Inception time DCNN for land cover classification by analyzing multi-temporal remotely sensed images

Indrajit Kalita, *Graduate Student Member, IEEE* and Moumita Roy, *Member, IEEE*

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884128

***Abstract*—In this work, a land cover classification methodology has been investigated using very high resolution multi-temporal remotely sensed images. The technique is developed considering pre-trained deep convolutional neural networks (DCNN) followed by an inception time model for the multi-temporal image dataset. Here, initially, the pre-trained DCNN is fine-tuned to obtain the representative features in reduced dimensions for each image. Thereafter, the extracted features along with the time dimension are fed to the inception-time network for multi-temporal classification. During training, the inception-time network learns the temporal relation between the input samples. To validate the efficiency of the proposed scheme, experiments have been conducted on the two new remotely sensed multi-temporal aerial image datasets obtained across the eastern and western subcontinent regions of India. The results obtained over these two different aerial image datasets provide superior performance for the proposed scheme as compared with the other state-of-the-art techniques.**

***Index Terms*—Land cover classification; Multi-temporal remote sensing images; Convolutional neural network; Inception time network**

I. Introduction

With the extensive development of remote sensing machinery over the past few decades, the massive volume of very high resolution (VHR) remotely sensed images are captured using different types of sensing devices such as satellite, manned, and unmanned aircrafts. Such images with spectral, spatial, and temporal features can aid in the creation of accurate land cover maps [1]. These maps can be utilized for precise monitoring of the Earth and ecosystem in real-time, which have a vast spectrum of applications like crop monitoring, deforestation control, soil contamination, water pollution, biogeochemical cycling, and thermal mapping [2]. Therefore, the demand for automated techniques is rising as the development of the land cover maps by field surveys is costly, time-consuming, and mostly inefficient. In literature, the supervised machine learning (ML) techniques are applied proficiently for the automatic generation of land-cover maps by mostly considering the spectral and spatial information of the images [1]. Moreover, due to the availability of VHR remotely sensed images (on large scale), the use of supervised deep learning (DL) based approaches is becoming increasingly important. However, the classification of remotely sensed images by considering the temporal information is one of the sparsely explored topics due to the unavailability of state-of-art models and proper datasets.

I. Kalita and M. Roy are with Indian Institute of Information Technology Guwahati, India - 781015 (e-mail: indrakalita09@gmail.com; moumita2009.roy@gmail.com). Corresponding author: M. Roy

In literature, the automatic land cover classification (LCC) has been commonly categorized in two different approaches: pixel-level [1] and scene-level approaches [3]. The former class of strategy focuses on the spectral information acquired from pixels, while the latter supports taking into consideration the spectral as well the spatial information from the entire scene. Here, the performance of both types of LCC relies on the quality and quantity of available training samples. However, at pixel-level, it is difficult to identify the informative samples to train the classifier. Furthermore, the pixel-level techniques fail for the VHR datasets where each pixel contains granular information and missed the spatial information.

The advent of higher resolution images (VHR) has facilitated the development of scene-level LCC techniques [3] using the deep neural network (DNN) architectures. Such techniques provide an alternative way to generate the land cover maps by considering the spectral as well as spatial features from the entire scene. This strategy is quite effective in case of very high resolution (VHR) datasets where ignoring the unnecessary information from all the data points becomes crucial for the elimination of noise for adequate training of classifiers. Under this scenario, the DNN architectures such deep belief network (DBN) [4], deep convolutional neural network (DCNN) [3], stacked autoencoder (SAE) [5], and recurrent neural network (RNN) [6] have been popular choices for land cover classification. However, the most popular deep learning architecture for image classification is undoubtedly the deep convolution neural network (DCNN) [3], [7]. The DCNN which is primarily a stacking of convolutional neural network (CNN) is implemented by Sameen *et al.* [3] with a rectified linear unit (ReLU) activation function for aerial image classification. Similarly, Postadjian *et al.* [7] also uses the CNN architecture for the classification of SPOT satellite images with 1.5m resolution. However, the availability of ground truth information for remotely sensed image datasets is limited. Thus, it is not effective to learn knowledge from limited data using a DCNN approach directly. Nonetheless, a pre-trained DCNN can be valuable to transfer knowledge for a dataset with constrained data [8], [9]. Scott *et al.* [8] have fine-tuned a pre-trained Alexnet with a remotely sensed image dataset for land cover classification. The same pre-trained model has been used for wet-land image data classification in [9].

