Inverse Coefficient of Variation Feature and Multilevel Fusion Technique for Hyperspectral and LiDAR Data Classification

Farah Jahan, Jun Zhou, *Senior Member, IEEE*, Mohammad Awrangjeb, *Senior Member, IEEE*, and Yongsheng Gao, *Senior Member, IEEE*

***Abstract*—Multisource remote sensing data contain complementary information on land covers, but fusing them is a challenging problem due to the heterogeneous nature of the data. This article aims to extract and integrate information from hyperspectral image (HSI) and light detection and ranging (LiDAR) data for land cover classification. As there is a scarcity of a large number of training samples for remotely sensed hyperspectral and LiDAR data, in this article, we propose a model, which is able to perform impressively using a limited number of training samples by extracting effective features representing different characteristics of objects of interest from these two complementary data sources (HSI and LiDAR). A novel feature extraction method named inverse coefficient of variation (ICV) is introduced for HSI, which considers the Gaussian probability of neighborhood between every pair of bands. We, then, propose a two-stream feature fusion approach to integrate the ICV feature with several features extracted from HSI and LiDAR data. We incorporate a fusion unit named canonical correlation analysis as a basic unit for fusing two different sets of features within each stream. We also incorporate the concept of ensemble classification where the features produced by two-stream fusion are distributed into subsets and transformed to improve the feature quality. We compare our method with the existing state-of-the-art methods, which are based on deep learning or handcrafted feature extraction or using both of them. Experimental results show that our proposed approach performs better than other existing methods with a limited number of training samples.**

***Index Terms*—Canonical correlation analysis (CCA), fusion, hyperspectral, light detection and ranging (LiDAR), multisource.**

I. Introduction

EXTRACTION and integration of useful information from multiple data sources in an effective way remains an open challenge for automatic remote sensing data interpretation. Fusion of hyperspectral image (HSI) and light detection and ranging (LiDAR) data provokes substantial research attention and motivates researchers to propose many effective techniques [1]–[4].

Most recent methods used raw spectral features from HSI, digital surface model (DSM) and intensity from LiDAR data, and spatial features from all of them for classification of complex land cover classes [1], [2], [4], [5]. Morphological attribute profiles (APs) from HSI and DSM already proved its effectiveness as spatial features [4]–[6]. Since an HSI contains a large number of bands, it is not feasible to extract AP from each band. An effective way for deriving morphological APs is to apply it on the first few principal components (PCs) of HSI [4], [6]. New spatial features such as extinction profile [3], [7], [8] and pseudo-waveforms [9] have also been used for land cover classification. When the efficiency of feature extraction is concerned, Bao *et al.* [10] extracted derivatives of spectral reflectance signatures from HSI, which captured sharp changes between neighboring bands with little cost of computation.

The recent success of deep learning has motivated the development of spectral-spatial features from HSI and spatial features from LiDAR data using convolutional neural networks [3], [11]–[13]. Compared with traditional classification methods, deep-learning-based classifiers have great potential to obtain high classification performance for mixed and complex inputs. Li *et al.* [12] used a three-stream CNN to extract informative spectral, spatial, and elevation features individually from both hyperspectral and LiDAR sources. To fuse heterogeneous features extracted by CNN, instead of a simple stacking strategy, a multisensor composite kernels is designed. Chen *et al.* [13] proposed a new feature fusion framework based on deep neural networks. In general, deep-learning models have many parameters to be learned in the training procedure, which means a large number of training samples are needed. There is a scarcity of a large number of training samples for remotely sensed hyperspectral and LiDAR data, so many works still focus on the innovation of feature extraction, fusion, and classification approach instead of using deep models.

Feature-level fusion is the most widely used method of multisource data fusion for land cover classification. Apart from layer stacking and principal component analysis (PCA)-based fusion [1], [2], Liao *et al.* [5] proposed graph-based fusion for dimensionality reduction and feature fusion. The edges of the generated fusion graph were weighted by the distance between the stacked feature points. In addition, Gu and Wang [14] proposed a discriminative graph-based method, which fused heterogeneous features from HSI and LiDAR data for urban area classification. The main idea of the method was to search for a projection matrix for preserving the local geometry of each

Manuscript received July 8, 2019; revised October 22, 2019 and December 2, 2019; accepted December 19, 2019. Date of publication January 16, 2020; date of current version February 12, 2020. *(Corresponding author: Jun Zhou.)*

The authors are with the Institute of Integrated and Intelligent Systems, Griffith University, Nathan, QLD 4111, Australia (e-mail: farah.jahan@griffithuni.edu.au; jun.zhou@griffith.edu.au; m.awrangjeb@griffith.edu.au; yongsheng.gao@griffith.edu.au).

Digital Object Identifier 10.1109/JSTARS.2019.2962659

class by maximizing between-class distance and minimizing within-class distance. Rasti *et al.* [8] introduced a sparse and low-rank technique to estimate the low-rank fused features from the derived spectral-spatial features of HSI and spatial features of LiDAR data. Kernel-based methods were also used for feature-level fusion, which are useful for separating different classes in a high-dimensional space. Zhang *et al.* [15] developed a novel framework for fusing spectral, spatial, and height information derived from HSI and LiDAR data by using a composite kernel method. Li *et al.* [16] constructed a new family of generalized composite kernels, which exhibited great flexibility when combining the spectral and the spatial information contained in the HSI data, without any weight parameters. Jahan *et al.* [17] proposed band grouping for both spectral and spatial feature extraction from HSI and fused the extracted features using discriminant correlation analysis (DCA). In addition, DCA was employed for the fusion of spatial features from both HSI and LiDAR.

Decision-level fusion is another fusion strategy. Xia *et al.* [18] introduced a novel method that iteratively runs feature subset creation, transformation, and classification in order to design an ensemble classifier. Finally, majority voting was applied on the classification results obtained from all iterations to produce final result. Bigdeli *et al.* [19] presented a fuzzy multiple classifier system to fuse HSI and LiDAR data based on decision template. Khodadadzadeh *et al.* [4] integrated multiple types of features from HSI and LiDAR using a multiple feature learning approach based on the subspace multinomial logistic regression classifier. Zhong *et al.* [20] proposed an optimal decision fusion method based on adaptive differential evolution where weighted voting was used to obtain the final classification result.

In general, looking for effective features from both data sources should be the first step for achieving high classification accuracy. In other words, the derived features must be discriminative so that samples from different classes can be identified accurately. Furthermore, effective combination of features from different sources is important for generating more discriminative feature than individual sources. Feature fusion is a process that reduces the dimensionality as well as increases the discriminative power of features without losing important information. Simple concatenation of inverse coefficient of variation (ICV) features with other hyperspectral and LiDAR features increases dimensionality in addition to redundancy. On the other hand, using canonical correlation analysis (CCA), we use spectral and spatial features with other hyperspectral and LiDAR features in one stream, and fuse ICV features, spatial features built on top of ICV features with other hyperspectral and LiDAR features in another stream. CCA fusion reduces dimensionality as well as increases between-class correlation and decreases within-class correlation. Besides feature fusion, we use decision fusion by incorporating random subspace, data transformation, and decision weighting to strengthen the system's discriminative power by minimizing misclassification. Experimental results show that the proposed fusion model performs well with a limited number of training samples.

In this article, we propose a novel feature extraction and fusion method for both HSI and LiDAR data. We first introduce an ICV feature extraction method that measures the similarity between a band and the rest of the bands in an HSI and, then, converts the detailed similarity measures into a set of three-dimensional (3-D) features. The ICV feature is integrated with several generic features (GF) from both HSI and LiDAR data using a two-stream feature fusion model. In one stream, CCA is proposed to fuse spectral with spatial features from HSI and every GF separately. In the other stream, ICV replaces HSI spectral feature while keeping other components the same as the first stream. The fused features are classified using an ensemble classification method, which divides the features into random subsets, reduces dimensions using a transformation process, and then concatenates them to form a feature vector as input to an RF classifier. In the classification stage, a random feature selection goes through an iterative process which helps the RF to return the weight of each class given a testing sample, so the class label can be determined by summing the weights of all iterations. This is a different strategy from the majority voting method by Xia *et al.* [18]. A summary of our method is given in Fig. 1.

The main contributions of this article are threefold given as follows.

1) We propose a novel feature extraction method by calculating the similarity between a band and the rest in terms of conditional probability. We calculate the ICV for converting the detailed similarity measure into a set of 3-D features. These ICV features provide complementary information for the HSI spectral features.
2) We propose a two-stream feature fusion model where CCA works as a basic fusion unit. This two-stream fusion generates discriminative features by fusing HSI, ICV, and other features from HSI and LiDAR data by maximizing within-class correlation and minimizing between-class correlation.
3) An ensemble classifier system is incorporated by randomly grouping, transforming, concatenating all transformed groups and, then, classifying the resultant features. Cumulative weight calculation technique is applied to iterative results to improve classification accuracy.

The rest of the article is organized as follows. Section II provides a detailed description on the proposed method. The experimental results on two real HSI and LiDAR datasets are presented and discussed in Section III. Finally, Section IV concludes this article with a few observations.

II. Methodology

In this section, we describe our proposed approach of multilevel fusion of heterogeneous features from HSI and LiDAR data for land cover classification. As shown in Fig. 1, our method has several key steps including ICV cube generation, generic and spatial feature extraction, two-stream CCA fusion, and ensemble classification.

A. Feature Extraction

Our feature extraction incorporates ICV cube generation, GF extraction, spatial feature extraction from ICV cube, HSI cube, and every GF. In the following section, we discuss each step in detail.

