# A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification

Ce Zhang [1,*], Xin Pan [2,3], Huapeng Li [2], Andy Gardiner [4], Isabel Sargent [4], Jonathon Hare [5], Peter M. Atkinson [1,*]

*[1] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK; [2] Northeast Institute of Geography and Agroecology, Chinese Academic of Science, Changchun 130102, China; [3] School of Computer Technology and Engineering, Changchun Institute of Technology, 130021 Changchun, China; [4] Ordnance Survey, Adanac Drive, Southampton SO16 0AS, UK; [5] Electronics and Computer Science (ECS), University of Southampton, Southampton SO17 1BJ, UK*

**Abstract** The contextual-based convolutional neural network (CNN) with deep architecture and pixel-based multilayer perceptron (MLP) with shallow structure are well-recognized neural network algorithms, representing the state-of-the-art deep learning method and the classical non-parametric machine learning approach, respectively. The two algorithms, which have very different behaviours, were integrated in a concise and effective way using a rule-based decision fusion approach for the classification of very fine spatial resolution (VFSR) remotely sensed imagery. The decision fusion rules, designed primarily based on the classification confidence of the CNN, reflect the generally complementary patterns of the individual classifiers. In consequence, the proposed ensemble classifier MLP-CNN harvests the complementary results acquired from the CNN based on deep spatial feature representation and from the MLP based on spectral discrimination. Meanwhile, limitations of the CNN due to the adoption of convolutional filters such as the uncertainty in object boundary partition and loss of useful fine spatial resolution detail were compensated. The effectiveness of the ensemble MLP-CNN classifier was tested in both urban and rural areas using aerial photography together with an additional satellite sensor dataset. The MLP-CNN classifier achieved promising performance, consistently outperforming the pixel-based MLP, spectral and textural-based MLP, and the contextual-based CNN in terms of classification accuracy. This research paves the way to effectively address the complicated problem of VFSR image classification.

Keywords: convolutional neural network; multilayer perceptron; VFSR remotely sensed imagery; fusion decision; feature representation

## 1. Introduction

With the rapid development of modern remote sensing technologies, a large quantity of very fine spatial resolution (VFSR) images is now commercially available. These VFSR images, typically acquired at sub-metre spatial resolution, have opened up many opportunities for new applications (Zhong et al., 2014), for example, urban land use retrieval (Mathieu et al., 2007; Shi et al., 2015), precision agriculture (Ozdarici-Ok et al., 2015; Zhang and Kovacs, 2012), and tree crown delineation (Ardila et al., 2011; Yin et al., 2015). However, despite the presence of a rich spatial data content (Huang et al., 2014), the information conveyed by the imagery is conditional upon the quality of the processing (Längkvist et al., 2016). With fewer spectral channels in comparison with coarse or medium spatial resolution remotely sensed data, it can be challenging to differentiate subtle differences amongst similar land cover types (Powers et al., 2015). Meanwhile, objects of the same class may exhibit strong spectral heterogeneity due to differences in age, level of maintenance and composition as well as illumination conditions (Demarchi et al., 2014), which might be further complicated by the scattering of peripheral ground objects (Chen et al., 2014). As a consequence, such high intra-class variability and low inter-class disparity make automatic classification of VFSR images a challenging task.

Ever since the advent of VFSR imagery, tremendous efforts have been made to develop robust and accurate, automatic image classification methods. Among these, machine learning is currently considered as the most promising and evolving approach (Zhang et al., 2015). Popular pixel-based machine learning algorithms, such as Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Random Forest (RF), have drawn considerable attention in the remote sensing community (Attarchi and Gloaguen, 2014; Yang et al., 2012; Zhang et al., 2015). The MLP, as a typical non-parametric neural network classifier, is designed to learn the nonlinear spectral feature space at the pixel level irrespective of its statistical properties (Atkinson and Tatnall, 1997; Foody and Arora, 1997; Mas and Flores, 2008). The MLP has been used widely in remote sensing applications, including VFSR-based land cover classification (e.g. Del Frate et al., (2007), Pacifici et al. (2009)). The MLP algorithm is mathematically complicated yet can be simple in model architecture (e.g., a shallow classifier with one or two feature representation levels). At the same time, a pixel-based MLP classifier does not consider, or make use of, the spatial patterns implicit in images, especially for VFSR imagery with unprecedented spatial detail. In essence, the MLP (and related algorithms, e.g. SVM, RF, etc.) is a pixel-based classifier with shallow structure (one or two layers) (Chen et al., 2016), where the membership association of a pixel for each class is predicted.

Recent advances in neuroscience have shown that deep feature representations can be learned hierarchically from simple concepts such as oriented edges to higher-level complex patterns such as textures, segments, parts and objects (Arel et al., 2010). This discovery motivated the breakthrough of the so-called “deep learning” methods that represent the state-of-the-art in a variety of domains, including target detection (Chen et al., 2016; Tang et al., 2015), image recognition (Farabet et al., 2013; Krizhevsky et al., 2012) and robotics (Bezak et al., 2014; Lenz et al., 2015; Yu et al., 2013), amongst others. The convolutional neural network (CNN), a well-established deep learning approach, has produced excellent results in the field of computer vision and pattern recognition (Schmidhuber, 2015), such as for visual recognition (Farabet et al., 2013; Krizhevsky et al., 2012), image retrieval (X. Yang et al., 2015) and scene annotation (Othman et al., 2016).

In the remote sensing domain, CNNs have been studied actively and shown to produce state-of-the-art results over the past few years, focusing primarily on object detection (Dong et al., 2015) or scene classification (Hu et al., 2015a; Zhang et al., 2016). Recent studies further explored the feasibility of CNNs for the task of remotely sensed image classification. For example, Yue et al., (2016) utilized spatial pyramid pooling to learn multi-scale spatial features from hyperspectral data, Chen et al. (2016) introduced a 3D CNN to jointly extract spectral–spatial features, thus, making full use of the continuous hyperspectral and spatial spaces. In terms of the classification of multi- and hyperspectral imagery, a deep CNN model was formulated through a greedy layer-wise unsupervised pre-training strategy (Romero et al., 2016), whereas an image pyramid was built up through upscaling the original image to capture the contextual information at multiple scales (Zhao and Du, 2016). For VFSR image classification, CNN models with varying contextual input size were constructed to learn multi-scale features while preserving the original fine resolution information (Längkvist et al., 2016). All of the above-mentioned work applied CNNs with contextual patches as their inputs, and demonstrated the robustness and effectiveness in spatial feature representations with excellent classification performance. However, the benefits and shortcomings of the

CNN as a classifier itself have not been studied thoroughly. In particular, the CNN, as a contextual classifier with deep structures (Szegedy et al., 2015), explores the complex spatial patterns hidden in the image that are not seen by representation in its shallow counterparts, whereas it may overlook certain information in spectral space observed by pixel-based classifiers. Moreover, uncertainties may appear in object boundaries due to the usage of convolutional filters of the CNN. These issues deserve further investigation.

Any single set of features (e.g., spectral only) or a specific classifier (e.g., pixel-based only) is unlikely to achieve the highest classification accuracies for VFSR imagery because the result is conditional upon both spectral and spatial information. In this context, two categories of spectral and spatial information were fused for classification or handled with a classifier ensemble. Information fusion can be realized by stacking the spatial and spectral information as feature bands. However, this does not allow the specification of the relative influence of the extracted features (Wang et al., 2016). Others proposed integrative algorithms considering the spatial and spectral features at the same time. For example, Fauvel et al., (2012) proposed a composite kernel-based SVM with spectral and spatial kernels applied simultaneously. However, the spatial kernel summarizes only basic information (e.g. median) of the spatial neighbourhood (Wang et al., 2016).

In terms of classifier ensemble technology, two strategies, namely “multiple classifier systems” (Benediktsson, 2009) and “decision fusion” (Fauvel et al., 2006) are employed. Multiple classifier systems are based on the manipulation of training sample sets, including boosting (Freund et al., 2003) and bagging (Breiman, 1996). This ensemble approach, however, usually requires a relatively large sample size and the computational complexity tends to be high. An alternative classifier ensemble is derived from decision fusion of the outputs of different classification algorithms according to a certain combination of approaches (Du et al., 2012; Löw et al., 2015). This decision fusion-based ensemble approach is preferable where the individual classifiers demonstrate complementary behaviour. For instance, different non-parametric classifiers are sometimes accurate in different locations in a classification map, thus, producing complementary results from the ensemble (Clinton et al., 2015; Löw et al., 2015). However, all the aforementioned fusion strategies are conducted using pixel-based classifiers with shallow structures, whose complementary behaviours are insufficient to address the challenges of VFSR image classification.

In this paper, a hybrid classification system was proposed that combines the CNN (a contextual-based classifier with deep architectures) and MLP (a pixel-based classifier with shallow structures) using a rule-based decision fusion strategy. The hypothesis is that both MLP and CNN classifiers can provide different views or feature representations with strong complementarity. Thus, the classifier ensemble has the potential to enhance the final classification performance. The decision fusion rules were built up at the post-classification stage, primarily based on the confidence distribution of the contextual-based CNN classifier, such that the classified pixels with low confidence can be rectified by the MLP at the pixel level. The effectiveness of the proposed method was tested on images of both an urban scene and a rural area. A benchmark comparison was provided by the standard pixel-based MLP, spectral-texture based MLP as well as contextual-based CNN classifiers.

