Earth Science Informatics
https://doi.org/10.1007/s12145-023-01190-6

RESEARCH

A deep learning-based technique for firm classification and domain adaptation in land cover classification using time-series aerial images

Indrajit Kalita[1] · Shounak Chakraborty[2] · Talla Giridhara Ganesh Reddy[3] · Moumita Roy[4]

Received: 1 September 2023 / Accepted: 8 December 2023

Abstract

In this manuscript, a novel framework has been presented for firm classification of a geographical area based on spatial as well as time-series analysis of multi-temporal very high resolution (VHR) satellite images. For this dual objective, an attention-based deep learning mechanism combined with the capabilities of convolutional-recurrent neural networks has been investigated for this purpose. The proposed classification strategy is introduced as 'firm' since it allows the classifier to assign only one class label to a multi-temporal image stack of co-registered images, as opposed to multiple. This technique ascertains the land-cover class by taking into consideration the geophysical changes on a landmass and thus outsmarting the conventional techniques relying on the visual interpretation of a single image. The attention mechanism focuses on the important portions of the image scene while the convolutional long short-term memory neural networks exploit the temporal dependencies on the time-series image scenes. Moreover, an adaptive land cover classification scheme, considering the features extracted from the proposed classification approach has been explored for more robust time-series based firm classification. To assess the performance of the proposed schemes, the experiments have been conducted on the two novels VHR multi-temporal land cover classification datasets. The investigated models have been shown to have the capacity to outperform the other state-of-the-art techniques under non-adaptive as well as adaptive scenarios using the multi-temporal images captured over disjoint geographical locations.

Keywords Time-series images · Attention mechanism · Convolutional LSTMs · Domain adaptation · Multi-temporal datasets

Introduction

The recent advancement in the remote sensing machinery presents an extensive scope for the development of interesting applications in Earth observation (EO) such as forest and crop monitoring, thermal mapping, land cover map generation, biogeochemical cycling, biodiversity to name a few (Meher et al. 2007). However, central to these EO activities is the land cover classification (LCC) which deals with the annotation of high-resolution images/pixels captured daily by the remote sensing instruments. To avoid the tedious process of manual annotation, machine learning-based approaches have been thoroughly investigated in the previous decade as an identification of multiple land cover objects through land cover classification (LCC) (Camps-Valls et al. 2008). Typically, these techniques exploit the spectral and spatial information from the remotely sensed satellite images to identify the land cover objects. In terms of machine learning, most of these approaches are supervised ones based on support vector machines (SVMs) (Ashourloo et al. 2018), kernel-based methods (Camps-Valls et al. 2008), and artificial neural networks (ANNs) (Chakraborty and Roy 2018) which require an extensive amount of labeled information for training (Imani and Ghassemian 2015). Furthermore, the inability to extract features automatically and subsequent classification of very high-resolution images have driven the researchers to use deep learning (DL) based on modern approaches (Postadjian et al. 2018). The objective of the technique investigated in this manuscript is two-folded (a) to firmly annotate a geographical region based on the satellite images taken on multiple (seasonally discrete) time-stamps, (b) to develop an adaptive

Communicated by: H. Babaie

T. Reddy and S. Chakraborty contributed equally to the manuscript (the joint second author).

✉ Indrajit Kalita
indrajit@bu.edu

Extended author information available on the last page of the article

Published online: 19 December 2023

LCC technique considering cross-domain multi-temporal images. The consequent paragraphs comprehensively bring out the motivation for each of these two problem statements.

In literature, sophisticated approaches have been investigated for land-cover classification using the spatial and spectral information of the remotely sensed satellite images (Postadjian et al. 2018; Othman et al. 2017). These classification techniques, however, try to mimic only the visual interpretation capacity of human beings on observation of a single image. In practicality, the land cover classes can be firmly ascertained only by analyzing the seasonal changes of a geographical region. In this way, an agricultural area (for example) can be identified despite its occasional changes to barren land in the one dry season. Deviating from the traditional LCC techniques, the proposed scheme explores a possibility where a classifier can outsmart a human annotator in scene classification using the information available from multi-temporal satellite images collected over multiple timestamps. As this classification paradigm analyses the time-series information corresponding to a geographical area to assign a major (singular) class label irrespective of its off-seasonal (minor) changes, it can be called a "firm" classification. More technically, the temporal characteristics are considered along with the conventional spectra-spatial features to develop an LCC technique which is later made adaptable to classify time-series geo-images across diverse domains. Such precise land cover classification mechanism considering the geographical changes (owing to natural phenomena or human activities) across multiple time frames has seldom been investigated in the literature mostly due to the unavailability of data and the underlying costs of collecting them until the recent times (Rußwurm and Körner 2018). As per the knowledge of the authors, only Zhu et al. 2021 have exploited the temporal information in updation of the land cover map through repeated retraining of the classifier on the acquisition of an updated land image. Instead of this exhaustive retraining, the technique presented here analyses the past changes in the land surface to assign a firm classification to the geographical region. Further, the scheme has also been made adaptive to cross-classify a new region having diverse class distribution as compared to the training images. The recent advancement of the remote sensing machinery owing to diverse data sources like USGS Earth Explorer, Sentinel Open Access Hub, NASA Earthdata Search, Google Earth, and many others provide the scope for such firm land cover classification of a geographical area by extracting spatial-spectral-temporal features from multiple images.

The methodology presented here has dual objectives for annotating images considering their changing dynamics as well as to develop an adaptive classification technique to handle trans-domain classifications. The latter stems from a problem frequent in a supervised classification where there is enough number of labeled temporal images in a known region (source domain). Simultaneously, there is a severe scarcity of the same in an unknown region (target domain); further, the difference in class distribution across the two domains limits the direct use of the source images as the training set for classification of the target images. Under this scenario, a classifier trained using available information from the source domain has to be made adaptive to generalize further over the images from the target domain. This process is well-known as transductive transfer learning (TTL) (Kalita and Roy 2020), or domain adaptation (DA) (Ammour et al. 2018; Tuia et al. 2016) in the literature. In such a situation, a classifier trained using the labeled temporal images from the source domain has to be adapted to be able to classify the temporal sequence of images from the target domain. Thus in addition to developing a generalized model over the temporal image sequences, the scheme is improvised in developing an adaptive one that can cross-classify temporal sequences from another region (domain) having related but different class-distribution with respect to the source.

In this manuscript, a novel attention-based mechanism has been introduced at the onset for identifying important regions from each of the time-series images. Thereafter, a hybrid of convolution and recurrent neural networks (specifically, the convolutional LSTM) has been investigated to identify the relation between the land cover scenes amongst the images in time. Further, a dense layers-based conventional classification mechanism has been incorporated to obtain the classes available in the time-series images. Though effective, the consequent investigations reveal that the technique fails when the training and test set of multi-temporal images follow different class distributions or in other words have a high divergence between them. As a solution, a domain adaptation (DA) technique has been further developed for reducing the cross-domain distribution differences amongst the features extracted out using the proposed mechanism consider the attention-based convolutional LSTM architecture. Finally, the DA technique investigated through a novel domain-neutralizer network performs a cross-domain (hetero-spatial) multi-temporal classification on a new intermediate feature space. The performance evaluation of both the multi-temporal classification technique and the DA methodology has been carried out through experimentation on two novel VHR datasets captured over the Indian sub-continent. In consolidation, the key novelties in the manuscript can be summarised in terms of the following aspects.

1. A new end-to-end firm land cover classification mechanism for geographical areas using multi-temporal images through hybrid integration of attention, CNN, and RNN deep learning technologies
2. A domain adaptation technique for multi-temporal LCC using a novel domain neutralizer network to reduce

the cross-domain distribution difference of the features extracted from the source-target time-series images
3. A solution to the data scarcity in multi-temporal LCC by providing two new easily accessible multi-temporal datasets captured over the Indian sub-continent using Google Earth freely available images.

The rest of the manuscript is organized with sections regarding related work, proposed methodology, experimental results, and conclusion.

Related works

The proposed work in this manuscript is centered around an LCC technique considering multi-temporal VHR images in non-adaptive as well as adaptive (cross-domain classification) scenarios. In this section, a concise literature survey has been presented for the establishment of the theoretical background of the proposed technology shouldering on the two aforementioned problem statements.

Land cover classification under pixel-level and scene-level categories

Traditionally, land cover classification (LCC) deals with annotating the remotely sensed images either on the pixel-level (Imani and Ghassemian 2015) or on the scene-level (Scott et al. 2017). The basic difference between these two is that the former category of the strategies assigns the class labels to each pixel (smallest unit) of an image; whereas, the latter deals with the assignment of class labels to a more abstract object or the image scene. At the pixel-level, the artificial neural networks, principal component analysis, and support vector machines (Haykin 2007) are familiar for the classification of remotely sensed image datasets. However, the advent of higher resolution images has facilitated the scene-level LCC techniques (Scott et al. 2017; Yang and Newsam 2010). Such techniques provide an alternative way to generate the land cover maps by extracting features at a higher level only from the entire scene. The literature on the stated research indicates that the scene-level classification is carried out by both manual and automatic feature extraction approaches (Postadjian et al. 2018; Yang and Newsam 2010; Liang et al. 2018). The classification has been carried out in the former category by utilizing the extracted features manually using bag-of-visual-words from the local patches of the images (Yang and Newsam 2010). However, the conventional techniques for feature extraction were soon replaced by deep neural networks (DNNs) based on automatic techniques. Among them, deep convolution neural network (DCNN) is convincingly the most popular one in LCC (Postadjian et al. 2018; Othman et al. 2017). Further, other DNNs like deep belief network (Lv et al. 2015), stacked autoencoder (Liang et al. 2018), and recurrent neural network (RNN) (Castro et al. 2018) have also been exploited for LCC. However, this manuscript is centered around an LCC approach using multi-temporal VHR images.

Land cover classification using multi-temporal images

Generally, the LCC techniques have been investigated on the spectral and spatial patterns collected manually (McClellan et al. 1989) or automatically (Scott et al. 2017) from a geographical area of interest. The community has realized the importance of temporal characteristics of the remotely sensed images only in crop classification where the seasonal crops were needed to be identified (Kussul et al. 2017). However, the investigated approach has been pixel-based analysis of the Landsat-8 and Sentinel-1A images using a hierarchical CNN-oriented (supervised) methodology. Another study (Lavreniuk et al. 2018) improves the same crop classification accuracy using SAR and Sentinel data captured over the same site on different timestamps. An autoencoder has been used for subspace alignment followed by fine-tuning of a CNN in a pixel-wise supervised fashion. Similarly, a recurrent neural network has been used for crop classification from time-sequential SAR data (Castro et al. 2018). Moreover, handcrafted physics-based features were extracted for temporal crop classification from Landsat images which have then been classified using SVM classifier (Ashourloo et al. 2018). On the same note, the phenotype cycles in crop cultivation have also been identified using spectral reflectance curve (Kim et al. 2018) and deep CNNs (Guo et al. 2018). A well-organized document for the mathematical formulations in multivariate time-series analysis has been presented by Zheng et al. (2014). Historically, the first record of such multi-temporal classification dates back to 2002 where Yang et al. (2002) quantified the loss of forest owing to the urban sprawl in Atlanta's accelerated spread. The time-series analysis of the satellite images caught its pace in 2017 with the availability of TiSeLaC dataset challenge (Ienco and Gaetano 2007). An end-to-end CNN model for extracting the spatio-temporal features from the TiSeLaC images has shown outstanding performance in this regard (Di Mauro et al. 2017). Summarising the contributions, it can be said that the sophisticated pixel-based implementations presented in the literature lack the generalization ability to exploit the large-scale spatio-temporal and spectral information that may be available from the scenes of the images of a VHR object-level dataset (mainly due to the unavailability of the

latter dataset). Further, the transfer-learning or adaptation of such approaches to cross-classify the images obtained from diverse domains is also yet to be considered.

Domain adaptation in land cover classification

Most of the existing LCC techniques are however supervised which perform well under certain conditions like (a) there are adequate high-quality samples in the training set and (b) the samples in the training and test set follow the same probability distribution. However, the primary bottleneck of automatic LCC techniques (both in pixel-level or scene-level) arises in the collection of new training samples for the classification of images captured from a new region (Ammour et al. 2018). Under such a scenario, a classifier trained using the training images from a known region (source domain) has to be adapted for the prediction of images from an unknown region (target domain), which is well-known as domain adaptation (DA) in the literature. In literature, the pixel-level DA techniques have been sub-divided into unsupervised, semi-supervised, and active learning-based approaches. In the case of unsupervised DA, the labeled samples are available only from the source region, whereas the samples from the target domain are unlabeled (Gopalan et al. 2014). However, by using a few labeled samples from the target domain (if available), along with all the labeled samples from the source domain, the semi-supervised scheme iteratively selects the 'most accurate' patterns from the unlabeled target samples (Chakraborty and Roy 2018; Bruzzone and Marconcini 2009). On the other hand, in the case of the active learning strategy, the supervisor effectively engages in the learning process by providing external labeling for some of the most informative samples selected from the target domain (Senthilnath et al. 2011).

Alternatively, in the scene-level, the DA problem can also be categorized viz. unsupervised, semi-supervised, and active learning-based strategy. Here, the unsupervised scheme follows the data transformation-based approaches to obtain the common subspace between the source and the target regions (Riz et al. 2016). Similarly, the semi-supervised approaches explore DA techniques by considering the limited amount of labeled samples from the target domain (Postadjian et al. 2018). In this regard, Postadjian et al. (2018) have trained a DCNN architecture with adequate labeled source data. Thereafter, this neural network architecture has been fine-tuned using a few target samples to obtain the final prediction for the rest of the unknown target samples. On the other hand, the active learning technique detects certain samples from the target area to be labeled manually (Kalita and Roy 2020). In this way, Kalita et al. (2020) explored a cost-effective active learning approach to obtain the label of few target samples using the supervisor. However, it is worth mentioning that the investigation in the manuscript is based on the unsupervised DA technique approaches using multi-temporal remotely sensed images.