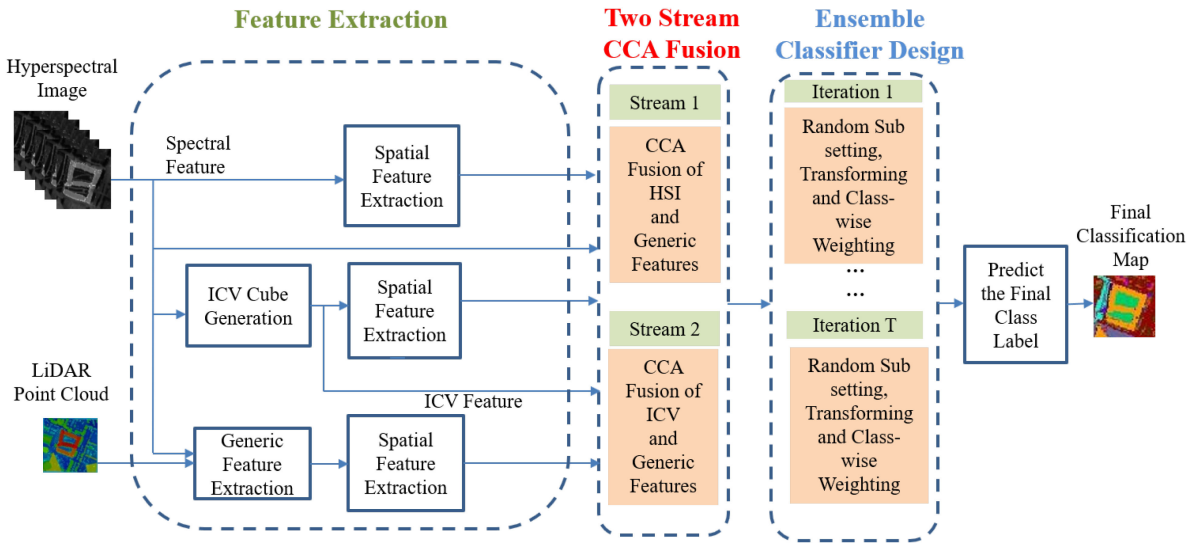

Fig. 1. Proposed architecture for feature and decision fusion.

1) ICV Cube Generation: Band to band relationship of HSI data is important for spectral feature extraction [10]. We develop a novel feature extraction method based on relationship of a band with other bands. For measuring the relationship between two bands, we extend the stochastic neighborhood embedding (SNE) [21] concept. SNE represents similarity between two points in terms of conditional probability.

Given an HSI cube $\mathcal{H} \in \mathbb{R}^{M\times N\times R}$, where M and N are height and width in spatial domain, respectively, and R is the number of bands. A pixel at location (x, y) can be represented as a spectral feature vector $< q^1_{x,y}, q^2_{x,y}, \ldots, q^R_{x,y} >$, where $q^i_{x,y}$ represents the intensity at location (x, y) in the ith band. The conditional probability $p_{i|j}$ for bands i and j at pixel location (x, y) represents the similarity between the bands and is calculated as

$$p_{i|j} = \frac{\exp\left(\frac{-||q^i_{x,y}-q^j_{x,y}||^2}{2\phi_i^2}\right)}{\sum_{k\neq i}\exp\left(\frac{-||q^i_{x,y}-q^k_{x,y}||^2}{2\phi_i^2}\right)} \tag{1}$$

where ϕ^i is the bandwidth of the Gaussian kernel and is centered at $q^i_{x,y}$. We derive band-pair relationship for all bands using (1). Since the HSI has R bands, a similarity matrix $S \in \mathbb{R}^{R\times R}$ is generated for each pixel, where $S^{i,j} = p_{i|j}$. In the matrix S, the values in the ith row contains the similarity between ith band and all bands. Since S has R^2 elements, which exponentially increases the dimensionality of features, we reduce the dimensions of the similarity matrix by employing the ICV concept, which is widely used in image processing [22]. ICV calculates the ratio of mean to standard deviation [23] as follows:

$$\mu^i = \frac{\sum_{j=1}^{R} S^{i,j}}{R} \tag{2}$$

$$\sigma^i = \sqrt{\frac{\sum_{j=1}^{R}\left(S^{i,j}-\mu^i\right)^2}{R-1}} \tag{3}$$

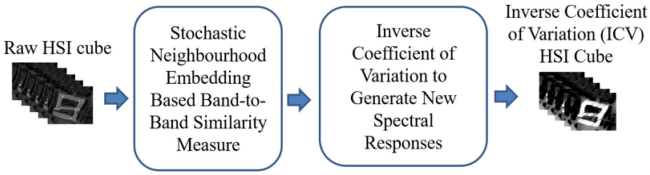

Fig. 2. ICV cube generation.

and

$$v^i = \frac{\mu^i}{\sigma^i} \tag{4}$$

where μ_i, σ_i, v_i represent the mean, the standard deviation, and the ICV for band i, respectively. The ratio in (4) estimates the strength of a band versus the noise in the rest of the bands. For a pixel in the HSI, ICV returns a scalar value for each band and, thus, produces a vector of length R by converting the similarity matrix S. The same operation is performed on each pixel in the HSI, leading to an ICV cube having the same size as the original HSI. Fig. 2 shows the steps of ICV cube generation.

To highlight the linear dependency between two features of the same type, e.g., raw hyperspectral features or ICV values from two planes, we calculate the Pearson correlation coefficient [24] between them. This is performed for every individual feature with respect to the rest of the feature set. Figs. 3 and 4 show such dependencies in Houston data [25] and GU data, respectively (details on these two datasets are presented in Section III). It is clearly seen that the spectral responses of HSI are more correlated than ICV. In other words, ICV reduces the dependency among features, thus improves discriminative capability with respect to the original spectral features. It is also clear that more structural information can be observed in the ICV plane than the corresponding HSI plane.

2) Generic Feature Extraction: We derive a total of eight general features from the HSI and LiDAR data [26]. Among them, normalized difference vegetation index (NDVI) [1], [27] and Entropy are extracted from HS. NDVI evaluates

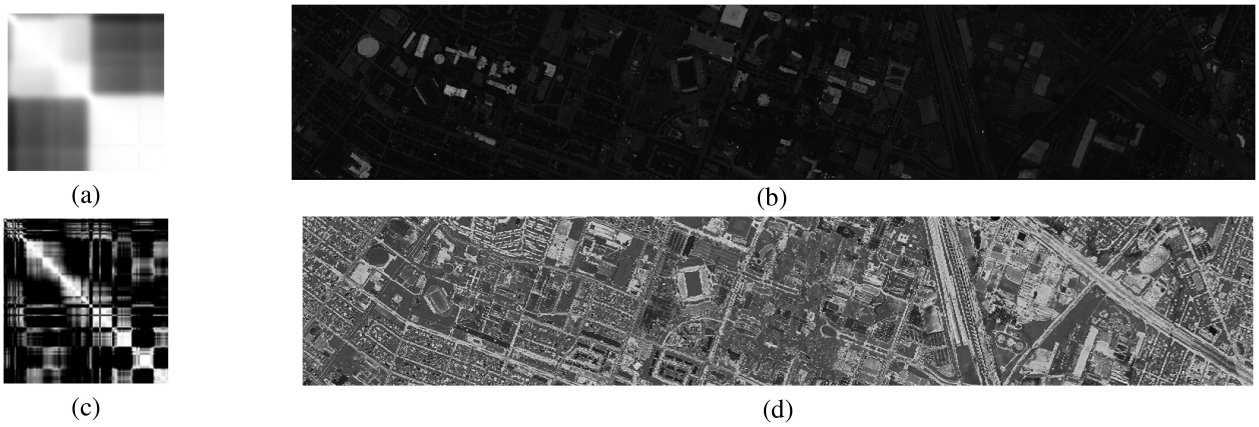

Fig. 3. Feature correlation on the Houston data. (a) Correlation coefficient matrix of spectral features. (b) Band 20 from the HSI. (c) Correlation coefficient matrix of ICV features. (d) Plane 20 of the ICV cube.

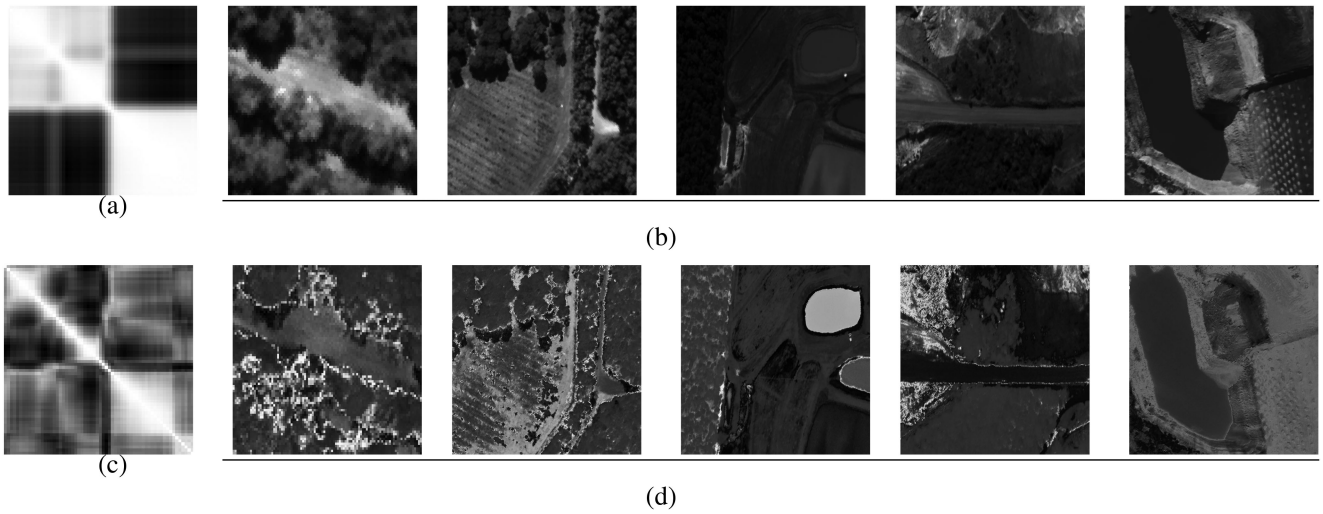

Fig. 4. Feature correlation on the GU data. (a) Correlation coefficient matrix of spectral features. (b) Band 20 from the HSI. (c) Correlation coefficient matrix of ICV features. (d) Plane 20 of the ICV cube.

whether the target being observed contains live green vegetation. $Entropy(E)$ characterizes the randomness of the texture. Note that the calculation of entropy requires a gray scale image. Given an HSI cube, we select three bands from the cube and generate a pseudo-color RGB image. Then, the RGB image is transformed to a gray scale image using a luminosity conversion function [28]. The entropy is calculated at each pixel centered in a patch of 9×9 of the gray scale image from a local neighborhood [28].