## 2. Methodology

### *2.1 Brief review of multilayer perceptrons (MLP)*

A multilayer perceptron (MLP) is a network that maps sets of input data onto a set of outputs in a feedforward manner (Atkinson and Tatnall, 1997). The typical structure is that the MLP is composed of interconnected nodes in multiple layers (input, hidden and output layers), with each layer fully connected to the preceding layer as well as the succeeding layer (Del Frate et al., 2007). The outputs of each node are weighted units followed by a nonlinear activation function to distinguish the data that are not linearly separable (Pacifici et al., 2009). Formally, the output activation $a^{(l+1)}$ at layer $l$+1 is derived by the input activation $a^{(l)}$:

$$a^{(l+1)} = \sigma(w^{(l)} a^{(l)} + b^{(l)}) \tag{1}$$

Where $l$ corresponds to a specific layer, $w^{(l)}$ and $b^{(l)}$ denote the weight and bias at layer $l$, and $\sigma$ represents the nonlinear activation operation (e.g. sigmoid, hyperbolic tangent, rectified linear units) function. For an $m$ layer multilayer perceptron, the first input layer is $a^{(1)} = x$ while the last output layer is:

$$h_{w,b}(x) = a^{(m)} \tag{2}$$

The weights $w$ and bias $b$ in equation (2) are learned by supervised training using a backpropagation algorithm to approximate an unknown input-output relation (Del Frate et al., 2007). The objective function is to minimize the difference between the predicted outputs and the desired outputs:

$$J(W,b;x,y) = \frac{1}{2}\left\| h_{w,b}(x) - y \right\|^2 \tag{3}$$

### *2.2 Brief review of Convolutional Neural Networks (CNN)*

A Convolutional Neural Network (CNN), is a variant of the multilayer feed forward neural networks, and is designed specifically to process large scale images or sensory data in the form of multiple arrays by considering local and global stationary properties (LeCun et al., 2015). Similar to the MLP, the CNN is a network stacked into a number of *layers*, where the output of the previous layer is connected sequentially to the input of the next one by a set of learnable weights and biases (Romero et al., 2016). The major difference is that each layer is represented as input and output feature maps by capturing different perspectives on features through the operation of convolution.

The CNN basically consists of three major operations: convolution, nonlinearity and pooling/subsampling (Schmidhuber, 2015). The convolutional and pooling layers are stacked together alternatively in the CNN framework, until obtaining the high-level features on which a fully connected classification is performed (LeCun et al., 2015). In addition, several feature maps may exist in each convolutional layer and the weights of convolutional nodes in the same map are shared. This setting enables the network to learn different features while keeping the number of parameters tractable. Mathematically, the output feature map ${y_{i,j}}^{(l)}$ at convolutional layer *l* is calculated as:

$$ {y_{i,j}}^{(l)} = \sigma^{(l)}\left(\sum_{n=1}^{k}\sum_{m=1}^{k} {w_{n,m}}^{(l)} \cdot {x_{i+n,j+m}}^{(l-1)} + b^{(l)}\right) \tag{4}$$

Where the ${w_{n,m}}^{(l)}$ denotes the convolutional filter with size $k \times k$ at layer *l*, and the ${x_{i+n,j+m}}^{(l-1)}$ represents the spatial position of the corresponding feature map at the preceding layer *l*-1. The algorithm passes the convolutional filter throughout the input feature map using the dot product $(\cdot)$ between them with an addition of a bias unit $b^{(l)}$ (Arel et al., 2010). Moreover, a nonlinear activation function $\sigma^{(l)}$ at layer $l$ is taken outside the dot product to strengthen the nonlinearity (Strigl et al., 2010).

The pooling/subsampling layer can generalize the convolved features through down-sampling and thereby reduce the computational complexity during the training process (Zhao and Du, 2016). Given a pooling/subsampling layer $q$, the feature output $F^q$ can be derived from the preceding layer $f^{(q-1)}$ through

$$F_{i,j}^{q} = \max( f_{1+p(i-1),1+p(j-1)}^{q-1},\ldots, f_{pi,1+p(j-1)}^{q-1},\ldots, f_{1+p(i-1),pj}^{q-1},\ldots f_{pi,pj}^{q-1}) \quad (5)$$

Where $p \times p$ is the size of the local spatial region, and $1 \le i, j \le (m-n+1)/p$, here the $m$ refers to the size of input feature map, while $n$ corresponds to the size of filter (Längkvist et al., 2016). The *max* simply summarizes the input features within local spatial region using the maximum value (Figure 1: Pooling). By doing this, the learnt features become robust and abstract with certain sparseness and translation invariance.

Once the higher level features are extracted, the output feature maps are flattened into a one-dimensional vector, followed by a fully connected output layer (Figure 1: fully connect). This operation is exactly a simple logistic regression, which is equivalent to the standard MLP discussed in section 2.1, but without any hidden layer.

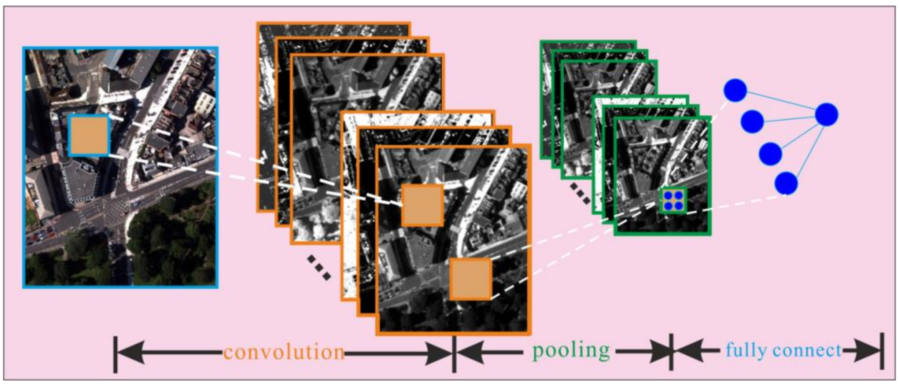


Figure 1 A schematic illustration of the three core layers within the CNN architecture, including the convolutional layer (convolution), pooling layer (pooling) and fully connected layer (fully connect).

### *2.3 Hybrid MLP-CNN Classification Approach*

Suppose the predictive outputs of the MLP and CNN at each pixel are $n$-dimensional vectors $C = (c_1, c_2, ..., c_n)$, where $n$ represents the number of classes and each dimension $i \in [1, n]$ corresponds to the probability of a specific class ($i$-th class) with certain membership association. Ideally, the probability of the classification prediction would be 1 for the target class and 0 for the others. However, due to the uncertainty in the process of remotely sensed image classification, the probability value $c$ is denoted as $f(x) = \{c_x \mid x \in (1,2,...,n)\}$, where $c_x \in [0,1]$ and $\sum_1^n c_x = 1$. The classification model simply takes the maximum membership association as the predicted output label (denoted as class($C$)):

$$class(C) = \arg\max(\{ f(x) = c_x \mid x \in (1,2,\ldots,n)\}) \quad (6)$$

The confidence $conf$ of such membership association is defined here as:

$$conf = Max(C) - Mean(C) \quad (7)$$

In equation (7), *Max*($C$) represents the maximum value of vector $C$, while *Mean*($C$) denotes the average of all the values of $C$. The *conf*, quantified by the difference between *Max*($C$) and *Mean*($C$), measures the confidence or reliability of the class membership allocation (i.e. classification confidence map). Since the CNN takes contextual image patches as its inputs instead of image pixels, it has the following properties:

(1). If the input image patch is located at the central homogeneous region, its class purity is relatively high with large difference between the membership association of the predicted class and those of the other classes, and the *conf* tends to be large (White regions in Figure 2(*c*)).

(2). If the image patch contains other land cover classes as contextual information, the resulting distinction between the membership association of prediction and those of the others is relatively low, and the *conf* tends to be small (Dark regions in Figure 2(*c*)).

However, the MLP (spectral feature only) is based on per-pixel spectral information, thereby ruling out the difference of membership association between central and boundary regions of the classified objects (Figure 1(*b*)). According to the aforementioned properties, the fusion decision rules are constructed primarily based on CNN confidence. To be more specific, the fusion output gives credit to the CNN when its confidence is larger than a predefined threshold ($\alpha_1$), while the MLP is trusted given that the CNN confidence is lower than another threshold ($\alpha_2$); once the confidence of the CNN lies in-between the two thresholds ($\in (\alpha_1, \alpha_2)$), the fusion output chooses the CNN or MLP classification result with a larger confidence. Therefore, for a given image pixel at location $(h, v)$, a rule-based decision fusion approach to determining the class label ($class(h, v)$) of the corresponding pixel is formulated as follows:

$$class(h,v) = \begin{cases} class_{mlp} & conf_{cnn} < \alpha_1 \\ class_{mlp} & (\alpha_1 \le conf_{cnn} < \alpha_2 \;\&\; conf_{cnn} < conf_{mlp}) \\ class_{cnn} & (\alpha_1 \le conf_{cnn} < \alpha_2 \;\&\; conf_{cnn} > conf_{mlp}) \\ class_{cnn} & conf_{cnn} \ge \alpha_2 \end{cases} \quad (8)$$

Where the $class_{mlp}$ and $class_{cnn}$ represent the classification results of the MLP and CNN respectively; the $conf_{mlp}$ and $conf_{cnn}$ denote the classification confidence of the MLP and CNN accordingly.