Domain adaptation using multi-temporal images

Coincidentally, the interesting idea of designing adaptive classifiers made its way into the remote sensing community with the advent of the multi-temporal images captured over the same site. Under such a situation, the land cover maps generated by a supervised classifier had to be updated which mandated retraining using manually acquired ground truth collected afresh for the updated images. To save the human effort therein, Bruzzone et al. have proposed various parametric (Bruzzone and Prieto 2001) and non-parametric (Bruzzone and Marconcini 2009) SVM based solutions to adapt to the new reference image following an unsupervised paradigm (under pixel-level category). Similarly, a temporal adaptive SVM has been investigated (Guo et al. 2017) by twerking the kernel function to regularise over the updated temporal image. Further, a centroid-based alignment has been proposed to handle the class distribution shifts in temporal images (Zhu and Ma 2016). The problem with these pixel-based approaches, however, is the non-utilization of spatial information in the adaptation process. On the other hand, some of the recent scene-level approaches comprehend the information from the complete image scenes obtained from the VHR satellite imagery (Othman et al. 2016). Othman et al. (2016) investigated a three-layer convex network to solve the DA between the two multi-temporal images of the same region. In other work, Othman et al. (2016) implements the combination of extreme learning machine and scale-invariant feature transform approaches for the same multi-temporal dataset. However, these approaches only consider adaptation on two images taken in discrete time-stamps over the same region which is far from the practical scenario where multiple time-series images may be available across the discrete sites. The latter problem intensifies as the probability distribution significantly changes in such situations instead of being a mere class-wise shift.

The previous works of the authors (Chakraborty and Roy 2018; Kalita and Roy 2020) have focused on developing adaptive methods for LCC with low interclass and high intraclass variation across training-test sets from discriminate domains. A further investigation has shown that the inherent changing nature of image scenes can open up the possibility of assigning multiple labels to it (Chakraborty et al. 2020). To further avoid this non-deterministic classification, firm decisions can be provided for the image scenes only if inherent temporal changes can also be taken into account along with the spectra-spatial characteristics. The details of the proposed methodology can be found in the consequent section.

Proposed methodology

At the onset, an attention mechanism has been utilized to capture the important areas of the image, crucial for the identification of land cover types like farmlands, beaches, forests, urban areas, and water bodies. The attention mechanism employs image interpolation (Rukundo and Maharaj 2014) based encoding-decoding methodology to capture the distinct land-class from the input image. The resulting image is a focused image obtained by blurring the irrelevant portion of the input image. Thereafter, a long short-term memory (LSTM) network has been used to capture the time-sequence information from each of the focused temporal images. This enables the proposed classification model to now identify a land cover class across the time-series geophysical changes. This generalization mechanism allows the classifier to see through an agriculture area (for example) which might have turned into a barren land due to seasonal changes. The experimentation carried out on novel Indian sub-continent datasets suggests the applicability of such an attention-based classification model. Further, an adaptive scheme has been developed by using the proposed model as a feature extractor to allow cross-classification across multiple multi-temporal datasets having diverse class distributions. This adaptation mechanism, using a novel domain neutralizer network, allows alleviating the need for building a fresh training set through manual processes while classifying the test (multi-temporal) images belonging to a disparate region. A schematic diagram has been presented in Fig. 1.

Symbolic representation Let us assume $I_s = \{(\overrightarrow{M_i^s}, N_i^s)\}_{i=1}^{n_s}$ denotes the training set having n_s ordered set of multi-temporal images. Each of the i^{th} multi-temporal set is associated with a single class label N_i^s; where, $N_i^s \epsilon \{1, 2, 3, ..., C\}$. Here, C denotes the total number of land cover classes associated with the multi-temporal dataset. It is to be noted that each set of $\overrightarrow{M_i^s}$ has t number of images denoted as $\overrightarrow{M_i^s} = \{\overrightarrow{m_{i_p}^s}\}_{p=1}^{t}$ collected at different time-stamps. Moreover, the test set is denoted as I_t where the set of multi-temporal images are unlabeled $I_t = \{(\overrightarrow{M_i^t})\}_{i=1}^{n_t}$. A further list of the selected symbols is presented in Table 1.

Image pre-processing

As already mentioned, the proposed scheme investigates multi-temporal (very high resolution) images captured over a survey site. These images having multiple spatio-temporal

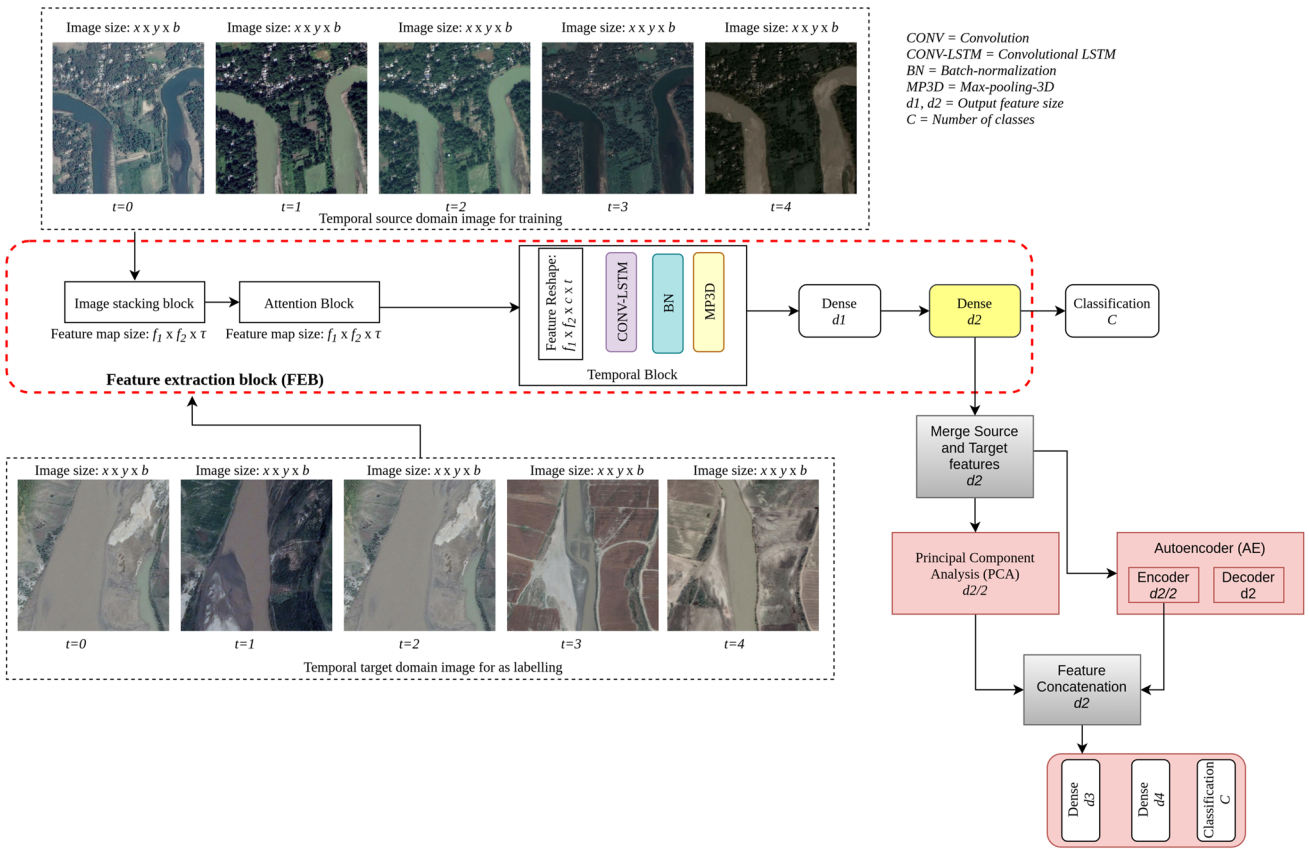

Fig. 1 Schematic diagram of the proposed model

Table 1 List of important symbols

Symbol	Significance
I_s	Dataset used for training
$\overrightarrow{M}_i^s$	i^{th} Set of multi-temporal images in I_s
N_i^s	Unique class label of the set $\overrightarrow{M}_i^s$
n_s	Number of $\overrightarrow{M}_i^s$ in I_s
C	Number of land cover classes
$\overrightarrow{m}_{i_p}^s$	Single image from the set $\overrightarrow{M}_i^s$ taken at p^{th} timestamp
t	Total numb of timestamps
I_t	Dataset used for testing
$\overrightarrow{M}_i^t$	i^{th} Set of multi-temporal images in I_t
n_t	Number of $\overrightarrow{M}_i^t$ in I_t
$x \times y$	Dimension of input images
b	Number of bands
$k_1 \times k_2 \times 1$	Kernel size of CNN filters used for Image pre-processing
$f_1 \times f_2$	Size of feature maps
F_t	Input to the CONV-LSTM layer at time t
$\overrightarrow{E}_i^s$	Extracted source features from FEB block
$\overrightarrow{E}_i^t$	Extracted target features from FEB block
$\overrightarrow{E}_i^{sp}$	Features extracted from the $\overrightarrow{E}_i^s$ using PCA
$\overrightarrow{E}_i^{sa}$	Features extracted from the $\overrightarrow{E}_i^s$ using AE
$\overrightarrow{E}_i^{tp}$	Features extracted from the $\overrightarrow{E}_i^t$ using PCA
$\overrightarrow{E}_i^{ta}$	Features extracted from the $\overrightarrow{E}_i^t$ using AE
$\overrightarrow{\hat{E}}_i^s$	Transformed source domain features by combining $\overrightarrow{E}_i^{sp}$ and $\overrightarrow{E}_i^{sa}$
$\overrightarrow{\hat{E}}_i^t$	Transformed target domain features by combining $\overrightarrow{E}_i^{tp}$ and $\overrightarrow{E}_i^{ta}$

bands require a special pre-processing before application of the attention mechanism as described in the consequent section. As a pre-processing technique, the t temporal images of sizes $x \times y$ having b number of bands are stacked one after another to be fed as an input to a convolutional neural network (CNN) to extract important spatial features. Mathematically, the i^{th} multi-temporal set of the images $\overrightarrow{M_i^s}$ has t number of time-series images $\{\overrightarrow{m_{i_p}^s}\}$ each of which has b number of spectral bands. This makes the input of the CNN as $x \times y \times b \times t$; however, to preserve the temporal identification of each of the images the size of the CNN filter is applied as $k_1 \times k_2 \times 1$. The third dimension is set to one so that the CNN can produce feature maps corresponding to each of the image bands b. Also, there is only one filter in the CNN that has its value tuned during the end-to-end training process. These maps preserve the individual identity of the image bands (corresponding to each temporal timestamp). As, already discussed, this preservation of temporal characteristics is essential to preserve for firm classification against the seasonal changes on the area. The feature maps containing isolated spectral and temporal information of the images are then input to the attention block of the proposed end-to-end scheme. To summarise, a feature map is obtained corresponding to each spectral bands of every timestamp image at the end of this pre-processing phase. Thus, the output of this pre-processing block is τ number of feature maps equal to number of bands available for each image (i.e., b) multiplied with the total number of timestamps (t). More precisely, the $x \times y \times b \times t$ input to this block is rendered as feature maps of the size $f_1 \times f_2 \times b \times t$ or $f_1 \times f_2 \times \tau$ where $\tau = b \times t$. Following the symbolic terminology, each feature map is generated for each of the b bands of $\{\overrightarrow{m_{i_p}^s}\}$, where p ranges from *1* to t. The feature maps generated here will be called 'feature bands' in the rest of the manuscript as these are maps corresponding to each of the spectral bands. A figure corresponding to show this stage has been presented in Fig. 2.

Attention mechanism to capture focused information

The image pre-processing block, from the previous section, has converted each of the input image samples to a set of feature bands while preserving their temporal and spectral identification. In the consequent stage, an attention mechanism will be investigated to focus on the actual land cover objects present in the feature maps, thereby, ignoring the irrelevant portions of the image scene. The maps obtained from the previous stage will be passed through a U-net like attention architecture as shown in Fig. 3. Here, the number of features bands input to the attention model is considered as τ (where $\tau = b \times t$), i.e., one feature band corresponds to a single band of an image obtained for a particular timestamp. The architecture contains multiple up-sampling and down-sampling sub-blocks to magnify and reduce the image sizes, respectively, to capture the important information from them.

Here, an image interpolation-based technique (Rukundo and Maharaj 2014) has been used for up-sampling; whereas, a max-pooling operation (Goodfellow et al. 2016) has been used for down-sampling. During the up-sampling process, a simple duplication of pixel intensity values is applied to the neighbouring pixels to scale up the image to a factor dependent on the scale by which the image is to be magnified. On the contrary, the down-sampling process involves the application of a max-pooling operation (Goodfellow et al. 2016) over the image where the size of the filter is determined by the scaling factor.