Six features are extracted from LiDAR, including DSM, digital elevation model (DEM), normalized DSM (nDSM), difference between the first and last LiDAR pulse returns (PD), intensity of the LiDAR return, and entropy of nDSM. DSM captures the natural elevation of the earth surface with objects on it. DEM is the representation of the bare-earth terrain without any object. nDSM is the difference between bare earth and first LiDAR return reflected by an object on earth. PD is the difference between the first and the last LiDAR pulse returns of a pixel. It gives important information related to the solidness of the surface, and can distinguish trees and buildings with the same height. Intensity is the strength of the LiDAR return pulse. nDSM entropy measures the entropy calculated from the nDSM data and is calculated at each pixel centered in a patch of 9×9 of the nDSM image from a local neighborhood. Fig. 5 shows the steps of generic feature extraction. DSM, DEM, and other eight features extract useful elevation and spatial features from LiDAR as well as spectral and spatial features from hyperspectral. They hold complementary information with each other.

3) Spatial Feature Extraction: For better classification of HSI and LiDAR data, it is important to analyze geometric properties of objects in a scene [29]. An effective technique for extracting spatial features from an image is mathematical morphological operation. Like morphological profiles [30], APs perform multiscale analysis of an image by using structuring element (SE) [30]. APs incorporate different types of attributes to produce different characterizations of the scale of the structures of a scene captured in an image. Since AP uses a sequence

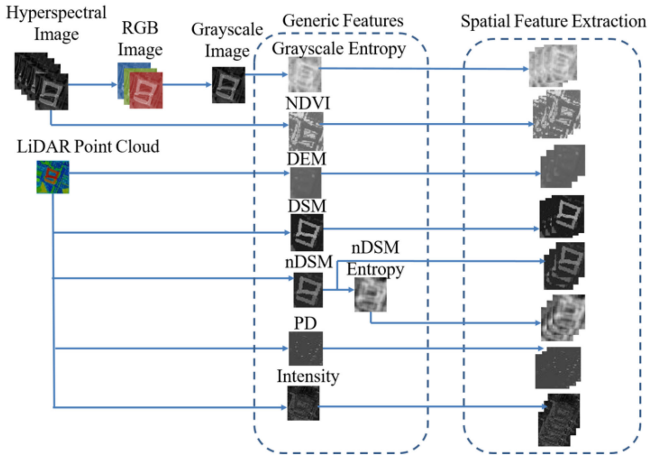

Fig. 5. Generic features extraction and spatial feature extraction from corresponding generic feature.

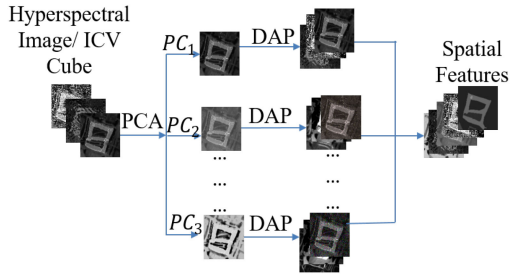

Fig. 6. Spatial feature extraction from HSI and ICV cube.

of opening and closing with SE of increasing size, it is able to extract properties invariant to scale. Given h attributes (e.g., area) for an image plane, AP produces a stack of $2h+1$ planes, in which h planes come from the closing profile, the original plane itself is a profile, and another h planes come from the opening profile. The differential attribute profile (DAP) [29], [30] stores the residuals of the subsequent increasing profiles from a plane. Since important components of the profiles are more evident in DAP, it is more practical for obtaining important information such as spatial features. Therefore, we extract spatial features using DAP from the planes of the HSI cube, ICV cube, and eight GF.

Since the ICV cube contains the same number of planes as the original HSI cube, which is still a large number, it is infeasible to apply DAP on each plane of the cube as it produces $R \times (2h+1)$ dimensional feature vector for every pixel of a cube with R planes. This may lead to the Hughes phenomenon [29]. In such a situation, it is more practical to reduce the bunch of planes without losing significant information. Thus, we select the few largest PCs with a cumulative variance of at least 99%. The feature space is, then, projected onto the selected PCs for generating a new feature space of reduced dimension. Then, we apply DAP on the reduced feature planes [4]. Similar procedure is applied for HSI cube. Fig. 6 shows steps of spatial feature extraction from HSI and ICV cubes. Fig. 5 shows spatial feature extraction from each generic feature.

B. Two-Stream CCA Fusion

Our proposed two-stream fusion (i.e., Stream 1 and Stream 2) involves fusing features from HSI and LiDAR data using CCA. CCA fuses two sets of features and produces a new set of features as an output after transformation and fusion. In the first stream, pair-wise CCA fusion takes place between spectral features from HSI and spatial features extracted from the original HSI and eight GF. In the second stream, pairwise CCA fusion takes place in the similar way of Stream 1 but replace the spectral features with ICV features.

CCA has been widely used to analyze associations between two sets of variables. In the proposed method, CCA calculates the correlation between two groups of feature vectors. It is not only suitable for information fusion, but also eliminates the redundant information within the features.

Given two feature sets $\mathbf{F}$ and $\mathbf{L}$ extracted from n samples, let $\mathbf{F} \in \mathbb{R}^{a\times n}$ and $\mathbf{L} \in \mathbb{R}^{b\times n}$ be two matrices each containing n feature vectors from two feature sets. We calculate center matrices $\mathbf{F}$ and $\mathbf{L}$ as follows:

$$\mathbf{F}_c = \mathbf{F} - \overline{\mathbf{F}} \tag{5}$$

$$\mathbf{L}_c = \mathbf{L} - \overline{\mathbf{L}} \tag{6}$$

where $\overline{\mathbf{F}}$ and $\overline{\mathbf{L}}$ are the means of feature sets $\mathbf{F}$ and $\mathbf{L}$, respectively. Within-set covariance matrices corresponding to $\mathbf{F}$ and $\mathbf{L}$ can be denoted by $\mathbf{S}_{FF} \in \mathbb{R}^{a\times a}$ and $\mathbf{L}_{KK} \in \mathbb{R}^{b\times b}$, which can be calculated as follows:

$$\mathbf{S}_{FF} = \mathbf{F}_c \times {\mathbf{F}_c}^T \tag{7}$$

$$\mathbf{L}_{KK} = \mathbf{L}_c \times {\mathbf{L}_c}^T \tag{8}$$

where F_c and L_c are the feature vectors of class c from the sets F and L, respectively. The eigen-decomposition of a matrix, which analyzes the structure of the matrix can be defined for $\mathbf{S}_{FF}$ as

$$\mathbf{S}_{FF}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{9}$$

where λ is the eigenvalue associated to the eigenvector $\boldsymbol{u}$. We ignore zero eigenvalues and associated eigenvectors and sort eigenvalues and their corresponding eigenvectors in descending order and, then, obtain the sorted eigenvector matrix as a projection matrix $\boldsymbol{u}_{FF}$. Similarly, we obtain the projection $\boldsymbol{v}_{KK}$ from $\mathbf{L}_{KK}$. The feature sets $\mathbf{F}_c$ and $\mathbf{L}_c$ are, then, projected to the spaces defined by $\boldsymbol{u}_{FF}$ and $\boldsymbol{v}_{KK}$ respectively using

$$\mathbf{F}_c^* = \mathbf{F}_c \boldsymbol{u}_{FF} \tag{10}$$

$$\mathbf{L}_c^* = \mathbf{L}_c \boldsymbol{\nu}_{KK}. \tag{11}$$

Our idea is to extract the canonical correlation between $\mathbf{F}_c^*$ and $\mathbf{L}_c^*$. We aim to compute the sample cannonical coefficient matrices $\mathbf{W}_F$ and $\mathbf{W}_L$ for the feature sets $\mathbf{F}_c^*$ and $\mathbf{L}_c^*$, respectively, such that the correlations of projections of the variables onto corresponding cannonical coefficient matrices are maximized for separate classes while the correlation among the projected variates within each class are minimized. The jth columns of $\mathbf{W}_F$ and $\mathbf{W}_L$ contain the canonical coefficients, i.e., the linear combination of variables making up the jth canonical variable

for $\mathbf{F}_c^*$ and $\mathbf{L}_c^*$, respectively. We transform the matrices $\mathbf{F}_c^*$ and $\mathbf{L}_c^*$ using

$$\mathbf{F}^* = (\alpha_1, \alpha_2, \ldots, \alpha_d)^T \mathbf{F}_c^* = \mathbf{W}_F^T \mathbf{F}_c^* \tag{12}$$

$$\mathbf{L}^* = (\beta_1, \beta_2, \ldots, \beta_d)^T \mathbf{L}_c^* = \mathbf{W}_L^T \mathbf{L}_c^* \tag{13}$$

where $\mathbf{W}_F = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ and $\mathbf{W}_L = (\beta_1, \beta_2, \ldots, \beta_d)$ and $\alpha_1, \alpha_2, \ldots, \alpha_d$ and $\beta_1, \beta_2, \ldots, \beta_d$ are correlation coefficients. We concatenate two features using $\mathbf{E} = [\mathbf{F}^* \mathbf{L}^*]$ to obtain the final feature generated by CCA from two feature sets.

C. Ensemble Classifier Design

Ensemble classifier design includes feature subsetting, data transformation, weighting the outputs of several random forest (RF) classifiers and predicting the final class label.

