HOLP-DF: HOLP Based Screening Ultrahigh Dimensional Subfeatures in Deep Forest for Remote Sensing Image Classification

Alim Samat, *Member, IEEE*, Erzhu Li, Wei Wang, Sicong Liu, *Senior Member, IEEE*, and Ximing Liu

***Abstract*—To overcome the high intramodel dimensionality and low ensemble diversity issues, which limit the classification performance of original deep forest (DF), a new version of DF, the high-ordinary least square projection (HOLP) DF, was proposed in this article by introducing model-based HOLP feature screening (FS), random subspace propagation, and reduced error pruning techniques. To evaluate the performance of the proposed HOLP-DF, total eleven popular FS algorithms and total six advanced deep learning methods are selected. Experimental results on three widely acknowledged hyperspectral and PolSAR image classification benchmarks showed that: 1) HOLP is an optimal choice for FS in contrast with other screeners in terms of high classification accuracy and execution efficiency; 2) HOLP-DF is capable of obtaining better results than the original DF, DF with confidence screening and feature screening; 3) optimum sets of model depth, propaganda ratio and screening ratio parameters are 30, 40%, and 40%, respectively; 4) performance of HOLP-DF can be further boosted by extra usage of patch-based pooling and morphological profiling techniques.**

***Index Terms*—Deep forest (DF), feature screening, high-ordinary least square projection (HOLP), hyperspectral, image classification, PolSAR.**

I. Introduction

LAND cover classification has always been a research hotspot in the fields of remote sensing (RS) image processing and applications, which has received general attentions from many socio-economic and environmental application fields [1], [2], [3]. Among the many methods, machine learning (ML) based RS image classification has always been an active topic in the RS image processing and application communities, mainly due to their superior robustness and efficiency compared to conventional model-based approaches [4], [5], [6]. And in contrast with the conventional shallow ML methods, neural networks (NNs) based deep learning (DL) methods has gradually outperformed shallow ones and become the mainstream solution for the most RS image classification problems owing to their strong intramodel feature extraction ability, complex model structure, and plentiful parameters [5], [7], [8], [9], [10].

Even though remarkable advances have been achieved, still neither of those deep NNS (DNNs) based models can serve as the one solution to solve all problems. Particularly when considering the scenarios of learning from small-scale samples, predetermination of network topology structure, and the well-known difficulty of theoretical analysis of black-box features [11]. After many attempts of theoretical analysis, the learning mechanism of NNs based DL are not completely clear, but it has also been preliminarily proved that layer-to-layer processing, in-model feature representation and sufficient model complexity are the three basic principles that may underpin the success of DNNs [11], [12]. Based on this, a novel non-NN style DL model named multigrained cascade forest (gcForest), which realized by nondifferentiable modulus without backpropagation training for constructing deep forest (DF) was proposed by Zhou and Feng [12]. In contrast with NNs based DL models, gcForest comes with the appealing properties of the following:

1) easy to deploy and train with much fewer parameters;
2) can achieve high predictive accuracy on datasets across different domains by using almost the same setting of hyperparameters;
3) capable of capturing contextual or structure features;
4) the model complexity can be determined automatically in a data independent way which enabling gcForest (represented by DF thereafter) to perform well even on small-scale datasets [11], [12].

Despite the remarkable advantages previously that have been proven from wide range of applications, original DF is also limited by the high time cost and memory requirement with owes much to the aspect of: 1) cascade structure of DF leading to a linear increase in price of time complexity as the numbers of level increases; and 2) multigrained scanning procedure significantly increase the number of training instances not only, also produces a high-dimensional input for the cascade

Manuscript received 5 July 2022; revised 7 September 2022; accepted 11 September 2022. Date of publication 22 September 2022; date of current version 30 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071424 and Grant 41801327 and in part by the Youth Innovation Promotion Association Foundation of the Chinese Academy of Sciences under Grant 2018476. *(Corresponding author: Alim Samat.)*

Alim Samat, Wei Wang, and Ximing Liu are with the State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China, also with the Research Center for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi 830011, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: alim_smt@ms.xjb.ac.cn; wangwei177@mails.ucas.ac.cn; liuximing20@mails.ucas.ac.cn).

Erzhu Li is with the Department of Geographical Information Science, Jiangsu Normal University, Xuzhou 221100, China (e-mail: lierzhu2008@126.com).

Sicong Liu is with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China (e-mail: sicong.liu@tongji.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3206886

procedure [11]. To address these issues, gcForest has been improved by introducing techniques like confidence screening (CS) [13], feature screening (FS) [14], feature pooling and error screening [15], patch-based pooling, morphological profiling, and pseudolabeling (PL) [16]. However, these solutions were conducted on sufficient sample training sets, and may not hold for small samples (*n*) with large numbers of features (*p*) training set scenarios, the so-called large-*p*-small-*n* problem in statistics. Because the classical ordinary least-squares (OLS) estimate used for linear regression is no longer applicable due to insufficient degrees of freedom. Additionally, the proven superior performances of patch based pooling and morphological profiling techniques could be further boosted and generalized by interacting with state-of-the-art FS techniques those designed for high and ultrahigh dimensional FS problems with sparsity.