The encoder and decoder sub-blocks are responsible for spatial expansion or compression of the feature bands based on the image interpolation or the max-pooling operations, respectively, as shown in Fig. 3. Moreover, each of

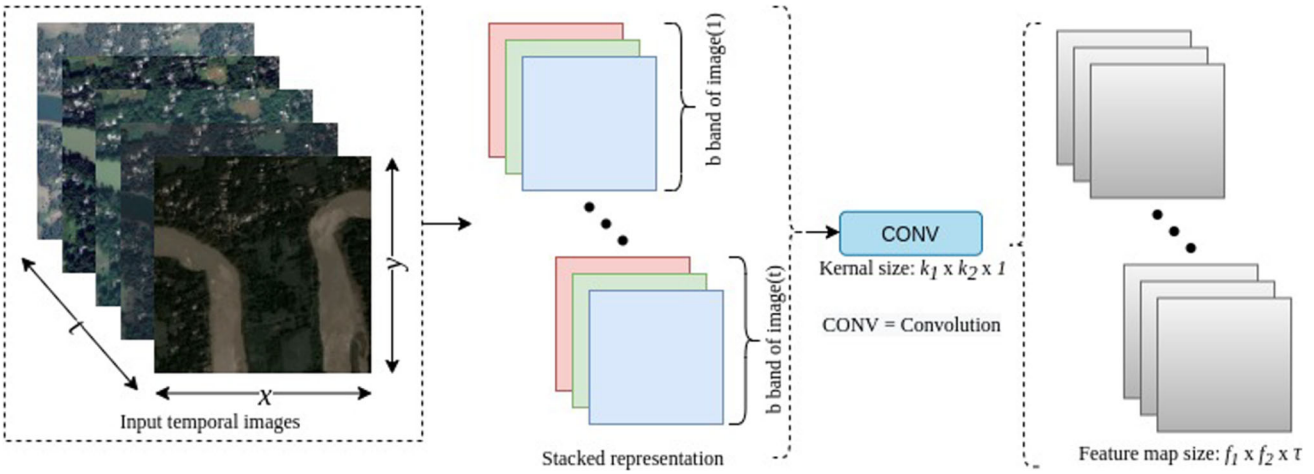

Fig. 2 Stacked representation of the image and feature generation of the proposed methodology

these sub-blocks performs 1 × 1 convolution and batch-normalization operations to extract meaningful information from the features (expanded or compressed) while preserving their dimensions. Moreover, the number of times for which the encoding-decoding cycle operates depends on a hyper-parameter called the encoder-depth. Here also, it is to be noted that the values of the filters get updated through training in the proposed end-to-end mechanism. Nevertheless, the output from this segment is the focused images corresponding to the temporal time-stamps. In consolidation,

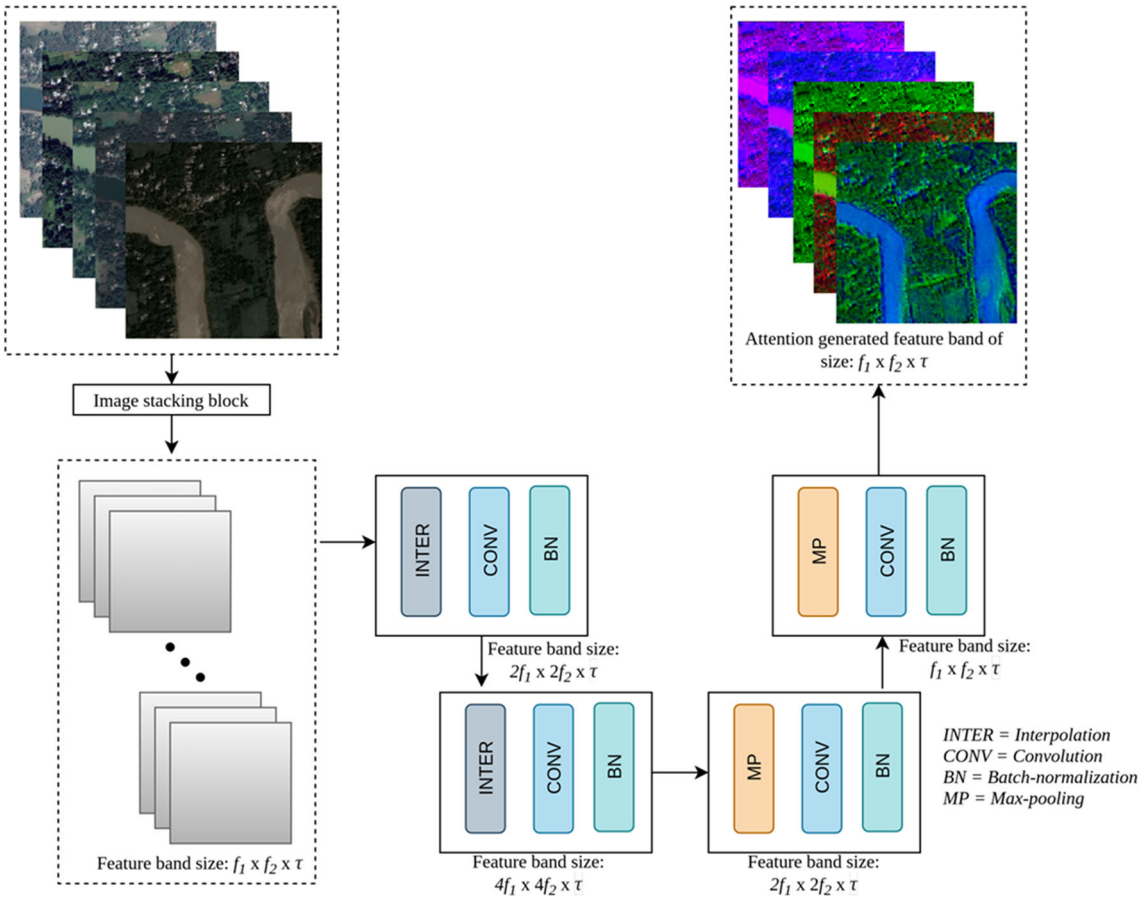

Fig. 3 Attention block of the proposed approach

the attention block generates a focused feature map of size $f_1 \times f_2 \times \tau$ corresponding to each of the feature bands obtained from the previous section. Disintegrating the representation, the focused feature map can be now represented as F_t having a size of $f_1 \times f_2 \times b$ forming t number of maps which is equal to the number original time-stamp images (since, $\tau = b \times t$). Further, a recurrent neural network, as described in the consequent sub-section, has been used to extract temporal information from these images.

Classification of temporal images using the features extracted by a convolutional LSTM

In the previous sections, the focused feature bands have been obtained corresponding to the temporal images and the bands therein. On application of the attention mechanism, the focused features have been carefully generated from the pre-processed image to identify the major land cover components present in the image scene. The next step is to extract the temporal information present across the multiple images captured over the same survey site in different timestamps. This will facilitate the classifier to know about the temporal variation in the land cover elements that may have occurred in time. For this important task, a temporal block (as shown in Fig. 1), containing a special type of recurrent neural network (RNN) along with batch-normalization and three-dimensional max-pooling layers, has been introduced in the proposed architecture. The RNN, specifically, the long short term memory (LSTM) neural network, utilized here has a special capability to solve the vanishing gradient problems inherent in the traditional RNNs (Goodfellow et al. 2016). An LSTM is a special type of neural network which contains multiple cells each corresponding to capture a temporal entity of the input. Each of the LSTM cells contain one memory cell (represented here as c_t) and three essential gates like input (as i_t), forget (as f_t) and output (as o_t) gates. The gates help in extracting and regulating the temporal flow of information within the network. More precisely, the input (as mentioned in the previous sub-section F_t) flows first through the input gate which decides the amount of input information to be passed on to the memory cell (c_t). Similarly, the forget gate controls the amount of previously stored information to be retained in the memory cell. Finally, the output gate has the authority to decide on the amount of information to be propagated along with the network from the memory cell. The mathematical formulation has been given in Eqs. 1 - 5. Here, the weights and biases are represented using W and b (Xingjian et al. 2015; Nguyen et al. 2019). It is to be noted that the '$\circ$' operator denotes the Hadamard product.

$$i_t = \sigma(W_{xi} \times F_t + W_{hi} \times h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} \times F_t + W_{hf} \times h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \quad (2)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_{xc} \times F_t + W_{hc} \times h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo} \times F_t + W_{ho} \times h_{t-1} + W_{co} \circ c_t + b_o) \quad (4)$$

$$h_t = o_t \circ tanh(c_t) \quad (5)$$

However, the classical structure of the LSTM does not allow the direct use of two-dimensional input such as images. To deal with this, the use of a convolutional LSTM (Shi et al. 2015) has been emphasized in this manuscript to preserve the spatio-temporal information. The convolutional LSTM (CONV-LSTM), which is investigated as having a many-to-many architecture, can capture the spatio-temporal features from the images. CONV-LSTM takes the temporal maps $F_1, F_2, ...F_t$ as inputs each having a size of $f_1 \times f_2 \times b$ as input to the corresponding cells to generate outputs $h_1, ..., h_t$. As in the case of LSTMs, here also, each of the CONV-LSTM cell p processes the input F_p having size $f_1 \times f_2 \times b$ and decides on the quantity of the information c_p to be forwarded to the consequent cell in the temporal timeline (here, p ranges from 1 to t). The difference with the LSTMs is only in terms of the convolution operation is denoted using '$\star$' in Eqs. 6 - 10. Instead of taking one-dimensional data as input, CONV-LSTM takes a single three-dimensional temporal image as a tensor (shown in Fig. 4) to extract the spatial information from the image matrix. The adopted mechanism of CONV-LSTM has been shown in Fig. 4.

$$i_t = \sigma(W_{xi} \star F_t + W_{hi} \star h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \quad (6)$$

$$f_t = \sigma(W_{xf} \star F_t + W_{hf} \star h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \quad (7)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_{xc} \star F_t + W_{hc} \star h_{t-1} + b_c) \quad (8)$$

$$o_t = \sigma(W_{xo} \star F_t + W_{ho} \star h_{t-1} + W_{co} \circ c_t + b_o) \quad (9)$$

$$h_t = o_t \circ tanh(c_t) \quad (10)$$

Coming back to the proposed technique, the temporal block (shown in Fig.1) is composed of a CONV-LSTM layer, a batch normalization (BN) layer, and a three-dimensional max-pooling (MP3D) layer. As already mentioned, the CONV-LSTM has been investigated to accommodate multi dimensional data instead of one, as in the case of plain LSTMs. The temporally arranged cells capture the input F_t at various time-stamps and produce an output at each of them in form of feature maps. The latter is then passed through batch-normalization and three-dimensional max-pooling layers before passing on through two dense layers for classification. Here, the number of neurons in the two dense layers are denoted as $d1$ and $d2$, respectively. Moreover, the number of neurons in the classification layer is C (equal to the number of available classes in the dataset). The scheme till here is used for the classification of multi-temporal remotely sensed images.

The proposed approach from the preceding section has also been exploited to handle the LCC by considering the

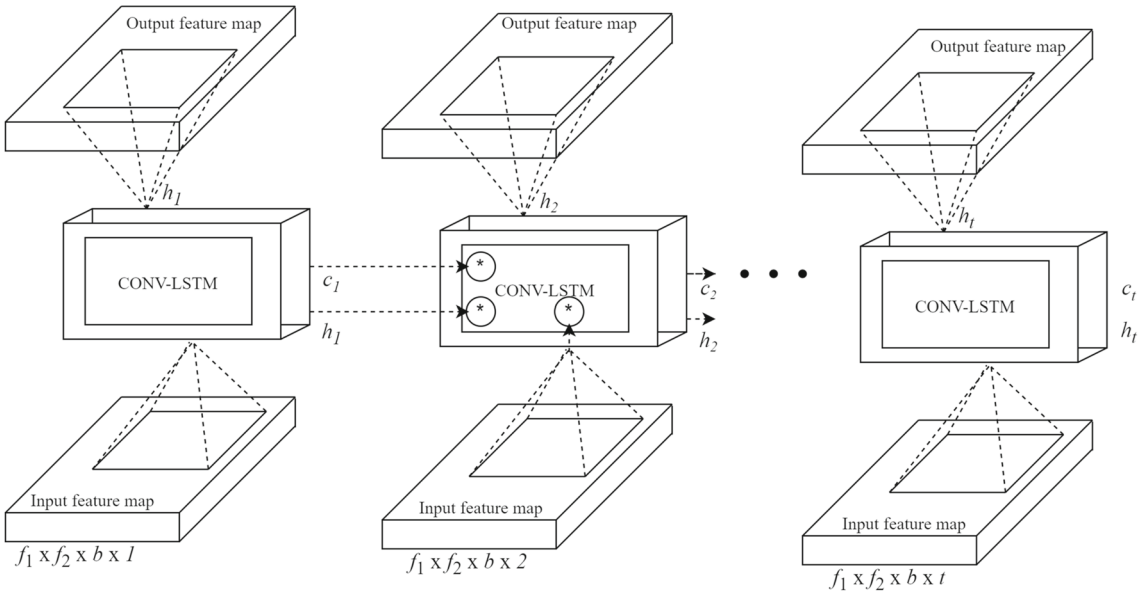

Fig. 4 Convolution LSTM

multi-temporal images of two different regions. During the investigation, the classification model has been improvised with some additional changes to incorporate adaptive scenarios of multi-temporal images.