1) Subsetting: From two-stream CCA fusion, we derive 18 groups of features, nine from each of Streams 1 and Stream 2. Random feature selection from each subset produces diverse combinations of features. Each CCA fusion block fuses spatial features from HS, ICV or GF with either HSI spectral or ICV features, thus produces distinctive features. To ensure the representation of features from all 18 groups, we apply random selection within each group and, then, concatenate them to form a new feature vector. The number of features selected for a subset from each group depends on the number of features available in that group and the total number of desired subsets. For example, if a group derived from a CCA block contains f features and the number of subset is m, then in each selection $\lceil \frac{f}{m} \rceil$ features are selected and they are moved to a subset. If f is not multiple of m, the number of features in the last selection is less than $\lceil \frac{f}{m} \rceil$. This process is applied to each CCA derived features. Suppose the ith CCA block generates f_i features and we aim to generate m subsets from g groups, then in the first selection we select $\lceil \frac{f_1}{m} \rceil$, $\lceil \frac{f_2}{m} \rceil$,..., $\lceil \frac{f_g}{m} \rceil$ features from group 1, 2,..., g, respectively. Then, they are concatenated to form the first subset. Similarly, we form subset 2,..., m.

2) Data Transformation: Data transformation plays an important role in the proposed ensemble classifier. In this article, minimum noise fraction (MNF) [31] is used as the transformation technique. MNF chooses new components to maximize the signal-to-noise ratio. Noise can be effectively removed from the data by transforming the original data to the MNF space, smoothing or rejecting the noisiest components, and, then, retransforming back to the original space. We concatenate the transformed components extracted from all subsets to generate a novel feature set and use the set to train an RF classifier. Besides MNF, the performances of two different data transformation strategies, i.e., linearity preserving projection (LPP) [32] and neighborhood preserving embedding (NPE) [33]), are analyzed in the experiment section.

3) Class Prediction: The proposed ensemble classification uses both RF and support vector machine (SVM) classifiers. The signature generated by the transformation process is sent to the classifier. Specially, the RF classifier provides out-of-bag (OOB) error for each class at the training stage. OOB is the mean prediction error on each training sample by aggregating the prediction error caused by the trees constructed by the bootstraps not containing the sample. Aggregating the errors for samples of a particular class, we get OOB error for each class. The higher is the OOB error is, the lower is the prediction accuracy.

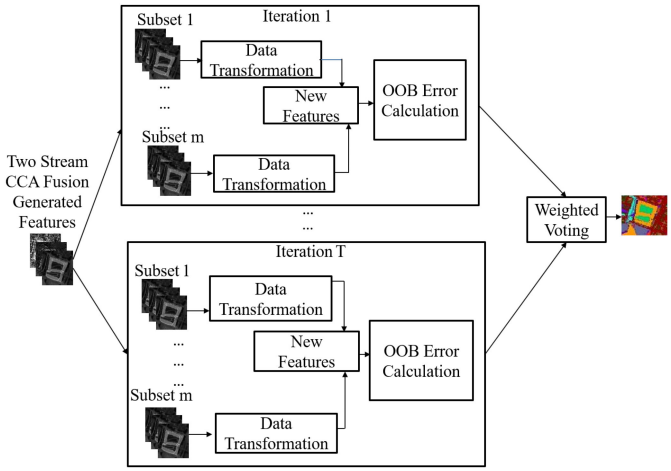

Fig. 7. Ensemble classifier design.

Given a new sample in testing stage, we get a probabilistic score of every class for the sample. Let the sample space has C classes and the system is iteratively trained and tested for T times. In each iteration $t \in \{1, 2, \ldots, T\}$, we get the OOB error e_c^t for class $c \in \{1, 2, \ldots, C\}$ at training stage of the tth iteration. Given a test sample, it achieves probability of P_c^t for class c at the testing stage of the tth iteration. We aim to use the OOB error for finalizing the class prediction by inversely weighting by the error. Thus, a test sample gets a total weight of $\sum_{t=1}^{T} \frac{P_c^t}{e_c^t}$ for class c through all iterations. Out of all T iterations, we determine the class having maximum weight $c_{\max} \in \{1, 2, \ldots, C\}$ by using the following equation:

$$c_{\max} = \max_c \sum_{t=1}^{T} \frac{P_c^t}{e_c^t}. \tag{14}$$

Finally, the given test sample is predicted as the class $c_{\max}$. We predict every test sample in this way. Fig. 7 shows steps of ensemble classifier design.

The computational complexity of the ensemble classifier is calculated using the number of samples and attributes of tree. Suppose that we have n samples and each tree has w attributes. The computational cost of building a tree is $O(wn \log n)$ where it is assumed that the depth of a tree is in the order of $\log n$ [34]. For a RF with τ decision trees, the complexity is $O(\tau(wn \log n)$ [34]. The total complexity of subtree lifting for a decision tree is $O(n(\log n)^2)$. So, the full complexity of the RF classifier is as follows:

$$O(\tau(wn \log n) + \tau(n(\log n)^2)). \tag{15}$$

For our ensemble classifier design, we iterate the classification process for t times so the final computational complexity is as

TABLE I
HOUSTON I DATA: CLASS-WISE TESTING SAMPLES FOR FIXED NUMBER OF TRAINING SAMPLES

	Number of Samples	
Class Name	**Training**	**Testing**
Grass healthy	20	1053
Grass stressed	20	1064
Synthetic grass	20	505
Trees	20	1056
Soil	20	1056
Water	20	143
Residential	20	1072
Commercial	20	1053
Road	20	1059
Highway	20	1036
Railway	20	1054
Parking lot1	20	1041
Parking lot 2	20	285
Tennis court	20	247
Running track	20	473
Total	300	12197

This distribution is used for the results in Table IV.

TABLE II
HOUSTON II DATA: CLASS-WISE SAMPLES PROVIDED BY HOUSTON DATA

	Number of Samples	
Class Name	**Training**	**Testing**
Grass healthy	198	1053
Grass stressed	190	1064
Synthetic grass	192	505
Trees	188	1056
Soil	186	1056
Water	182	143
Residential	196	1072
Commercial	191	1053
Road	193	1059
Highway	191	1036
Railway	181	1054
Parking lot1	192	1041
Parking lot 2	184	285
Tennis court	181	247
Running track	187	473
Total	2832	12197

This distribution is used for the results in Tables V and VI.

TABLE III
GU DATA: DISTRIBUTION OF TRAINING AND TESTING SAMPLES

	Number of Samples	
Class Name	Training	Testing
Road	30	822
Tree	30	1355
Grass	30	821
Water	30	379
Soil	30	780
Total	150	4157

This distribution is used for the results in Tables VII and VIII.

below

$$O\left(\sum_{i=1}^{t}\left(\tau_i(w_i n_i \log n_i) + \tau_i(n_i(\log n_i)^2)\right)\right). \quad (16)$$

III. EXPERIMENTS AND ANALYSIS

In this section, we present the experimental results on two real-world datasets, and compare our method with several state-of-the-art approaches.

A. Data Description

1) Houston Data: The Houston data consists of an HSI and a LiDAR-derived DSM. This dataset was distributed at the 2013 GRSS data fusion contest [25]. Both HSI and LiDAR data were collected over the University of Houston campus and the neighboring urban area. The LiDAR data were acquired on June 22, 2012 and the average height of the sensor above ground was 2000 ft. The HSI was acquired by a compact airborne spectrographic imager on June 23, 2012 and the average height of the sensor above ground was 5500 ft. The size of both HSI and LiDAR data is 349 $\times$ 1905 pixels with the spatial resolution of 2.5 m. The HSI consists of 144 spectral bands ranging from 0.38 to 1.05 μm. The 15 classes of interests are Grass Healthy, Grass Stressed, Grass Synthetic, Tree, Soil, Water, Residential, Commercial, Road, Highway, Railway, Parking Lot 1, Parking Lot 2, Tennis Court, and Running Track. For Houston data, we use two distributions of samples in the experiments, which are given in Tables I and II. Table I is used to examine the performance of the proposed model for a limited number of training samples. We use both distribution to make the results fully comparable with other works in the literature.

2) GU Data: The second dataset is provided by the Spectral Imaging Lab at Griffith University. It consists of an HSI and corresponding LiDAR point cloud collected from Yarraman State Forest, Queensland, Australia, and its adjacent area. The total area was around 8 km^2. The data were captured in June 2015. The HSI consists of 62 bands ranging from 408.54 to 990.62 nm in wavelength, where the spectral resolution is 8.94 to 9.81 nm. The spatial resolution of the dataset is 0.5 m. For the LiDAR point cloud, the number of returns is 6 and the average point spacing is 0.2 m. We collected training and testing samples from 33 different areas of the image, where the HSI is cropped using an open source tool Barista and the LiDAR data are cropped using MARS 7. We use ENVI 5.3 to generate DSM and DEM, respectively. The selected areas cover five classes of interests, e.g., Road, Tree, Grass, Water, and Soil. Out of the 33 areas, we arbitrarily select the points from five areas for training and the remaining 28 areas for testing. The distribution of training and testing samples are given in Table III.

B. Experimental Setup

For selecting suitable bandwidth of Gaussian kernel ϕ for (1), we analyze the overall accuracy (OA) curve with respect to perplexity. Fig. 8(a) and (b) shows OA for different values of perplexity for the Houston and the GU data, respectively. Observing the best OA, we consider the perplexity in the range of 111 to 120 for the Houston data. On the other hand, we consider perplexity in the range of 40 to 50 for the GU data.