Following them throughout the investigation, it is observed that the state-of-the-art LCC approaches using VHR images are developed by considering the spectral and spatial information only. However, it is important to obtain certain approaches to develop some sophisticated classification model by considering not only the spectral, spatial information but

also the temporal information of the satellite data as the temporal information can improvise the performance of the LCC technologies. Moreover, the frequent visits of the extensive Earth observation satellites alleviated the limitation of VHR temporal data availability. In literature, the time series data has been handled by exploring the RNN-based architectures. Furthermore, training of these architectures requires a large amount of data and high processing resources. Nonetheless, the pre-trained DCNN-based methods work better for the images with limited data. Therefore, it motivates us to work with DCNN architecture while handling the temporal information for the classification of multi-temporal VHR remotely sensed images.

In this manuscript, the proposed algorithm goes through two major stages: fine-tuning of pre-trained DCNN architecture and training of inception time-series model. Under this scenario, initially, the pre-trained DCNN architecture is fine-tuned using the temporal dataset. Thereafter, the features of all the images are extracted using the fine-tuned network. Thereafter, a time series architecture based on the inception module is trained using extracted features. Here, the inception time series model is responsible to handle the temporal information. In this regard, the proposed work's contribution can be listed out in the following two aspects:

1) A noval solution to the practical problem of LCC by analyzing multi-temporal VHR remotely sensed images is investigated.
2) An inception-based time series DCNN architecture has been explored to extract the spectral, spatial, and temporal information of the data.

II. Proposed Methodology

As already mentioned, the proposed methodology has been carried out in two phases. The details of both phases are explained in the following sub-sections. Moreover, the graphical representation of the overall methodology is presented in Figure 1.

A. Fine-tuning of Pre-trained DCNN model to extract features

The deep CNN is a multi-layer stage (convolution, activation, and pooling layer) of the CNN architecture developed hierarchically. The output of each stage generates a feature map. In the convolution layer, a convolution operation is performed over the input, while the activation layer added non-linearity to the output of the convolution operation. Thereafter, the pooling layer reduces the spatial dimension over the activation map. After executing these three operations in multiple stages, fully connected layers followed by a classification layer are added to combine all feature maps and to obtain the class labels. However, due to the availability of limited training samples for remote sensing data, the pre-trained deep CNN model has been used here, which is fine-tuned in this phase.

In this work, the pre-trained VGG16 [10] model has been fine-tuned using the samples available in the training set. Let assume, $I = \{(\overrightarrow{x_i^j}, y_i)\}_{i=1\ to\ m}^{j=1\ to\ t}$ denotes the set of m number of train samples with t timestamp. C represents the number of different class labels, where $y_i \epsilon \{1, 2, 3, ..., C\}$. Here, the set I has been utilized to fine-tune the VGG16 pre-trained model. Under this scenario, the number of output neurons in the last layer of VGG16 has been updated corresponding to the number of available classes (i.e., C). Moreover, two dense layers with sizes $d1$ and $d2$ are placed before the classification layer. Finally, the trained fine-tuned model is used to extract the $d2$ dimensional features from the last dense layer of the network (before the classification layer). In this regard, the extracted $d2$ dimensional features $\overrightarrow{f_i^j}$ has been derived from the sample $\overrightarrow{x_i^j}$ using the fine-tuned model.