Estimation of the two thresholds ($\alpha_1$,$\alpha_2$) is conducted using grid search with cross-validation (Min and Lee, 2005; Zhang et al., 2015) based on the CNN classification confidence map (as illustrated by Figure 2(*c*)). Specifically, the $\alpha_1$ was searched from 0.1 to 0.5 to detect those regions with low confidence as predicted by the CNN, while the $\alpha_2$ was chosen from 0.5 to 0.9 to discover the high confidence regions. By initially fixing $\alpha_1$ as 0.1, $\alpha_2$ was tuned with step size of 0.05 (i.e. $\alpha_2$=0.5, 0.55, 0.6, ..., 0.9) to cross-validate the classification accuracy influenced by the selected thresholds; $\alpha_1$ was then increased to further tune $\alpha_2$ in a similar way until the optimal $\alpha_1$ and $\alpha_2$ were found with the best classification accuracy.

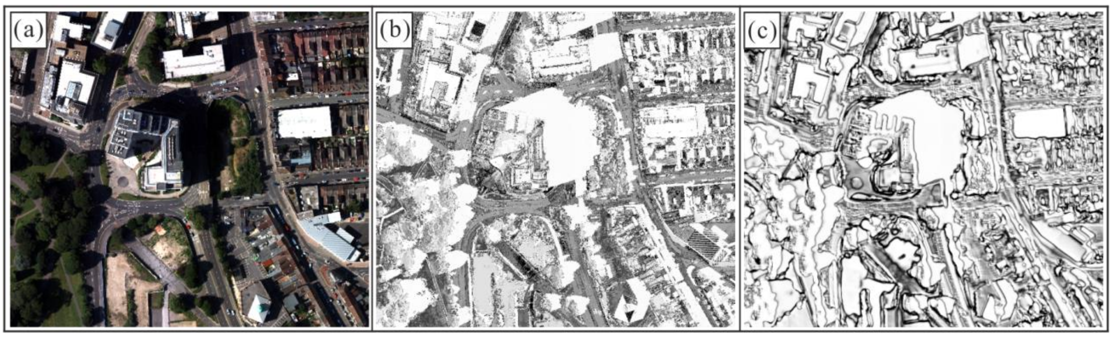

Figure 2 (a) A subset of the original imagery with RGB spectral bands, (b) the classification confidence of the MLP and (c) the classification confidence of the CNN. The dark pixels represent low confidence, while white pixels signify high confidence.

# 3. Experiment

## *3.1 Study area and data source*

For this study, the city of Southampton, UK and its surrounding environment, which lies on the south coast of England, was chosen as a case study area (Figure 3). The urban and suburban areas in Southampton are strongly heterogeneous with a mixture of anthropogenic urban surface (e.g. roof materials, asphalt, concrete) and semi-natural environment (e.g. vegetation, bare soil), thereby representing a good test for classification algorithms.

A scene of aerial imagery of Southampton was captured on 22 July 2012 using a Vexcel UltraCam Xp digital aerial camera with 50 cm spatial resolution and four multispectral bands (Red, Green, Blue and Near Infrared). Two study sites S1 (3087×2750 pixels) and S2 (2022×1672 pixels) were selected to investigate the effectiveness of the proposed algorithm. S1 is located in the city centre of Southampton, which consists of eight dominant land cover classes, including Clay roof, Concrete roof, Metal roof, Asphalt, Grassland, Trees, Bare soil and Shadow, with detailed descriptions listed in Table 1. S2, on the other hand, is situated in a suburban and rural area of Southampton comprised of large patches of forest, grassland and bare soil speckled with small buildings and roads. There are six land cover categories in this study site, namely, Buildings, Road-or-track, Grassland, Trees, Bare soil and Shadow (Table 1).


Figure 3 Southampton, UK Location of study area and aerial imagery with two study sites S1 and S2.

Sample points were collected using a stratified random scheme from ground data provided by local surveyors at Southampton, and split into 50% training samples and 50% testing samples for each class (Table 1). Field land cover survey was conducted throughout the study area on July 2012 to further check the validity and precision of the selected samples. In addition, a highly detailed vector map from Ordnance Survey, namely the MasterMap Topographic Layer (Regnauld and Mackaness, 2006), was fully consulted and cross-referenced to gain a comprehensive appreciation of the land cover and land use within the study area.

Table 1 Detailed description of land covers at two study sites with training and testing sample size per class.

| Study Sites | Class | Train | Test | Description |
|---|---|---|---|---|
| S1 | Clay roof | 144 | 144 | Predominantly residential buildings in red clay tiles |
| | Concrete roof | 132 | 132 | Predominantly residential buildings in grey clay tiles |
| | Metal roof | 134 | 134 | Predominantly industrial buildings in white metal panels |
| | Asphalt | 136 | 136 | Urban road or cark parks covered by asphalt |
| | Grassland | 126 | 126 | Areas of grass covering the urban park or lawn |
| | Trees | 137 | 137 | Patches of tree species |
| | Bare soil | 118 | 118 | Open areas covered by bare soil |
| | Shadow | 123 | 123 | Areas of shadow cast from buildings and trees |
| S2 | Building | 82 | 82 | Predominantly small buildings at rural areas |
| | Road-or-track | 85 | 85 | Asphalt road or small path |
| | Grassland | 86 | 86 | Large areas of wild grass or lawn |
| | Trees | 98 | 98 | Large patches of deciduous trees |
| | Bare soil | 84 | 84 | Open areas covered by bare soil |
| | Shadow | 86 | 86 | Areas of shadow cast from buildings and trees |

To further test the applicability of the proposed method, another scene of Worldview-2 satellite sensor imagery was acquired on 24 July 2013 in the same region of Southampton with urban (S1') and rural (S2') study sites close to the Northwest of S1 and S2. The Worldview-2 image was geometrically and atmospherically corrected, and pan-sharpened at 50 cm spatial resolution to be consistent with the aerial imagery. Figure 4 demonstrates the WorldView-2 satellite sensor image together with two subsets S1' and S2'.

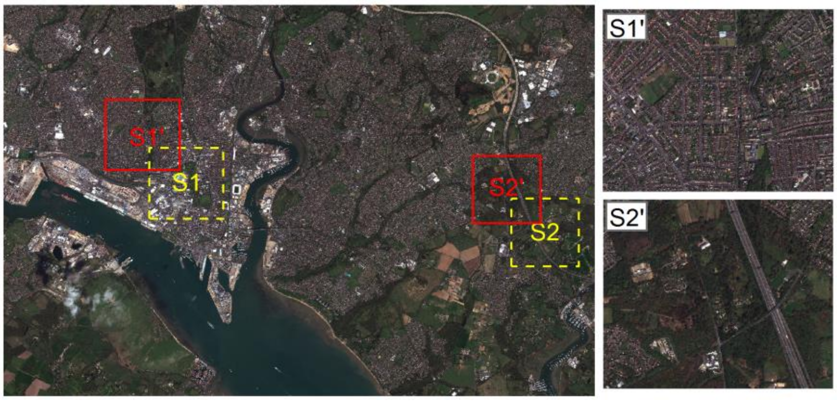


Figure 4 Additional WorldView-2 satellite sensor image covering the same region of Southampton with the S1' and S2' study sites to the northwest of S1 and S2, respectively.

### *3.2 Model input variables and parameters*

Model inputs: the standard pixel-based MLP (hereafter, MLP) and CNN take only the four spectral bands as their input variables, whereas the pixel-based texture MLP based on the standard Grey Level Co-occurrence Matrix (hereafter, GLCM-MLP) simultaneously makes use of both the four spectral bands and the texture features derived from GLCM textural features including the Mean, Variance, Homogeneity, Contrast, Dis-similarity, Entropy, Second moment and Correlation (Haralick et al., 1973; Rodriguez-Galiano et al., 2012; Xia et al., 2010; Zhang et al., 2003). Three window sizes for each spectral band, including $3\times3$ ($1.5\times1.5$ m), $5\times5$ ($2.5\times2.5$ m), and $7\times7$ ($3.5\times3.5$ m), were optimally chosen to perform multi-scale texture feature representation, thus generating 96 GLCM texture features in total. It should be noted that both the MLP and the CNN as well as the GLCM-MLP were trained to predict all pixels within the images. Although the CNN was designed to predict a single label from a small image patch, the sliding window was densely overlapping to cover the entire image at the inference phase.

Both the MLP (also including GLCM-MLP) and CNN models require a series of predefined parameters to optimize the learning accuracy and generalization capability. Following the recommendations of Mas and Flores, (2008), the MLPs with one, two and three hidden layers were tested, using a varying number of {4, 8, 12, 16, 20, and 24} nodes in each layer. The learning rate was chosen optimally as 0.2 and the momentum factor was set as 0.7. In addition, the number of iterations was set as 1000 to fully converge to a stable state. Through cross-validation with different numbers of nodes and hidden layers, the best predicting MLP was found using two hidden layers with 8 nodes in each layer. Similar parameters were also set for the GLCM-MLP, except that two hidden layers with 20 nodes in each layer were found to be the optimal solution in this case.