To calculate NDVI, we choose 797.90 nm (band 92) for NIR and 678.79 nm (band 30) for red on the Houston data, and

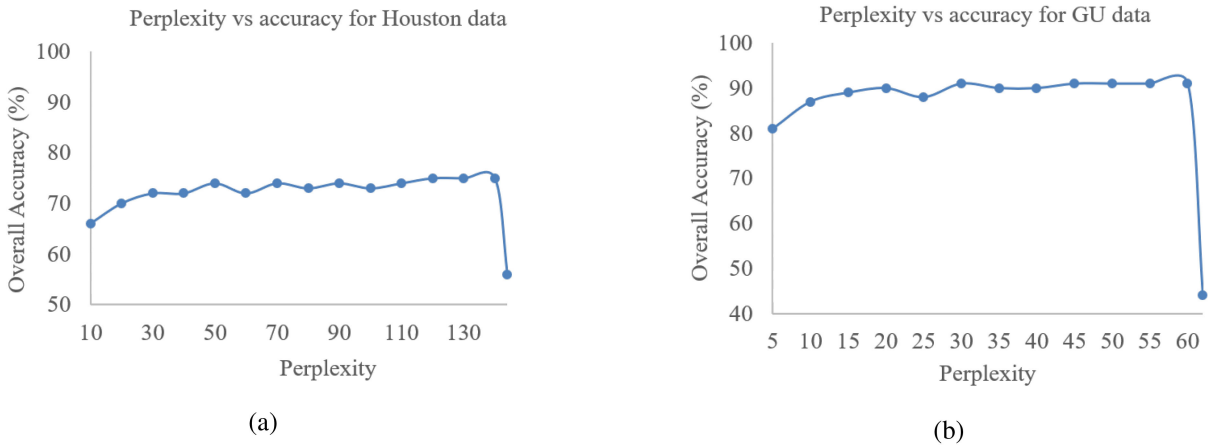

Fig. 8. OA for different values of perplexity. (a) Houston Data (b) GU Data.

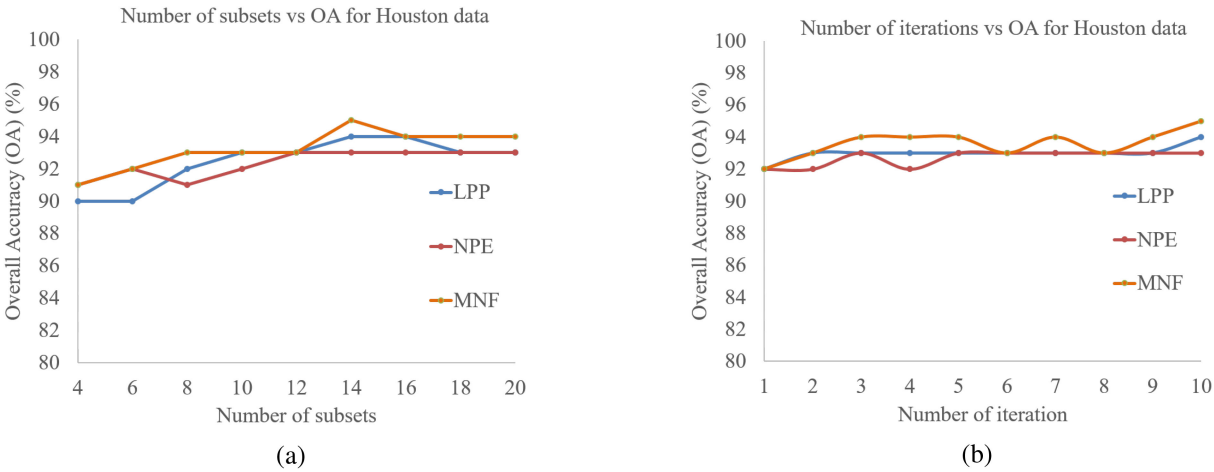

Fig. 9. Houston data: OA obtained by applying RF for (a) different number of subsets and (b) different number of iterations for ensemble classifier design.

795.16 nm (band 42) for NIR and 679.46 nm (band 30) for red on the GU data [26].

The calculation of Entropy from HSI requires a gray scale image. We generate the gray scale image from a pseudo-color RGB image. For Houston data, we use 693.10 nm (band 70) for red, 597.80 nm (band 50) for green, and 454.80 nm (band 20) for blue, respectively. For the GU data, we choose 650.84 nm (band 27) for red, 536.64 nm (band 15) for green, and 472.19 nm (band 8) for blue [26]. We choose two different wavelengths for red to compute NDVI and RGB image. According to [1], we prefer far-red rather than red to derive NDVI.

We apply DAP technique for spatial feature extraction. Before this step, the input HSI is normalized in the range of [0 1]. Prior to applying the DAP on HSI and ICV, PCA is applied to the HSI and ICV cubes. In both cases, the first few PCs having at least 99% of their cumulative variance are selected. Similarly, we apply DAP on DSM, DEM, nDSM, intensity, PD, nDSM entropy, and gray entropy of HSI and NDVI. In order to generate DAP, the areas of 10, 15, and 20 and bounding box diagonals of 50, 100, and 500 are used. A total of 13 profiles are generated for these areas and diagonal parameters.

In ensemble classifier design, we randomly generate subsets of features. Figs. 9(a) and 10(a) show how OA varies for different number of subsets for Houston and GU data, respectively. For Houston data, the highest OA is obtained when the number of subsets is 13 and for GU data the highest OA is obtained when the number of subsets is 6. Since Figs. 9(b) and 10(b) show the best OA in ten iterations for both the Houston and GU data, we use ten iterations for our experiments.

In classification, we separately use two widely used classifiers, i.e., RF and SVM. For RF, the number of trees is set to 300. The number of the prediction variable is set approximately to the square root of the number of input bands. The SVM is trained with a Gaussian kernel by automatically setting the scale with a Bayesian optimization model and penalty parameter (e.g., box constraint) in the range of 10^{-2}–10^{4} during optimization. For Houston and GU data, the penalty parameter is set to the best performing box constraints of 100 and 26.8, respectively. The classifier models are validated with ten-fold cross validation.

For SVM classification, we also iteratively train and test the system. In this case, we record the classification result, e.g., predicted label, for a sample throughout all T iterations. The final

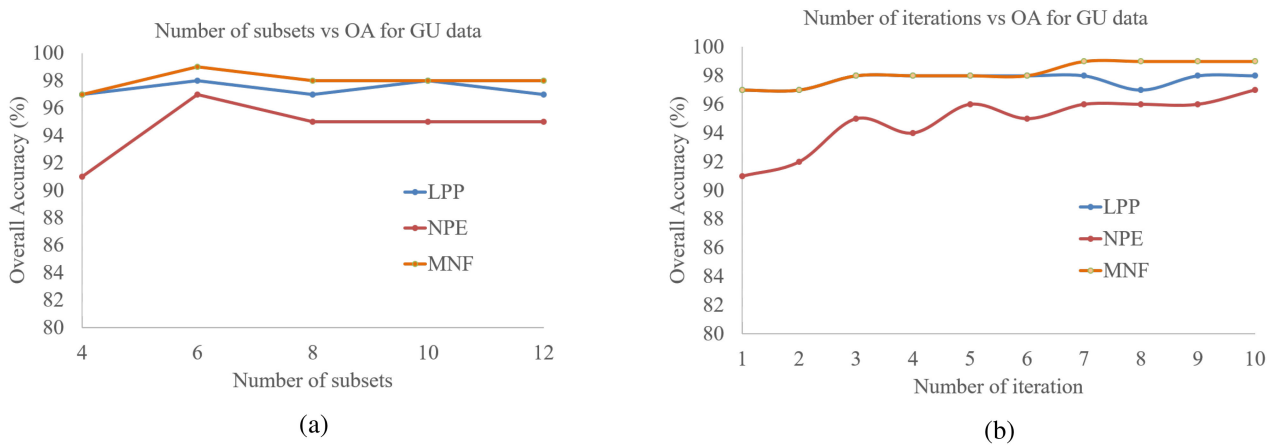

Fig. 10. GU data: OA obtained by applying RF for (a) different number of subsets and (b) different number of iterations for ensemble classifier design.

TABLE IV
RF ON HOUSTON I: CLASSIFICATION ACCURACIES BY DIFFERENT FEATURES AND FUSION OPTIONS FOR THE SAMPLE DISTRIBUTION OF TABLE I

	HS+ DAP(HS) (183)	ICV+ DAP(ICV) (183)	DAP(GF) (104)	Two-Stream Fusion			Ensemble Classifier Design		
				Stream 1 (252)	Stream 2 (248)	Combined (500)	LPP (390)	NPE (390)	MNF (390)
Grass Healthy	83.57	92.50	85.00	83.86	91.93	87.84	84.14	89.93	87.94
Grass Stressed	98.40	98.68	98.87	99.15	98.59	98.78	98.50	98.59	98.87
Grass Synthetic	99.80	99.60	99.41	100.00	99.60	99.60	100.00	100.00	100.00
Tree	96.97	85.89	99.81	99.24	99.72	99.15	100.00	99.81	100.00
Soil	98.20	97.92	98.39	99.15	98.77	99.05	99.91	99.62	99.91
Water	83.22	93.71	97.20	86.71	96.50	92.31	93.01	100.00	84.62
Residential	83.96	80.50	93.00	93.28	88.53	90.02	89.55	89.55	95.06
Commercial	56.98	51.28	82.91	78.63	84.43	84.05	78.82	83.76	90.22
Road	74.79	62.70	74.41	89.52	88.29	87.63	84.42	82.34	82.81
Highway	40.93	67.28	84.27	77.61	82.92	88.32	68.92	68.82	75.39
Railway	89.09	72.49	91.56	97.53	87.00	94.59	80.93	74.29	93.83
Parking Lot 1	56.48	60.71	70.03	71.57	75.50	70.99	74.16	70.51	86.94
Parking Lot 2	75.44	55.09	76.84	84.91	87.37	87.37	89.47	89.47	89.82
Tennis Court	99.60	99.60	98.79	100.00	100.00	100.00	100.00	99.60	100.00
Running Track	99.58	91.75	99.79	100.00	98.73	99.79	100.00	99.15	100.00
OA	80.23	78.70	88.88	90.00	90.61	91.01	87.57	87.43	**91.92**
AA	82.47	80.65	90.02	90.74	91.86	91.97	89.45	89.70	**92.36**
Kappa	0.79	0.77	0.88	0.89	0.90	0.90	0.87	0.86	**0.91**

predicted class is determined by applying majority voting on the predicted labels of all iterations. For instance, suppose for a test sample, the predicted class label at the tth iteration is denoted by l^t, then the final prediction label y for the sample is determined by applying majority voting as given by the following equation:

$$y = \text{mode}\{l^t | t = 1, \ldots, T\}. \tag{17}$$

The classification results are represented by *OA*, *average accuracy* (AA), and the *kappa* coefficient [1]. The metrics OA and AA are in percentage. Kappa coefficient does not have a unit. The best OA, AA, and Kappa results are shown in bold-face in Tables IV, V, VI, VII, and VIII. The number of features is put in parentheses.