Adaptive classification model

The proposed approach from the preceding section has been exploited to handle the LCC considering the multi-temporal images of two different regions. During the investigation, the classification model has also been improvised with some additional changes to incorporate adaptive scenarios. In this scenario, it is assumed that the training set ($I_s = \{(\overrightarrow{M_i^s}, N_i^s)\}_{i=1}^{n_s}$) and testing set ($I_t = \{(\overrightarrow{M_i^t}, N_i^t)\}_{i=1}^{n_t}$) of images are similar but follow different class-distributions and thus are from different domains. In another way, the training and the test set are from diverse datasets. The applicability of such systems lies in situations where a classifier trained with labeled images of a dataset can be used for predicting labels of images from an unlabeled dataset. This alleviates the human effort required for building a new training set through manual annotation of images from the new domain. In this regard, the classification model trained using the available labeled multi-temporal images from the source domain is used as a feature extractor. The layers through which the features are extracted are marked as feature extraction block (FEB) in Fig. 1. As highlighted, the FEB super-block consists of image stacking, attention, and temporal blocks followed by two dense layers. Thus, the dimension of the extracted features obtained using FEB is $d2$. In this regard, the $d2$-dimensional features have been extracted from source and target sample $\overrightarrow{E_i^s} = [e_1^s, e_2^s, ..., e_{d2}^s]$ and $\overrightarrow{E_i^t} = [e_1^t, e_2^t, ..., e_{d2}^t]$ from $\overrightarrow{M_i^s}$ and $\overrightarrow{M_i^t}$ source and target images, respectively has been obtained using the FEB. This feature extractor isolates features from the training (source) and the testing (target) images belonging to both domains. Here, it is to be noted that the class labels for the target domain images are not necessary for this extraction. The extracted features from images belonging to both the domains are concatenated and then presented to the domain neutralizer network for minimization of cross-domain distribution differences. Finally, the cross-domain adaptive classification takes place in the new combined feature space having reduced distribution difference. The adaptation mechanism has been presented in Fig. 5.

At the onset, the multi-temporal images from the source and the target domains are input to the feature extraction block to get the features $\overrightarrow{E}_i^s$ and $\overrightarrow{E}_i^t$, respectively. Consequently, the features are sent to a domain neutralizer network which minimizes the distribution difference across the features using an ensemble of principal component analysis (Jolliffe and Cadima 2016) and autoencoder neural network (Goodfellow et al. 2016). The task of a domain neutraliser network is to perform representation learning where the aim is to find a transformation function to map the features from source and target domains into an intermediate feature space where the distribution difference is neutralized (Bengio et al. 2013). The transformed features from the intermediate space are then used for classification typically under a supervised framework.

In literature, the task of representation learning is typically carried out using an autoencoder or through principal component analysis (PCA) (Tuia et al. 2016). The two approaches

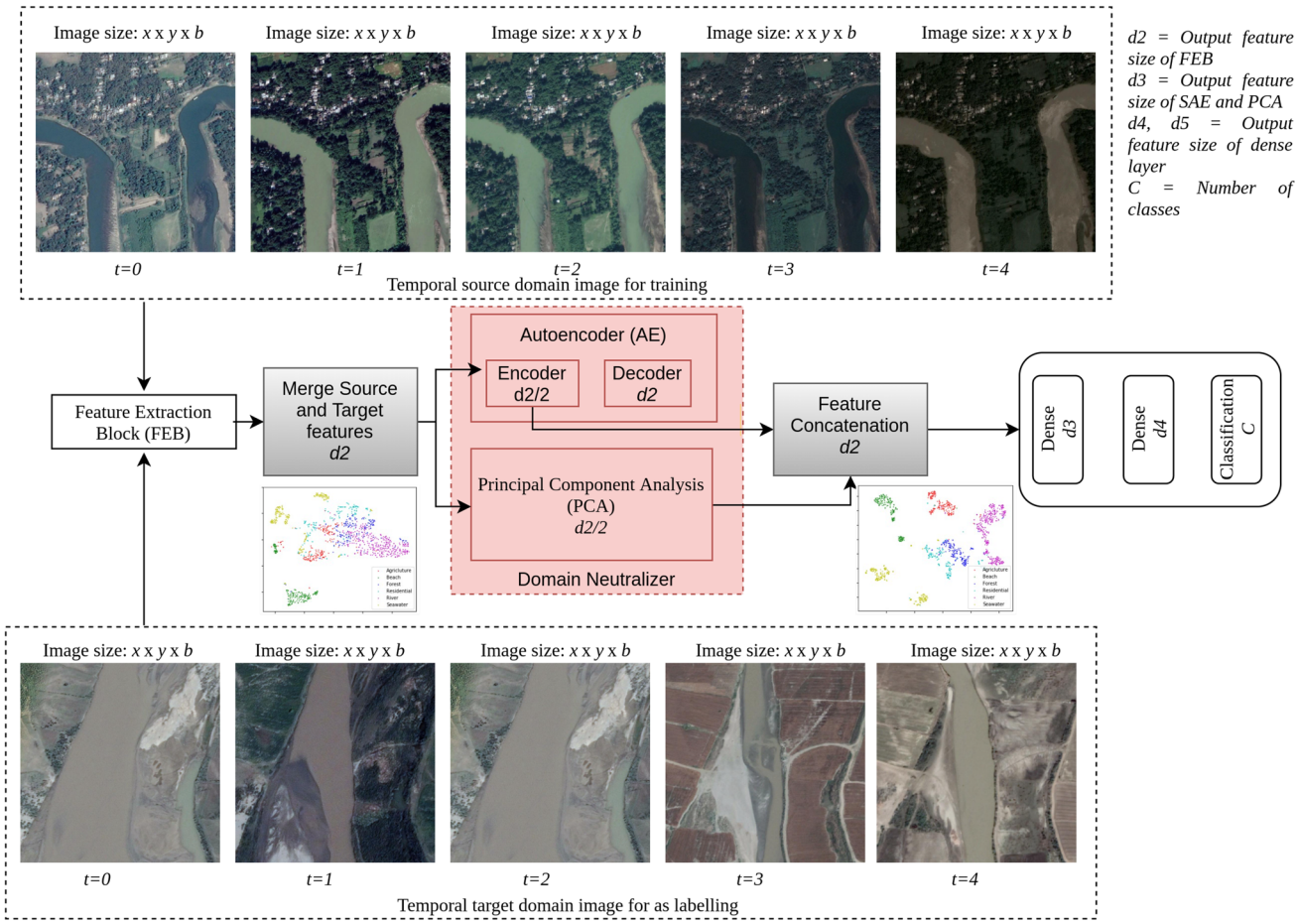

Fig. 5 Proposed domain neutraliser in multi-temporal images

mitigate the DA problem in two different directions. The autoencoder technologies solve the DA problem through non-linear mapping functions to find out an intermediate feature space using neural training. It uses iterative training to reconstruct the composite source-target input features on the hidden layer using the back-propagation algorithm (Haykin 2007). However, tt does not guarantee the linearly independent variables in the latent representation space. This means there can still be some redundancy in the latent space. On the other hand, PCA can shave away some of that redundancy as it is a feature extraction technique that disintegrates the existing features into linearly uncorrelated and orthogonal principal components. There have been numerous DA techniques investigated through the PCA mechanism like transfer component analysis (TCA) (Pan et al. 2011), metric learning (Geng et al. 2011) and semi-supervised TCA (Matasci et al. 2015) to name a few. As a result, the concatenation of the new features derived using autoencoder and PCA will contain both linear and non-linear properties.

In the proposed domain neutraliser network, the $d2$ features have been extracted individually from the i^{th} source (i.e., $\overrightarrow{E}_i^s$) and i^{th} target domain images (i.e., $\overrightarrow{E}_i^t$) using the proposed FEB. The domain-wise samples have then been concatenated vertically meaning that the total number of samples is n, where $n = n_s + n_t$. These union of samples have been presented to PCA and autoencoder to form a new representation of the source and target samples of size $\frac{d_2}{2}$ as $\overrightarrow{E}_i^{sp}$ and $\overrightarrow{E}_i^{sa}$, respectively. Similarly, the transformed target samples from PCA and AE is denoted as $\overrightarrow{E}_i^{tp}$ and $\overrightarrow{E}_i^{ta}$, respectively. Specifically, a feature-level ensemble has been used to concatenate the transformed features $\overrightarrow{\hat{E}}_i^s$ and $\overrightarrow{\hat{E}}_i^t$, where, $\overrightarrow{\hat{E}}_i^s = \overrightarrow{E}_i^{sp} \cup \overrightarrow{E}_i^{sa}$ and $\overrightarrow{\hat{E}}_i^t = \overrightarrow{E}_i^{tp} \cup \overrightarrow{E}_i^{ta}$. The dimension of the transformed source and target samples are an addition of $\frac{d_2}{2}$ features obtained from both PCA and AE thus making d_2 features again. Finally, a supervised classifier, comprising of two dense (with $d3$ and $d4$ neurons) and one output layer, has been trained using the labeled source (transformed) samples $\overrightarrow{\hat{E}}_i^s$ to predict the class labels for the unlabeled target (transformed) samples $\overrightarrow{\hat{E}}_i^t$. A detailed schematic diagram of the proposed multi-temporal DA approach is presented in Fig. 5.

Description of datasets

As elaborated in the previous sections, an investigation has been carried out for the classification and adaptation of multi-temporal remotely sensed images. The objective here is to identify the land cover classes devoid of the seasonal changes in the land surface. For this, the classifier is trained with images captured over multiple timestamps for better generalization. The first part of the work classifies the multi-temporal images for an annotation irrespective of the seasonal changes in a particular land area. The second part is concerned with using the extracted information to classify the temporal images captured over another region (area) similar to the first but having different class distribution. The experimentation is carried out using the satellite images captured over the Indian sub-continent. Here, the images are captured using Google earth pro software from the entire subcontinent which is then zone-wise divided into two sets. The images collected over Eastern India and Bangladesh are grouped to form the Eastern sub-continent dataset (ESD); whereas, those from Western India and Pakistan are grouped into the Western sub-continent dataset (WSD). Both of the datasets have images from six land cover classes like agriculture, forest, river, beach, residential area, and seawater. Each of the entries corresponds to five temporal images taken at a different time (seasons) of a year (between 2016-18). The total number of survey sites in ESD and WSD datasets are 981 and 1215, respectively, and the details are given in Table 2. It is to be noted that five time-stamp images are corresponding to each of these sites and some sample temporal images have been presented in Fig. 6.

An interesting mention in this aspect is in the data collection of images from multiple timestamps across diverse survey sites. At the onset, the images of resolution 4800 $\times$ 2791 $\times$ 3 have been collected on a survey site and then patches of size 600 $\times$ 600 $\times$ 3 have been extracted out of this. Later, one such patch has been selected for which cloud (or noise)-free images could be acquired across multiple time-frames. As already mentioned, the datasets have been prepared from the freely available images through Google Earth software to solve the dataset scarcity in the remote sensing community once and for all. About the Google Earth images, they are acquired using Worldview-3 and Landsat-8 satellites and the spatial resolution varies around 50 cm - 3 m. The number of bands on the VHR images, however, are three corresponding to the red, green, and blue bands.

Table 2 Description of the datasets

Land cover class	Number of survey sites in ESD	Number of survey sites in WSD
Agriculture	132	162
Forest	142	180
River	224	164
Beach	108	164
Residential	188	324
Sea	187	215
Total	981	1209

Details of the experimentation

As already mentioned, the input images are of size 300×300 multi-spectral images having three bands (i.e., $b = 3$) each corresponding to red, green, and blue channels. Moreover, each of the i^{th} site has $t = 5$ number of time-series images taken at different times of a year, the bundle of which is represented as $\overrightarrow{M}_i^s$. Further, each of the i^{th} bundle $\overrightarrow{M}_i^s$ is associated with one of the $C = 6$ land cover classes. At the onset, the images corresponding to a single site has been stacked to form an input of size $300 \times 300 \times 3 \times 5$ to a CNN with a kernel of size $5 \times 5 \times 1$ to preserve the individual identity of the image bands across the time-stamps. The output is 15 ($\tau = 3 \times 5$) number of feature bands is of size 148×148 (i.e., $f_1 \times f_2$). Consequently, the attention-based mechanism is applied to the output of these feature bands. The encoder-depth is set to 2 experimentally (meaning the process iterates through two up-sampling and two down-sampling blocks). The series of up and down-sampling results each of the feature bands to have sizes of 296×296, 592×592, 296×296 and 148×148, to take it back to the original size (as shown in Fig. 3). This focus feature bands $148 \times 148 \times 3 \times 5$ serves as input to the convolutional LSTMs having $t = 5$ cells and therefore the input to each of CONV-LSTM cells are of size $148 \times 148 \times 3$. The number of filters are three sizes 3×3 and stride $(2, 2)$; the output is then passed through the max-pooling layer of pool size $(2, 1, 1)$. The output is then flattened and passed through two dense layers having 1000 ($d = 1000$) and 256 ($d2 = 256$) neurons; this makes the final dimension of input features as 1×256 (output of feature extraction block). Next is a classification layer having six (since, $C = 6$) neurons corresponding to each class of the multi-temporal dataset. For experimentation, the images in each of the two datasets have been split into 70-10-20 percentage for training-validation-testing; further, the reported results are those obtained through 5-fold cross-validation.