For the CNN, a range of parameters including the number of layers, the input image patch size, the number and size of convolutional filter, as well as other predefined parameters, such as the learning rate and number of epochs (iterations), need to be tuned (Romero et al., 2016). Following the discussion by Längkvist et al., (2016), the input image size was chosen from {$8\times8$, $10\times10$, $12\times12$, $14\times14$, $16\times16$, $18\times18$, $20\times20$, $22\times22$ and $24\times24$} to evaluate the influence of context area on classification performance. In general, a small-sized contextual area results in overfitting of the model, whereas a large one often leads to under-segmentation. In consideration of the image object size and contextual relationship coupled with a small amount of trial and error, the optimal input image patch size was set to $16\times16$ in this research. Besides, as discussed by Chen et al., (2014) and Längkvist et al., (2016), the depth plays a key role in classification accuracy because the quality of learnt feature is highly influenced by the level of abstraction and representation. As suggested by Chen et al. (2016), the number of CNN layers was chosen as four to balance the network complexity and robustness. Other parameters were set based on standard practice in the field of computer vision. For example, the filter size was set to $5\times5$ for the first convolution layer and $3\times3$ for the rest with stride of 1, and the number of the filters was set to 24 to extract multiple convolutional features at each level. The fully connected layer was tuned as 12 nodes followed by a softmax classification. The learning rate was set to 0.01 and the number of epochs (iterations) was chosen as 600 to fully learn the features through backpropagation. The detailed architecture of the CNN and its parameter configurations is illustrated in Figure 5.

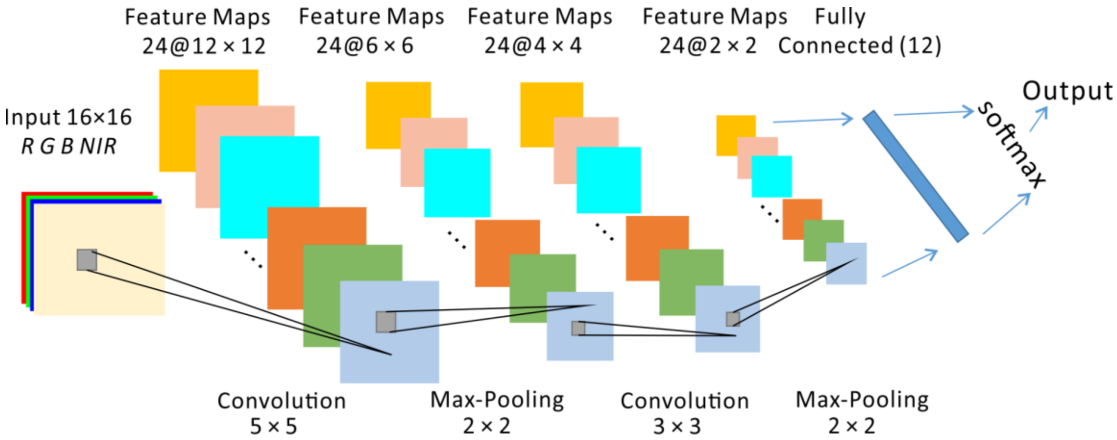


Figure 5. The architecture of the CNN and its configurations.

### *3.3 Decision Fusion Parameter Setting and analysis*

A rule-based decision fusion approach was implemented based on the classification confidence maps of the CNN (e.g. Figure 2(*b*)) and MLP (e.g. Figure 2(*c*)). The parameters of decision fusion, including two thresholds $\alpha_1$ and $\alpha_2$, were determined by grid search with cross-validation using 10% of the randomly chosen samples. In this study, the optimal thresholds $\alpha_1$=0.4 and $\alpha_2$=0.6 were found that reported the greatest classification accuracy.

For the sake of visual interpretation, the confidence distribution of the CNN and MLP influenced by the chosen thresholds is shown in Figure 6. Clearly, the CNN and MLP demonstrated individually consistent, but mutually converse distribution patterns in the two study sites: along with the increase in the CNN's confidence, the MLP inversely exhibited a decreasing trend. Specifically, for low CNN confidence (<0.4), the MLP confidence was around 0.75, significantly higher than that of the CNN, thus outputting the results of MLP in the final decision; once the CNN confidence ranged from 0.4 to 0.6, no significant difference was shown between the two classifiers, thereby, optimally choosing the classification results based on the competitive "winner-takes-all" approach; while for large CNN confidence (>0.6), the MLP was, in contrast, much less reliable (<0.45), thus, taking the classification results of the CNN only in this situation.

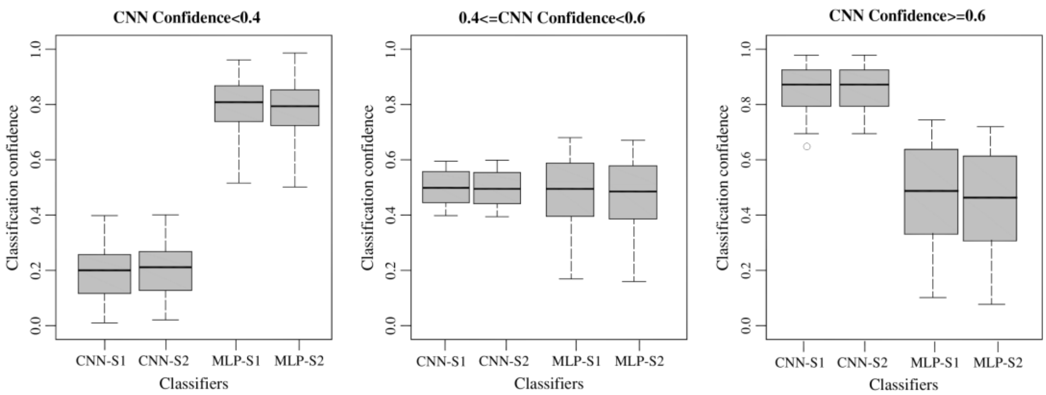


Figure 6 Classification confidence distributions of the CNN and MLP at two study sites (S1 and S2) under different fusion thresholds.

## 3.4 Classification results and analysis

### *3.4.1 Classification results and visual assessment*

By integrating the classification results of the MLP and CNN using the above-mentioned fusion parameters, the final classification of the proposed MLP-CNN was obtained at both study sites, S1 (city centre with complex urban scene) and S2 (rural areas with natural phenomena). To provide a better visualization, Figure 7 (three subsets of S1) and Figure 8 (three subsets of S2) highlights the correct or incorrect classification results of different classifiers marked in yellow or red circles, respectively.

From Figure 7, it can be seen that the MLP classification results consist of undesirable noise (marked in red circle), such as a severe salt-and-pepper effect in Figure 7(*a*) and 7(*b*), and linear noisy textures in Figure 8(*c*). Besides, Trees and Grassland are seriously confused with each other as illustrated by Figure 7(*c*) and Figure 8(*a*) and 8(*b*). However, as shown by Figure 7(*b*), the MLP has certain advantages over CNN in identifying the Clay roof class with spectrally distinctive features (marked in yellow circle). With the addition of the GLCM textures, the GLCM-MLP achieved certain improvements in both spectral and spatial pattern differentiation. For example, Trees and Grassland are better distinguished to some extent compared with the pixel-based MLP results, as illustrated in Figure 7(*c*) and Figure 8(*b*). Besides, the clear linear noisy textures in Figure 8(*c*) are much reduced, and primarily turned into small speckles due to the introduction of texture features. Yet, the GLCM-MLP falsely identifies some edges or boundaries as Clay Roof, as shown in Figure 7(*c*) and Figure 8(*a*) and 8(*b*) (marked in red circle). Additionally, some geometrical distortions of building roof tops, e.g. the Metal Roof and Concrete Roof in Figure 7(*b*), are shown in the GLCM-MLP classification results caused by the GLCM texture filters.

In contrast to the pixel-based MLP and the GLCM-MLP, the classification results of the CNN in both study sites exhibit smoothed visual effects with the least speckle noise as shown by Figure 7 and 8. Additionally, the classes of green vegetation including Grassland and Trees are accurately distinguished as demonstrated by the yellow circles in Figure 7(*c*) and Figure 8(*a*) and 8(*b*) in spite of their spectral similarity. Moreover, the CNN is able to discriminate the Concrete roof from Asphalt with a moderate accuracy, as highlighted by the yellow circle in Figure 7(*a*). Nevertheless, the CNN delivers some uncertainties in partitioning object boundaries. For example, the regular shapes of some buildings (e.g. the geometries of some Clay roof and Concrete roof areas) are distorted with false boundary partitions, as marked by the red circle in Figure 7(*b*). In addition, small or linear features are either merged into a large object or discarded by over-smoothness. For instance, some Clay roof buildings (small objects) are falsely connected together, while Asphalt is sometimes misclassified as Clay roof (Figure 7(*c*)) and the small paths covered by Bare soil are discarded (Figure 8(*b*)).

With respect to the results of the MLP-CNN, all of the aforementioned misclassifications produced by MLP or CNN are resolved with a higher resulting accuracy. Thus, the incorrect classifications (marked by red circles) which appeared in Figure 7 and 8 are revised accordingly, with no red circles appearing in the classification results of MLP-CNN. The MLP-CNN modifies the classification errors of the CNN for Asphalt, as illustrated by the red circles in Figure 7(*c*) and Figure 8(*b*), thanks to the correct classification results of the MLP. Moreover, the linear-shaped Bare Soil area missed by the CNN in Figure 8(*a*) is brought back correctly without losing useful information. In addition, the original shapes of the Clay roof and Concrete roof areas shown in Figure 7(*b*) are accurately restored. Most importantly, some mutual misclassifications between the MLP and CNN are successfully rectified. For example, the MLP-CNN correctly differentiates some Asphalt (with spectrally distinctive but spatially confusing characteristics) and Concrete roof (distinctive in texture and geometry but vague in spectrum) areas that are mutually misclassified by the MLP and CNN respectively (see the regions marked by red circles in Figure 7(*a*)).