To represent the performance metrics of different methods, we use the following short-hand terms:

1) *HS+DAP (HS):* stacked spectral responses of HSI and DAP applied to the PCs (having cumulative variance of $\geq 99\%$) of HS;
2) *ICV+DAP(ICV):* stacked responses of ICV and DAP applied to the PCs (having cumulative variance of $\geq 99\%$) of ICV;
3) *DAP(GF):* DAP applied on GF;
4) *Two-Stream Fusion (Stream 1):* each of the pairs of HSI and DAP(HS), HSI and DAP(DSM), HSI and DAP(DEM), HSI and DAP(nDSM), HSI and DAP(Intensity), HSI and DAP(PD), HSI and DAP(nDSM entropy), HSI and DAP(gray entropy), and HSI and

TABLE V
RF ON HOUSTON II: CLASSIFICATION ACCURACIES BY DIFFERENT FEATURES AND FUSION OPTIONS FOR THE SAMPLE DISTRIBUTION OF TABLE II

	HS+ DAP(HS) (183)	ICV+ DAP(ICV) (183)	DAP(GF) (104)	Two-Stream Fusion			Ensemble Classifier Design		
				Stream 1 (258)	Stream 2 (256)	Combined (514)	LPP (390)	NPE (390)	MNF (390)
Grass Healthy	83.29	94.02	84.71	83.48	93.45	90.31	83.86	84.33	84.52
Grass Stressed	96.99	96.99	98.59	99.15	98.40	98.50	99.25	99.15	99.15
Grass Synthetic	100.00	99.60	99.41	100.00	100.00	100.00	100.00	100.00	100.00
Tree	98.30	91.48	99.91	99.81	99.15	99.43	100.00	99.91	100.00
Soil	98.77	97.92	97.44	99.81	98.58	99.05	99.53	99.62	99.72
Water	97.90	99.30	100.00	100.00	98.60	100.00	98.60	100.00	100.00
Residential	87.50	83.02	86.94	91.42	93.66	92.26	91.98	93.47	93.66
Commercial	71.51	61.54	87.94	95.92	87.08	96.58	92.40	95.35	97.34
Road	81.11	76.96	83.19	88.57	90.75	90.84	95.47	93.86	95.47
Highway	65.35	71.91	73.55	72.88	73.75	72.78	81.66	78.28	76.93
Railway	87.57	82.64	97.53	98.58	94.12	97.72	95.45	95.92	97.72
Parking Lot 1	65.80	67.15	78.58	91.21	85.88	88.57	88.66	87.42	95.20
Parking Lot 2	77.89	76.49	68.77	78.95	84.56	83.86	89.82	86.67	87.02
Tennis Court	100.00	99.60	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Running Track	99.37	94.93	99.79	99.15	98.52	99.58	98.73	99.15	100.00
OA	85.34	84.02	89.63	93.18	92.24	93.25	93.52	93.41	**94.52**
AA	87.42	86.24	90.42	93.66	93.10	93.97	94.36	94.21	**95.11**
Kappa	0.84	0.83	0.89	0.93	0.92	0.93	0.93	0.93	**0.94**

TABLE VI
SVM ON HOUSTON: CLASSIFICATION ACCURACIES BY DIFFERENT FEATURES AND FUSION OPTIONS FOR SAMPLE DISTRIBUTION OF TABLE II

	HS+ DAP(HS) (183)	ICV+ DAP(ICV) (183)	DAP(GF) (104)	Two-Stream Fusion			Ensemble Classifier Design		
				Stream 1 (258)	Stream 2 (256)	Combined (514)	LPP (390)	NPE (390)	MNF (390)
Grass Healthy	81.77	69.80	83.86	81.96	90.69	89.27	89.17	89.36	83.67
Grass Stressed	97.56	95.77	100.00	99.53	98.31	98.97	98.97	98.87	100.00
Grass Synthetic	100.00	99.80	100.00	100.00	100.00	100.00	100.00	100.00	99.60
Tree	95.45	86.84	98.39	97.06	96.59	98.30	97.82	98.67	97.16
Soil	97.06	97.92	99.24	99.05	98.77	98.96	98.96	99.15	99.34
Water	96.50	97.20	99.30	83.92	95.10	92.31	81.12	93.01	99.30
Residential	90.21	78.64	90.30	86.85	82.74	85.07	82.56	83.86	87.69
Commercial	51.28	57.45	79.11	59.64	74.74	57.64	61.54	61.73	78.82
Road	77.90	34.28	83.95	90.08	90.84	95.18	95.94	95.85	85.84
Highway	65.93	69.59	86.97	94.31	85.62	93.53	87.45	88.42	85.33
Railway	88.33	54.36	97.06	93.83	59.28	91.94	92.22	93.93	97.72
Parking Lot 1	76.75	63.69	88.28	89.82	84.73	87.90	89.53	87.90	91.64
Parking Lot 2	65.96	64.56	88.07	69.12	71.23	69.82	70.53	71.23	88.77
Tennis Court	100.00	99.60	100.00	100.00	99.19	100.00	99.60	100.00	100.00
Running Track	100.00	98.94	100.00	98.52	98.73	98.73	98.73	98.73	100.00
OA	83.86	73.90	**91.70**	89.71	89.91	90.24	89.89	90.37	**91.70**
AA	85.65	77.90	92.97	89.58	90.44	90.51	89.61	90.71	**92.99**
Kappa	0.82	0.72	**0.91**	0.89	0.89	0.89	0.89	0.90	**0.91**

DAP(NDVI) are fused with CCA and then the results are concatenated;

5) *Two-Stream Fusion (Stream 2):* each of the pairs of ICV and DAP(ICV), ICV and DAP(DSM), ICV and DAP(DEM), ICV and DAP(nDSM), ICV and DAP(Intensity), ICV and DAP(PD), ICV and DAP(nDSM entropy), ICV and DAP(gray entropy), and ICV and DAP(nDVI) are fused with CCA and then the results are concatenated;
6) *Two-Stream Fusion (Combined):* two-Stream Fusion (Stream 1) and two-stream fusion (Stream 2) are stacked;
7) *Ensemble Classifier Design (LPP):* LPP is used for data transformation in ensemble classifier design;

TABLE VII
RF ON GU: CLASSIFICATION ACCURACIES BY DIFFERENT FEATURES AND FUSION OPTIONS

	HS+DAP(HS) (101)	ICV+DAP(ICV) (101)	DAP(GF) (104)	Two-Stream Fusion			Ensemble Classifier Design		
				Stream 1 (184)	Stream 2 (184)	Combined (370)	LPP (180)	NPE (180)	MNF (180)
Road	94.23	91.29	93.50	98.53	96.93	98.28	96.07	94.11	96.44
Tree	86.64	88.36	92.39	98.28	97.63	98.13	99.07	98.64	99.86
Grass	88.75	85.08	95.45	94.69	95.83	95.70	94.69	93.68	97.35
Water	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Soil	95.13	95.51	99.49	95.13	100.00	100.00	99.62	98.85	100.00
OA	91.34	90.71	95.21	97.21	97.81	98.22	97.83	96.97	**98.75**
AA	92.95	92.05	96.16	97.32	98.08	98.42	97.89	97.05	**98.73**
Kappa	0.89	0.88	0.94	0.96	0.97	**0.98**	0.97	0.96	**0.98**

TABLE VIII
SVM ON GU: CLASSIFICATION ACCURACIES BY DIFFERENT FEATURES AND FUSION OPTIONS

	HS+DAP(HS) (101)	ICV+DAP(ICV) (101)	DAP(GF) (104)	Two-Stream Fusion			Ensemble Classifier Design		
				Stream 1 (184)	Stream 2 (184)	Combined (370)	LPP (180)	NPE (180)	MNF (180)
Road	66.26	94.97	89.57	91.66	94.11	95.58	95.58	96.32	96.93
Tree	81.54	79.38	93.46	90.37	87.43	90.66	88.43	89.15	94.40
Grass	76.99	81.67	95.32	94.82	84.58	91.02	95.07	86.60	96.46
Water	100.00	100.00	100.00	100.00	100.00	100.00	89.71	100.00	100.00
Soil	87.69	95.00	96.03	99.23	99.49	99.36	99.87	99.87	100.00
OA	80.51	87.68	94.13	94.01	91.60	94.18	93.36	93.07	**96.85**
AA	82.50	90.20	94.88	95.22	93.12	95.33	93.73	94.39	**97.56**
Kappa	0.75	0.84	0.92	0.92	0.89	0.92	0.91	0.91	**0.96**

8) *Ensemble Classifier Design (NPE):* NPE is used for data transformation in ensemble classifier design;
9) *Ensemble Classifier Design (MNF):* MNF is used for data transformation in ensemble classifier design.