The consequent investigation focuses on a graver problem that may occur while applying the trained classification model in the prediction of multi-temporal images from a discriminate dataset. However, it is found that there is a failure of the model as is described in the consequent sections. To understand the reason behind the deterioration in performance for the cross-dataset classification a KL

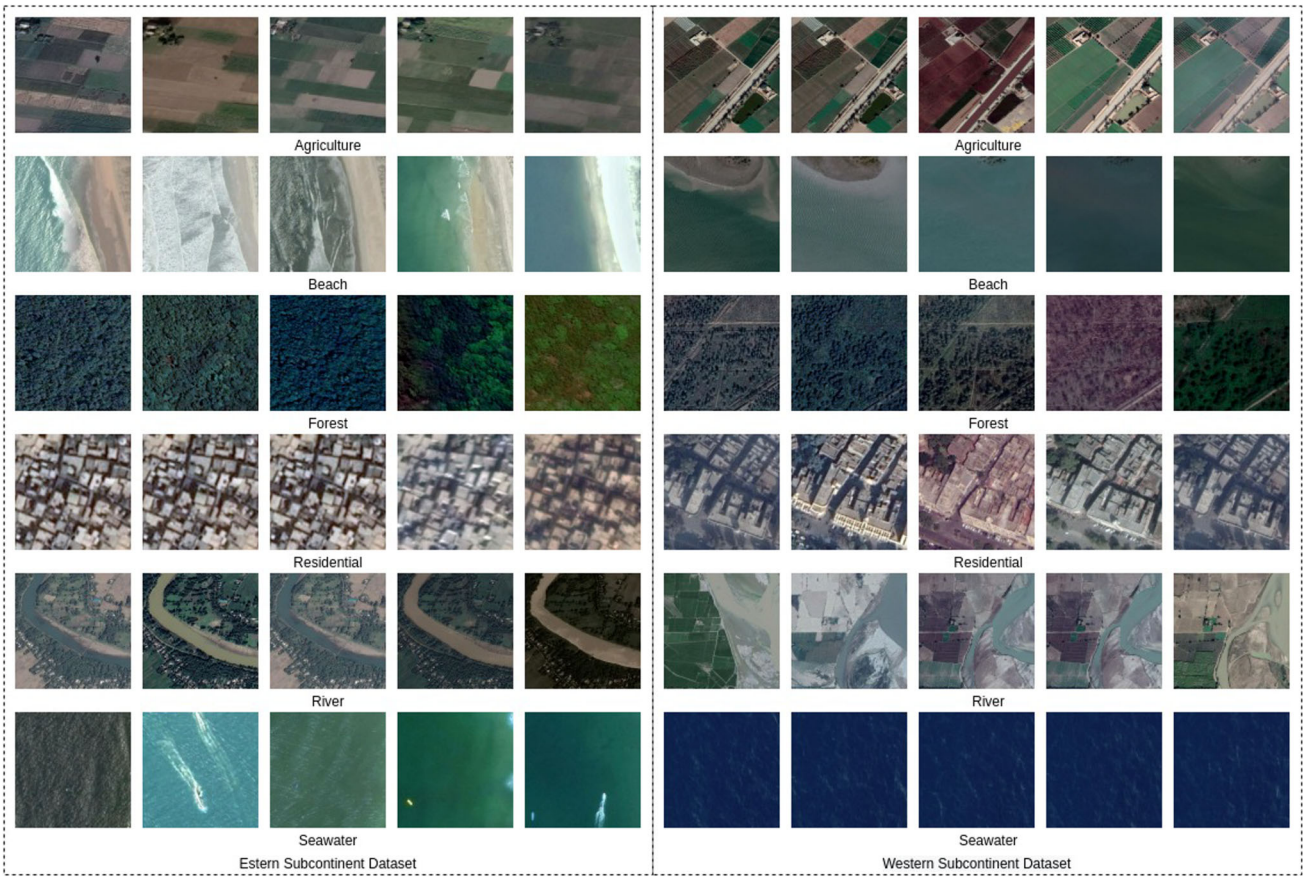

Fig. 6 Sample multi-temporal images

divergence (Kullback and Leibler 1951) among the extracted features for each of the training-test classes has been calculated and presented in Table 3. From this table, a low inter-class and high intra-class distribution divergence are evident among the source-target domains (the closest target class for each of the source classes has been boldfaced and an ideal scenario would have been boldface across the diagonal elements). Thereafter, the training and testing images are called source and target domains having abundance and scarcity of ground-truth information, respectively, for the multi-temporal images. For evaluation purposes, the experimentation has been carried out using the ESD dataset as the source and the WSD dataset as target domains. Thus, the number of the multi-temporal bundles associated with a single class label is $n_s = 981$ and $n_t = 1209$ for source and target domains, respectively. Later, the source-target domains have been reversed for a comprehensive validation of the problem statement. Next, to solve this problem a new block called the domain neutralizer block has been introduced to the system.

The domain neutralizer block takes the consolidated features from both domains as input. In this regard, the features of size 1×256 are extracted using the previous block from both the source and target images separately. Finally, their union i.e., a total number of 2190 ($n_s + n_t$) features (each of size 1×256) are together placed from source-target

Table 3 KL divergence amongst the extracted features from each class

Class	Agricluture	Beach	Forest	Residential	River	Seawater
Agricluture	1.87	**1.84**	2.01	2.06	2.01	1.90
Beach	1.76	1.86	**1.75**	1.76	1.75	1.68
Forest	1.75	1.81	1.76	1.77	1.75	**1.63**
Residential	2.31	2.31	2.32	2.41	2.39	**2.24**
River	1.77	1.78	1.79	1.82	1.82	**1.64**
Seawater	**1.72**	1.79	1.77	1.83	1.74	1.94

domains as input to the neutralizer block. The neutralizer block has an autoencoder and PCA sub-blocks which find out the intermediate features having the reduced dimension of 1×128. More technically, the autoencoder (AE) has 256-128-256 neurons in the input-hidden-output layers using which the training occurs for the $n = 2190$ input features from both the domains. The activation function for both encoding and decoding layers is the exponential linear unit and the learning rate is 0.0001. Finally, the input features ($\overrightarrow{E}_i^s$ and $\overrightarrow{E}_i^t$) from both the domains are passed through the trained AE to get an transformed representations ($\overrightarrow{E}_i^{sa}$ and $\overrightarrow{E}_i^{ta}$) each of sizes 1×128. A similar transformed representations ($\overrightarrow{E}_i^{sp}$ and $\overrightarrow{E}_i^{tp}$) is obtained using PCA. Consecutively, the transformed features obtained from PCA and AE are concatenated to obtain the final (transformed) representations of each domain. The size of each transformed source sample ($\overrightarrow{E}_i^s = \overrightarrow{E}_i^{sp} \cup \overrightarrow{E}_i^{sa}$) is 1×256 taking 128 features each from AE and PCA. Similarly, the size of the transformed target features is also 1×256. The cross-dataset classification has been conducted for the features in this transformed space and reported in the appropriate sections. For classification, three dense layers with 256-64-32 neurons have been used to classify the cross-domain transformed features into one of the six classes (equal to the neurons in the output layer). All the experiments have been performed using the Keras (backend as TensorFlow) framework on Intel Xeon processor with 128 GB DDR4 RAM and 32 GB NVIDIA Tesla V100 graphics.

Finally, the details of the hyperparamters involved in all the above phases are presented in Table 4 for comprehensive understanding of the tuning procedure.

Results and analysis

As already mentioned, the effectiveness of the proposed methodology has been assessed through experimentation using two very high-resolution multi-temporal aerial image datasets. Moreover, the probability distribution difference between the two datasets can be observed using the KL divergence (Kullback and Leibler 1951) measure among the two datasets, and results are presented in Table 3.

In this manuscript, a land cover classification technique has been proposed at the onset to identify the land classes from a set of images captured over a site across multiple time-frames. A deep attention-based convolutional LSTM model has been investigated for this purpose which has later been used as a feature extractor to develop an adaptive model capable of cross-classifying a set of multi-temporal images from two different domains. The performance evaluations of the two models have been carried out using two novel datasets consisting of multi-temporal VHR images captured over the Indian sub-continent region. As already mentioned, these images have been grouped to form Eastern and Western Indian sub-continent datasets which are abbreviated as ESD and WSD, respectively. Here, the performance has been quantified in terms of statistical measures like precision, recall, F-score, average accuracy (AA) and overall accuracy (OA) (Chakraborty et al. 2020).

Performance evaluation of the proposed end-to-end multi-temporal classification model

This section is dedicated to the evaluation of the developed attention-based multi-temporal classification technique on the two investigated datasets, as shown in Table 5. In recapitulation, a novel deep learning-based classifier has been proposed here which can exploit the temporal information from the remotely sensed images captured over a survey site. The input images are a bundle of multi-temporal images captures over the same region and the plausible output is a single class label that best annotates them. For experimentation, 5-fold cross-validation is performed over the datasets by splitting the samples into $70 : 20 : 10$ ratio, i.e., 70% samples as the training set, 20% samples as the testing set, and 10% samples as a validation set. A consolidated average of 5-fold cross-validation results has been presented to evaluate the performance of the models. Here, the larger training data (70% samples) helps to efficiently train the models for the given classification task. However, to neutralize the overfitting issue during training, a validation set of 10%samples has been used to identify such adverse situations. Here, the evaluation statistics for the proposed classification model using both the datasets (i.e., ESD and WSD) have been depicted in Table 5. A comparison in performance for the proposed approach against the state-of-the-art approaches has been presented in Table 6. Further, graphical representations for the performance parameters have been demonstrated graphically in Figs. 7, 8, 9 and 10.

An interesting observation from Table 5 and Fig. 7 can be attributed to the fact that the proposed technique has a near-perfect classification performance for the classes like agriculture and seawater in the ESD dataset. On the other hand, the performance on the beach and residential classes of the ESD dataset have different interpretations in terms of precision and recall. The Beach class has a perfect recall but poor precision which gives a hint on the presence of false alarms; whereas, the opposite for the residential class indicates the presence of missed alarms therein. The indices for the river class show satisfactory performance; however, there is considerable scope for improvement in the forest class. In consolidation, the proposed methodology has shown a high success rate in aggregate performance measures like overall accuracy (90.56%) along with average precision (91%), average recall (87%), and average F-score (87%). Turning

Table 4 The hyperparameters and their tuning procedure used in the proposed methodology

Name of the parameter	Model	Range/ fixed value	Optimal value	Tuning Procedure
Epoch	TLCC	100	50	The model is trained for a fixed number of epochs multiple times by observing the early stopping criteria. The epoch that satisfies early stopping criteria is considered the optimal one.
	TDA: Autoencoder	5000	2000	
	TDA: Supervised classifier	1000	500	
Learning rate	TLCC	0.1-0.000001	0.000075	The model is trained with varying learning rates in the considered range with a decremental heuristic fashion. The learning rate producing minimum loss function is selected as the optimal one
	TDA: Autoencoder	0.1-0.00001	0.0001	
	TDA: Supervised classifier	0.1-0.0001	0.001	
Minibatch size	TLCC	50-15	15	The model is attempted to be trained for varying mini-batch sizes by observing the memory bound of system. The mini-batch size that suits the memory bound of the system is selected for training
	TDA: Autoencoder	4096-2048	2048	
	TDA: Supervised classifier	4096-2048	2048	

Table 5 Class-wise performance evaluation of the proposed end-to-end multi-temporal classification technique

Dataset	**ESD**			**WSD**		
	Precision	Recall	F-score	Precision	Recall	F-score
Agricluture	1.00	1.00	1.00	1.00	1.00	1.00
Beach	0.71	1.00	0.83	0.95	1.00	0.97
Forest	0.88	0.61	0.72	0.33	0.87	0.48
Residential	1.00	0.66	0.80	1.00	0.50	0.67
River	0.86	0.95	0.90	0.84	0.79	0.81
Seawater	1.00	1.00	1.00	0.94	0.94	0.94
Average	0.91	0.87	0.87	0.84	0.85	0.81
Overall	90.56			88.97		

Table 6 Performance comparison of the proposed end-to-end multi-temporal classification technique with benchmark technique (A), Zhu et. al. (2021) (B), Scott et. al. (2017) (C), Bakhti et. al. (2019) (D), and proposed approach (P)

Dataset	Techniques	Precision	Recall	F-score	OA (in percentage)
ESD					
	A	0.64	0.61	0.62	62.56
	B	0.77	0.81	0.79	83.26
	C	0.70	0.75	0.72	76.83
	D	0.77	0.79	0.78	81.69
	P	0.91	0.87	0.87	90.56
WSD					
	A	0.61	0.62	0.61	62.37
	B	0.75	0.80	0.77	82.64
	C	0.69	0.71	0.70	72.44
	D	0.77	0.73	0.75	75.69
	P	0.84	0.85	0.81	88.97

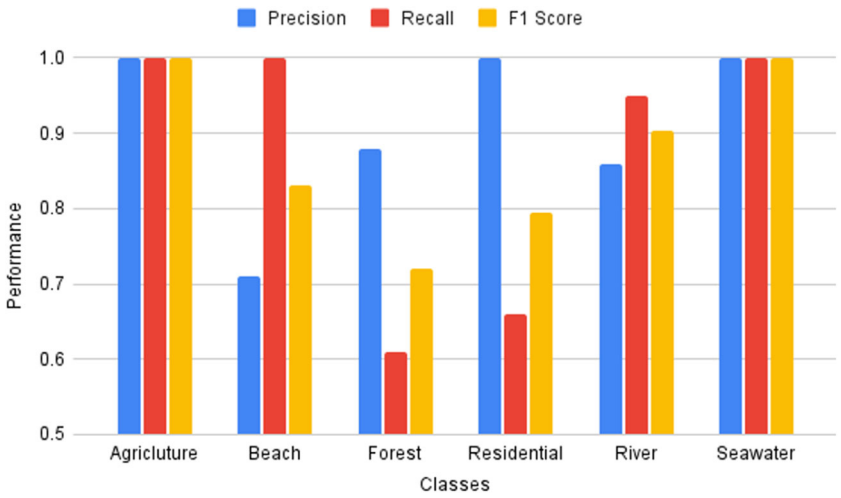

Fig. 7 Class-wise performance of the proposed classification scheme in terms of precision, recall, and F-score for ESD dataset

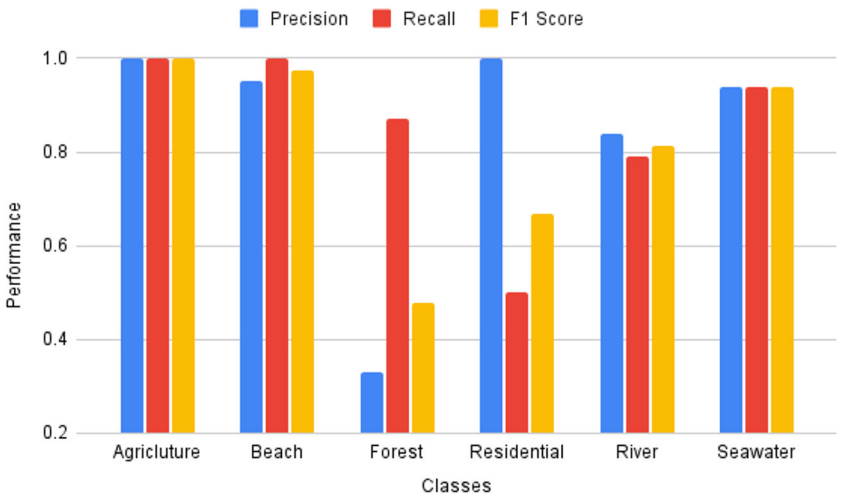

Fig. 8 Class-wise performance of the proposed classification scheme in terms of precision, recall, and F-score for WSD dataset