Thanks to the rapid advances in earth observation technology, now it has brought us an unprecedented array of large, enrich, diverse and complex RS data. In the meantime, challenging issues from dimensionality that exists in statistics, ML, RS data processing, storage, and access arose more severely in this era of big data. Although, an explosion of developing approaches for handling large data sets with high dimensionality have been witnessed in recent years, the common assumption underlying these approaches, variables that affect the response is relatively small, may not hold for large-*p*-small-*n* problem, and the computation cost for large-scale LASSO-based optimization becomes a serious concern as well [17], [18]. Additionally, current common feature selection and extraction methods may also not work well due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability, which often requires sophisticated estimation techniques, strong model assumptions, and advanced computing algorithms [19], [20]. For example, traditional variable selection methods like Akaike information criterion, Bayesian information criterion (BIC), and extended BIC always involve a combinatorial NP-hard optimization problem with computational time increasing exponentially with the data dimensionality. Hence, it is truly desirable if one can rapidly reduce the ultrahigh dimensionality before conducting a refined analysis. And a practical approach is to use a screening procedure to reduce the dimension of feature space to a moderate scale, and then apply variable selection methods in the second phase.

In contrast with the consistent variable selection, FS (also known as variable screening) deals with a much less ambitious goal of sure screening could be achieved by using some both conceptually and computationally simple method [22]. Thereby, marginal FS becomes indispensable for linear, generalized linear, and robust linear models on ultrahigh dimensional data and has received much attention, especially since the seminal work of sure independent screening (SIS) method was proposed by Fan and Lv for linear regression [22], [23], [24]. However, all these approaches based on sure screening protocol require the specification of a particular model structure, which is usually an impossible task under the ultrahigh dimensional setting. Thus, model-free FS methods are naturally more appealing. Toward this direction, Ball correction SIS (BcorSIS) [25], distance correlation SIS (DC-SIS) [26], sure independence ranking and screening (SIRS) [27], mean-variance SIS (MVSIS) [28], martingale difference correlation SIS (MDCSIS) [29], and Henze–Zirkler SIS [20] are the most undertaken ones in diverse studies.

Although the model-free FS methods are capable of avoiding the impossible task of a particular model structure specification, current methods are still based on some assumptions for the predictor and response variables on the one hand. For example, correlation metric-based SIRS and DCSIS are not robust to the predictors whose distributions are heavy tail, and DC-SIS requires both the predictors and response variables to satisfy the subexponential tail probability uniformly. On the other hand, model-free FS methods cannot simultaneously satisfy the following two important demands in designing a screening operator: 1) straightforward and efficient to compute; and 2) the resulting estimator must possess the sure screening property under reasonable assumption to assure that the most discriminate subfeatures are remained. In this sense, model-based screeners like high-ordinary least square projection (HOLP) [18] and sparsity-restricted maximum likelihood estimator (SMLE) [30], which possesses the sure screening property, gives consistent variables selection without strong marginal correlation assumption and computationally efficient, still will be practically appealing. And in contrast with the latter one which is in the context of ultrahigh dimensional generalized linear models, HOLP is more simple, easy to implement, and more flexible for both linear and generalized linear models with Ridge-regression. Thus, it is of great interest to adopt HOLP to solve the aforementioned issue of dimensionality in DF via screening out the most irrelevant features in the cascade structure construction phase.

Essentially being as novel decision trees based EL method, multigrained cascade structure based in-model feature representation and layer-to-layer processing enable DF with high predictive accuracy that competitive to diverse DNNs in wide range of tasks [11], [12], [13], [14], [15], [16]. However, this property not only limit the original DF by high time cost and memory requirement, as discussed earlier, but also can limit the predictive accuracy by using layer-to-layer processing with lower diversity and average fusion strategy by underfitting and/or overfitting. This is mainly due to the fact that original DF and its variants simply passing the original input features for concatenate with features from next layers [11], [13], [14], [15], [16]. In this case, the variance of the input will typically get smaller as learners get better and better at predicting the output and the remaining errors become increasingly difficult to correct. As a result, this multi-co-linearity can significantly limit the ability of the ensemble to improve upon the best score of the subsequent layer as there is too little variation in predictions for the ensemble to learn useful combinations. And this is true in small ensemble size with high classifier diversity and large ensemble size with low classifier diversity scenarios that particularly use average and majority voting fusion strategies [4], [31]. Noteworthy, there are only four learners, two random forest and two ExtraTrees, to enhance the diversities in the cascade layers of DF and have equal contributions to produce the final prediction. This is why there was no obvious improvement in accuracy and the increased depth of DF being larger than ten layers [16]. Moreover, this nondate-driven manual hard-definition of diversity from forests may raise the risk of overfitting and/or underfitting on small-scale or class-imbalanced data [32].