B. Training of Inception time model

The inception time model [11] is the state-of-the-art DL approach based on the inception network [12] for time series classification (TSC). The primary aim of the model is to handle the temporal information present in the input. The inception module which is the heart of the inception time network includes filters of varying lengths. It enables the network to extract multiple features from the same input. In the inception module, initially, the input is passed through a bottleneck layer to reduce the dimensionality of the time series data as well as the complexity of the model. This layer aims to mitigate the overtting problems for small datasets. Thereafter, the output of the bottleneck is fed to the three convolution layers with multiple kernels and one max-pooling layer. Furthermore, the output of all the convolution layers and max-pooling layer are combined using the depth concatenation layer. Here, the output of the depth concatenation layer is the output for the inception module. However, these outputs are forwarded by the residual block of the network. The residual block is a combination of two inception modules and a convolution layer. Following these residual blocks, a Global Average Pooling (GAP) layer is applied to averages the output of time series over the whole time dimension. In the end, a traditional classification layer with the number of neurons equal to the number of classes (C) available in the dataset is employed. In the current investigation, the inception time network has been trained using the extracted features $\overrightarrow{f_i^j}$ of the multi-temporal dataset. Finally, the trained model has been used to obtain the class label of the final test data.

III. Experimental Results

A. Descriptions of datasets

The experimentation using two temporal image data sets from different regions evaluates the performance of the proposed approach. Here, the images are captured using Google earth pro software from the entire subcontinent which is then divided into two sets zone-wise. 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 datasets are a combination of 6 land cover classes. The total number of images in ESD and WSD are 956 and 1215, respectively. The descriptions of the classes and the number of samples available under each class have been listed in Table I.

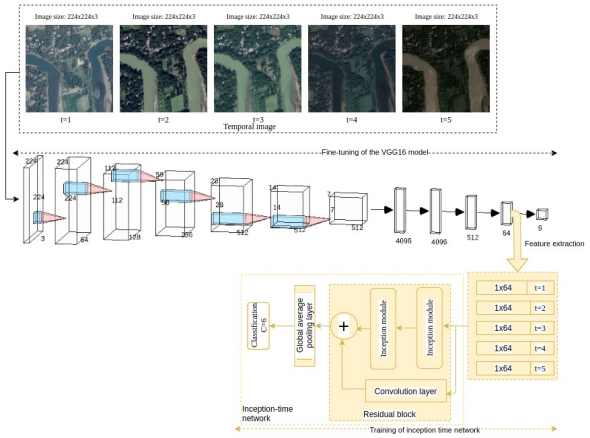


Fig. 1: Graphical representation of the proposed model

TABLE I Description of the ESD and WSD datasets

Name of the Class	Number of time series samples	
	ESD	WSD
Agriculture	132	162
Forest	142	180
River	224	164
Beach	108	164
Residential	188	324
Sea	187	215
Total	**981**	**1209**

B. Details of experimentation

1) Setup of fine-tuned DCNN: As already mentioned, VGG16 has been used as a pre-trained model in the current investigation.During the fine-tuning phase, the number of output neurons in the last dense layer of the pre-trained model has been changed based on the number of classes available in the cross-domain scenario. Moreover, the two dense layers of size $d1 = 512$ and $d2 = 64$ are placed before the classification layer. For experimentation in the first phase, the minibatch size and initial learning rate are fixed at 16 and 0.0001, respectively, and the model has been iterated for 50 number of times. As already mentioned, the trained fine-tuned model is used as a feature extractor for the multi-temporal dataset. In this regard, the extracted $d2 = 64$ dimensional features $\overrightarrow{f_i^j}$ has been derived from the sample $\overrightarrow{x_i^j}$ available in I. Here, the number of timestamps i.e., $j = 5$ for all the images.

2) Setup of inception time network: In the proposed investigation, the inception time network is developed using two inception time modules along with one residual connection and one average pooling layer. Here, the input of the extracted features of size $1 \times 64 \times 5$ (where 5 represents the number of timestamps)have been forwarded for the bottleneck layer (in the inception module). Thereafter, the bottleneck layer operates on sliding 32 kernels of length 1 with a stride equal to 1. Moreover, the three convolutional layers with a kernel of size 10, 20, and 30 are incorporated on the output of the bottleneck layer. Furthermore, the max-polling operation is also performed with three kernels on the bottleneck output. Finally, the features of the convolution and max-pooling layers are merged to obtain the 128 feature maps. On the other hand, the residual block connects the output of two inception modules and one convolution layer. Here, the convolution layer generates 128 feature maps. For experimentation, the inception time network is trained by keeping the learning rate, and momentum as 0.01 and 0.09, respectively. The model has been iterated for 20 number of times with a fixed batch of size 1024.

