TIME SERIES LAND COVER CLASSIFICATION BASED ON SEMI-SUPERVISED CONVOLUTIONAL LONG SHORT-TERM MEMORY NEURAL NETWORKS

S. Jing [1], T. Chao [2, *]

[1] School of Geosciences and Info-physics, Central South University, Changsha, China - sjing069@csu.edu.cn
[2] School of Geosciences and Info-physics, Central South University, Changsha, China – kingtaochao@csu.edu.cn

Commission III, WG III/6

KEY WORDS: Time series imagery, Land cover classification, Semi-supervised learning, LSTM, deep learning

ABSTRACT:

Time series imagery containing high-dimensional temporal features are conducive to improving classification accuracy. With the plenty accumulation of historical images, the inclusion of time series data becomes available to utilize, but it is difficult to avoid missing values caused by cloud cover. Meanwhile, seeking a large amount of training labels for long time series also makes data collection troublesome. In this study, we proposed a semi-supervised convolutional long short-term memory neural network (Semi-LSTM) in long time series which achieves an accurate and automated land cover classification with a small proportion of labels. Three main contributions of this work are summarized as follows: i) the proposed method achieve an excellent classification via a small group of labels in long time series data, and reducing dependence of training labels; ii) it is a robust algorithm in accuracy for the influence of noise, and reduces the requirements of sequential data for cloudless and lossless images; and iii) it makes full advantage of spectral-spatial-temporal features, especially expanding time context information to enhance classification accuracy. Finally, the proposed network is validated on time series imagery from Landsat 8. All quantitative analyses and evaluation indicators of the experimental results demonstrate competitive performance in the suggested modes.

1. INTRODUCTION

Remote sensing image classification is extensively used in various areas, like change detection. With the growing development of remote sensing technology, multi-source and multi-temporal Earth Observations (EO) data have easier to access. The increasing demand for products encouraged scholars to research approaches of remote sensing image classification that can take full advantage of the rich information, incorporated spatial, spectral, and temporal data, aimed to improve the accuracy of classification and meet a wide range of information requirements and applications.

The current classification methods mainly comes from the task of singe-phase image, which can be roughly divided into three categories: unsupervised, supervised and semi-supervised classification methods. Unsupervised classification algorithms cluster elements by similar attributes without any priori human intervention, like ISODATA (Boles et al., 2004). Despite the automatic extraction in the classified process without any previous knowledge and samples, they become time consuming when high dimension or large volume of data (Chen, Gong, 2013), and interpreting clusters correctly is a major challenge. At the same time, supervised classification algorithms (e.g. Random Forests, RF (Belgiu, Dragut, 2016); Support Vector Machines, SVM (Mountrakis et al., 2011); and Artificial Neural Networks, ANN (Bagan et al., 2005)) identify other unknown categories of pixels by learning priori knowledge. For these methods, selecting a representative and abundant training samples is crucial. However, the statistical distribution of various types of objects in remote sensing images is complex and random. The training samples are artificially selected through limited experience and knowledge whether it is field exploration or reference data. There is no guarantee that the selected classification samples have a valid representation of corresponding land cover classes. The reliance on selected samples also hinders the application and development of different spatial and temporal image classifications. Therefore, semi-supervised learning from the field of data mining is applied to various classifiers, like the self-learning (Wang et al., 2015) and the graph-based method (Jamshidpour et al., 2016). It mines the inherent structural features of object types in unlabelled samples to correct fitting classifiers that may be caused by the poor representation of labelled samples. The semi-supervised classifiers can effectively improve the problem of poor representative known samples, and can solve the classification problem of small samples or areas where effective classification samples cannot be obtained in practice.

As the development of a variety of satellite technologies, the accumulation of multi-source and historical data enables more abundant information to be exploited, not limited to the basis of a single-phase image. It turns out that inclusion of time and ancillary data improved the accuracy of the classification (Khatami et al., 2016). More and more researchers explore novel approaches incorporated temporal element which utilize the differences of various ground object categories in time context in classification and other fields. For example, (Jia et al., 2014) explored the time series NDVI data to improve land cover classification, and especially phenological features had a significant effect on accuracy.

With the heavy attention of deep learning, numerous studies have investigated that recurrent neural networks (RNNs) have been widely employed in time series analysis and applications. When an image is divided into sequence data by row, long

* Corresponding author

short-term memory (LSTM) (Greff et al., 2017) which is a variant of RNN can also implement a high-precision result of classification. Then, (Russwurm and Korner, 2017) used LSTM to identify crops by temporal vegetation modelling, and proved that the LSTM-based method performed better than the classical RNN. In addition, convolutional neural networks (CNNs) are well adapted to cope with spatial autocorrelation, and thus reach a high accuracy of semantic segmentation and classification in images (Scott et al., 2017; Shelhamer et al., 2017). On this basis, several network variants were proposed which combines convolutional and recurrent neural network components. Convolutional LSTM network (ConvLSTM) was first proposed for the goal of spatiotemporal sequence precipitation nowcasting (Shi et al., 2015). After that, various models combining convolutional and recurrent neural networks were trained and applied in other fields, like (Mou et al., 2019) utilized a recurrent convolutional neural network (ReCNN) for change detection in biphasic multispectral images.

However, deep neural networks usually require the quantity and quality of training samples, which undoubtedly increases the workload and computation (Gomez et al., 2016). On the one hand, obtaining training labelled data for classification tasks of long time series imagery is a challenge. It is easier to get the label at a certain moment actually, instead of all labels of time series data. The more labels you require, the larger cost and workload you pay. On the other hand, the selection of training samples is a major challenge. In terms of the representation of samples in supervised classification mentioned above, people have defaulted that the selected samples have a good representation for individual classes because of the subjective judgment. It is also not conductive to the further development of image classification methods, because the unsatisfactory classification results always attribute to the inapplicability of the algorithm or the parameter selection.