Diversity is the first key component of constructing an effective EL system, usually can be achieved by applying specific techniques on sample, label, feature, and model parameters in separate or hybrid ways [33], [34]. For the techniques from feature space, the most popular approaches are random rotation [35], regularized random rotation [36], random projection [37], random partition [38], and random subspace [39]. Theoretically, any of these approaches can be adopted to increase variations between propagated features from the original input and/or earlier layers in DF. However, compared with the random subspace approach, the most other existing works do not have strong theories support [40]. Most importantly, the random subspace method leverages the idea that neighborhoods of feature space have a specific local structure via sticking to the original features, and the minimal discriminative subsets could appear in some of the subspaces. This property can be very powerful when the local structure or the discriminative subset first needs to be extracted, before an estimator learning to generalize for features ranking, screening, and classification [39], [40]. In fact, most ML algorithms with convex optimization objectives are ill-equipped to solve the problem of multimodal probability distribution estimation from all feature spaces. The random subspace method can overcome this issue by allowing base estimators to fit one mode of the distribution at a time. Thus, random subspace based feature propagation technique is selected to increase the variations from features before screening to the next layers of our modified version of DF.

However, propagating and screening all generated random subspaces may not be a wise idea in practice. Especially in the sparse classification case as many random subspaces could contain low discriminative and even corrupted signals, and adopting an average voting-based fusion strategy in the original DF. Furthermore, the unnecessarily large random subspaces-based ensemble can lead to extra memory usage, computational costs, and may occasionally degrade the generalized performance [41]. A straightforward way to alleviate these shortcomings is the selection of a fraction of the base learners before combination, which is commonly called as ensemble pruning, ensemble selection, ensemble thinning, and selective ensemble [42], [43]. Some theoretical and empirical evidences have also shown that performance of an ensemble consisting of small ensembles could better than all [44], [45]. Motivated by these reasons, many ensemble pruning algorithms have been proposed in the last decades. However, while those algorithms report to greedy search not have theoretical or empirical quality guarantee, those based on swarm intelligence optimization algorithms are sensitive to noise, and those stand on Bayesian probabilistic distribution are usually require advanced computation techniques and may not applicable for ensemble pruning [41], [43]. Hence, due to the straightforward, effective, efficient, and easy to implement reasons, we select the reduced error pruning (REP) [46] before fusion in the proposed DF algorithm.

The main contributions of this article are summarized as follows.

1) A new version of DF was proposed for pixelwise RS image classification by adoption of HOLP based feature screening, random subspace propagation, and reduced error running techniques.
2) Total of twelve popular feature screening methods for high and ultrahigh dimensional settings were studied to solve the issues of memory requirement from the original DF.
3) Optimum choices for random subspace propagation and screening ratios were recommended for the proposed HOLP-DF algorithm.

II. Related Work

In ultrahigh dimensional setting, SIS procedure was first introduced to significantly reduce the dimensionally by strongly rely on the assumption that the valuable features in the data have large marginal correlations with the response. But this assumption is often violated in reality, as predictors are often correlated. In further, valuable features that are jointly correlated to the response can be screened out simply because they are marginally uncorrelated to the response [18]. Nevertheless, as a seminal work for FS, SIS is still appealing in practice due to its sure screening, straightforward, and computationally efficient properties. And more appealing solutions might be reachable by loosening the restrictive marginal correlation assumption based on the OLS estimator and the Ridge regression, which are the HOLP and Ridge-HOLP [18].

Consider the familiar linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{1}$$

where $\mathbf{X} \in R^{n\times p}$ is the design matrix composed of number of n samples with p variables, $\mathbf{Y} \in R^n$ is the response vector, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a p-vector of parameters, and $\varepsilon \in R^n$consists of independently identical distribution errors with ε_i follows a distribution with zero mean and variance σ^2. A general class of linear estimates of $\boldsymbol{\beta}$ can be formed as