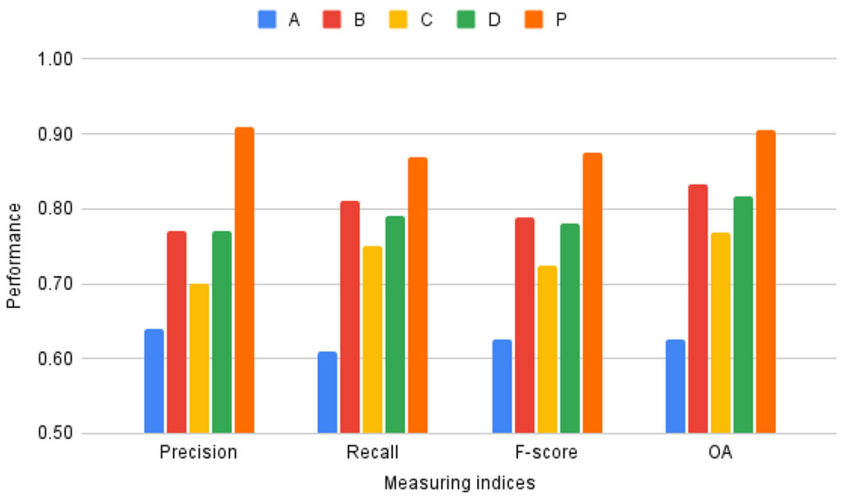

Fig. 9 Bar graph comparison of the classification results in terms of average Precision, average Recall, and average F-score (overall classes) using ESD obtained using benchmark technique (A), Zhu et. al. (2021) (B), Scott et. al. (2017) (C), Bakhti et. al. (2019) (D), and proposed approach (P)

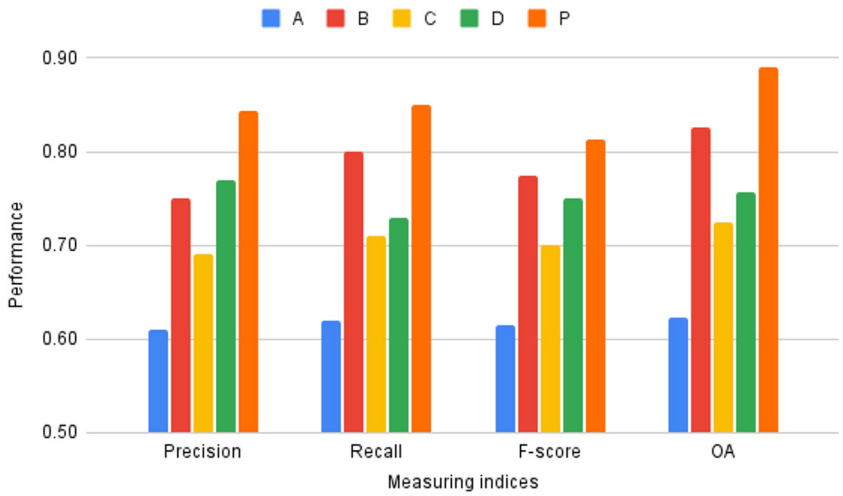

Fig. 10 Bar graph comparison of the classification results in terms of average Precision, average Recall, and average F-score (overall classes) using WSD obtained using benchmark technique (A), Zhu et. al. (2021) (B), Scott et. al. (2017) (C), Bakhti et. al. (2019) (D), and proposed approach (P)

towards the WSD dataset (shown in Table 5 and visually interpreted in Fig. 8), agriculture, seawater, and the beach classes have shown exceptional performance (94-100%) in terms of precision, recall, and F-score. Similar to the observation as in the case of the ESD dataset, deterioration can be observed in the performance of the WSD forest class. The diverse nature of the residential class has once again shown to have missed alarms in the identification of the same class in WSD too. Meanwhile, the indices show satisfactory performance in the river class; finally, an aggregated value of 84%, 85%, and 81% has been observed for average precision, average recall, and average F-score, respectively. Summarising the observations, the overall accuracies combining all the classes in ESD and WSD is observed as 88-90% which comprehensively shows the effectiveness of the proposed methodology.

Further, the performance of the proposed approach has been compared with that of a benchmark technique and three other state-of-the-art approaches as shown in Table 6. For visual understanding, the values from the statistical measures have been plotted in Figs. 9, and 10. In the benchmark technique (marked as A in the Table 6), features extracted from multi-temporal images using a pre-trained Alexnet (Krizhevsky et al. 2012) have been used to train a simple multi-layer perceptron. Further, the performance comparisons have been carried out with state-of-the-art approaches as investigated in (Zhu et al. 2021) (B), (Scott et al. 2017) (C) and (Bakhti et al. 2019) (D). To maintain homogeneity with the proposed approach (P), the same training ecosystem in terms of training-test ratio and cross-validation has been presented to all the compared techniques. As seen in Table 6 and Fig. 9, all the intelligent techniques developed over the last few years have improved on the results in comparison to the benchmark technique. Scott et al. (2017) uses a pre-trained deep CNN to classify the VHR (non-temporal) images under the transfer learning paradigm; whereas, Zhu et al. (2021) have emphasised classifier retraining to update the classification maps through freshly collected temporal images. On the other hand, Bakhti et al. (2019) have investigated a hybrid neural network model (combining CNN and RNN) for the mapping and monitoring of multi-temporal vegetation using VHR datasets. It is also imperative to note that the technique from Zhu et al. (2021) has been the better approach for multi-temporal classification by a significant margin; however, it also has a scope of improvement. Based on these premises, the proposed classification technique has been developed by taking into consideration the changes in the land surface in the attention-based classifier. Looking at the numbers in the case of the ESD dataset (as also shown in Table 6 and Fig. 10), the proposed classification system (P) has outperformed (Zhu et al. 2021) (B) by a comprehensive margin of 14%, 6%, 8%, and 7% in terms of precision, recall, F-score and OA. Meanwhile, on comparing the performance of the proposed technique with (Scott et al. 2017) (Bakhti et al. 2019), the improvement is 21% (14%), 12% (8%), 15% (9%) and 30% (8%) in the same parameters in the same order. The same numbers as compared to the benchmark technique are 27%, 26%, 25%, 28%.

Similarly, the improvement for the proposed approach is approximately 9%, 5%, 4%, and 6% on the same parameters in WSD dataset, as compared to (Zhu et al. 2021). On investigation over the same dataset, Bakhti et al. (2019) had been the previous best in terms of the precision where the proposed approach has just superseded by 7%, as can be seen in Table 6 and Fig. 10. On the other parameters like recall, F-score, and OA the improvement is 12%, 6%, and 13% for the proposed approach. The improvement in terms of precision is 15% as compared to (C) same margins are 14%, 11%, 16 % in terms of recall, F-score, and OA. The improvements in terms of 23% (precision), 23% (recall), 20% (F-score), 26% (OA) as compared to the benchmark technique clearly show the efficiency of the proposed scheme. Based on the

observations, it can be concluded that the proposed approach has outperformed the previous approaches in the literature in terms of classifying multi-temporal images. These observations corroborate with the graphical interpretations shown in Fig. 10.

Performance evaluation of the adaptive methodology

As already mentioned, the proposed multi-temporal classification technique has been made adaptive to cross-classify the images having a difference in class distribution. In this regard, the labeled temporal images from one dataset are used to train the new adaptive model which then predicts the class labels for the multi-temporal images from another dataset. As intuitive, the performance deteriorates significantly in such a cross-dataset classification environment which is reflected in the non-adaptive row in Table 8. The performance in the individual classes is further shown in Table 7. Here, the accuracy for the classes having discriminate distributions across the source-target datasets shows poor performance in terms of precision, recall, and F-score, as shown in the case of non-adaptive classification (in Fig. 14). Here, the ESD has been used as the source domain and WSD as the target domain. Non-adaptive classification depicts results for cross-dataset classification when no adaptive algorithm is used and is thus necessary to analyze for establishing the motivation for the investigated DA scheme. Examples of such difficult classes are agriculture, residential, and forest where the performance parameters show less than 0.5. On the contrary, beach, seawater, and river show considerably higher indices due to their distribution similarity across the domains. Due to this mixed class-wise performance, the overall and average accuracies have also deteriorated as shown in the first row of Table 8. The performance of the adaptability in the proposed approach has been cross-checked by using both ESD and WSD datasets as a source and target domains turn-by-turn. For an easier interpretation to the readers, the event when ESD is used as source and WSD is used as the target is shown as ESD→WSD and vice-versa.

In the proposed scheme for adaption in the multi-temporal datasets, the classes like beach, river, and seawater have shown considerable improvement in terms of precision (73% - 93%), recall (87% - 96%), and F-score (83% - 90%) for the ESD→WSD dataset. In the same dataset, the classes like residential have shown an improvement of 54%, 32%, and 40% on the same parameters owing to the diversification of the samples on the increase of the training data. However, there is a scope for improvement in classes like agriculture and forest where the performance is merely 18-35%, 6-28%, and 10-22% in terms of precision, recall, and F-score, respectively. The interpretation and validation of these numbers can be seen in Table 7, Fig. 11. Now analyzing for the WSD → ESD dataset, a superlative performance can be observed in the classes like beach, river, and seawater where the accuracy is 68-95%, 40-71% and 56-76% in terms of precision, recall, and F-score. Similar to the previous observation, the residential class shows moderate performance which is 37%, 67%, and 48% on the same parameters. On the other hand, agriculture and forest have again shown a disappointing performance in 27-31%, 29-33%, and 30% accuracy in the investigated parameters in the same order. The visual interpretations can be found in Table 7 and Fig. 12 for the dataset WSD → ESD. A further assessment shows a considerable improvement in average precision (57-59 %), average recall (52-57%), average F-score (51-55%), and overall accuracy (53-63%) while investigating both the dataset combinations. For further validation of the proposed DA approach, the scatter plots (using t-SNE (Maaten and Hinton 2008) have been shown (in Fig. 13) before and after the adaptation process using the extracted features from the FEB block.

The performance of the adaptive algorithm is compared with some well-known state-of-the-art domain adaptation schemes like MIDA (Yan et al. 2017), DAN (Othman et al. 2017), AANN (Ammour et al. 2018) and CS-DDA (Zhang et al. 2020). These sophisticated techniques have been evaluated on the same parameters of precision, recall, F-score, and overall accuracy (OA). Amongst the compared schemes, Yan et al. (2017) have introduced maximum independence DA where the features having the least cross-domain similarities

Table 7 Performance evaluation of the proposed multi-temporal DA technique

Dataset	ESD→WSD			WSD→ESD		
	Precision	Recall	F-score	Precision	Recall	F-score
Agriculture	0.35	0.06	0.10	0.27	0.33	0.30
Beach	0.81	0.94	0.87	0.84	0.69	0.76
Forest	0.18	0.28	0.22	0.31	0.29	0.30
Residential	0.54	0.32	0.40	0.37	0.67	0.48
River	0.73	0.96	0.83	0.68	0.71	0.69
Seawater	0.93	0.87	0.90	0.95	0.40	0.56
Average	0.59	0.57	0.55	0.57	0.52	0.51
Overall	63.85			53.49		

Table 8 Performance comparison of the proposed DA scheme with other benchmark and existing DA techniques

Techniques	Dataset: ESD→WSD				Dataset: WSD→ESD			
	Precision	Recall	F-score	OA	Precision	Recall	F-score	OA
Techniques	Precision	Recall	F-score	OA	Precision	Recall	F-score	OA
Non-adaptive	0.46	0.44	0.45	47.35	0.49	0.46	0.47	46.25
MIDA	0.44	0.41	0.42	43.19	0.41	0.39	0.40	41.11
DAN	0.49	0.46	0.47	50.58	0.51	0.50	0.50	50.23
AANN	0.51	0.46	0.48	52.94	0.54	0.48	0.51	51.26
CS-DDA	0.60	0.49	0.54	56.48	0.55	0.51	0.53	51.28
Proposed	0.59	0.57	0.58	63.85	0.57	0.52	0.54	53.49

OA: Overall accuracy in percentage

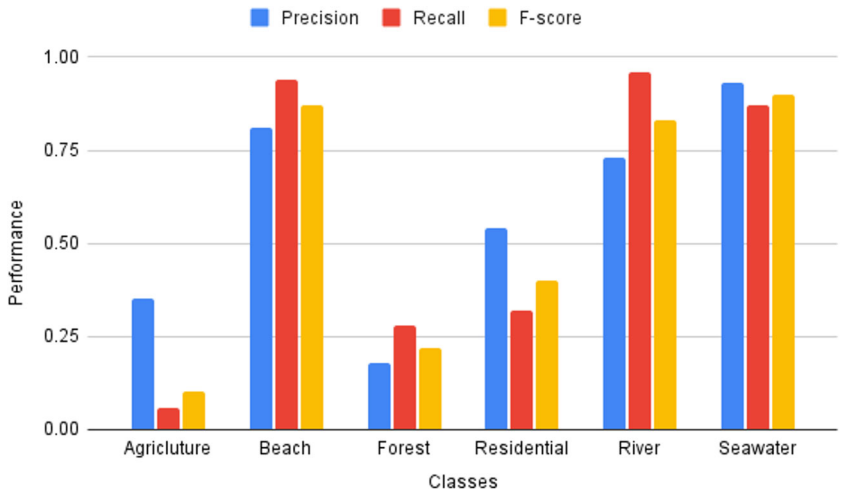

Fig. 11 Class-wise performance of the adaptive technique corresponding to precision, recall, and F-score for ESD → WSD scenario