Besides, for optical remote sensing images, it is difficult to avoid cloud, snow and shadow cover. Even though historical satellite images are easy to seek, there are a few inevitable conditions of discontinuous time series data due to these noisy observations. This problem will affect accuracy of results to some extent, so a sort of pre-processing will be employed in advance, such as cloud and shadow detection (Zhu, Woodcock, 2012), or fitting the multi-temporal curve to fill in missing values (Brooks et al., 2012; Yuan et al., 2015).

In order to address these issues mentioned above, we propose a semi-supervised convolutional long short-term memory neural network (Semi-LSTM) in long time series imagery. Different from traditional deep learning classifiers, this end-to-end Semi-LSTM achieves an accurate and automated land cover classification with a small proportion of labels, and weakens the dependence on the selected samples. The main contributions of this work are as follows: i) it achieves an excellent classification via a small group of labelled samples in long time series data, and reducing the cost of labels; ii) it is a robust algorithm to a certain degree for classification of continuous time series data with the influence of noise (especially clouds and shadows), which even can be learned and identified as noisy features. It decreases the requirements of time series imagery without cloud cover and missing values, and can be widely used in areas that are often obscured by clouds, such as subtropical areas; iii) due to the integration of the convolutional neural network and the recurrent neural network, it makes full use of spectral-spatial-temporal characteristics to classify time series satellite images, especially expanding the time context information to enhance classification accuracy.

2. METHODOLOGY

In this study, we designed a semi-supervised convolutional long short-term memory neural network (Semi-LSTM), which models the spectral, spatial and temporal information for classifying the long time series imagery with a small number of labels. Figure 1 illustrates the basic framework of our proposed method, which can be decomposed into two parts: a pretrained model and a semi-supervised ConvLSTM model. In the following, we first detailed a pre-trained CNN model to enhace spatial, texture and spectral features for each image of time series data. Then, we focused on modeling the time context information with a small account of training labeled samples by the semi-supervised ConvLSTM. Furthermore, the time series data will be divided into two pieces: one for training is all images except the last moment, and the other is the time series subdata containing the last time image for testing and retrieving the final predicted label.

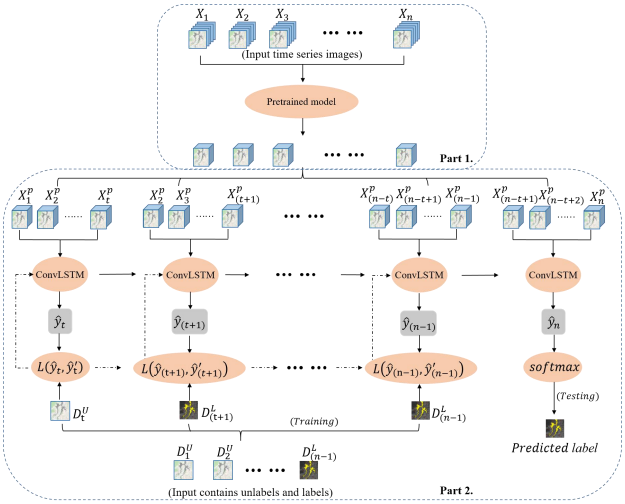

Figure 1. The framework overview of our proposed Semi-LSTM model for land cover classification in time series images. The inputs containing images (X_1,X_2,...,X_n) and corresponding labels (which consist of labelled D_i^L and unlabelled data D_i^U, i=1,2,...(n-1) here) produce the final predicted label. X_i^p (i=1,2,...,n) represents the feature map after pretraining. $\hat{y}$ and $\hat{y}'$ denote predictions from the same time series subdata set due to random squential variation, and $L(\hat{y}, \hat{y}')$ is the loss function.

2.1 Enhancing spatial features via the pre-trained model

Previous works (Marmanis et al., 2016; Zhou et al., 2017) in remote sensing fields have demonstrated that feature extractors based on CNNs can effectively generate powerful feature representations from the complexity of satellite images. There are also no sufficient remote sensing datasets to train a network from scrath, so the pretrained CNN designed for a completely different classification task is used as a feature extractor. Here, we choose the Residual Network (ResNet) (He et al., 2016) as the pre-trained CNN model, which had been trained by a large database named ImageNet (Deng et al., 2009) containing approximately 15 million high-resolution natural images and labels with twenty thousand classes.

The input is a time series data set T which is composed of multispectral images $X_i \in \mathbb{R}^{(h\times w\times c)}$ in order of time (i=1,2,3,$\cdots$,n). As shown in Figure 2, a new image $X_i^p \in \mathbb{R}^{(h\times w\times c')}$ (c'>c) with powerful feature representations is output for each multi-spectral image. Specifically, because the ResNet is trained from three-channel images, it is necessary to perform principal

component analysis (PCA) on the original multispectral image. Then, three principal components are extracted and input them into the pretrained CNN. Due to the low- and middle-resolution remote sensing images, the deeper convolution layer will lost the more spatial and texture features. This is the reason why we choose the feature map after the first convolutional layer. After the convolution kernel with the size of 7×7, the height and width of current feature map are reduced to $h/2$ and $w/2$, and the number of channels is increased to 64. In order to restore the same h and w as the original image, it is upsampled by bilinear interpolation F_{up}. At this time, the enhanced spatial and texture information has been received. To further retain the multi-spectral information, the multiple channels of feature maps are concatenated with the multispectral of remote sensing images (F_{cat}). In the end, we get a series of images with enhanced feature extraction ($X_1^p, X_2^p, X_3^p, ..., X_n^p$). The pretrained process can be summarized as (1).

$$X_i^p = F_{cat}\left\{F_{up}\left(ResNet^1\left(X_i\right)_{PCA}\right),\ X_i\right\} \tag{1}$$

where $ResNet^1$ = the first convolutional layer of *ResNet*

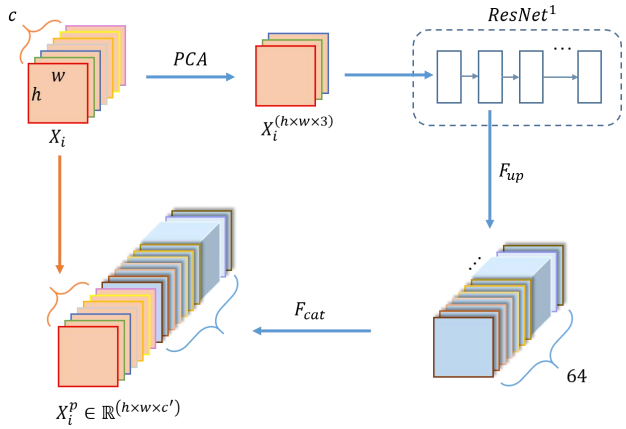

Figure 2. The pre-training process, and X_i^p is the output.