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \mathbf{AY} = \mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \varepsilon)\\ &= (\mathbf{AX})\boldsymbol{\beta} + \mathbf{A}\varepsilon\end{aligned} \tag{2}$$

where $\mathbf{A} \in R^{n\times p}$ maps the response to an estimate and the SIS set $\mathbf{A} = \mathbf{X}^{\perp}$ [24], $\perp$represents the matrix transpose, $\mathbf{A}\varepsilon$consist of linear combinations of zero mean random noises, and $(\mathbf{AX})\boldsymbol{\beta}$ is the signal. In order to preserve the signal part as much as possible, an ideal choice of $\mathbf{A}$ is it should satisfy $\mathbf{AX} = \mathbf{I}$. And if this choice is possible, the signal part would dominate the noise part $\mathbf{A}\varepsilon$under suitable conditions, which leads naturally to the OLS estimate where $\mathbf{A} = (\mathbf{X}^{\perp}\mathbf{X})^{-1}\mathbf{X}^{\perp}$ only if $p < n$. However, when p is larger than n, $\mathbf{X}^{\perp}\mathbf{X}$ is degenerate and $\mathbf{AX}$ cannot be an identity matrix $\mathbf{I}$. Fortunately, $(\mathbf{X}^{\perp}\mathbf{X})^{-1}\mathbf{X}^{\perp}$ and $\mathbf{X}^{\perp}(\mathbf{XX}^{\perp})^{-1}$ can be seen as the Moore–Penrose inverse of X for $p < n$ and $p > n$, respectively [18]. In $p > n$ case, nevertheless the $\mathbf{AX} = \mathbf{X}^{\perp}(\mathbf{XX}^{\perp})^{-1}\mathbf{X}$ is no longer an identity matrix, $\tilde{\beta}_i(i \notin S)$ can take advantage of the large diagonal terms of $\mathbf{AX}$ to dominate $\tilde{\beta}_i(i \notin S)$ that is just a linear combination of OFF-diagonal terms, as long as $\mathbf{AX}$ is diagonally dominant. Where $S = \{j : \beta_j \neq 0, j = 1, 2, \ldots, p\}$ is the index set of the nonzero β_j's with cardinality $s = |S|$ from the true model $\mathbf{M}_S \in \mathbf{M} = \{x_1, x_2, \ldots, x_p\}$. By rewriting $\mathbf{X}$ via singular value decomposition as $\mathbf{X} = \mathbf{VDU}^{\perp}$, where $\mathbf{V}$ is an $n \times n$ orthogonal

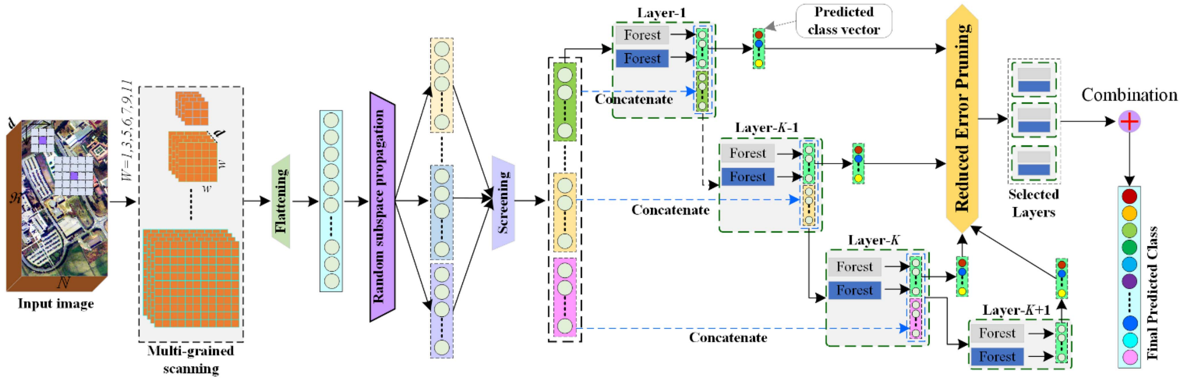


Fig. 1. Architecture of the HOLP-DF.

matrix, $\mathbf{D}$ is an $n \times n$ diagonal matrix, and $\mathbf{U}$ is an $p \times n$ matrix that belongs to the Stiefel manifold $\mathbf{V}_{n,p}$ [47]. Then by

$$\mathbf{X}^{\perp}(\mathbf{X}\mathbf{X}^{\perp})^{-1}\mathbf{X} = \mathbf{U}\mathbf{U}^{\perp}$$
$$\mathbf{X}^{\perp}\mathbf{X} = \mathbf{U}\mathbf{D}^{2}\mathbf{U}^{\perp}. \tag{3}$$

It can be proven that $\mathbf{A}\mathbf{X} = \mathbf{X}^{T}(\mathbf{X}\mathbf{X}^{T})^{-1}\mathbf{X}$ can reduces the impact from the high correlation of $\mathbf{X}$ by removing the random diagonal matrix $\mathbf{D}$, will be diagonal dominating with overwhelming probability as well. In this regard, a very simple variable screening method can be obtained by rewriting the

$$\mathbf{M}_{\gamma} = \{x_j : |\tilde{\beta}_j| \geq \gamma\}$$
$$\tilde{\boldsymbol{\beta}} = \mathbf{X}^{\perp}(\mathbf{X}\mathbf{X}^{\perp})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{X}^{\perp}(\mathbf{X}\mathbf{X}^{\perp})^{-1}\varepsilon \tag{4}$$