C. Results and Discussion on Houston Data

1) RF Houston I: Tables IV and V show the classification accuracies obtained by using RF on Houston data given in Tables I and II, respectively. Table IV shows the performance of feature combination, two-stream CCA fusion and ensemble classifier performance using a small training set (20 samples from each class). The results endorse that our feature combination, fusion and ensemble method is robust enough to correctly classify complex classes with a small number of training samples. In Stream 1 of two-stream fusion, the OA is increased by 9.77% compared to the combination HS+DAP(HS) when CCA fusion is done between HSI spectral responses and DAP features derived from HSI and LiDAR. Similarly, in Stream 2 of two-stream fusion, CCA fusion between ICV responses and other DAP features derived from ICV and LiDAR improves OA by 11.91% compared to the combination ICV+DAP(ICV). Two-stream fusion increases complementary information between both streams and provides OA of 91.01%, which is 10.78% higher than HS+DAP(HS), 12.31% higher than ICV+DAP(ICV), and 2.13% higher than DAP(GF). With a small training set, we achieve OA of 91.92% using our approach. Ensemble classifier using MNF increases OA by 11.69% compared to HS+DAP(HS), 13.22% compared to ICV+DAP(ICV), and 3.04% compared to DAP(GF). Our approach of CCA fusion between two feature sets identifies the linear relationship between two modalities by maximizing interclass covariance. Pair-wise CCA fusion creates variation in feature fusion in comparison with layer stacking as well as reduces the dimensionality. Random subsetting, data transformation using MNF, and accumulating decisions of several RF classifiers give better classification accuracy with a small amount of training samples and also reduce the dimensionality of the feature vector.

2) RF Houston II: Table V uses the distribution of training and testing samples of Houston data given in Table II. As per Table V, two-stream CCA fusion achieves OA of 93.25%, which is 7.91% higher than HS+DAP(HS), 9.23% higher than ICV+DAP(ICV), and 3.62% higher than DAP(GF). Ensemble classifier using MNF shows OA of 94.52%, which is the highest among the three. The proposed method increases OA of individual class such as "Grass Stressed," "Tree," "Soil," "Residential," "Commercial," "Road," "Highway," "Parking Lot 1," "Parking Lot 2," and "Running Track" by 0.65%, 0.57%, 0.67%, 1.4%, 0.76%, 4.63%, 4.15%, 6.63%, 3.16%, and 0.42%, respectively from two-steam fusion.

3) SVM Houston: Table VIII shows the classification results obtained by the SVM classifier, which performs well when the number of features increases. SVM outperforms RF for

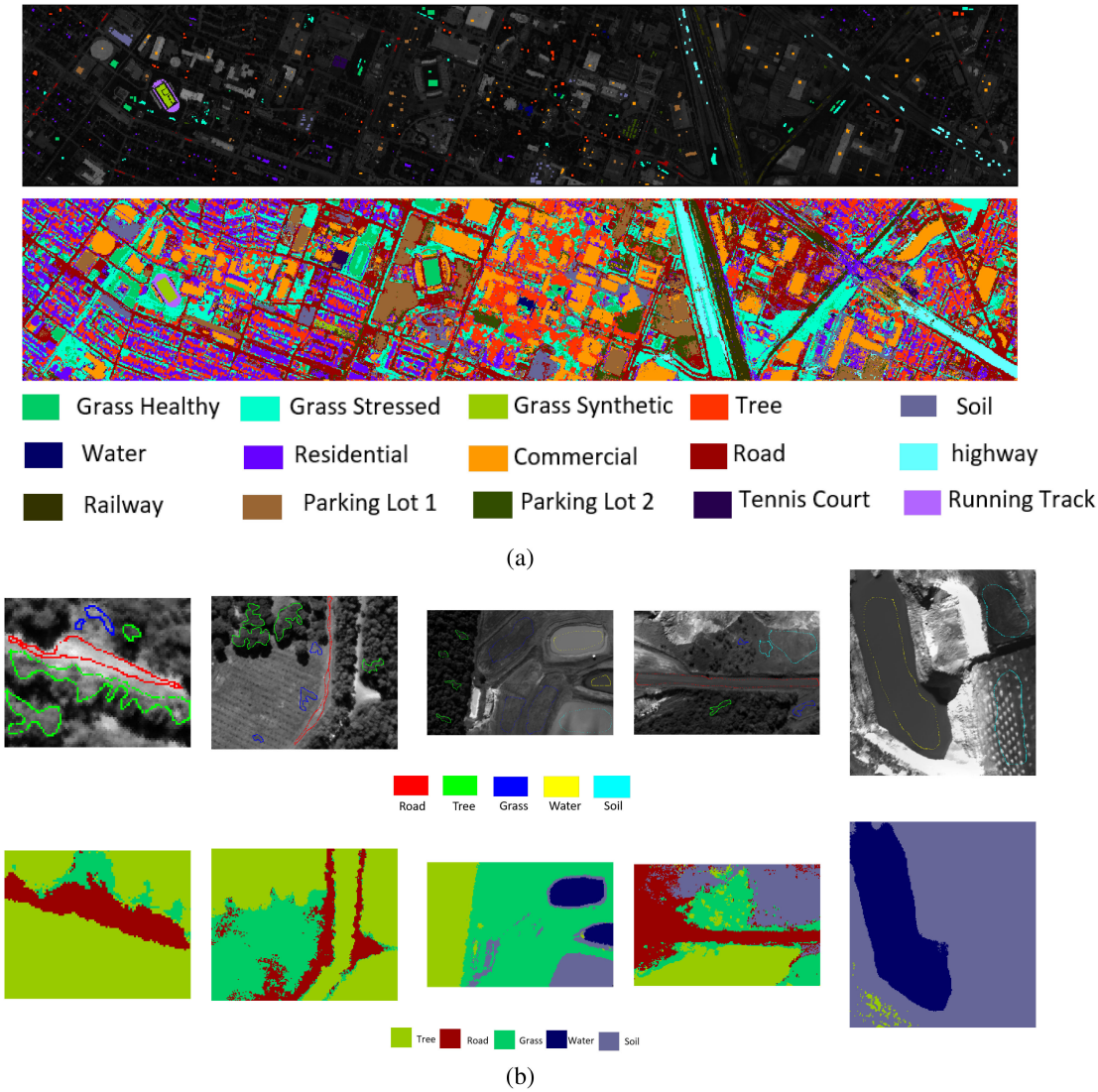

Fig. 11. Ground truth and classification maps (generated by the output of the proposed method using RF). (a) Houston data. (b) GU data.

DAP(GF). In consideration of overall system performance, RF outperforms SVM, which is also confirmed by other works in the literature [8]. The highest classification accuracy achieved by SVM for our proposed approach is 91.70%.

ICV accumulates complementary information by analyzing band to band relation with HSI spectral features. The two-stream fusion fuses ICV and HSI spectral and spatial features with other spatial features from LiDAR. Stream 2 in two-stream fusion uses the proposed ICV feature instead of HSI and achieves higher accuracy than Stream 1 that uses HSI. As a result, ICV contributes in increasing over all performance.

D. Results and Discussion on GU Data

1) RF GU: Table VII shows the classification results obtained by the RF classifier on the GU data. The OA on GU data is 98.75%. As the number of classes of GU data is one-third of the number of Houston data, the confusion between classes in the GU data is lower than Houston data. Ensemble classifier improves OA by 7.41% compared to HS+DAP(HS), 8.04% compared to ICV+DAP(ICV), and 3.54% compared to DAP(GF). In two-stream fusion, Stream 2 uses ICV features and achieves higher accuracy than Stream 1 that uses HSI. This is because the correlation between features is lower in ICV than in HSI.

2) SVM GU: Table VIII shows the classification results obtained by the SVM classifier. The highest classification accuracy obtained by SVM is 96.85% using ensemble classifier, which improves OA by 2.67% compared to two-stream fusion, 16.34% compared to HS+DAP(HS), 9.17% compared to ICV+DAP(ICV), and 2.72% compared to DAP(GF).

Fig. 11 shows classification maps of the proposed method and ground truth for Houston and GU data.

E. Performance Comparison

Finally, we compare our method with six state-of-the-art approaches. Xia *et al.* [18] derived spectral and morphological

TABLE IX
PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART STUDIES IN THE LITERATURE

	Xu et al. [11]	Khodadadzadeh et al. [4]	Ghamisi et al. [3]	Rasti et al. [8]	Liao et al. [5]	Xia et al. [18]	Jahan et al. [17]	Proposed Method
OA	87.98	90.65	91.02	91.30	94.00	**95.11**	93.05	94.52
AA	90.11	91.11	91.82	91.95	93.79	94.57	93.86	**95.11**
Kappa	0.87	0.90	0.90	0.91	**0.94**	**0.94**	0.92	**0.94**

TABLE X
PERFORMANCE COMPARISON WITH A STUDY IN THE LITERATURE THAT USED THE DISTRIBUTION OF TRAINING AND TESTING SAMPLES GIVEN IN TABLE I

	Xia el al. [18]	Proposed Method
OA	85.48	**91.92**
AA	86.75	**92.36**
Kappa	0.84	**0.91**

TABLE XI
PERFORMANCE COMPARISON WITH OTHER STUDIES IN THE LITERATURE USING THE DISTRIBUTION OF TRAINING AND TESTING SAMPLES GIVEN IN TABLE III FOR GU DATA

	Liao et al. [5]	Xia et al. [18]	Jahan et al. [17]	Proposed Method
OA	90.26	92.54	96.82	98.75
AA	92.56	93.67	97.39	98.73
Kappa	0.90	0.93	0.96	0.98

TABLE XII
COMPARISON OF ELAPSED TIME (m: MINUTES) OF FEATURE EXTRACTION, FEATURE FUSION TRAINING AND TESTING USING RF OF OUR PROPOSED METHOD WITH A STUDY IN THE LITERATURE

	Xia et al. [18]	Liao et al. [5]	Jahan et al. [17]	Proposed Method
Houston Data (m)	636	63.68	61.35	603
GU Data (m)	219	31.98	30.33	201

features from HSI and morphological features from LiDAR DSM. The extracted features were split into several disjoint subsets, data transformation was applied on each subset, and the transformed subsets were concatenated and applied as the input to an RF classifier. The processes of subsetting, transformation, and classification were repeated several times and the classification results achieved at each iteration were integrated by a majority vote. Xu *et al.* [11] developed two-tunnel CNN for deriving spectral-spatial features from HSI and cascade block CNN was designed to extract features from LiDAR. The extracted features were combined and classified. Khodadadzadeh *et al.* [4] integrated different types of features extracted from HSI and LiDAR without any regularization parameters. Ghamisi *et al.* [3] fused spatial-spectral features from HSI with spatial features from LiDAR and provided the fused feature vector as an input to a CNN. Rasti *et al.* [8] introduced a new sparse and low-rank technique for the fusion of hyperspectral and LiDAR features. Liao *et al.* [5] introduced one of the baseline fusion techniques called generalized graph-based fusion, which incorporated both feature fusion and dimensionality reduction. Jahan *et al.* [17] proposed DCA for feature extraction and fusion from both HSI and LiDAR. They also incorporated band grouping for spectral and spatial feature extraction from HSI.