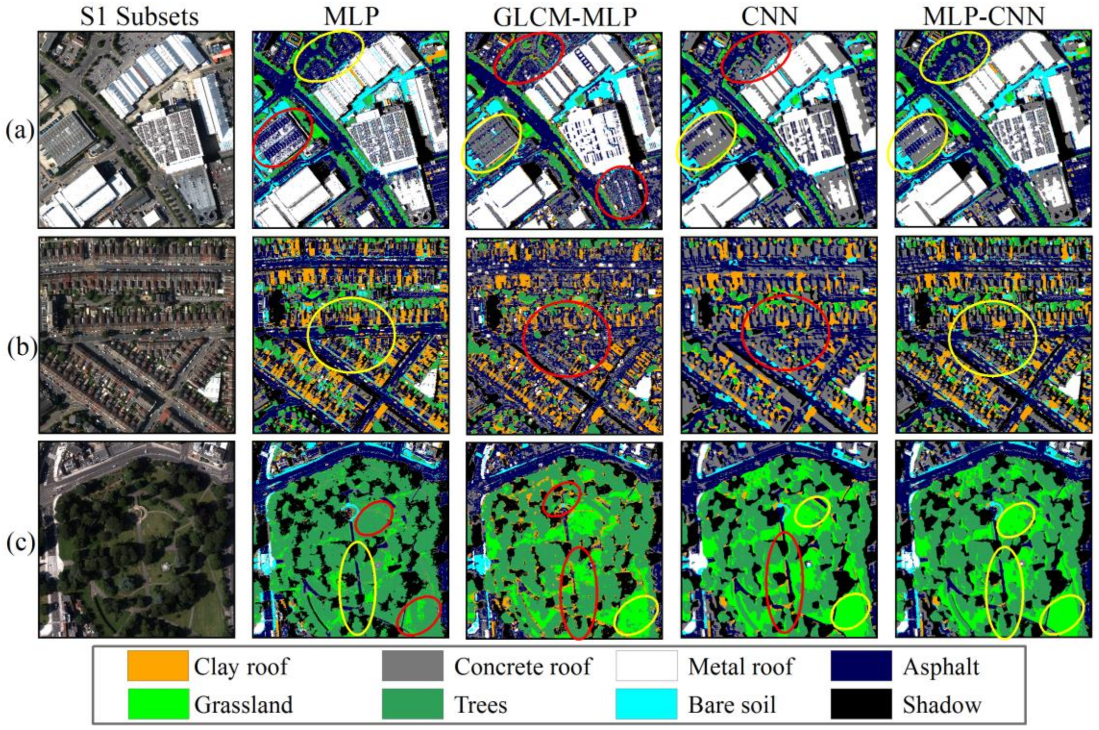


Figure 7 Three typical image subsets (a, b and c) in study site S1 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP classification, the GLCM-MLP classification, the CNN classification and the MLP-CNN classification correspondingly. The red and yellow circles denote incorrect and correct classification, respectively.

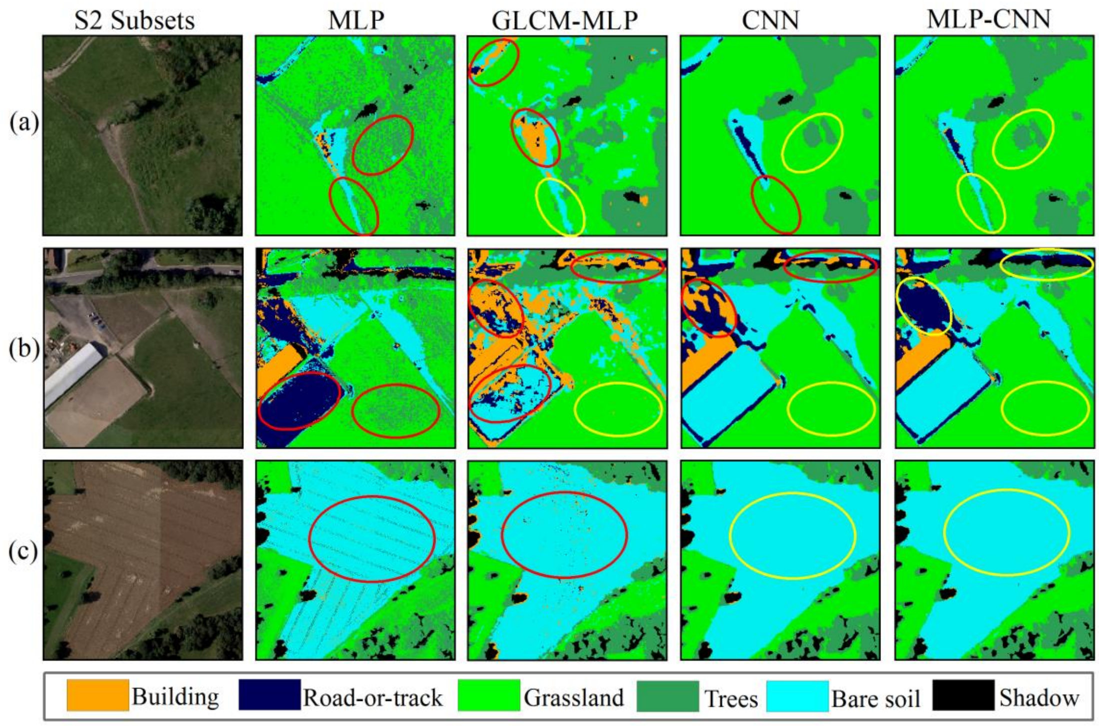


Figure 8 Three typical image subsets (a, b and c) in study site S2 with their classification results. Columns from left to right represent the original images (R G B bands), the MLP classification, the

GLCM-MLP classification, the CNN classification and the MLP-CNN classification correspondingly. The red and yellow circles denote incorrect and correct classification, respectively.

### *3.4.2 Classification accuracy assessment*

The classification performance of the proposed MLP-CNN approach was further investigated through benchmark comparison with the MLP, GLCM-MLP and the CNN. Table 2 lists the classification accuracy assessment, including the overall accuracy (OA), Kappa coefficient (κ), and the class-wise mapping accuracy. From the table, it can be seen that the decision fusion approach (MLP-CNN) consistently reports the best classification OA with up to 90.93% for S1 and 89.64% for S2, higher than that of the CNN (85.39% and 86.56%, respectively) and GLCM-MLP (83.12% and 82.63%, respectively) as well as MLP (81.62% and 80.73%, respectively) (Table 2). Moreover, a Kappa $z$-test for pair-wise comparison also shows that a significant increase in classification accuracy has been achieved by the proposed MLP-CNN classifier over the MLP, GLCM-MLP and CNN in S1, with $z$-value=3.68, 3.12 and 2.25, respectively. For S2, the MLP-CNN also revealed a significant increase over the MLP with $z$-value=3.71 as well as GLCM-MLP with $z$-value=3.18, but no significant difference in comparison with the CNN ($z$ = 1.59, smaller than 1.96 at 95% confidence level) (Congalton, 1991), despite the obvious improvement shown in Table 2.

The increase in classification accuracy was also checked by class-wise accuracy assessment (Table 3). As illustrated by the table, MLP-CNN outperforms CNN for all classes at both study sites in terms of classification accuracy. The largest increase is up to 9.77% for the class of Concrete roof in S1 and 7.16% for the class of Road-or-track in S2. Similar patterns were found such that the MLP-CNN was constantly superior to GLCM-MLP at the class-wise level, where the greatest increase in accuracy was shown up to 11.56% for the class of Concrete Roof in S1 and 11.74% for the class of Grassland in S2. When compared with the MLP, most classes in the two sites except for Asphalt and Shadow in S1 are classified with higher accuracy by the MLP-CNN. Here, Grassland exhibits the highest increase in classification accuracy, up to 33.51% and 18.83% for S1 and S2, respectively. For the classes of Asphalt and Shadow, the accuracy of the MLP is slightly larger than that of the MLP-CNN, but without a statistically significant difference. Thus, they can be regarded as similar to each other.

With respect to the three benchmark classifiers themselves (i.e. MLP, GLCM-MLP and CNN), it can be seen from Table 2 that their classification accuracies are ordered as: MLP <GLCM-MLP < CNN. While the accuracy of CNN is remarkably higher (3%-5%) than that of the MLP and GLCM-MLP, the GLCM-MLP is just slightly higher (<2%) than the MLP. The Kappa *z*-tests (Table 3) further demonstrate that the CNN is statistically significantly more accurate than MLP and GLCM-MLP in both urban and rural areas, whereas a significant increase in accuracy of the GLCM-MLP over the MLP appears only in the rural area rather than the urban area.