where $\mathbf{M}_{\gamma}$ is a submodel of full model $\mathbf{M}$, γ is a user specified threshold. This estimator also named as the HOLP, where the first term indicates that it can be seen as a projection $\boldsymbol{\beta}$. And from the standpoint of matrix $\mathbf{X}\mathbf{X}^{\perp}$ is of full rank whenever $p > n$, HOLP is unique to high and ultrahigh dimensional data analysis. Furthermore, HOLP is easy to implement and can be efficiently computed with the complexity of $O(n^2 p)$, and it is scale invariance in the signal part $\mathbf{X}^{\perp}(\mathbf{X}\mathbf{X}^{\perp})^{-1}\mathbf{X}\boldsymbol{\beta}$ [18].

III. Proposed Method

As a novel non-NNs based DL model, original DF has gained increasing attentions in recent years. Although its remarkable performances have been proven from wide range of studies and applications, also many modified versions have been proposed to overcome the high time cost and memory requirement limitations, which owes much to the aspect of cascade structure and multigrained scanning procedure, limited classification performance caused by high intramodel dimensionality and low ensemble diversity is still remains open. To solve this issue and also enlightened by the earlier works of [13], [14], and [15], an improved version of DF, the HOLP-DF, was proposed by adoption of HOLP based FS algorithm to solve the issue of high intramodel dimensionality on the one hand, adopting random subspace propagation, and REP techniques to increase the ensemble diversity on the other hand.

The architecture of the proposed HOLP-DF for RS image classification method is shown in Fig. 1, which consist of three major steps of: 1) contextual features extraction with multigrained scanning (overlapped image patching); 2) HOPL based screening the propagated random subfeatures of flattened features from step 1; and 3) layer-by-layer training DFs with screened features by following the basic structure, as shown in Fig. 1, but using a different features concatenation strategy and using REP ensemble selection strategy before fusion.

For a given RS image $\overset{\leftrightarrow}{\mathbf{I}}^{\Re\times\mathbb{N}\times d'}$, where $\Re$, $\mathbb{N}$, and d' represents the numbers of rows, columns, and channels, respectively, we first obtain the overlapped neighboring image patches $\mathbf{P} = \{\mathbf{P}_{r=1}^{w\times w\times d'}, \mathbf{P}_{r=2}^{w\times w\times d'}, \dots, \mathbf{P}_{r=\Re * N}^{w\times w\times d'}\}_{r=1}^{\Re\times N}$ with the specified patch size of *w*. Then flattening procedure will be executed on image patches $\mathbf{P}$ to obtain vectorized image $\overset{\leftrightarrow}{\mathbf{I}}^{*} = \{\mathbf{P}_{i}^{1\times w\times w\times d'}\}_{i=1}^{\Re\times\mathbb{N}}$ for random subspace propagation to the next level for screening by using HOLP according to (4). Let δ is the random subspace propagation ratio, φ is the screening ratio using HOLP, K is the number of total random subspaces, K independent random subspaces after screening are generated as $\{\overset{\leftrightarrow}{\mathrm{I}}_{rs1}^{*}\}_{\Re\times\mathbb{N}}^{d}, \{\overset{\leftrightarrow}{\mathrm{I}}_{rs2}^{*}\}_{\Re\times\mathbb{N}}^{d}, \dots, \{\overset{\leftrightarrow}{\mathrm{I}}_{rsK}^{*}\}_{\Re\times\mathbb{N}}^{d}$, where the dimensionality determined by $d = \lfloor K\delta\varphi \rfloor$, and for any subspace we have $\forall \overset{\leftrightarrow}{\mathrm{I}}_{rs}^{*} \subseteq \overset{\leftrightarrow}{\mathbf{I}}^{*}$. Subsequently, number of K general gcForest classifiers are obtained, and normally we can aggregate the outputs of classifiers to form the decision function by taking a simple average via

$$G(x) = \frac{1}{K+1}\sum_{k=1}^{K+1} g(x_{rs}^{k})$$
$$= \frac{1}{K+1}\sum_{k=1}^{K+1} \underset{c\in\{1,\dots,C\}}{\arg\max} \left[f_T(x_{rs}^{k})\right]_c \tag{5}$$

where $g(x_{rs}^{k})$ is the prediction function of original gcForest model using data x_{rs}^{k}, x_{rs}^{k} is the training set from the *k*th random subspace after HOLP based screening of original training set x, C is the number of class labels, $[f_T(x_{rs}^{k})]_c$ is the cth element of the label vector $f_T(x_{rs}^{k})$. At level $t \in \{1, \dots, T\}$, f_t is the cascade of elements of forests $\mathbf{f} = \{f_i, \dots, f_T\}$ up to level t. Notably, when stacking several layers of forests in original gcForest, the variance of input $[x, f_{t-1}(x)]$ will typically get smaller as the next layers of forests get better and better at predicting the output and the remaining errors become increasingly difficult to correct for. In result, ability of the ensemble to improve upon the best score of the subsequent layers can be significantly limited because there is too little variation in predictions for the ensemble to learn useful combinations. One way to increase this variation is to propagate features from the subset of original input and/or earlier layers. In this sense, the cascade of elements of forests $\mathbf{f} = \{f_i, \ldots, f_T\}$ up to level t that defined in [14] will be rewritten as