TABLE II The class-wise precision (P), recall (R), F1-score (F), and overall accuracy (OA) for ESD and WSD. Results are in percentage

Classes	ESD			WSD		
	P	R	F	P	R	F
Agricluture	96.02	98.05	97.02	**100.00**	**100.00**	**100.00**
Beach	99.21	90.00	94.38	96.21	97.80	97.00
Forest	95.60	96.13	95.86	92.00	99.17	95.45
Residential	91.54	97.29	94.33	**100.00**	94.99	97.43
River	99.67	99.97	99.82	**100.00**	97.60	98.79
Seawater	**100.00**	**100.00**	**100.00**	96.91	**100.00**	98.43
Average	97.01	96.91	96.90	97.52	98.26	97.85
Overall		97.40			98.39	

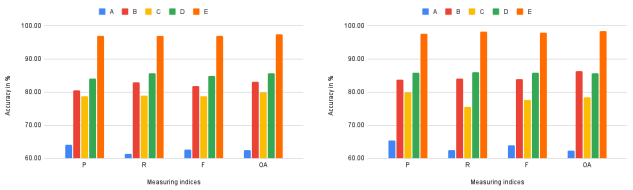
C. Result and analysis

1) Classwise analysis of results using two datasets: To access the effectiveness of the proposed methodology, the experiments have been conducted on two remotely sensed datasets (ESD and WSD). In this regard, the precision (P), recall (R), F1-score (F), and overall accuracy(OA) of the two datasets are depicted in Tables II. It is worth mentioning that the results are obtained by taking the average of the 5-fold cross-validation. In the case of ESD, it is found that the P, R, and F reached the accuracy 100% for the class 'Seawater'. However, the average of P, R, and F is reported as 97.01%, 96.91%, and 96.90% respectively, considering all the 6 classes. Moreover, the overall accuracy (OA) is reported as 97.40%. Similarly, for WSD, it is found that three classes (i.e., Agriculture, Residential, River) have been successfully reached the maximum precision (100%), while two classes (i.e., Agriculture and Seawater) have been effectively distinguished by the proposed methodology as R reached the maximum value (i.e., 100%). Moreover, the average of P, R, and F is reported as 97.52%, 98.26%, and, 97.85%, respectively. Further, the overall accuracy (OA) is reported as 98.39%.

2) Comparative analysis of the result obtained using the proposed approach with the state-of-the-art approach in remote sensing literature: The effectiveness of the proposed methodology for multi-temporal land-cover classification has been evaluated using two remote sensing datasets (i.e., ESD and WSD). The comparison of the performance of the investigated methodology (E) has been carried out against the four different scenarios: (a) situation where pre-trained Inception-V3 [13] is implemented to extract features of the data and then the features are used to train a multi-layer classifier, (b) situation where the Inception-V3 [13] is fine-tuned to extract the relevant features for the datasets and then trained a long short term memory (LSTM) network for time-series classification, (c) situation where Scott *et al.,* [8] have fine-tuned a pre-trained Alexnet with a remotely sensed image dataset for land-cover classification, (d) situation where Bakhti *et al.,* [14]

TABLE III Comparision of results (in percentage) of proposed approach (E) in terms of precision (P), recall (R), F1-score (F), and overall accuracy (OA) using ESD and WSD obtained using Inception-V3 with MLP technique (A), Inception-V3 with LSTM model (B), techniques implemented by Scott *et al.,* [8] (C), and technique explored by Bakhti *et al.,* [14] (D)

Method	ESD				WSD			
	P	R	F	OA	P	R	F	OA
A	64.02	61.40	62.68	62.56	65.31	62.48	63.86	62.37
B	80.56	82.97	81.75	83.09	83.67	84.10	83.88	86.31
C	78.67	78.93	78.80	79.83	79.92	75.57	77.68	78.44
D	84.11	85.69	84.89	85.69	85.82	85.94	85.88	85.69
E	**97.01**	**96.91**	**96.96**	**97.40**	**97.52**	**98.26**	**97.89**	**98.39**



(a) ESD dataset (b) WSD dataset

Fig. 2: Bargraph representation for the comparision of results of proposed apprach (E) (in percentage) in terms of precision (P), recall (R), F1-score (F), and overall accuracy (OA) using ESD and WSD obtained using inception-V3 with MLP technique (A), inception-V3 with LSTM model (B), techniques implemented by Scott *et al.,* [8] (C), and technique explored by Bakhti *et al.,* [14] (D)

implemented a hybrid DNN architecture (CNN+LSTM) for multi-temporal vegetation modeling using Sentinel-2A images.