2.2 Learning temporal features via the semi-supervised ConvLSTM model

In this section, we detail the semi-supervised ConvLSTM model and how to deal with the classification task of time series images. Inspired by the convolutional LSTM network for precipitation nowcasting (Shi et al., 2015), we adopted the ConvLSTM network to process time series images. In the following, we recall the standard ConvLSTM unit in Figure 3. Owing to the convolutional operations both in input-to-state and state-to-state transitions, it can better preserve spatial information and reduce the redundancy of spatial data compared with the traditional LSTM. It learns time context information in a core memory cell that is jointly modulated by three gates: input (i_t), forget (f_t) and output (o_t) gate. As an input X_t feeds into, the current cell state $\hat{S}_t$ gathers the current information and the previous state $H_{(t-1)}$. The long-term memory S_t will be accumulated from both $\hat{S}_t$ controlled by the input gate i_t and the past cell state value $S_{(t-1)}$ controlled by the forget gate f_t. S_t is further controlled by the output gate o_t and propagates to the final hidden output H_t (the short-term memory). The key formulas of ConvLSTM are shown in (2) below.

After the pre-training model, the new time series data set needs to be divided into numerous time series subdata sets with the same time series length (t), that is $T_1=(X_1, X_2, \ldots, X_t)$, $T_2=(X_2, X_3, \ldots, X_{(t+1)})$… It is beneficial to the computational efficiency and the training effect of the temporal model. We take the sequence subdata T_1 as an example (Figure 3). The self-looping structure based on RNN can be regarded as a connection with multitudinous neural units, as shown in Figure 4 where the pink box represents a neural unit. As each image inputs in sequence, the weights and memory state values are continuously updated and saved. The weights among the hidden layers are shared, resulting of the network has the ability to remember. Owing to the many-to-one network form, each sequence input has only one output, namely the sequence data T_1 outputs the latest state H_t. For the classification task of time series imagery, each input image pass sequentially to the ConvLSTM encoder, and then the prediction $\hat{y}_t$ outputs by a softmax function.

$$\begin{aligned}
i_t &= \sigma(W_{i,X} * X_t + W_{i,H} * H_{t-1} + W_{i,S} \cdot S_{t-1} + b_{in}) \\
f_t &= \sigma(W_{f,X} * X_t + W_{f,H} * H_{t-1} + W_{f,S} \cdot S_{t-1} + b_f) \\
o_t &= \sigma(W_{o,X} * X_t + W_{o,H} * H_{t-1} + W_{o,S} \cdot S_{t-1} + b_o) \\
\hat{S}_t &= \tanh(W_{S,X} * X_t + W_{S,H} * H_{t-1} + b_S) \\
S_t &= f_t \cdot S_{t-1} + i_t \cdot \hat{S}_t \\
H_t &= o_t \cdot \tanh(S_t)
\end{aligned} \tag{2}$$

where σ = sigmoid function
tanh = hyperbolic tangent function
W = weights matrices
b = bias coefficients
'*' = the convolution operator
'·' = a Hadamard product

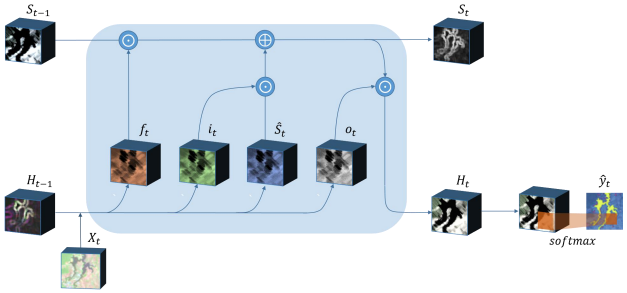

Figure 3. A neural unit of ConvLSTM network when X_t is input.

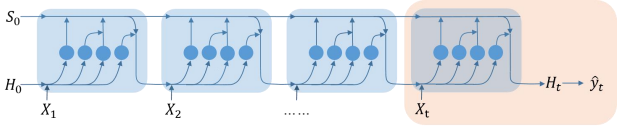

Figure 4. The illustration of ConvLSTM for time series data.

For traditional ConvLSTM, the loss function is defined as the standard cross-entropy which is calculate by the predicted label and respective reference label from ground truth data during the training process. This is essentially a supervised classification that needs large enough labels for training. Actually for the classification task of long time series remote sensing data, corresponding labels are mixtures with labelled D_i^L and un-labelled data D_i^U, rather than labels for every image. Thus, semi-supervised learning inspired by Π-model (Laine and Aila, 2018) is utilized. Figure 5 illustrates the procedure of semi-supervised learning and the calculation of loss function L. As shown in (3), L consists of supervised and unsupervised loss components. One is the standard cross-entropy between models' predictions $\hat{y}_i$ and reference labels D_i^L, evaluated for labels only. Because of the class imbalance, inspired by the Focal Loss (Lin et al., 2020), the tunable focusing parameter γ is adopted to add a modulating factor $(1-\hat{y}_i * D_i^L)^\gamma$ to the cross entropy loss. The

other is evaluated for all inputs with labelled and un-labelled data. Because of random sequential variation, when the same time series data is input, two different predicted vectors from hidden layers ($\hat{y}_i$ and $\hat{y}_i'$) are obtained. Then the mean square difference between these two values is made. Besides, to combine the two loss terms, the latter is scaled by time-dependent loss weighting function $w(t)$ for a more accurate training model. The initial value of $w(t)$ is set to 0, that is, the loss value of the unsupervised part is not calculated. With the new input and continuous iterative calculation, the value of $w(t)$ ramps up. The *Adam* optimizer is utilized to converge the loss function. In the end, the optimization of the model parameters is completed. It should be noted that one-hot coding (Chren, 1995) is a smart way to demote values of different categories for multi-class tasks. In testing, the prediction $\hat{y}_n$ at the time n is produced by the trained network. Finally, a softmax function is utilized to generate the predicted label, where each pixel corresponds to a category.

