., A., S., Lal, S., 2021. Deep learning ensemble method for classification of satellite hyperspectral images. Remote. Sens. Appl.: Soc. Environ. 23, 100580. http://dx.doi.org/10.1016/j.rsase.2021.100580, URL https://www.sciencedirect.com/science/article/pii/S2352938521001166.

Jeevan, A., Shanthi, S.A., 2024. Remote sensing revolution- A critical analysis of change detection on satellite images. In: 2024 1st International Conference on Trends in Engineering Systems and Technologies. ICTEST, pp. 1–6. http://dx.doi.org/10.1109/ICTEST60614.2024.10576102.

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4 (9), 100804. http://dx.doi.org/10.1016/j.patter.2023.100804, URL https://www.sciencedirect.com/science/article/pii/S2666389923001599.

Kaur Buttar, P., Sachan, M.K., 2022. Semantic segmentation of clouds in satellite images based on U-Net++ architecture and attention mechanism. Expert Syst. Appl. 209, 118380. http://dx.doi.org/10.1016/j.eswa.2022.118380, URL https://www.sciencedirect.com/science/article/pii/S0957417422014932.

Khan, A.A., Chaudhari, O., Chandra, R., 2024. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Syst. Appl. 244, 122778. http://dx.doi.org/10.1016/j.eswa.2023.122778, URL https://www.sciencedirect.com/science/article/pii/S0957417423032803.

Kheddam, R., Tahraoui, A., 2024. Unsupervised multiclass change detection and mapping using deep neural network. In: 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing. ATSIP, 1, pp. 290–295. http://dx.doi.org/10.1109/ATSIP62566.2024.10638894.

Kucharski, D., Wijata, A.M., Fu, L., Lin, W., Xue, Y., Kawa, J., Zheng, Y., Gregory Lip, Y.H., Nalepa, J., 2024. Giraffe: A genetic programming algorithm to build deep learning ensembles for ecg arrhythmia classification. In: 2024 IEEE International Conference on Image Processing. ICIP, pp. 3070–3076. http://dx.doi.org/10.1109/ICIP51287.2024.10647780.

Kuck, T.N., Silva Filho, P.F.F., Sano, E.E., Bispo, P.d.C., Shiguemori, E.H., Dalagnol, R., 2021. Change detection of selective logging in the Brazilian Amazon using X-Band SAR data and pre-trained convolutional neural networks. Remote. Sens. 13 (23), http://dx.doi.org/10.3390/rs13234944, URL https://www.mdpi.com/2072-4292/13/23/4944.

Lee, Y.-S., Lee, S., Jung, H.-S., 2020. Mapping forest vertical structure in Gong-ju, Korea using Sentinel-2 satellite images and artificial neural networks. Appl. Sci. 10 (5), http://dx.doi.org/10.3390/app10051666, URL https://www.mdpi.com/2076-3417/10/5/1666.

Leenstra, M., Marcos, D., Bovolo, F., Tuia, D., 2021. Self-supervised pre-training enhances change detection in Sentinel-2 imagery. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII. Springer-Verlag, Berlin, Heidelberg, pp. 578–590. http://dx.doi.org/10.1007/978-3-030-68787-8_42.

Li, X., Xiong, H., Chen, Z., Huan, J., Xu, C.-Z., Dou, D., 2021. "In-network ensemble": Deep ensemble learning with diversified knowledge distillation. ACM Trans. Intell. Syst. Technol. 12 (5), http://dx.doi.org/10.1145/3473464.

Lim, K., Jin, D., Kim, C.-S., 2018. Change detection in high resolution satellite images using an ensemble of convolutional neural networks. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 509–515. http://dx.doi.org/10.23919/APSIPA.2018.8659603.

Liu, T., Yang, L., Lunga, D., 2021. Change detection using deep learning approach with object-based image analysis. Remote Sens. Environ. 256, 112308. http://dx.doi.org/10.1016/j.rse.2021.112308, URL https://www.sciencedirect.com/science/article/pii/S0034425721000262.

Mao, Y., He, Q., Li, J., Yang, B., 2024. TACD: A novel 3-D swin transformer with enhanced feature aggregation for change detection in image time series. IEEE Trans. Geosci. Remote Sens. 62, 1–16. http://dx.doi.org/10.1109/TGRS.2024.3496994.