Table 2 Classification accuracy comparison amongst MLP, GLCM-MLP, CNN and the proposed MLP-CNN approach for study sites S1 and S2 using the per-class mapping accuracy, overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

| Study Sites | Class | MLP | GLCM-MLP (Benchmark) | CNN | MLP-CNN |
|---|---|---|---|---|---|
| S1 | Clay roof | 92.26% | 91.43% | 90.11% | **95.03%** |
| | Concrete roof | 67.06% | 62.44% | 64.23% | **74.00%** |
| | Metal roof | 91.13% | 90.36% | 94.19% | **94.63%** |
| | Asphalt | **92.72%** | 88.67% | 85.98% | 91.26% |
| | Grassland | 60.51% | 82.58% | 90.73% | **94.02%** |
| | Trees | 63.88% | 78.46% | 82.28% | **88.83%** |
| | Bare soil | 79.63% | 83.05% | 86.16% | **92.49%** |
| | Shadow | **92.33%** | 91.06% | 91.14% | 91.52% |
| | Overall Accuracy (OA) | 81.62% | 83.12% | 85.39% | **90.93%** |
| | Kappa Coefficient (κ) | 0.78 | 0.81 | 0.84 | **0.89** |
| S2 | Building | 82.83% | 80.79% | 83.08% | **88.48%** |
| | Road or track | 83.02% | 80.14% | 82.42% | **89.58%** |
| | Grassland | 71.11% | 78.20% | 88.34% | **89.94%** |
| | Trees | 79.31% | 84.55% | 90.70% | **92.86%** |
| | Bare soil | 74.07% | 76.32% | 81.36% | **86.86%** |
| | Shadow | 89.41% | 88.25% | 88.37% | **90.17%** |
| | Overall Accuracy (OA) | 80.73% | 82.63% | 86.56% | **89.64%** |
| | Kappa Coefficient (κ) | 0.78 | 0.79 | 0.84 | **0.87** |

Table 3 Kappa *z*-test (*p*-value) comparing the performance of the three classifiers for two study sites S1 and S2. Significantly different accuracies with confidence of 95% (z-value > 1.96 with p-value < 0.05) are indicated by *.

| Study sites | Classifiers | Kappa Z-test (p-value) | | | |
|---|---|---|---|---|---|
| | | MLP | GLCM-MLP (Benchmark) | CNN | MLP-CNN |
| S1 | MLP | — | | | |
| | GLCM-MLP | 1.56 (0.1188) | — | | |
| | CNN | 2.64* (0.0083) | 2.44* (0.0147) | — | |
| | MLP-CNN | 3.68* (0.0002) | 3.12* (0.0018) | 2.25* (0.0244) | — |
| S2 | MLP | — | | | |
| | GLCM-MLP | 2.05* (0.0404) | — | | |
| | CNN | 2.51* (0.0121) | 2.36* (0.0183) | — | |
| | MLP-CNN | 3.71* (0.0002) | 3.18* (0.0015) | 1.59 (0.1118) | — |

The proposed MLP-CNN method and the other three benchmarks (MLP, GLCM-MLP and the CNN) were also validated using an additional WorldView-2 satellite sensor dataset at the S1’ and S2’ study sites. The OA and κ of both study sites are in accordance with the results of aerial photo classification, where the decision fusion approach (MLP-CNN) acquires the largest OA of 90.56% at S1’ and 89.77% at S2’, consistently higher than the CNN (86.15% and 86.39%), the GLCM-MLP (83.26% and 82.52%) and the MLP (81.42% and 80.32%) (Table 4). Such coherency of classification results further demonstrates the wide applicability of the proposed method with different datasets.

Table 4 Classification accuracy comparison amongst MLP, GLCM-MLP (Benchmark), CNN and the proposed MLP-CNN approach for study sites S1’ and S2’ from the WorldView-2 satellite sensor image using overall accuracy (OA) and Kappa coefficient (κ). The bold font highlights the greatest classification accuracy per row.

| WorldView-2 | Classification | MLP | GLCM-MLP (Benchmark) | CNN | MLP-CNN |
|---|---|---|---|---|---|
| S1’ | OA | 81.42% | 83.26% | 86.15% | **90.56%** |
| | $\kappa$ | 0.77 | 0.80 | 0.82 | **0.89** |
| S2’ | OA | 80.32% | 82.52% | 86.39% | **89.77%** |
| | $\kappa$ | 0.77 | 0.79 | 0.83 | **0.87** |

## 4. Discussion

In this research, a rule-based decision fusion approach (MLP-CNN) was proposed to integrate classifiers of the pixel-based MLP with shallow structures and the contextual-based CNN with deep architectures for the classification of VFSR remotely sensed imagery. The MLP-CNN takes advantage of the merits of the two classifiers and overcomes their individual shortcomings as discussed below.

### *4.1 Characteristics of MLP and GLCM-MLP classification*

In principle, the MLP builds the decision boundaries among classes in feature space based on per-pixel spectral information (Mokhtarzade and Zoej, 2007). Such classification boundaries are very sensitive to the class with salient spectral properties that are spectrally distinctive from other classes (Berberoglu et al., 2000). For example, classes like Clay roof, Asphalt and Shadow in Site 1 are spectrally exclusive to other classes, leading to high classification accuracies, up to 92.26%, 92.72% and 92.33%, respectively (Table 2). However, the MLP relies on the pixel-based spectral information in the classification process without exploiting the abundant spatial information appearing in the VFSR imagery (e.g. texture, geometry or contextual relationship) (Wang et al., 2016). These limitations often result in unsatisfactory classification performance; for example, confusion and misclassification between the Trees and Grassland classes that are spectrally similar. Even for those correctly identified objects, severe salt and pepper effects still exist (Dark and Bram, 2007), for example, the linear texture noise appearing for Bare soil in Figure 8(*c*). For these reasons, the classification accuracy of MLP is generally statistically significantly lower than that of the CNN and the proposed MLP-CNN. However, objects in VFSR imagery are mostly depicted by pure pixels, especially for human-made features with crisp boundaries, such as buildings, residential houses and cultivated land. The membership association of a pixel deduced by MLP is, therefore, not affected by its relative position (e.g. lying on or close to boundaries), as long as the corresponding spectral space is separable.

The inclusion of GLCM texture features in the GLCM-MLP classifier enables the model to process spectral and spatial information simultaneously. Those GLCM texture descriptors are handcrafted features that are designed to capture statistical co-occurrence information (Xia et al., 2010). However, the GLCM textures are essentially first or second order feature transformations instead of feature learning. Such hand-coded features might be effective for a particular region and/or season, but are often challenging to generalize to other domains and datasets. Besides, the addition of 96 GLCM textures results in a dramatically increased number of input variables, which leads to a relatively high dimensional feature space. The so-called "curse of dimensionality" (Hughes, 1968) and collinearity make the GLCM-MLP hard to parameterize and potentially leads to texture overfitting. That is why the GLCM-MLP cannot substantially increase the classification accuracy compared to the MLP. That is, the spectral and spatial information cannot be effectively exploited by the GLCM-MLP. For example, some spectrally different classes but with similar textures such as Clay Roof, Concrete Roof and Asphalt are confused to some degree.

### *4.2 Characteristics of CNN classification*

Spatial features in remotely sensed data like VFSR imagery are intrinsically local (especially in lower layers) and spatially invariant (Masi et al., 2016). The MLP, however, assumes that the location of the data in the input is irrelevant to the model construction and it is, thus, incapable of learning spatial features of remote sensing data. In contrast, by using multiple convolution and pooling operations, CNN models the way that the human visual cortex works and enforces weight sharing with translation invariance that enables the extraction of high-level spatial features from image patches. It should be mentioned that the pooling operations play an important role in dimension reduction, thus, avoiding "the curse of dimensionality" present in the GLCM-MLP classifier. Thanks to these superior characteristics, the CNN classifier outperforms the MLP and GLCM-MLP classifiers in both the urban scene and rural areas. Especially, classes like Concrete roof and Road-or-track that are difficult to distinguish from their backgrounds with only spectral or low-level features (e.g. distance between the prediction and the target class at spectral space), are identified with relatively high accuracies. In addition, classes with heavy spectral confusion in both study sites (e.g. Trees and Grassland), are accurately differentiated due to their obvious spatial pattern differences; for example, the texture of tree canopies is generally much rougher than for grassland. As a contextual classifier with deep architectures, the CNN could reveal the spatial patterns hidden in the image data that cannot be perceived by its shallow counterparts (e.g. MLP classifier or even the GLCM-MLP classifier). The higher layers in CNN models provide more semantically meaningful information concentrating on global semantics rather than local or pixel-level information, making the CNN classification work well for classes with spectral confusion (Hu et al., 2015a, 2015b;

Yang et al., 2015). Therefore, the CNN shows an impressive stability and effectiveness in spatial feature representation, which is crucial for VFSR image classification (Zhao and Du, 2016).