$$L(\hat{y}_i, \hat{y}_i') = -\frac{\sigma}{|B|}\left(1 - \hat{y}_i * D_i^L\right)^{\gamma} \sum_{i \in (B \cap L)} \log\left(\hat{y}_i, D_i^L\right) + w(t)\frac{1}{C|B|}\sum_{i \in B}\left\|\hat{y}_i - \hat{y}_i'\right\|^2 \quad (3)$$

where σ = the balanced factor
B = each minibatch
L = training input with known labels
C = the number of different classes

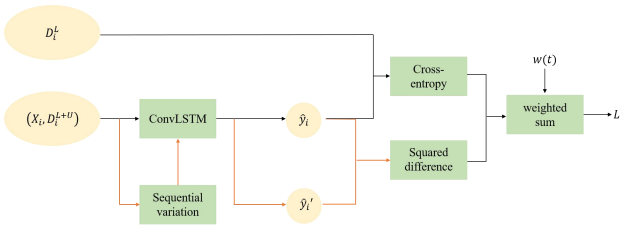

Figure 5. The process of semi-supervised learning and loss function L

3. TIME SERIES DATA

The study area is located in a town in Jiamusi City, Heilongjiang Province of China, which is in the hinterland of Sanjiang Plain between 45°42′-48°31′N latitude and 129°22′-129°42′E longitude. We select three typical study areas with significant land cover changes which are mainly based on urban expansion and natural phenology of various crops from 2015 to 2016. Each area has the same size (256×256 pixels), with an area of about 60 km^2. All images are acquired by the Landsat 8 from OLI sensor with nine bands and a spatial resolution of 30m, but we only keep 7 bands, including costal aerosol, blue, green, red, NIR, and two SWIR bands.

To generate time series data with approximately equal intervals, we adopt all available Landsat data during two years with cloud-free or low cloud coverage. Due to a 16-day revisit cycle and free resources of satellite, 26, 38 and 36 images are collected in three areas (named Data_1, Data_2 and Data_3) respectively (shown in Figure 6) after pre-processing of images (including radiometric calibration, atmospheric correction, relative image registration and cropping). Figure 7 illustrates the different time distributions of images in various regions. Each dataset has an average of 1 to 2 scenes per month.

In order to train and verify classification model, we manually label the multi-temporal data with 6 covered classes according to the actual situation, involving cultivated field, forest, construction (incorporated buildings and roads), water, cloud cover and shadow. However, the distribution among the categories is unbalanced due to human activities. Cloud cover and shadows appear instantaneously and vary widely. In the remaining four categories, cultivated land accounts for the largest proportion of all study areas (around 58%, 66% and 73% in three datasets respectively), followed by construction and forest, and the water area accounts for the least proportion. In particular, there is no water in Data_3. We prepared the corresponding labels of all images, but the actual situation is not ideal for time series imagery. Therefore, in the subsequent experiments, we will simulate the actual situation to remove part of the label data. For example, the labels corresponding to the images blocked by clouds, or when exploring the influence of the number of labels on the training network, some labels will be removed chronologically.

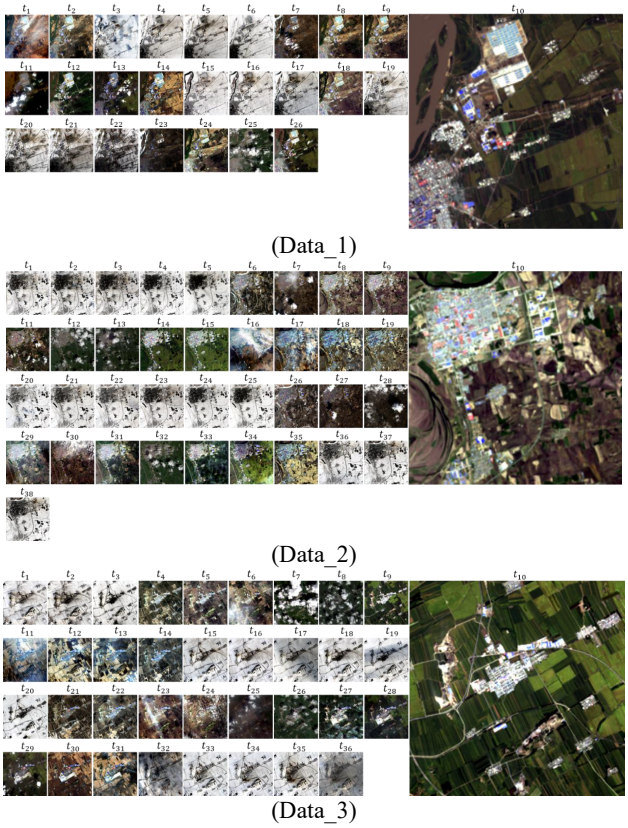

Figure 6. Three typical study areas (Data_1, Data_2 and Data_3) which are mainly based on urban expansion and natural phenology of various crops from 2015 to 2016.

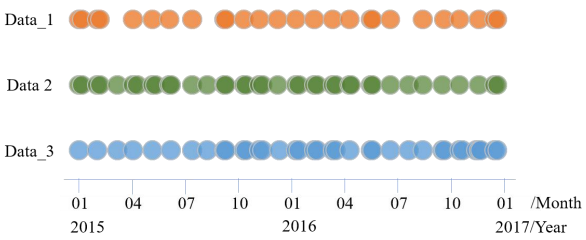

Figure 7. The time distributions of three experimental data sets. Each data set has an average of 1 to 2 scenes per month.