All of these approaches use Houston data for experiments. To be fully comparable with other approaches, we use the same settings of the training and testing samples as used in these approaches. Table IX shows performance comparison of six state-of-the-art techniques and our proposed approach. Our ICV features, two-stream fusion and ensemble classification system improve the performance of our proposed approach, which is also evident from Tables IV to VIII. The performance of method of Xia *et al.* [18] is similar to our approach for the standard settings of Houston data. However, our method performs very well with a limited number of training samples compared to Xia *et al.* [18] (see Table X). The performance of our proposed approach is competitive with the other state-of-the-art methods.

We also examine the performance of our approach using a small number of training samples as did in the work of Xia *et al.* [18], which uses only 20 samples per class for training, keeping the number of testing samples the same as given with the original dataset. Table X shows the comparison between our approach and the approach of Xia *et al.* [18].

We further compare our proposed approach with the approaches proposed by Liao *et al.* [5], Xia *et al.* [18], and Jahan *et al.* [17] based on GU data. Table XI shows how our proposed method improves the performance of GU data applied on the same training and testing setting. Table XII compares the running time of our proposed method for feature extraction, fusion and classification on both Houston and GU data with another ensemble classifier-based method proposed by Xia *et al.* [18]. As our two-stream CCA fusion runs faster than graph-based feature fusion method used by Xia *et al.* [18], the total execution time of our proposed method is 33 and 18 min less than from Xia *et al.* [18] for Houston and GU data, respectively. Table XII compares the running time of our proposed method on both Houston and GU data with other two methods proposed by Liao *et al.* [5] and Jahan *et al.* [17]. Ensemble classifier requires several iterations of RF classifier and RF takes longer time for training and testing. For that reason, the elapsed time for Xia *et al.* [18] and our proposed method is longer than Liao *et al.* [4] and Jahan *et al.* [2]. All the experiments are implemented using MATLAB 2018a on the Windows 10, Intel Core i7-3687 U CPU @2.10 GHz, 8-GB memory.

IV. CONCLUSION

In this article, we introduce a novel multilevel fusion framework of heterogeneous features for land cover classification. This framework uses a novel stochastic approach for signal enhancement-based feature extraction from hyperspectral data.

The proposed feature adds significant complementary information to the hyperspectral data and improves the land cover classification accuracy. The proposed ICV feature preserves the local structure of the raw spectral responses by analyzing band to band similarity. The spatial features derived from GF have significant discriminative capability. Our two-stream CCA fusion strengthens the discriminative ability of both HSI and LiDAR derived features by maximizing within-class correlation and minimizing between-class correlation. Finally, the ensemble classifier design achieves OA of 94.52% on Houston data, which is competent with the existing literature. In our future work, we aim to devise an effective method for fusing data from multiple sources by incorporating material specific feature fusion followed by an advanced decision fusion approach.

REFERENCES

[1] Q. Man, P. Dong, and H. Guo, "Pixel- and feature-level fusion of hyperspectral and LiDAR data for urban land-use classification," *Int. J. Remote Sens.*, vol. 36, no. 6, pp. 1618–1644, 2015.

[2] S. Luo *et al.*, "Fusion of airborne discrete-return LiDAR and hyperspectral data for land cover classification," *Remote Sens.*, vol. 8, no. 1, pp. 3–22, 2016.

[3] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[4] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[5] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.

[6] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and LiDAR data," *Int. J. Image Data Fusion*, vol. 6, no. 3, pp. 189–215, 2015.

[7] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, "Extinction profiles for the classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5631–5645, Oct. 2016.

[8] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6354–6365, Nov. 2017.

[9] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Feb. 2014.

[10] J. Bao, M. Chi, and J. A. Benediktsson, "Spectral derivative features for classification of hyperspectral remote sensing images: Experimental evaluation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 594–601, Apr. 2013.

[11] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[12] H. Li, P. Ghamisi, U. Soergel, and X. Zhu, "Hyperspectral and lidar fusion using deep three-stream convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, 2018, Art. no. 1649.

[13] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[14] Y. Gu and Q. Wang, "Discriminative graph-based fusion of HSI and LiDAR data for urban area classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 906–910, Jun. 2017.

[15] M. Zhang, P. Ghamisi, and W. Li, "Classification of hyperspectral and LiDAR data using extinction profiles with feature fusion," *Remote Sens. Lett.*, vol. 8, no. 10, pp. 957–966, 2017.

[16] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013.

[17] F. Jahan, J. Zhou, M. Awrangjeb, and Y. Gao, "Fusion of hyperspectral and LiDAR data using discriminant correlation analysis for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ Remote Sens.*, vol. 11, no. 10, pp. 3905–3917, Oct. 2018.

[18] J. Xia, N. Yokoya, and A. Iwasaki, "Fusion of hyperspectral and LiDAR data with a novel ensemble classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 957–961, Jun. 2018.

[19] B. Bigdeli, F. Samadzadegan, and P. Reinartz, "Fusion of hyperspectral and LiDAR data using decision template-based fuzzy multiple classifier system," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 38, pp. 309–320, 2015.

[20] Y. Zhong, Q. Cao, J. Zhao, A. Ma, B. Zhao, and L. Zhang, "Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and Lidar data," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 868.

[21] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 857–864.

[22] L. M. Kegelmeyer, P. W. Fong, S. M. Glenn, and J. A. Liebman, "Local area signal-to-noise ratio (LASNR) algorithm for image segmentation," in *Proc. Appl. Digit. Image Process.*, 2007, vol. 6696, pp. 66962–66970.

[23] K. K. Sharma and H. Krishna, "Asymptotic sampling distribution of inverse coefficient-of-variation and its applications," *IEEE Trans. Rel.*, vol. 43, no. 4, pp. 630–633, Dec. 1994.

[24] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 66–70.

[25] D. F. Contest, "Fusion of Hyperspectral and LiDAR Data Using Discriminant Correlation Analysis for Land Cover Classification," 2013. [Online]. Available: http://hyperspectral.ee.uh.edu/. Accessed on: Jun. 2, 2017.

[26] F. Jahan and M. Awrangjeb, "Pixel-based land cover classification by fusing hyperspectral and LiDAR data," *Int. Archives Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W7, pp. 711–718, 2017.

[27] A. M. Holm, D. G. Burnside, and A. A. Mitchel, "The development of a system for monitoring trend in range condition in the arid shrublands of western Australia," *Australian Rangeland J.*, vol. 9, no. 1, pp. 14–20, 1987.

[28] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[29] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[30] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[31] G. Lixin, X. Weixin, and P. Jihong, "Segmented minimum noise fraction transformation for efficient feature extraction of hyperspectral images," *Pattern Recognit.*, vol. 48, no. 10, pp. 3216–3226, 2015.

[32] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 153–160, 2004.

[33] X. He, D. Cai, S. Yan, and H. J. Zhang, "Neighborhood preserving embedding," in *Proc. 10th IEEE Int. Conf. Comput. Vis.*, 2005, vol. 2, pp. 1208–1213.

[34] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. New York, NY, USA: Elsevier, 1999.

Farah Jahan received the B.Sc.(Hons.) degree in computer science and engineering from the University of Chittagong, Chittagong, Bangladesh, in 2008, the M.E. degree in information and telecommunication engineering from Korea Aerospace University, Goyang, South Korea, in 2011 and the Ph.D. degree from Griffith University, Nathan, QLD, Australia, in 2019.

She is an Associate Professor with the Department of Computer Science and Engineering, University of Chittagong. Her research interests include pattern recognition, machine learning, computer vision, and data fusion in the field of remote sensing.

Jun Zhou (Senior Member, IEEE) received the B.S. degree in computer science and the B.E. degree in international business from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1998, respectively. He received the M.S. degree in computer science from Concordia University, Montreal, QC, Canada, in 2002, and the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2006.

He is an Associate Professor with the School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia. Previously, he had been a Research Fellow with the Research School of Computer Science, Australian National University, Canberra, ACT, Australia, and a Researcher with the Canberra Research Laboratory, NICTA, Canberra, ACT, Australia. His research interests include pattern recognition, computer vision, and spectral imaging with their applications to remote sensing and environmental informatics.

Mohammad Awrangjeb (Senior Member, IEEE) received the Ph.D. degree in 2009 from Monash University, Clayton, VIC, Australia.

Before joining Griffith University as a Lecturer, he was a Senior Research Fellow with the Federation University Australia, Monash University, and University of Melbourne. His research interests include automatic feature extraction and matching, multimedia security and image processing, automatic building extraction, and 3-D modeling from remote sensing data.

Dr. Awrangjeb received the Discovery Early Career Researcher Award by the Australian Research Council for the period 2012–2015.

Yongsheng Gao (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 1985 and 1988, respectively, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore in 2000.

He is currently a Professor with the School of Engineering, Griffith University, Nathan, QLD, Australia, and Director of Australian Research Council Research Hub for Driving Farming Productivity and Disease Prevention. He had been the Leader of Biosecurity Group, Queensland Research Laboratory, National ICT Australia (ARC Centre of Excellence), a Consultant of Panasonic Singapore Laboratories, and an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include smart farming, intelligent agriculture, biosecurity, face recognition, biometrics, image retrieval, computer vision, pattern recognition, environmental informatics, and medical imaging.