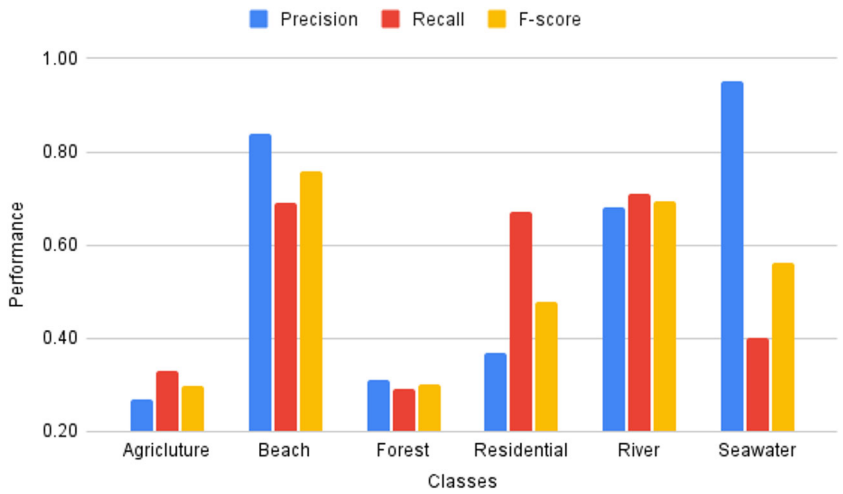

Fig. 12 Class-wise performance of the adaptive technique corresponding to precision, recall, and F-score for WSD → ESD scenario

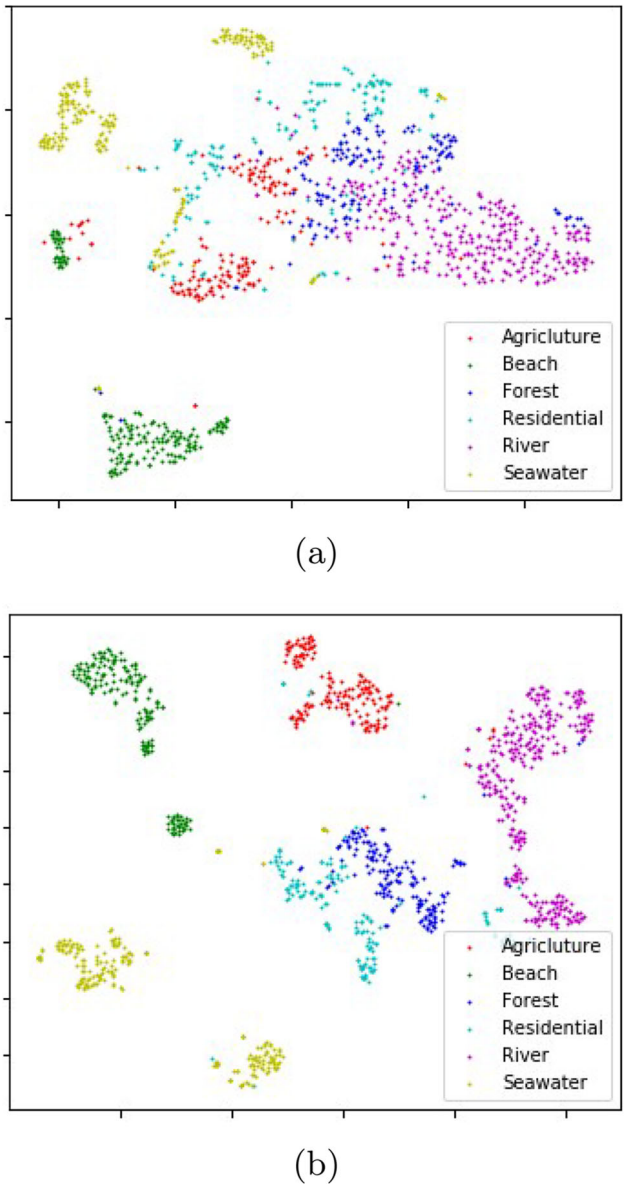

Fig. 13 Scatter plot amongst the extracted features (a) without adaptation, where the data-shift problem is prominent, (b) after adaptation. Note: two features are generated from 256 features using t-SNE [51]

are identified before learning a new feature sub-space using some augmented data; whereas, DAN (Othman et al. 2017) have used the transfer learning abilities of pre-trained CNNs to generate the intermediate features. Moreover, AANN (Ammour et al. 2018) is also a deep learning-based scheme that uses an autoencoder to transfer the labeled features to the unlabeled space; CS-DDA (Zhang et al. 2020) is class sensitive in minimizing the data discrepancy amongst the two domains while preserving the inter-class separability. Since no other DA techniques have been investigated previously for multi-temporal land cover classification, the evaluation is carried out on these spectral (non-temporal) DA techniques by setting up a multi-temporal environment. As introduced in the previous section, the non-adaptive scheme is included for the comparison in terms of the scenario where no adaptive algorithm is in place. The findings are presented in Table 8 and Figs. 14, and 15. At the onset, the comparative analysis of the ESD→WSD dataset is discussed in the consequent paragraph; later, the reverse will also be critically scrutinized.

As observed from Table 8 and Fig. 14, the proposed adaptive approach comprehensively outperforms all the compared schemes in terms of recall (8 -16%), F-score (4-16%) and OA (7-20%); however, it is only comparable to others in terms of precision for the ESD→WSD dataset. In greater detail, the improvement in the recall is 14%, 16%, 11%, 11% and 8% as compared to non-adaptive (benchmark scheme), MIDA, DAN, AANN, CS-DDA schemes, respectively. The same indices of F-score are 13%, 16%, 11%, 10%, 4% in comparison with all the schemes in the same order; similarly, for OA it is 16%, 20%, 13%, 10%, and 7%. Focusing on the precision, the proposed technique could establish its superiority with all other techniques like non-adaptive (13%), MIDA (15%), DAN (10%), and AANN (8%). However, the precision performance is comparable with that of CS-DDA. A visual interpretation can be found in Fig. 14 for easy understanding.

Now, the time is to analyze the results obtained for the WSD→ESD scenario as shown in the same Table 8. The first observation is in terms of precision where the proposed methodology has outperformed the non-adaptive scheme by 8%, MIDA by 16%, DAN by 6%, AANN by 3%, and CS-DDA by a margin of 2%. Secondly, the enhancement is 6%, 13%, 2%, 4%, 1% (with all the schemes following the same order) in terms of recall. Consequently, the DA scheme investigated in this manuscript has overpowered other schemes like non-adaptive (by 7%), MIDA (14%), DAN (4%), AANN (3%), and CS-DDA (1%) in terms of F-score. The same values for OA are 7%, 12%, 3%, 2%, 2% with the compared approaches. A graph can be found in Fig. 15.

Error analysis for the adaptive methodology

As already mentioned, the proposed multi-temporal classification technique has been made adaptive to cross-classify the images having a difference in class distribution. In this regard, the labeled temporal images from one dataset are used to train the new adaptive model for the prediction of the class labels of the multi-temporal images from another dataset. As intuitive, the performance deteriorates significantly in such a cross-dataset classification environment which is reflected in the non-adaptive row in Table 8. The non-adaptive classification depicts results for cross-dataset classification when no adaptive algorithm is used and is thus necessary to analyze and establish the motivation for the investigated DA scheme. The last row in Table 8 reflected the result of the cross-dataset

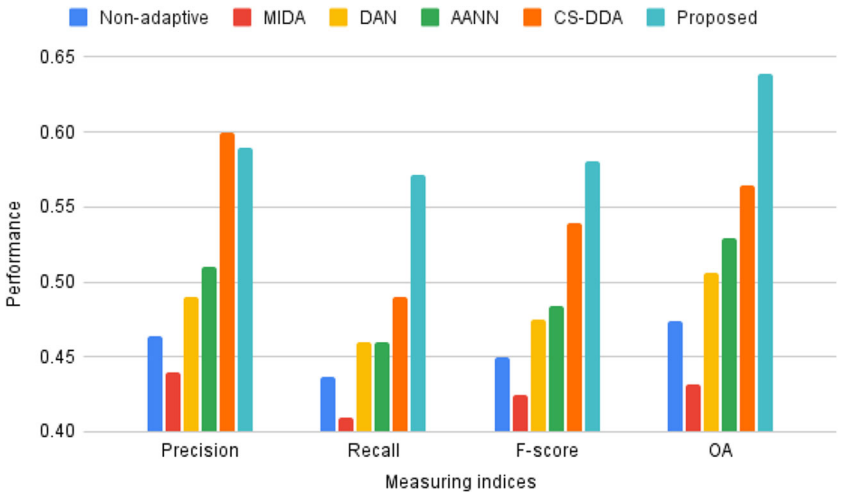

Fig. 14 Bar graph comparison of the adaptation results in terms of average Precision, average Recall, and average F-score, and over-all accuracy (OA) (overall classes), using ESD → WSD scenario obtained using Non-adaptive technique, MIDA [52], DAN [7], AANN [11], CS-DDA [53], and proposed approach

classification environment after handling the domain adaptation problem. To validate the importance of the proposed approach, the scatter plots (using t-SNE Maaten and Hinton (2008)) have been shown (in Fig. 13) considering the before and after the DA situation using the WSD→ESD scenario. To generate the scatter plots, the source-target features (of size 256 each) are extracted from the Domain Neutralizer block. Thereafter, the two features are generated from 256 features using t-SNE Maaten and Hinton (2008). In the plot (Figure 13), the six classes 'Agriculture', 'Beach', 'Forest', 'Residential', 'River', and 'Seawater' are represented by red, green, blue, cyan, magenta, and yellow color, respectively. Moreover, the KL divergence Kullback and Leibler (1951) (considering the extracted features) for WSD → ESD scenarios has been calculated and presented in Tables 3. In ideal instances, the KL divergence among the same classes should be minimum and the features of a class must be distinct from the other classes. By analyzing Fig. 13 and Table 3, the following cases are observed:

A: During investigation, it has been found that the minimum value of the same-class divergence is not held by the poor performing class. Furthermore, this value is relatively high in comparison to the minimum value. However, the features can be distinguished from the other classes with minimal overlapping. In this scenario, the model suffers to produce the good result (due to the high difference) for the classes falls under this case.
For example, the same class divergence value for 'Agriculture' is 1.87, and the difference with the lowest value is 0.15. Furthermore, the scatter plot revealed that the class is separated with the least amount of overlap. The model may become confused and perform badly for this class due to the large discrepancy between the lowest and same class divergence value. A similar observation may be found in the 'Seawater' class.

B: It has been observed during experimentation that the same class divergence value is not the least, but it is not far from it. Furthermore, the samples are well separated from the other classes. In this situation, the model may

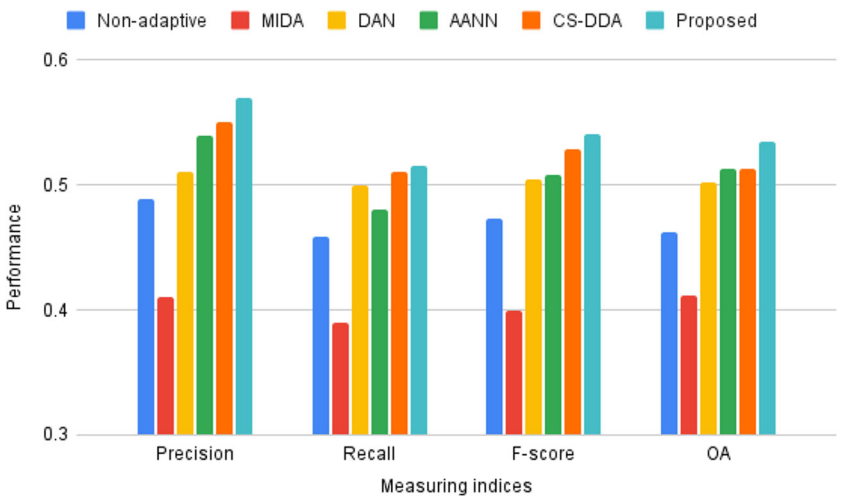

Fig. 15 Bar graph comparison of the adaptation results in terms of average Precision, average Recall, and average F-score, and over-all accuracy (OA) (overall classes), using WSD → ESD scenario obtained using Non-adaptive technique, MIDA [52], DAN [7], AANN [11], CS-DDA [53], and proposed approach

produce better results than the classes (considered in Case A).

For example, the class divergence values for 'Beach' and 'River' are 1.86 and 1.75, respectively, with a difference of 0.08 from the minimum value (for both the classes). Moreover, the scatter plot, shows that the samples with derived features are well separated from the other classes. This may enable the model to deliver a superior outcome as compared to the classes (considered in Case A).

C: During experimentation, it has been observed that the same class divergence value is not the minimum and may or may not be closer (nearer) to the minimum. However, the patterns of these classes are close to each other (high chances of inter-class overlapping in scatter plot). In this case, the performance of the model is either poor or average.

For example, the class divergence values for 'Forest' and 'Residential' are 1.76 and 2.41, respectively, while the difference with the minimum value is 0.01 (extremely minor) and 0.65. (very large). Furthermore, the scatter plot shows that both classes are not well separated from one another, and there is a considerable risk of misclassification (due to high chances of overlapping). As a result, the performance of these classes is either poor or mediocre.