4. RESULT

4.1 Experimental setup

In order to verify the feasibility and effectiveness of proposed method, other RNN-based methods (ConvLSTM and LSTM) and non-deep learning methods (SVM and RF) were used in time series classification task for comparison. All models with various parameter values were repeatedly tested many times in the Window 10 platform with a single NVIDIA RTX 2080 Ti GPU (memory 11 GB) and a CORE i7-7800X CPU. Each model were configured with appropriate hyper-parameters to achieve the best classification under the limited hardware conditions. There were several identical parameters and setting. Due to the small size of the three experimental data sets in the training of complex deep neural networks, some missing feature information appeared to be sensitive to the classification accuracy. As a result, pretrained models were used in all models.

Deep learning networks based on RNN were implemented in Tensorflow. The learning rate was set to 0.001, and the optimal length of time series was 20 (t=20) which performed well and did not take too much time. More details and analysis on it are expanded in section 5.1. For ConvLSTM and our Semi-LSTM models, thanks to the convolutional kernel size of 3×3, we fed original 3D image into the network of 128 neural units. To utilize spatial and spectral information effectively, we set the same patch size to *128×128* and sample the input at intervals of *64* pixels stride on every image. Differently, for LSTM, due to the pixel-level classification, each 3D image was reshaped to a vector with two dimensions (h×w, c), and then *1024* pixels as a batch fed into the network with *256* units. For non-deep-learning classifier, SVM and RF, we employed the Scikit-learn framework. To deal with unbalanced samples, *class weight = 'balanced'*, so that each class had a weight based on the size of training samples (Zhong et al., 2019). The random search strategy (Bergstra, Bengio, 2012) was also adopted to automatically pick the major optimal parameter values of model. For SVM with RBF kernel, the candidate C ∈ {0.01,0.03,0.1,0.3,1,3,5,7,10,15,30,100,300}, and gamma ∈ {0.1,1,2,4,10,'auto'}. For RF classifier, the candidate parameters: n_estimators ∈ {120, 300,500,800,1200}, max_depth ∈ {5,8,15,25,30,None}, min_samples_split ∈ {2,5,10,15,100}, min_samples_leaf ∈ {1,2,5, 10}, and max_features ∈ {'log2', 'sqrt', None}. Finally, the optimal parameters of SVM are $\theta_{SVM}^{Data\_1}$ = (C=0.1, gamma= 0.1), $\theta_{SVM}^{Data\_2}$ = (10, 0.1), and $\theta_{SVM}^{Data\_3}$ = (10,0.1) respectively. The RF preformed best with $\theta_{RF}^{Data\_1}$ = (n_estimators=1200, max_depth=8, min_samples_split=2, min_samples_leaf=1, max_features = 'log2'), $\theta_{RF}^{Data\_2}$ = (500,25,15,1,'log2'), and $\theta_{RF}^{Data\_3}$ = (800,25,10,10, 'log2') respectively.

4.2 Accuracy assessment of classification

To evaluate the performance of various classifiers, the following indicators are utilized: overall accuracy (OA), kappa coefficients (K) and the weighted F1 score (W-F1). All the evaluation indexes are employed in Scikit-learn package of Python. We tested many times on three long time series data sets to compare with our proposed model. The results for testing of different models are displayed in Figure 8, and more detailed accuracy assessments of classification are list in Table 1.

SVM and RF as a non-deep learning method performed well which have been widely used in classification (Carrao et al., 2008; Zhang et al., 2014). However, selecting corresponding parameters of these classifiers will still be repeated when targeting a new data set, even if the help of Random Search strategy. As input features are larger or more complex, the classification accuracy decreases and the training process is more time consuming. The best F1 score only reaches 75.56% for SVM and 75.52% for RF. As for data with the non-uniform class distribution, they are prone to omission and commission errors in class with a small number of samples, like forest and water cover in Figure 8.

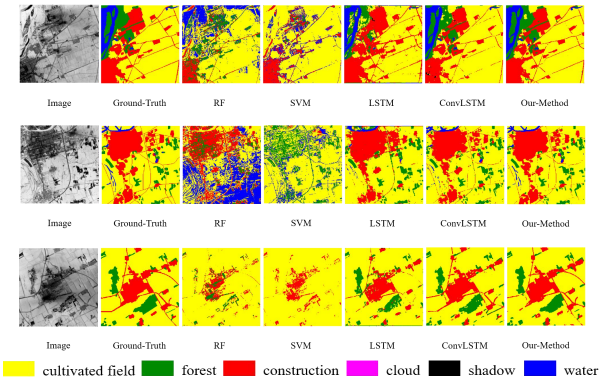

Figure 8. The testing results of different models are displayed. From top to bottom is Data_1, Data_2 and Data_3.

Classifiers	OA	K	W- F1
		Data_1	
RF	67.10%	0.29	0.58
SVM	64.65%	0.24	0.56
LSTM	88.21%	0.80	0.88
ConvLSTM	92.45%	0.87	0.92
Semi-LSTM (ours)	97.95%	0.97	0.98
		Data_2	
RF	75.52%	0.50	0.74
SVM	73.27%	0.37	0.69
LSTM	86.65%	0.74	0.86
ConvLSTM	90.01%	0.80	0.90
Semi-LSTM (ours)	94.31%	0.89	0.94
		Data_3	
RF	80.07%	0.40	0.78
SVM	75.56%	0.31	0.70
LSTM	81.67%	0.59	0.80
ConvLSTM	89.82%	0.77	0.88
Semi-LSTM (ours)	97.14%	0.93	0.97

Table 1. Overall accuracy (OA), kappa coefficients (K) and weighted F1 score (W-F1) achieved by various classifiers in three study areas.