$$f_t(x_{rs}^k) = \begin{cases} h_1(x_{rs}^k) & t = 1 \\ h_t([x_{rs}^k, \forall x_{rs}^{\neq k}, f_{t-1}(x_{rs}^k)]) & t > 1 \end{cases} \tag{6}$$

where $[x_{rs}^k, \forall x_{rs}^{\neq k}, f_{t-1}(x_{rs}^k)]$ is the input of the ensemble of forest h_t at level t.

Propagating portions of features at random could enhance the ability of ensemble by increasing diversities from the variations on feature space. Meanwhile, it could also limit and even degrade the generalized ability of ensemble in sparse classification and small ensemble cases, could also limited the performances on memory usage, computational cost, and occasionally degrade the generalized ability in unnecessarily large random subspaces scenarios. To solve aforementioned issues, we can use the REP technique, which was inspired by the decision tree pruning algorithm with the goal of choosing the set of Ψ^* classifiers that give the best voted performance on the pruning set [54]. In general, REP use a sophisticated search method called back-fitting as follows:

1) initialize the set of classifiers Ψ to contain the one classifier h_t^1 that has the lowest error on the pruning set;
2) add the classifier h_t^2 such that the voted combination h_t^1 and h_t^2 has the lowest pruning set error;
3) adds the classifier $h_t^{k\subseteq K, k\neq 1,2}$ such that the voted combination of all classifiers in Ψ has the lowest pruning set error;
4) revisit earlier decisions and deleting previously chosen classifiers and replacing them with best classifier continues until none of the classifiers changes or reaches the number of iterations.

Then, the objective function presented in (5) can be rewritten as

$$\begin{aligned} G(x)^* &= \frac{1}{\Psi^*} \sum_{\Psi=1}^{\Psi^*} g(x_{rs}^{\Psi}) \\ &= \frac{1}{\Psi} \sum_{\psi=1}^{\Psi^*} \arg\max_{c\in\{1,\ldots,C\}} \left[f_T(x_{rs}^{\Psi})\right]_c \end{aligned} \tag{7}$$

where $\Psi^* \subseteq \{f_T(x_{rs}^k)\}_{k=1}^{K+1}$ and number of classifiers in Ψ^* is much smaller than K.

IV. Datasets and Setup

A. Datasets

1) Pavia University: This data was captured over the Engineering School, University of Pavia, Pavia, Italy, by the reflective optics spectrographic image system (ROSIS) sensor, which provides 103 spectral channels with a spectral coverage ranging from 0.43–0.86 μm and with the spatial resolution of 1.3 m. Color images shown in Fig. 3(a) has 610 × 340 pixels size and the validation data refer to 9 land cover classes are shown in Table I with details about the number of samples and the legends.

TABLE I
Class Name, Color, and Samples' Details for Considered Datasets

Data	No.	Name & color	Test	Training	Data	No.	Name & color	Test	Training
A	1	Asphalt	6631	200	B	9	Peas	96473	32160
	2	Meadows	18649	200		10	Grass	107707	35905
	3	Gravel	2099	200		11	Water	118941	39650
	4	Trees	3064	200	C	1	Healthy grass	1053	198
	5	Metal	1345	200		2	Stressed grass	1064	190
	6	Baresoil	5029	200		3	Synthetic grass	505	192
	7	Bitumen	1330	200		4	Trees	1056	188
	8	Bricks	3682	200		5	Soil	1056	186
	9	Shadows	947	200		6	Water	143	182
B	1	Stem beans	6601	2200		7	Residential	1072	196
	2	Forest	17835	5945		8	Commercial	1046	191
	3	Potatoes	29069	9690		9	Road	1053	193
	4	Lucerne	40303	13435		10	Highway	1036	191
	5	Wheat	51537	17180		11	Railway	1050	181
	6	Bare soil	62771	20925		12	Parking Lot 1	1041	192
	7	Beet	74005	24670		13	Parking Lot 2	285	184
	8	Rape seed	85239	28415		14	Tennis Court	247	181
						15	Running Track	473	187