However, according to the "no free lunch" theorem (Wolpert and Macready, 1997), any elevated performance in one aspect of a problem will be paid for through others, and the CNN is no exception. Using contextual image patches as inputs and learning deep spatial features, the CNN demonstrates power in spatial pattern recognition but also weakness in spatial partition. Boundary uncertainties (over-smoothness) often appear in the classified object and small useful features are erased, somewhat similar to morphological or Gabor filter methods (Pingel et al., 2013; Reis and Tasdemir, 2011). For example, the human-made objects in urban scenes like buildings and asphalt are often geometrically enlarged with distortion to some degree (See Figure 7(*b*)). As for natural objects in rural areas (S2), edges or porosities of a landscape patch are simplified or ignored, and even worse, linear features like river channels or dams that are of ecological importance, are erroneously erased. One may argue that the reduction of image patch size might be able to detect small features by multiple CNNs by varying the contextual filter size as adopted in Längkvist et al. (2016). However, objects, whether large or small in size, all have boundaries, thus, retaining the problem of smoothing edges. In addition, the adoption of convolution and pooling operations intrinsically reduces the image contextual size but strengthens the spatial feature representation. Thus, a far too small initial image patch size can limit the network depth of a CNN model. In fact, the currently used 16×16 window size is close to the minimum requirements for a deep CNN with four hidden layers in total. Moreover, certain spectrally distinctive features without obvious spatial patterns are poorly differentiated. For example, some Asphalt pixels are wrongly identified as Concrete roofs as illustrated in Figure 7(*a*). This further demonstrates the necessity of introducing spectral features for VFSR image classification.

### *4.3 fusion decision of MLP-CNN classification*

Huge uncertainty and inconsistency exists inherently in any remotely sensed data (including VFSR imagery), and this runs through the training and the testing samples. In fact, different classification algorithms vary in terms of remote sensing data processing strategies. Thus there is no 'one-algorithm-fits-all' solution (Löw et al.,

2015) to various applications of VFSR image classification, even for the powerful CNN classifier with deep spatial feature representations. It is therefore especially important to make use of the complementarities of different classifiers. It should be mentioned that, the more heterogeneous the classification algorithms' behaviours, the more that different places might be accurately classified by each individual classifier, and the more accurate the ensemble classifier might be (Löw et al., 2015). An ideal ensemble classifier, thereby, should be established using individual classifiers that are very differently behaved.

The experimental results show that the pixel-based MLP classifier with shallow structures and the contextual-based CNN classifier with deep architectures can provide complementary information, leading to a more accurate classification result than either classifier alone. In addition to the elimination of heavy noise, the CNN can accurately identify classes with rich spatial information implicit in VFSR data. Such characteristics of the CNN emphasize the limitations of the MLP classifier for VFSR image classification. At the same time, the CNN might lose some useful details, and it has difficulties in utilizing spectral information and delineating object boundaries and is, thus, incapable of maintaining geometric fidelity. The MLP classifier, however, compensates directly with regard to the limitations of the CNN. The aforementioned complementary properties between the CNN and MLP are well reflected from the inverse confidence trends of the two classifiers (Figure 2). Specifically, in the case of the CNN with the highest confidence, the MLP has the least confidence and *vice versa*, which further indicates that the proposed MLP-CNN ensemble classifier can take advantage of the MLP and CNN.

The proposed fusion decision rules were derived primarily on the basis of the CNN's confidence distribution, in consideration of the superiority of CNN classification performance and the regularity of its confidence distribution. Such a decision fusion strategy captures the patterns of the complementarities between the two individual classifiers in general, thus, achieving a desirable classification result. At the same time, the MLP-CNN classifier demonstrates great utility and wide applicability for both aerial photography and WorldView-2 satellite sensor imagery with consistent and competitive classification performance. However, in comparison with MLP, the classification accuracies of Asphalt and Shadow were slightly higher than for the proposed MLP-CNN. This means that there is still room for improvement of the decision fusion rules at the class-wise level for VFSR image classification. It might be better to incorporate the spectral separability differentiated by MLP to achieve the best classification performance at class level. Besides, no significant improvement was acquired for rural areas (S2) by the MLP-CNN compared with the CNN. This is mainly due to the ineffectiveness of the MLP in classifying natural features that dominate in the rural environment. This shortcoming might be overcome by the replacement of the MLP by other non-parametric machine learning classifiers (e.g. SVM, RF, etc.). Moreover, incorporating other data sources (e.g. digital surface model) might be needed to increase the accuracy of the MLP-CNN for both the CNN and MLP with very low confidence simultaneously. These aforementioned issues will be investigated in future research.

## 5. Conclusion

Due to its high intra-class variability and low inter-class disparity, VFSR image classification poses great challenges to any single machine learning algorithm, even for the powerful deep learning convolutional neural network (CNN). In this paper, two neural network classifiers with strong heterogeneous behaviours (i.e. pixel-based MLP with shallow structures and contextual-based CNN with deep architectures), were integrated in a concise and effective way using a rule-based decision fusion strategy. The decision fusion rules, designed primarily on the basis of the classification confidence of the CNN, reflect the general complementary patterns of both the MLP and CNN. In consequence, the proposed ensemble classifier MLP-CNN harvests the complementary results acquired from the CNN with deep spatial feature representations (CNN) and from the MLP based on spectral discrimination. Meanwhile, limitations of the CNN such as uncertainty in object boundary partition and loss of useful fine resolution detail were compensated. The effectiveness of the new MLP-CNN algorithm was tested in both urban and rural areas using aerial and satellite sensor images. The MLP-CNN algorithm consistently outperformed both of the individual classifiers (MLP and CNN) as well as the GLCM-MLP that includes the GLCM texture features, with a statistically significant difference in the majority of cases. This research paves the way to an effective solution to the complicated problem of automatic VFSR image classification.

## Acknowledgement

This research was funded by PhD studentship “Deep Learning in massive area, multi-scale resolution remotely sensed imagery” (NO. EAA7369), sponsored by Ordnance Survey and Lancaster University. The authors thank to the staff from the Ordnance Survey for the supply of aerial imagery and supporting ground data. The authors also thank to the two anonymous referees for their constructive comments on this manuscript.

## Reference

Ardila, J.P., Tolpekin, V.A., Bijker, W., Stein, A., 2011. Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images. ISPRS J. Photogramm. Remote Sens. 66, 762–775. doi:10.1016/j.isprsjprs.2011.08.002

Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning - A new frontier in artificial intelligence research. IEEE Comput. Intell. Mag. 5, 13–18. doi:10.1109/MCI.2010.938364

Atkinson, P.M., Tatnall, A.R.L., 1997. Introduction Neural networks in remote sensing. Int. J. Remote Sens. 18, 699–709. doi:10.1080/014311697218700

Attarchi, S., Gloaguen, R., 2014. Classifying complex mountainous forests with L-Band SAR and landsat data integration: A comparison among different machine learning methods in the Hyrcanian forest. Remote Sens. 6, 3624–3647. doi:10.3390/rs6053624

Benediktsson, J.A., 2009. Ensemble classification algorithm for hyperspectral remote sensing data. IEEE Geosci. Remote Sens. Lett. 6, 762–766. doi:10.1109/LGRS.2009.2024624

Berberoglu, S., Lloyd, C.D., Atkinson, P.M., Curran, P.J., 2000. The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. Comput. Geosci. 26, 385–396. doi:10.1016/S0098-3004(99)00119-3

Bezak, P., Bozek, P., Nikitin, Y., 2014. Advanced robotic grasping system using deep learning. Procedia Eng. 96, 10–20. doi:http://dx.doi.org/10.1016/j.proeng.2014.12.092

Breiman, L., 1996. Bagging Predictors. Mach. Learn. 24, 123–140.

Chen, S., Member, S., Wang, H., Xu, F., Member, S., 2016. Target classification using the deep Convolutional Networks for SAR images. IEEE Trans. Geosci. Remote Sens. 54, 4806–4817.

Chen, Y., Jiang, H., Li, C., Jia, X., Member, S., 2016. Deep feature extraction and classification of hyperspectral images based on Convolutional Neural Networks. IEE Trans. Geosci. Remote Sens. 54, 6232–6251. doi:10.1109/TGRS.2016.2584107

Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y., 2014. Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 2094–2107. doi:10.1109/JSTARS.2014.2329330

Clinton, N., Yu, L., Gong, P., 2015. Geographic stacking: Decision fusion to increase global land cover map accuracy. ISPRS J. Photogramm. Remote Sens. 103, 57–65. doi:10.1016/j.isprsjprs.2015.02.010

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46.

Dark, S.J., Bram, D., 2007. The modifiable areal unit problem (MAUP) in physical geography. Prog. Phys. Geogr. 31, 471–479. doi:10.1177/0309133307083294

Del Frate, F., Pacifici, F., Schiavon, G., Solimini, C., 2007. Use of neural networks for automatic classification from high-resolution images. IEEE Trans. Geosci. Remote Sens. 45, 800–809. doi:10.1109/TGRS.2007.892009

Demarchi, L., Canters, F., Cariou, C., Licciardi, G., Chan, J.C.W., 2014. Assessing the performance of two unsupervised dimensionality reduction techniques on hyperspectral APEX data for high resolution urban land-cover mapping. ISPRS J. Photogramm. Remote Sens. 87, 166–179. doi:10.1016/j.isprsjprs.2013.10.012

Dong, Z., Pei, M., He, Y., Liu, T., Dong, Y., Jia, Y., 2015. Vehicle type classification using unsupervised Convolutional Neural Network. IEEE Trans. Intell. Transp. Syst. 16, 2247–2256. doi:10.1109/ICPR.2014.39

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., Liu, S., 2012. Multiple classifier system for remote sensing image classification: A review. Sensors 12, 4764–4792. doi:10.3390/s120404764

Farabet, C., Couprie, C., Najman, L., Lecun, Y., 2013. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929. doi:10.1109/TPAMI.2012.231