Maurya, A., Akashdeep, Mittal, P., Kumar, R., 2023. A modified U-net-based architecture for segmentation of satellite images on a novel dataset. Ecol. Informatics 75, 102078. http://dx.doi.org/10.1016/j.ecoinf.2023.102078, URL https://www.sciencedirect.com/science/article/pii/S1574954123001073.

Mertes, C., Schneider, A., Sulla-Menashe, D., Tatem, A., Tan, B., 2015. Detecting change in urban areas at continental scales with MODIS data. Remote Sens. Environ. 158, 331–347. http://dx.doi.org/10.1016/j.rse.2014.09.023, URL https://www.sciencedirect.com/science/article/pii/S003442571400368X.

Mulverhill, C., Coops, N.C., Achim, A., 2023. Continuous monitoring and sub-annual change detection in high-latitude forests using Harmonized Landsat Sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 197, 309–319. http://dx.doi.org/10.1016/j.isprsjprs.2023.02.002, URL https://www.sciencedirect.com/science/article/pii/S0924271623000424.

Nalepa, J., Myller, M., Cwiek, M., Zak, L., Lakota, T., Tulczyjew, L., Kawulok, M., 2021a. Towards on-board hyperspectral satellite image segmentation: Understanding robustness of deep learning through simulating acquisition conditions. Remote. Sens. 13 (8), http://dx.doi.org/10.3390/rs13081532, URL https://www.mdpi.com/2072-4292/13/8/1532.

Nalepa, J., Myller, M., Tulczyjew, L., Kawulok, M., 2021b. Deep ensembles for hyperspectral image data classification and unmixing. Remote. Sens. 13 (20), http://dx.doi.org/10.3390/rs13204133, URL https://www.mdpi.com/2072-4292/13/20/4133.

Ning, X., Zhang, H., Zhang, R., Huang, X., 2024. Multi-stage progressive change detection on high resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 207, 231–244. http://dx.doi.org/10.1016/j.isprsjprs.2023.11.023, URL https://www.sciencedirect.com/science/article/pii/S0924271623003404.

Okamoto, N., Hirakawa, T., Yamashita, T., Fujiyoshi, H., 2022. Deep ensemble learning by diverse knowledge distillation for fine-grained object classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022. Springer Nature Switzerland, Cham, pp. 502–518.

Papadomanolaki, M., Vakalopoulou, M., Karantzalos, K., 2021. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. IEEE Trans. Geosci. Remote Sens. 59 (9), 7651–7668. http://dx.doi.org/10.1109/TGRS.2021.3055584.

Pegia, M.E., Moumtzidou, A., Gialampoukidis, I., Jónsson, B.T., Vrochidis, S., Kompatsiaris, I., 2022. BiasUNet: Learning change detection over Sentinel-2 image pairs. In: Proceedings of the 19th International Conference on Content-Based Multimedia Indexing. CBMI '22, Association for Computing Machinery, New York, NY, USA, pp. 142–148. http://dx.doi.org/10.1145/3549555.3549574.

Peng, D., Liu, X., Zhang, Y., Guan, H., Li, Y., Bruzzone, L., 2025. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges. Int. J. Appl. Earth Obs. Geoinf. 136, 104282. http://dx.doi.org/10.1016/j.jag.2024.104282, URL https://www.sciencedirect.com/science/article/pii/S1569843224006381.

Pomente, A., Picchiani, M., Del Frate, F., 2018. Sentinel-2 change detection based on deep features. In: IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 6859–6862. http://dx.doi.org/10.1109/IGARSS.2018.8519195.

Prasad, J., Sreelatha, M., SuvarnaVani, K., 2023. V-BANet: Land cover change detection using effective deep learning technique. Ecol. Informatics 75, 102019. http://dx.doi.org/10.1016/j.ecoinf.2023.102019, URL https://www.sciencedirect.com/science/article/pii/S1574954123000481.

Qiqi Zhu, Z.L., Li, D., 2024. A review of multi-class change detection for satellite remote sensing imagery. Geo-Spatial Inf. Sci. 27 (1), 1–15. http://dx.doi.org/10.1080/10095020.2022.2128902.

Reddy, P.L., Pawar, S., 2020. Multispectral image denoising methods: A literature review. Mater. Today: Proc. 33, 4666–4670, International Conference on Nanotechnology: Ideas, Innovation and Industries. http://dx.doi.org/10.1016/j.matpr.2020.08.313. URL https://www.sciencedirect.com/science/article/pii/S221478532036185X.