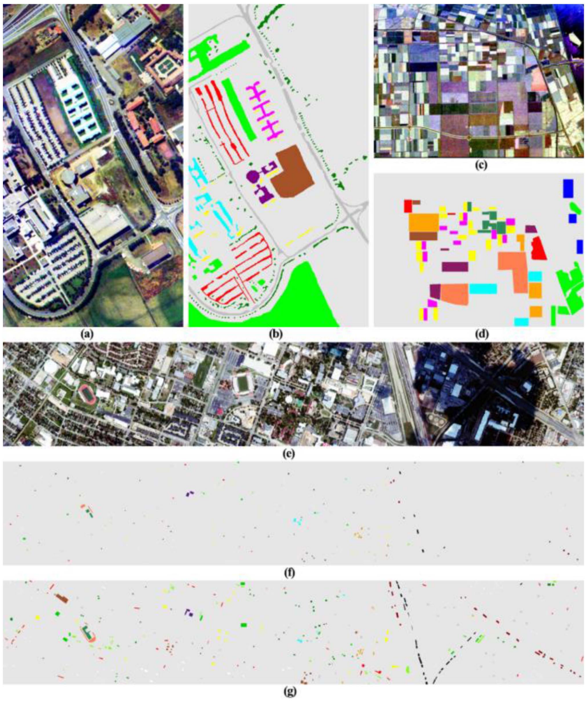
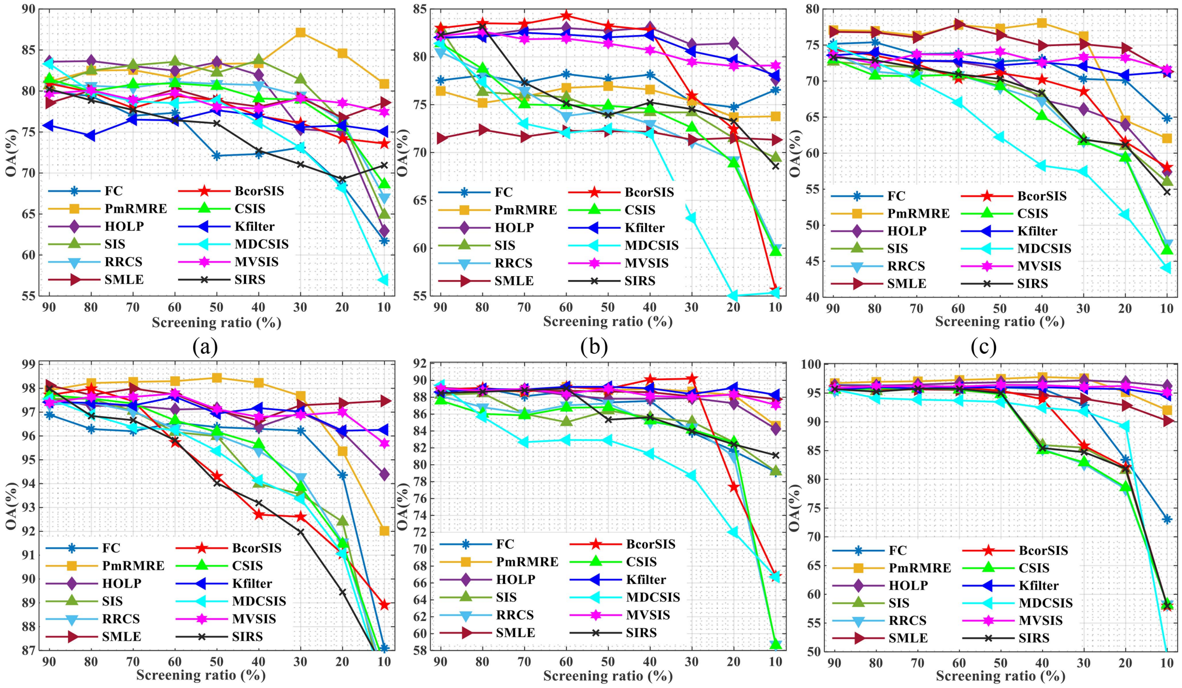


Fig. 2. Test images (a), (c), (e) with ground truth maps (b), (d), (f), (g).

2) AirSAR Flevoland: This data was obtained by the Airborne Synthetic Aperture Radar (AirSAR) over the Flevoland region (The Netherlands) in 1989. As part of National Aeronautics and Space Administration Earth Science Enterprise project, AirSAR was designed and built by the Jet Propulsion Laboratory and operating in full polarimetric mode L-band. The scene shown in Fig. 2(c) was extracted from the SIR-C education program, it has spatial resolution of 6.60 m in the slant range direction and 12.10 m in the azimuth direction with the size of 705 rows × 1024 columns size, and covers a large agricultural area of flat topography and homogeneous soils. The ground truth map shown in Fig. 2(d) refer to 11 land cover classes are shown



(a) (b) (c)

(d) (e) (f)

Fig. 3. OA values versus screening ratio from DF with various screeners for ROSIS Pavia (a), (d), DFC2013 Houston (b), (e) and AirSAR Flevoland (c), (f) datasets with 10 samples per class (row 1) and all in Table I (row 2).

also shown in Table I with details about the number of samples and the legends.