Fauvel, M., Chanussot, J., Benediktsson, J.A., 2012. A spatial-spectral kernel-based approach for the classification of remote-sensing images. Pattern Recognit. 45, 381–392. doi:10.1016/j.patcog.2011.03.035

Fauvel, M., Chanussot, J., Benediktsson, J.A., 2006. Decision fusion for the classification of urban remote sensing images. IEEE Trans. Geosci. Remote Sens. 44, 2828–2838. doi:10.1109/TGRS.2006.876708

Foody, G.M., Arora, M.K., 1997. An evaluation of some factors affecting the accuracy of classification by an artificial neural network. Int. J. Remote Sens. 18, 799–810. doi:10.1080/014311697218764

Freund, Y., Iyer, R., Schapire, R.E., Singer, Y., 2003. An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. 4, 933–969.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural Features for Image Classification. IEEE Trans. Syst. Man. Cybern. 3, 610–621. doi:10.1109/TSMC.1973.4309314

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015a. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. Remote Sens. 7, 14680–14707. doi:10.3390/rs71114680

Hu, F., Xia, G.S., Wang, Z., Huang, X., Zhang, L., Sun, H., 2015b. Unsupervised Feature Learning Via Spectral Clustering of Multidimensional Patches for Remotely Sensed Scene Classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 2015–2030. doi:10.1109/JSTARS.2015.2444405

Huang, X., Lu, Q., Zhang, L., 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. ISPRS J. Photogramm. Remote Sens. 90, 36–48. doi:10.1016/j.isprsjprs.2014.01.008

Hughes, G.F., 1968. On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Trans. Inf. Theory 14, 55–63. doi:10.1109/TIT.1968.1054102

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep Convolutional Neural Networks, in: NIPS2012: Neural Information Processing Systems. Lake Tahoe, Nevada, pp. 1–9.

Längkvist, M., Kiselev, A., Alirezaie, M., Loutfi, A., 2016. Classification and segmentation of satellite orthoimagery using Convolutional Neural Networks. Remote Sens. 8, 1–21. doi:10.3390/rs8040329

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. doi:10.1038/nature14539

Lenz, I., Lee, H., Saxena, A., 2015. Deep learning for detecting robotic grasps. Int. J. Rob. Res. 34, 705–724. doi:10.1177/0278364914549607

Löw, F., Conrad, C., Michel, U., 2015. Decision fusion and non-parametric classifiers for land use mapping using multi-temporal RapidEye data. ISPRS J. Photogramm. Remote Sens. 108, 191–204. doi:http://dx.doi.org/10.1016/j.isprsjprs.2015.07.001

Mas, J.F., Flores, J.J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. Int. J. Remote Sens. 29, 617–663. doi:10.1080/01431160701352154

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., Wang, L., Zhou, G., Thenkabail, P.S., 2016. Pansharpening by Convolutional Neural Networks. Remote Sens. 8, 1–22. doi:10.3390/rs8070594

Mathieu, R., Freeman, C., Aryal, J., 2007. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. Landsc. Urban Plan. 81, 179–192. doi:10.1016/j.landurbplan.2006.11.009

Min, J.H., Lee, Y.-C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst. Appl. 28, 603–614.

Mokhtarzade, M., Zoej, M.J.V., 2007. Road detection from high-resolution satellite images using artificial neural networks. Int. J. Appl. Earth Obs. Geoinf. 9, 32–40.

doi:10.1016/j.jag.2006.05.001

Othman, E., Bazi, Y., Alajlan, N., Alhichri, H., Melgani, F., 2016. Using convolutional features and a sparse autoencoder for land-use scene classification. Int. J. Remote Sens. 37, 2149–2167. doi:10.1080/01431161.2016.1171928

Ozdarici-Ok, A., Ok, A., Schindler, K., 2015. Mapping of agricultural crops from single high-resolution multispectral images—Data-driven smoothing vs. Parcel-based smoothing. Remote Sens. 7, 5611–5638. doi:10.3390/rs70505611

Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113, 1276–1292. doi:10.1016/j.rse.2009.02.014

Pingel, T.J., Clarke, K.C., McBride, W.A., 2013. An improved simple morphological filter for the terrain classification of airborne LIDAR data. ISPRS J. Photogramm. Remote Sens. 77, 21–30. doi:10.1016/j.isprsjprs.2012.12.002

Powers, R.P., Hermosilla, T., Coops, N.C., Chen, G., 2015. Remote sensing and object-based techniques for mapping fine-scale industrial disturbances. Int. J. Appl. Earth Obs. Geoinf. 34, 51–57. doi:10.1016/j.jag.2014.06.015

Regnauld, N., Mackaness, W. a., 2006. Creating a hydrographic network from its cartographic representation: a case study using Ordnance Survey MasterMap data. Int. J. Geogr. Inf. Sci. 20, 611–631. doi:10.1080/13658810600607402

Reis, S., Tasdemir, K., 2011. Identification of hazelnut fields using spectral and gabor textural features. ISPRS J. Photogramm. Remote Sens. 66, 652–661. doi:10.1016/j.isprsjprs.2011.04.006

Rodriguez-Galiano, V.F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P.M., Jeganathan, C., 2012. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens. Environ. 121, 93–107. doi:10.1016/j.rse.2011.12.003

Romero, A., Gatta, C., Camps-valls, G., Member, S., 2016. Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote

Sens. 54, 1349–1362. doi:10.1109/TGRS.2015.2478379.

Schmidhuber, J., 2015. Deep Learning in neural networks: An overview. Neural Networks. doi:10.1016/j.neunet.2014.09.003

Shi, H., Chen, L., Bi, F., Chen, H., Yu, Y., 2015. Accurate urban area detection in remote sensing images. IEEE Geosci. Remote Sens. Lett. 12, 1948–1952.

Strigl, D., Kofler, K., Podlipnig, S., 2010. Performance and scalability of GPU-based Convolutional Neural Networks, in: 2010 18th Euromicro Conference on Parallel, Distributed and Network-Based Processing. pp. 317–324. doi:10.1109/PDP.2010.43

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1–9. doi:10.1109/CVPR.2015.7298594

Tang, J., Deng, C., Huang, G.-B., Zhao, B., 2015. Compressed-domain ship detection on spaceborne optical image using Deep Neural Network and Extreme Learning Machine. IEEE Trans. Geosci. Remote Sens. 53, 1174–1185. doi:10.1109/TGRS.2014.2335751

Wang, L., Shi, C., Diao, C., Ji, W., Yin, D., 2016. A survey of methods incorporating spatial information in image classification and spectral unmixing. Int. J. Remote Sens. 37, 3870–3910. doi:10.1080/01431161.2016.1204032

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82. doi:10.1109/4235.585893

Xia, G.S., Delon, J., Gousseau, Y., 2010. Shape-based invariant texture indexing. Int. J. Comput. Vis. 88, 382–403. doi:10.1007/s11263-009-0312-3

Yang, W., Dai, D., Triggs, B., Xia, G.S., 2012. SAR-based terrain classification using weakly supervised hierarchical Markov aspect models. IEEE Trans. Image Process. 21, 4232–4243. doi:10.1109/TIP.2012.2199127

Yang, W., Yin, X., Xia, G.S., 2015. Learning high-level features for satellite image classification with limited labeled samples. IEEE Trans. Geosci. Remote Sens.

53, 4472–4482. doi:10.1109/TGRS.2015.2400449

Yang, X., Qian, X., Mei, T., 2015. Learning salient visual word for scalable mobile image retrieval. Pattern Recognit. 48, 3093–3101. doi:10.1016/j.patcog.2014.12.017

Yin, W., Yang, J., Yamamoto, H., Li, C., 2015. Object-based larch tree-crown delineation using high-resolution satellite imagery. Int. J. Remote Sens. 36, 822–844. doi:10.1080/01431161.2014.999165

Yu, J., Weng, K., Liang, G., Xie, G., 2013. A vision-based robotic grasping system using deep learning for 3D object recognition and pose estimation, in: 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 1175–1180. doi:10.1109/ROBIO.2013.6739623

Yue, J., Mao, S., Li, M., 2016. A deep learning framework for hyperspectral image classification using spatial pyramid pooling. Remote Sens. Lett. 7, 875–884. doi:10.1080/2150704X.2016.1193793

Zhang, C., Kovacs, J.M., 2012. The application of small unmanned aerial systems for precision agriculture: A review. Precis. Agric. 13, 693–712. doi:10.1007/s11119-012-9274-5

Zhang, C., Wang, T., Atkinson, P.M., Pan, X., Li, H., 2015. A novel multi-parameter support vector machine for image classification. Int. J. Remote Sens. 36, 1890–1906. doi:10.1080/01431161.2015.1029096

Zhang, F., Du, B., Zhang, L., 2016. Scene classification via a gradient boosting random convolutional network framework. IEEE Trans. Geosci. Remote Sens. 54, 1793–1802. doi:10.1109/TGRS.2015.2488681

Zhang, Q., Wang, J., Gong, P., Shi, P., 2003. Study of urban spatial patterns from SPOT panchromatic imagery using textural analysis. Int. J. Remote Sens. 24, 4137–4160. doi:10.1080/0143116031000070445

Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 113, 155–165. doi:10.1016/j.isprsjprs.2016.01.004

Zhong, Y., Zhao, J., Zhang, L., 2014. A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 52, 7023–7037. doi:10.1109/TGRS.2014.2306692