Rincy, T.N., Gupta, R., 2020. Ensemble learning techniques and its efficiency in machine learning: A survey. In: 2nd International Conference on Data, Engineering and Applications. IDEA, pp. 1–6. http://dx.doi.org/10.1109/IDEA49133.2020.9170675.

Roy, S.K., Jamali, A., Chanussot, J., Ghamisi, P., Ghaderpour, E., Shahabi, H., 2025. SimPoolFormer: A two-stream vision transformer for hyperspectral image classification. Remote. Sens. Appl.: Soc. Environ. 37, 101478. http://dx.doi.org/10.1016/j.rsase.2025.101478, URL https://www.sciencedirect.com/science/article/pii/S235293852500031X.

Sadel, J., Tulczyjew, L., Wijata, A.M., Przeliorz, M., Nalepa, J., 2025. Monitoring forest changes with foundation models and Sentinel-2 time series. IEEE Geosci. Remote. Sens. Lett. 22, 1–5. http://dx.doi.org/10.1109/LGRS.2025.3556601.

Santos, L.A., Ferreira, K.R., Camara, G., Picoli, M.C., Simoes, R.E., 2021. Quality control and class noise reduction of satellite image time series. ISPRS J. Photogramm. Remote Sens. 177, 75–88. http://dx.doi.org/10.1016/j.isprsjprs.2021.04.014, URL https://www.sciencedirect.com/science/article/pii/S0924271621001155.

Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10 (5), http://dx.doi.org/10.3390/agronomy10050641, URL https://www.mdpi.com/2073-4395/10/5/641.

Serief, C., Ghelamallah, Y., Bentoutou, Y., 2023. Deep-learning-based system for change detection onboard earth observation small satellites. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 8115–8124. http://dx.doi.org/10.1109/JSTARS.2023.3284919.

Serva, F., Sebastianelli, A., Saux, B.L., Ricciuti, F., 2025. MUMUCD: A multi-modal multi-class change detection dataset. IEEE Geosci. Remote. Sens. Lett. 1. http://dx.doi.org/10.1109/LGRS.2025.3585797.

Silva, C.A., Guerrisi, G., Frate, F.D., and, E.E.S., 2022. Near-real time deforestation detection in the Brazilian Amazon with Sentinel-1 and neural networks. Eur. J. Remote. Sens. 55 (1), 129–149. http://dx.doi.org/10.1080/22797254.2021.2025154.

Singh, A., 1989. Review Article Digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. 10 (6), 989–1003. http://dx.doi.org/10.1080/01431168908903939.

Somvanshi, M., Chavan, P., Tambade, S., Shinde, S.V., 2016. A review of machine learning techniques using decision tree and support vector machine. In: 2016 International Conference on Computing Communication Control and Automation. ICCUBEA, pp. 1–7. http://dx.doi.org/10.1109/ICCUBEA.2016.7860040.

Song, K., Jiang, J., 2021. AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 4816–4831. http://dx.doi.org/10.1109/JSTARS.2021.3077545.

Song, L., Xia, M., Weng, L., Lin, H., Qian, M., Chen, B., 2023. Axial cross attention meets CNN: Bibranch fusion network for change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16, 21–32. http://dx.doi.org/10.1109/JSTARS.2022.3224081.

Stofa, M.M., Abdani, S.R., Moubark, A.M., Zainuri, M.A.A.M., Ibrahim, A.A., Kamari, N.A.M., Zulkifley, M.A., 2025. Recent developments of artificial intelligence methods for sea ice concentration monitoring using high-resolution imaging datasets. Ecol. Informatics 87, 103132. http://dx.doi.org/10.1016/j.ecoinf.2025.103132, URL https://www.sciencedirect.com/science/article/pii/S1574954125001414.

Suresh, S., M., R.R., C.S., A., Dell'Acqua, F., 2024. RDC-UNet++: An end-to-end network for multispectral satellite image enhancement. Remote. Sens. Appl.: Soc. Environ. 36, 101293. http://dx.doi.org/10.1016/j.rsase.2024.101293, URL https://www.sciencedirect.com/science/article/pii/S2352938524001575.

Wang, D., Hu, M., Jin, Y., Miao, Y., Yang, J., Xu, Y., Qin, X., Ma, J., Sun, L., Li, C., Fu, C., Chen, H., Han, C., Yokoya, N., Zhang, J., Xu, M., Liu, L., Zhang, L., Wu, C., Du, B., Tao, D., Zhang, L., 2025. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 47 (8), 6427–6444. http://dx.doi.org/10.1109/TPAMI.2025.3557581.