The precision (P), recall (R), f1-score (F), and overall accuracy (OA) over all the classes for the proposed approach and the other state-of-the-art techniques are stated in Table III. Furthermore, the graphical representation for the same is presented in Figure 2. For the ESD dataset, the proposed approach (E) has outperformed technique A by a margin of $\approx 32\%$, $\approx 35\%$, $\approx 34\%$, and $\approx 34\%$ considering the parameter P, R, F, and OA. Similarly, it defeats the scheme B and C by a margin of $\approx 16\%$ (in P), $\approx 13\%$ in (in R), $\approx 15\%$ (in F), $\approx 14\%$ (in OA) and $\approx 18\%$ (in P and F), $\approx 17\%$ in (in R and OA), respectively. Moreover, the proposed scheme (E) brings an improved result of $\approx 12\%$ in P, F and $\approx 11\%$ in R, OA as compared to technique D. In the case of WSD, It is observed that the proposed approach (E) has outperformed technique A by a margin of $\approx 32\%$, $\approx 35\%$, $\approx 33\%$, and $\approx 36\%$ considering the parameter P, R, F, and OA. Similarly, it defeats the scheme B and C by a margin of $\approx 13\%$ (in P and F), $\approx 14\%$ in (in R), $\approx 12\%$ (in OA) and $\approx 17\%$ (in P), $\approx 22\%$ (in R), $\approx 20\%$ (in F), $\approx 19\%$ (in OA), respectively. Moreover, the proposed scheme (E) brings an improved result of $\approx 12\%$ in P, F and $\approx 11\%$ in R, OA as compared to technique D.

IV. Conclusion

In the present work, an inception time deep neural network has been developed for automated land cover classification using multi-temporal remotely sensed images. The entire strategy is a combination of pre-trained DCNN and inception time networks. The findings of the investigation over the collection of two different satellite image datasets confirm its potential to produce promising efficiency relative to the other state-of-the-art techniques. In addition, it has been observed that the inception time model can learn the temporal relationships from a set of the time-series images.

References

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

[2] S. K. Meher and D. A. Kumar. Ensemble of adaptive rule-based granular neural network classifiers for multispectral remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(5):2222–2231, 2015.

[3] M. I. Sameen, B. Pradhan, and O. S. Aziz. Classification of very high resolution aerial photos using spectral-spatial convolutional neural networks. *Journal of Sensors*, 2018, 2018.

[4] V. Mnih and G. E. Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. IEEE, 2010.

[5] P. Liang, W. Shi, and X. Zhang. Remote sensing image classification based on stacked denoising autoencoder. *Remote Sensing*, 10(1):16, 2018.

[6] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1685–1689, 2017.

[7] T. Postadjian, A. Le-Bris, C. Mallet, and H. Sahbi. Superpixel partitioning of very high resolution satellite images for large-scale classification perspectives with deep convolutional neural networks. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1328–1331. IEEE, 2018.

[8] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis. Training deep convolutional neural networks for landcover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters*, 14(4):549–553, 2017.

[9] M. Mahdianpari, M. Rezaee, Y. Zhang, and B. Salehi. Wetland classification using deep convolutional neural network. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 9249–9252. IEEE, 2018.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P. Muller, and F. Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 2015.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *International Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[14] K. Bakhti, K. Djerriri, M. E. A. Arabi, S. Chaib, and M. S. Karoui. Improvememt of multi-temporal vegetation modeling using hybrid deep neural networks of multispectral remote sensing images. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 1–4. IEEE, 2019.