Then, since LSTM is a variant RNN network with three gates, it is possible to obtain temporal context information and achieve better than non-deep learning methods. Because of the loss of information in the conversion of images, the classification of LSTM model cannot express the details well, such as omission errors in narrow roads. The ConvLSTM model further improves the classification accuracy (which OA of three data sets can reach around 90%). That is because the convolutional structures replace the fully connections of LSTM, and spatial information is effectively encoded. The ConvLSTM method is good for dealing with temporal and spatial features, and its expression of spatial information has more advantages than LSTM. As for our Semi-LSTM model which incorporates semisupervised learning based on ConvLSTM model, it performs best in classification and OA, K and W-F1 increased by an average of 5.6%, 0.1 and 0.06 respectively. It proves that the semi-supervised classifier improves the dependence on selected samples. Owing to the

addition of semi-supervised learning, it is a robust classifier especially in the case of a small number of labelled training samples (further discussed in Section 5.2), and solves the issue of poor representation in training samples to some extent.

5. DISCUSSION

5.1 The importance of temporal context information

To explore the effect of the time series length on classification, we implemented our Semi-LSTM model with same parameters except the length of time series data for comparison, t={2,6,10, 14,18,20,22}. It is worth noting that a fixed length (t) is used and moves it step by step to the last training image as an epoch to traverse all data for training. The results are illustrated in Figure 9. Since the length of time series imagery is limited, the corresponding number of iterations will be reduced when the length of time series input t is enlarged. Therefore, as the t increases, the accuracy of classification generally improves to a stable trend, or even decreases slightly. Meanwhile, the larger length of time series, the memory of our deep neural network may be lost during the continuous updating process, just as the memory of human beings for the longer time is blurred. That is the reason why the overall accuracy of t=22 is lower than that of t=20. In this comparison experiment, it is also found that the larger t, the more time it takes. Thus, we chose an optimal time step (t=20), which could not only ensure high-precision classification, but also save time on excessive consumption.

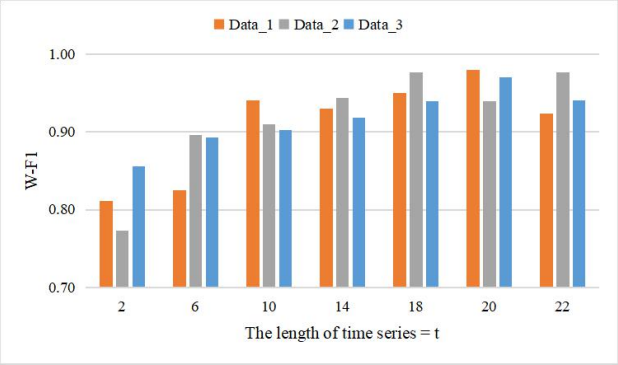

Figure 9. The results of comparative experiments with various temporal length.

5.2 Appropriate number of labelled training data

Supervised methods such as LSTM and ConvLSTM require a large number of known labels for training, resulting in the cost of increased computational intensity and labelled workload. Actually, it is hard to obtain enough labels, especially for long time series data. In this paper, owing to the idea of semi-supervised learning, our Semi-LSTM method makes it possible to use a small amount of training labels and receive excellent classification results (Table 1). This method greatly reduces the requirements and workload of training labels, and it makes long time series analysis easier to apply in various fields.

To find the optimal number of labelled training samples, we set up comparative experiments which labels account for 100%, 50%, 20% and 10% of the all training samples in three areas. The number of labels in different proportions are rounded down in practice. The results, weighted F1 score are shown in Figure 10 (other metrics have similar trends). As explained in Section 5.1, the length of time series input (t) is 20. With continuous time series input, the first 19 labels are not applied actually due to the many-to-one network. And the number of 50% labels in three areas is 18, 13 and 19 respectively (exactly no more than 19). This is the reason why metrics can hold relatively stable before removing half of labelled training samples.

As the proportion of labels decreases, the W-F1 metrics show similar trends in different data sets. The classification accuracy of LSTM and ConvLSTM with a small number of labels (training labels account for 20% and 10% of total samples) is sharply reduced, while our Semi-LSTM is steadily maintained with a slight decrease. That is to say, our method still performs well in the case of a small number of labelled training samples for classification. In this way, it benefits the collection of training data that the cost and requirement for long time series labelled data are greatly reduced.

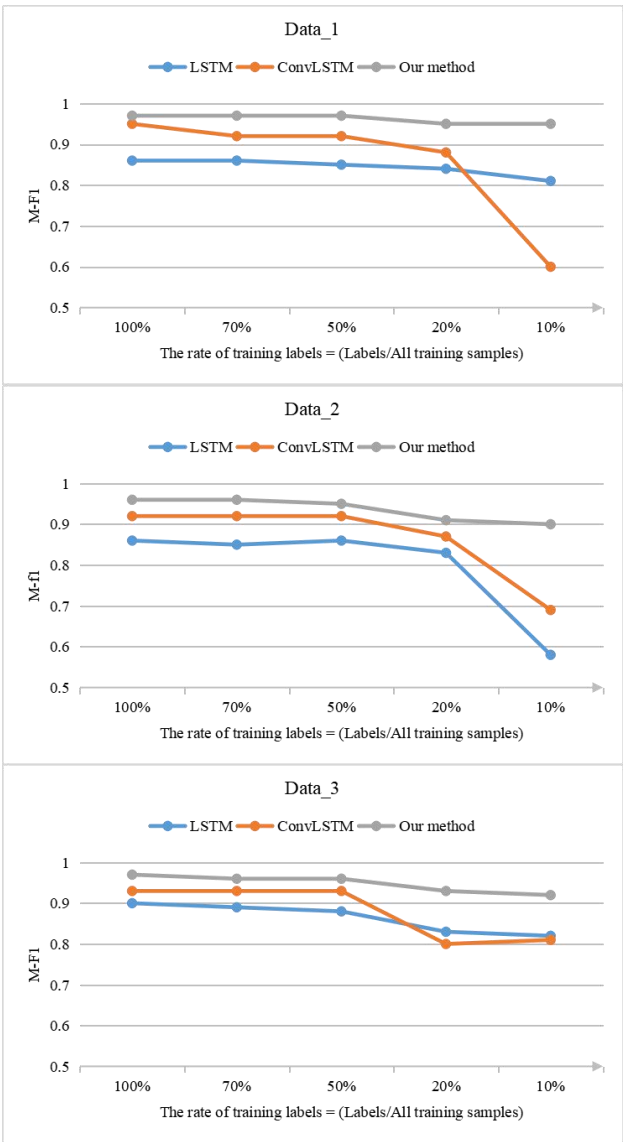

Figure 10. The results of weighted F1 score of comparative experiments in the various proportion of labelled training samples in three study areas.