Wang, X., Jing, S., Dai, H., Shi, A., 2023. High-resolution remote sensing images semantic segmentation using improved UNet and SegNet. Comput. Electr. Eng. 108, 108734. http://dx.doi.org/10.1016/j.compeleceng.2023.108734, URL https://www.sciencedirect.com/science/article/pii/S0045790623001581.

Wijata, A.M., Foulon, M.-F., Bobichon, Y., Vitulli, R., Celesti, M., Camarero, R., Di Cosimo, G., Gascon, F., Longépé, N., Nieke, J., Gumiela, M., Nalepa, J., 2023. Taking artificial intelligence into space through objective selection of hyperspectral Earth observation applications: To bring the "brain" close to the "eyes" of satellite missions. IEEE Geosci. Remote. Sens. Mag. 11 (2), 10–39. http://dx.doi.org/10.1109/MGRS.2023.3269979.

Yang, Y., Lv, H., Chen, N., 2023. A survey on ensemble learning under the era of deep learning. Artif. Intell. Rev. 56 (6), 5545–5589. http://dx.doi.org/10.1007/s10462-022-10283-5.

Yao, H., Zhu, D.-l., Jiang, B., Yu, P., 2020. Negative log likelihood ratio loss for deep neural network classification. In: Arai, K., Bhatia, R., Kapoor, S. (Eds.), Proceedings of the Future Technologies Conference (FTC) 2019. Springer International Publishing, Cham, pp. 276–282.

Ye, Y., Wang, M., Zhou, L., Lei, G., Fan, J., Qin, Y., 2023. Adjacent-level feature cross-fusion with 3-D CNN for remote sensing image change detection. IEEE Trans. Geosci. Remote Sens. 61, 1–14. http://dx.doi.org/10.1109/TGRS.2023.3305499.

You, Y., Cao, J., Zhou, W., 2020. A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. Remote. Sens. 12 (15), http://dx.doi.org/10.3390/rs12152460, URL https://www.mdpi.com/2072-4292/12/15/2460.

Zaigrajew, V., Baniecki, H., Tulczyjew, L., Wijata, A.M., Nalepa, J., Longépé, N., Biecek, P., 2024. Red teaming models for hyperspectral image analysis using explainable AI. CoRR http://dx.doi.org/10.48550/ARXIV.2403.08017, arXiv:2403.08017.

Zhang, C., Wang, L., Cheng, S., Li, Y., 2022. SwinSUNet: Pure transformer network for remote sensing image change detection. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2022.3160007.

Zhao, F., Gong, W., Bianchini, S., Yang, Z., 2024. Linking glacier retreat with climate change on the Tibetan Plateau through satellite remote sensing. Cryosphere 18 (12), 5595–5612. http://dx.doi.org/10.5194/tc-18-5595-2024, URL https://tc.copernicus.org/articles/18/5595/2024/.

Zhao, K., Wulder, M.A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., Brown, M., 2019. Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm. Remote Sens. Environ. 232, 111181. http://dx.doi.org/10.1016/j.rse.2019.04.034, URL https://www.sciencedirect.com/science/article/pii/S0034425719301853.

Zhou, W., Troy, A., Grove, M., 2008. A comparison of object-based with pixel-based land cover change detection in the Baltimore Metropolitan area using multitemporal high resolution remote sensing data. In: IGARSS 2008 IEEE International Geoscience and Remote Sensing Symposium, vol. 4. pp. IV – 683–IV – 686. http://dx.doi.org/10.1109/IGARSS.2008.4779814.

Zhu, Q., Guo, X., Li, Z., Li, D., 2024. A review of multi-class change detection for satellite remote sensing imagery. Geo-Spatial Inf. Sci. 27 (1), 1–15. http://dx.doi.org/10.1080/10095020.2022.2128902.

Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171. http://dx.doi.org/10.1016/j.rse.2014.01.011, URL https://www.sciencedirect.com/science/article/pii/S0034425714000248.

Ziaja, M., Bosowski, P., Myller, M., Gajoch, G., Gumiela, M., Protich, J., Borda, K., Jayaraman, D., Dividino, R., Nalepa, J., 2021. Benchmarking deep learning for on-board space applications. Remote. Sens. 13 (19), http://dx.doi.org/10.3390/rs13193981, URL https://www.mdpi.com/2072-4292/13/19/3981.