Ablation analysis of the proposed model

During the experimentation, the investigation revealed the significant role played by the novel attention-based mechanism, illustrated in the attention block in Fig. 1. This mechanism is crucial for identifying important regions within each of the time-series images. Simultaneously, the temporal block, integrated with a CONV-LSTM architecture (as shown in Fig. 1), was utilized to capture interrelationships among land cover scenes across the temporal sequence of images. To assess the impact of these two blocks, an ablation analysis was conducted as an integral part of the investigation. In this context, the proposed model (referred to as 'P' in Table 9) was compared against two alternative architectures. In the first architecture (Ablation1), the attention block was removed from the proposed model, and the output of the image stacking block was directed straight to the temporal block. The performance of this configuration in both end-to-end multi-temporal classification and multi-temporal domain adaptation (DA) cases is presented in Table 9 (designated as Ablation1). In the second architecture (Ablation2), the conventional LSTM architecture was employed to replace the temporal block, and the dimensions of the output features derived from the attention block were adjusted to $f \times c \times t$ (where $f = f_1 \times f_2$). This modification was made to enable the seamless feeding of these features into the LSTM architecture. The performance of this architecture is presented in Table 9 (designated as Ablation2). It is noteworthy that the ESD dataset was employed for end-to-end multi-temporal classification, and adaptive classification (multi-temporal DA) was evaluated by considering ESD→WSD scenarios.

Table 9 Ablation analysis of the proposed model

Techniques	Precision	Recall	F-score
End-to-End multi-temporal classification			
Ablation1	0.76	0.75	0.75
Ablation2	0.83	0.85	0.84
P	0.91	0.87	0.87
Multi-temporal DA technique			
Ablation1	0.52	0.39	0.45
Ablation2	0.58	0.55	0.56
P	0.59	0.57	0.58

Here Ablation1 and Ablation2 are the two alternative architectures and P is the proposed method for both End-to-End multi-temporal classification (for ESD dataset) and Multi-temporal DA techniques (for ESD→ WSD scenario)

The results presented in the table indicate that the proposed method (P) outperforms Ablation1 by a considerable margin of 15% (precision), 12% (recall and F-score), and Ablation2 by a narrower margin of 8% (precision), 2% (recall), and 3% (F-score) in the context of end-to-end multi-temporal classification strategy. Similarly, in adaptive scenarios (multi-temporal DA), the proposed model (P) has demonstrated superior performance compared to the other two counterparts (Adaptive1 and Adaptive2). Specifically, the proposed method outperforms Adaptive1/Adaptive2 by 7%/1% in precision, 18%/2% in recall, and 13%/2% in F-score. Throughout the experiments, it was observed that Ablation1 required the least amount of time to complete (as compared to proposed method P and method Ablation2), while Ablation2 consumed the maximum time for both end-to-end multi-temporal classification and multi-temporal DA cases.

Conclusion

The primary purpose of the study presented here is to firmly classify a geographical region based on the time-series images captured across multiple seasons. For the naive classification, an attention-based deep learning mechanism has been used at the onset to identify the predominant class of an image scene; later, the convolutional LSTMs capture the spatio-temporal dependencies existing amongst the time-series images. Thereafter, a transfer learning-based adaptive methodology has been adapted for cross-classification of

multi-temporal images captured over discriminate survey sites. More specifically, a domain neutralizer network has been introduced in the proposed end-to-end training mechanism which uses the attention-based convolutional LSTMs as a mere feature extractor. This improvisation has led to the development of the multi-temporal adaptive classification system using remotely sensed satellite images. Experimentation on two novel Indian sub-continent multi-temporal VHR datasets reveals the effectiveness of the proposed classification as well as the adaptive approaches showing exemplary performances in cross-domain identification of land areas corresponding to agriculture, beach, residential, river, etc. despite their geophysical changes across time.

Acknowledgements S. Chakraborty expresses gratitude for the project grant (with No. TDP/DRISHTI CPS/L2M/SL/2023/0007) received from the IITI DRISHTI CPS Foundation under the National Mission on Interdisciplinary Cyber-Physical System (NMICPS), Department of Science and Technology, Government of India. However, this grant is not a funding source for the manuscript, as this work has not received financial support from any external sources.

Author Contributions I. Kalita, as the first and corresponding author, played a central role in methodological development, code implementation, and manuscript composition. S. Chakraborty, as the joint second author (with T. Reddy), made significant contributions to the manuscript by actively engaging in its writing and proofreading. T. Reddy, serving as the joint second author (alongside S. Chakraborty), made substantial contributions to both code development and methodology formulation. M. Roy, serving as the third author, played a crucial role in verifying the methodology and diligently proofreading the manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of data and material The datasets generated during and/or analyzed during the current study are available and will be shared via GitHub (Link: https://github.com/indrakalita).

Code availability The code developed during and/or analyzed during the current study will be made available and will be shared via GitHub (Link: https://github.com/indrakalita).

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

Competing interests The authors declare no competing interests.

References

Ammour N, Bashmal L, Bazi Y, Al Rahhal MM, Zuair M (2018) Asymmetric adaptation of deep features for cross-domain classification in remote sensing imagery. IEEE Geosci Remote Sens Lett 15(4):597–601

Ashourloo D, Shahrabi HS, Azadbakht M, Aghighi H, Matkan AA, Radiom S (2018) A novel automatic method for alfalfa mapping using time series of landsat-8 oli data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(11):4478–4487

Bakhti K, Djerriri K, Arabi MEA, Chaib S, Karoui MS (2019) Improvememt of multi-temporal vegetation modeling using hybrid deep neural networks of multispectral remote sensing images. In: IEEE International Geoscience and Remote Sensing Symposium, pp 1–4. IEEE

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1798–1828

Bruzzone L, Marconcini M (2009) Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. IEEE Trans Geosci Remote Sens 47(4):1108–1122

Bruzzone L, Prieto DF (2001) Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. IEEE Trans Geosci Remote Sens 39(2):456–460

Camps-Valls G, Gomez-Chova L, Munoz-Mari J, Rojo-Alvarez JL, Martinez-Ramon M (2008) Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. IEEE Trans Geosci Remote Sens 46(6):1822–1835

Castro J, Feitosa R, Happ PN (2018) An hybrid recurrent convolutional neural network for crop type recognition based on multitemporal sar image sequences. In: In proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp 3824–3827

Chakraborty S, Roy M (2018) A neural approach under transfer learning for domain adaptation in land-cover classification using two-level cluster mapping. Appl Soft Comput 64:508–525

Chakraborty S, Agarwal N, Roy M (2020) A deep semi-supervised approach for multi-label land-cover classification under scarcity of labelled images. In: International Conference on Soft Computing for Problem Solving (SocProS), vol. 10, pp (in press)

Di Mauro N, Vergari A, Basile TMA, Ventola FG, Esposito F (2017) End–to–end learning of deep spatio-temporal representations for satellite image time series classification. In: DC@ PKDD/ECML

Geng B, Tao D, Xu C (2011) Daml: Domain adaptation metric learning. IEEE Trans Image Process 20(10):2980–2989

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press

Gopalan R, Li R, Chellappa R (2014) Unsupervised adaptation across domain shifts by generating intermediate data representations. IEEE Trans Pattern Anal Mach Intell 36(11):2288–2302

Guo Y, Jia X, Paull D (2017) A domain-transfer support vector machine for multi-temporal remote sensing imagery classification. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp 2215–2218

Guo Y, Jia X, Paull D (2018) Mapping of rice varieties with sentinel-2 data via deep cnn learning in spectral and time domains. In: Digital Image Computing: Techniques and Applications (DICTA), pp 1–7

Haykin S (2007) Neural Networks: A Comprehensive Foundation. New Delhi, Prentice-Hall of India

Ienco D, Gaetano R (2007) Tiselac: time series land cover classification challenge. TiSeLaC: Time Series Land Cover Classification Challenge 2

Imani M, Ghassemian H (2015) Feature extraction using weighted training samples. IEEE Geoscience and Remote Sensing Letters 12(7):1387–1386

Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065):20150202

Kalita I, Roy M (2020) Deep neural network-based heterogeneous domain adaptation using ensemble decision making in land cover classification. IEEE Trans Artif Intell 1(2):167–180

Kim M, Lee J, Han D, Shin M, Im J, Lee J, Quackenbush LJ, Gu Z (2018) Convolutional neural network-based land cover

classification using 2-D spectral reflectance curve graphs with multitemporal satellite imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(12):4604–4617

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems - Vol 1, pp 1097–1105. ACM

Kullback S, Leibler RA (1951) On information and sufficiency. The Annals of Mathematical Statistics 22(1):79–86

Kussul N, Lavreniuk M, Skakun S, Shelestov A (2017) Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci Remote Sens Lett 14(5):778–782

Lavreniuk M, Kussul N, Novikov A (2018) Deep learning crop classification approach based on sparse coding of time series of satellite data. In: In proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp 4812–4815

Liang P, Shi W, Zhang X (2018) Remote sensing image classification based on stacked denoising autoencoder. Remote Sens 10(1):

Lv Q, Dou Y, Niu X, Xu J, Xu J, Xia F (2015) Urban land use and land cover classification using remotely sensed sar data through deep belief networks. J Sensors 2015

Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605

Matasci G, Volpi M, Kanevski M, Bruzzone L, Tuia D (2015) Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. IEEE Trans Geosci Remote Sens 53(7):3550–3564

McClellan, DeWitt, Hemmer, Matheson, Moe (1989) Multispectral image-processing with a three-layer backpropagation network. In: The Proceedings of the IEEE International Joint Conference on Neural Networks, pp 151–153

Meher SK, Shankar BU, Ghosh A (2007) Wavelet-Feature-Based classifiers for multispectral remote-sensing images. IEEE Trans Geosci Remote Sens 45(6):1881–1886

Nguyen PL, Ji Y et al (2019) Deep convolutional lstm network-based traffic matrix prediction with partial information. In: IFIP/IEEE Symposium on Integrated Network and Service Management (IM), pp 261–269. IEEE

Othman E, Bazi Y, Alajlan N, AlHichri H, Melgani F (2016) Three-layer convex network for domain adaptation in multitemporal vhr images. IEEE Geosci Remote Sens Lett 13(3):354–358

Othman E, Bazi Y, Melgani F, Alhichri H, Alajlan N, Zuair M (2017) Domain adaptation network for cross-scene classification. IEEE Trans Geosci Remote Sens 55(8):4441–4456

Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks 22(2):199–210

Postadjian T, Le Bris A, Sahbi H, Malle C (2018) Domain adaptation for large scale classification of very high resolution satellite images with deep convolutional neural networks. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp 3623–3626

Riz E, Demir B, Bruzzone L (2016) Domain adaptation based on deep denoising auto-encoders for classification of remote sensing images. In: The Proceedings of the SPIE Image and Signal Processing for Remote Sensing

Rukundo O, Maharaj BT (2014) Optimization of image interpolation based on nearest neighbour algorithm. International Conference on Computer Vision Theory and Applications (VISAPP) 1:641–647

Rußwurm M, Körner M (2018) Multi–temporal land cover classification with sequential recurrent encoders. ISPRS International Journal of Geo-Information 7 (4)

Scott GJ, England MR, Starms WA, Marcum RA, Davis CH (2017) Training deep convolutional neural networks for land-cover classification of high-resolution imagery. IEEE Geosci Remote Sens Lett 14(4):549–553

Senthilnath J, Omkar SN, Mani V, Tejovanth N, Diwakar PG, Shenoy A (2011) Multi-spectral satellite image classification using glowworm swarm optimization. In: The Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pages 47–50

Shi X, Chen Z, Wang H, Yeung D, Wong W, Woo W (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1, NIPS'15, page 802–810, Cambridge, MA, USA. MIT Press

Tuia D, Persello C, Bruzzone L (2016) Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geoscience and Remote Sensing Magazine 4(2):41–57

Xingjian S, Chen Z, Wang H, Yeung D, Wong W, Woo W (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810

Yan K, Kou L, Zhang D (2017) Learning domain-invariant subspace using domain features and independence maximization. IEEE Trans Cybern 48(1):288–299

Yang X, Lo CP (2002) Using a time series of satellite imagery to detect land use and land cover changes in the atlanta, georgia metropolitan area. Int J Remote Sens 23(9):1775–1798

Yang Y, Newsam S (2010) Bag–of–visual–words and spatial extensions for land-use classification. In: SIGSPATIAL International Conference on Advances in Geographic Information Systems, pages 270–279. ACM

Zhang J, Liu J, Pan B, Shi Z (2020) Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing 58(11):7920–7930

Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL (2014) Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. Front Comput Sci 10(1):96–112

Zhu L, Ma L (2016) Class centroid alignment based domain adaptation for classification of remote sensing images. Pattern Recogn Lett 83:124–132

Zhu Y, Geiß C, So E, Jin Y (2021) Multitemporal relearning with convolutional lstm models for land use classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14:3251–3265

Authors and Affiliations

Indrajit Kalita[1] · Shounak Chakraborty[2] · Talla Giridhara Ganesh Reddy[3] · Moumita Roy[4]

Shounak Chakraborty
shounak@iiitk.ac.in

Talla Giridhara Ganesh Reddy
gtalla@buffalo.edu

Moumita Roy
moumita2009.roy@gmail.com

[1] Computing & Data Sciences (CDS), Boston University, Boston 02215, Massachusetts, USA

[2] Computer Science and Engineering, Indian Institute of Information Technology Design and Manufacturing, Kurnool 518007, Andhra Pradesh, India

[3] Computer Science and Engineering, University at Buffalo, Buffalo 14260, NewYork, USA

[4] Computer Science and Engineering, Indian Institute of Information Technology Guwahati, Guwahati 781015, Assam, India