5.3 Interference of clouds and continuous missing data

More and more optical satellites monitor the dynamic spatio-temporal processes of the Earth's surface in a regular time with a few days' intervals. However, satellites images are inevitably

missing as the surface is usually completely or partially covered by clouds. It has limited the extensive research and application of majority remote sensing approaches, and it poses an omnipresent challenge for the methods that are designed with cloud-free imagery in mind.

In this experiment, we trained the network on three datasets with varying degrees of cloud coverage. The ratio of cloud coverage containing clouds and shadows is calculated in each study area by the number of pixels in the specific area of interest. Besides, we supplemented images of full cloud coverage in the experimental areas from 2015 to 2016. In other words, there are two scenes per month on average, or even three scenes a month. Based on this, several subdata sets have been created, containing all observations, cloud coverage less than 50%, 25%, 10% and 1% (completely cloud-free images). The number of images in subdata sets is displayed in Table 2. It is noted that the training labels in all subdata sets here are labelled data without cloud cover, namely the number of labels is 28, 29 and 30 respectively with only 4 covered classes (cultivated field, forest, construction and water).

Cloud coverage	Data_1	Data_2	Data_3
<1%	28	29	30
<10%	31	34	33
<25%	34	36	36
<50%	38	39	38
All images	44	46	46

Table 2. The number of images with different degrees of cloud coverage in all subdata sets.

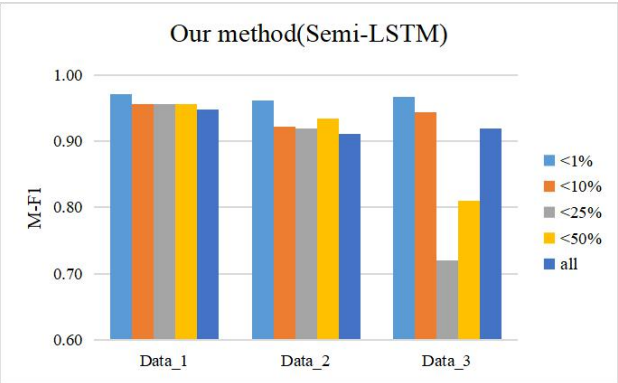

Figure 11. Overall accuracy of subdata sets with different degrees of cloud coverage via our proposed model.

Figure 11. demonstrates that the classification accuracy remains similar between all cloudy and cloud-free images. However, it is special to find that Data_3 with cloud coverage less than 25% and 50% have a poor classification. This is because the lack of a large number of consecutive temporal data. The proportion of missing data in three areas is similar, but only Data_3 lost 5 months of sequential data caused by heavy cloud cover. For cloud-free images, it can be accurately classified due to no interference from clouds and cloud shadows, even without sequential images. So from this comparative experiment, we find that the classification accuracy of time series imagery mainly depends on two factors, namely the lack of continuous time series data and cloud noise. But in practice, the noise caused by cloud and shadow coverage is always inevitable, so that our method is more beneficial. This model not only reduces the requirements for time-series imagery acquisition, but also resists the influence of cloud noise to a certain extent, and still has excellent performance in classification.

6. CONCLUSION

In this study, we developed a novel deep neural network Semi-LSTM to classify land cover by learning spectral-spatial-temporal features from time series imagery. It performs robust in classification tasks with a small amount of training labels. According to overall accuracy and the ability to identify individual class, our optimal model outperformed popular classifier like SVM as well as models based on recurrent neural networks (LSTM and ConvLSTM). We verified that appropriately increasing the length of time series input can improve the land cover classification effectively. Thanks to the advantages of semi-supervised learning, the model can attain excellent results via a small group of training samples, which simplifies a lot of artificial labelling work and reduces the dependence of training samples. In addition, it is difficult to avoid noise caused by clouds and shadows in remote sensing images. After replenishing a large number of cloudy data, this method is robust to processing time series images with clouds and shadows when there is no shortage of large amounts of consecutive data. The model does not require long time series satellite imagery with cloud-free and a lot of labels, which provides a flexible and automated way for land cover mapping applications. The study suggests that these benefits reduce the requirements for collection of dataset and make the classification tasks easier.

REFERENCES

Bagan, H., Wang, Q.X., Watanabe, M., Yang, Y.H., Ma, J.W., 2005: Land Cover Classification from Modis Evi Times-Series Data Using Som Neural Network. *International Journal of Remote Sensing* 26(22), 4999-5012. doi.org/10.1080/01431160500206650.

Belgiu, M., Dragut, L., 2016: Random Forest in Remote Sensing: A Review of Applications and Future Directions. *Isprs Journal of Photogrammetry and Remote Sensing* 11424-31. doi.org/10.1016/j.isprsjprs.2016.01.011.

Bergstra, J., Bengio, Y., 2012: Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13281-305. doi.org/10.1016/j.chemolab.2011.12.002

Boles, S.H., Xiao, X.M., Liu, J.Y., Zhang, Q.Y., Munkhtuya, S., Chen, S.Q., Ojima, D., 2004: Land Cover Characterization of Temperate East Asia Using Multi-Temporal Vegetation Sensor Data. *Remote Sensing of Environment* 90(4), 477-489. doi.org/10.1016/j.rse.2004.01.016.

Brooks, E.B., Thomas, V.A., Wynne, R.H., Coulston, J.W., 2012: Fitting the Multitemporal Curve: A Fourier Series Approach to the Missing Data Problem in Remote Sensing Analysis. *Ieee Transactions on Geoscience and Remote Sensing* 50(9), 3340-3353. doi.org/10.1109/tgrs.2012.2183137.

Carrao, H., Goncalves, P., Caetano, M., 2008: Contribution of Multispectral and Multiternporal Information from Modis Images to Land Cover Classification. *Remote Sensing of Environment* 112(3), 986-997.doi.org/10.1016/j.rse.2007.07.002.

Chen, Y.L., Gong, P., 2013: Clustering Based on Eigenspace Transformation - Cbest for Efficient Classification. *Isprs Journal of Photogrammetry and Remote Sensing* 8364-80. doi.org/10.1016/j.isprsjprs.2013.06.003.

Chren, W., 1995: One-Hot Residue Coding for High-Speed

Non-Uniform Pseudo-Random Test Pattern Generation. *Journal* 1(Issue), 401-404.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F., IEEE, 2009: Imagenet: A Large-Scale Hierarchical Image Database. CVPR: 2009 Ieee Conference on Computer Vision and Pattern Recognition, Vols 1-4, New York. doi.org/10.1109/CVPR.2009.5206848

Gomez, C., White, J.C., Wulder, M.A., 2016: Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. *Isprs Journal of Photogrammetry and Remote Sensing* 11655-72. doi.org/10.1016/j.isprsjprs.2016.03.008

Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J., 2017: Lstm: A Search Space Odyssey. *Ieee Transactions on Neural Networks and Learning Systems* 28(10), 2222-2232. doi.org/10.1109/tnnls.2016.2582924

He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J., IEEE, 2016: Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition, New York. doi.org/10.1109/cvpr.2016.90.

Jamshidpour, N., Homayouni, S., Safari, A., 2016: Graph-Based Semi-Supervised Hyperspectral Image Classi-fication Using Spatial Information. 2016 Ieee Conference on Computer Vision and Pattern Recognition, New York.

Jia, K., Liang, S.L., Wei, X.Q., Yao, Y.J., Su, Y.R., Jiang, B., Wang, X.X., 2014: Land Cover Classification of Landsat Data with Phenological Features Extracted from Time Series Modis Ndvi Data. *Remote Sensing* 6(11), 11518-11532. doi.org/10.3390/rs61111518.

Khatami, R., Mountrakis, G., Stehman, S.V., 2016: A Meta-Analysis of Remote Sensing Research on Supervised Pixel-Based Land-Cover Image Classification Processes: General Guidelines for Practitioners and Future Research. *Remote Sensing of Environment* 17789-100. doi.org/10.1016/j.rse.2016.02.028.

Laine, S.M., Aila, T.O., ICLR, 2017: Temporal Ensembling for Semi-Supervised Learning. 2017 ICLR conference, Paris.

Lin, T.Y., Goyal, P., Girshick, R., He, K.M., Dollar, P., 2020: Focal Loss for Dense Object Detection. Ieee Transactions on Pattern Analysis and Machine Intelligence 42(2), 318-327. doi.org/10.1109/tpami.2018.2858826

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016: Deep Learning Earth Observation Classification Using Imagenet Pretrained Networks. Ieee Geoscience and Remote Sensing Letters 13(1), 105-109. 10.1109/lgrs.2015.2499239

Mou, L.C., Bruzzone, L., Zhu, X.X., 2019: Learning Spectral-Spatial-Temporal Features Via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *Ieee Transactions on Geoscience and Remote Sensing* 57(2), 924-935. doi.org/10.1109/tgrs.2018.2863224.

Mountrakis, G., Im, J., Ogole, C., 2011: Support Vector Machines in Remote Sensing: A Review. *Isprs Journal of Photogrammetry and Remote Sensing* 66(3), 247-259. doi.org/10.1016/j.isprsjprs.2010.11.001.

Russwurm, M., Korner, M., Ieee, 2017: Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-Spectral Satellite Images. 2017 Ieee Conference on Computer Vision and Pattern Recognition Workshops, New York. doi.org/10.1109/cvprw.2017.193.

Scott, G.J., England, M.R., Starms, W.A., Marcum, R.A., Davis, C.H., 2017: Training Deep Convolutional Neural Networks for Land-Cover Classification of High-Resolution Imagery. *Ieee Geoscience and Remote Sensing Letters* 14(4), 549-553. doi.org/10.1109/lgrs.2017.2657778.

Shelhamer, E., Long, J., Darrell, T., 2017: Fully Convolutional Networks for Semantic Segmentation. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 39(4), 640-651. doi.org/10.1109/tpami.2016.2572683.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., Woo, W.-c., 2015: Convolutional Lstm Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems* 802-810. doi.org/10.1007/978-3-319-21233-3_6

Yuan, Y., Meng, Y., Lin, L., Sahli, H., Yue, A.Z., Chen, J.B., Zhao, Z.M., Kong, Y.L., He, D.X., 2015: Continuous Change Detection and Classification Using Hidden Markov Model: A Case Study for Monitoring Urban Encroachment onto Farmland in Beijing. *Remote Sensing* 7(11), 15318-15339. doi.org/10.3390/rs71115318.

Zhang, J.H., Feng, L.L., Yao, F.M., 2014: Improved Maize Cultivated Area Estimation over a Large Scale Combining Modis-Evi Time Series Data and Crop Phenological Information. *Isprs Journal of Photogrammetry and Remote Sensing* 94102-113. doi.org/10.1016/j.isprsjprs.2014.04.023.

Zhong, L.H., Hu, L.N., Zhou, H., 2019: Deep Learning Based Multi-Temporal Crop Classification. *Remote Sensing of Environment* 221430-443. doi.org/10.1016/j.rse.2018.11.032.

Zhou, W.X., Newsam, S., Li, C.M., Shao, Z.F., 2017: Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval. Remote Sensing 9(5), 20. 10.3390/rs9050489

Zhu, Z., Woodcock, C.E., 2012: Object-Based Cloud and Cloud Shadow Detection in Landsat Imagery. *Remote Sensing of Environment* 11883-94. doi.org/10.1016/j.rse.2011.10.028.