3) DFC2013 Houston: This data were acquired by the NSF-funded Center for Airborne Laser Mapping over the University of Houston campus and the neighboring urban area, on June 23, 2012. The hyperspectral consists of 144 spectral bands in the 0.380 μm to 1.05 μm region and has been calibrated to at-sensor spectral radiance units. The RGB image shown in Fig. 2(e) has 340 × 1900 pixels size with the spatial resolution of 2.5 m. Total 15 classes of interest are presented in ground truth maps [see Fig. 2(f) and (g)] whereas Table I reports the corresponding number of samples for both the training and validation sets.

B. Experimental Setup

To comparatively investigate the performance of HOLP screener on handling high and ultrahigh dimensionality issue of original DF in RS image classification, popular screeners including fusion clustering (FC) [48], SIS [24], conditional SIS (CSIS) [49], MDCSIS [29], SMLE [30], BcorSIS [33], Kolmogorov filter (Kfilter) [50], SIRS [27], robust rank correlation screening (RRCS) [23], MVSIS [28], and feature selection method parallelized minimum redundancy maximum relevance ensemble (PmRMR) [51] were considered.

As for the features, while six upper OFF-diagonal features of coherence matrix T3 stacked with Span feature from AirSAR Flevoland data was used, the first 10 principal component, which contain the most information at much lower volume size, were selected for Pavia University and DFC2013 Houston high dimensional hyperspectral datasets for all considered methods to avoid the out of memory issues in the running of experiment from high intermodel feature dimensional. Because even for 10 principal components, dimensionality of the original input features will be at $10 \times (1 + 3 \times 3 + 5 \times 5 + 7 \times 7 + 9 \times 9 + 11 \times 11) = 2860$ at the first round of flattening procedure.

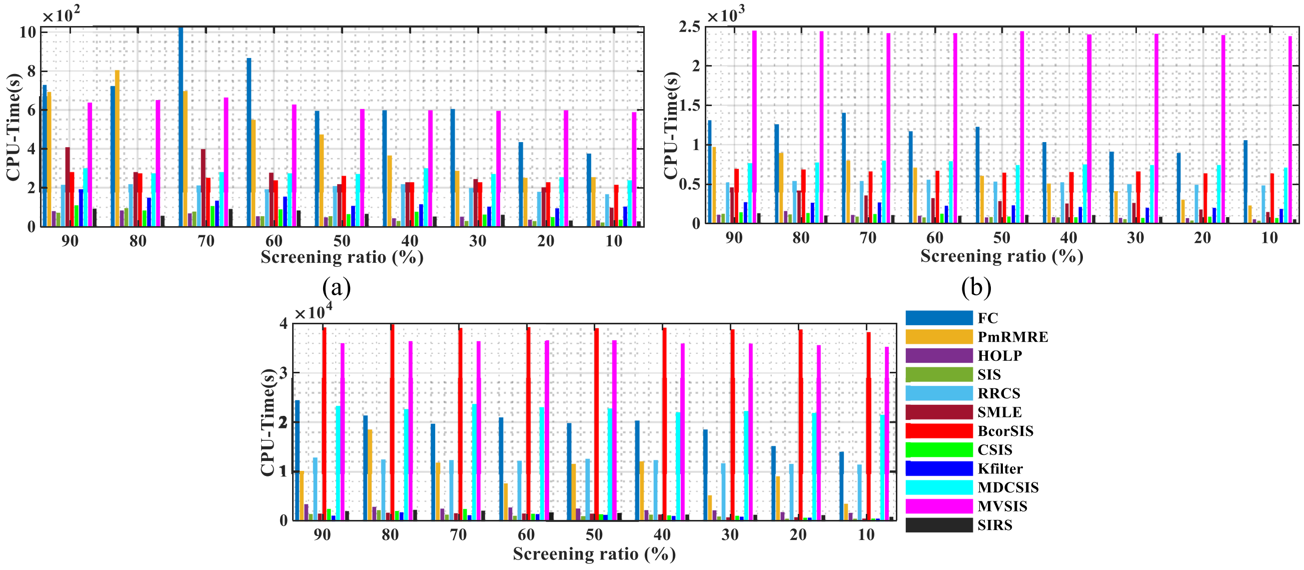
In all experiments, the overall accuracy (OA), average accuracy (AA), kappa coefficient (Ka), algorithm running time in seconds are used to evaluate the classification performances of the adopted classifiers. All the experiments are conducted by using Python 3.9.7 and PyTorch 1.9.0 installed on a machine with 64-bit Windows 10 system use an Intel (R) Core (TM) i7-7820X 3.60-GHz CPU and 128 GB RAM, and with an NVIDIA Quadro RTX8000 GPU card equipped with 4608 CUDA parallel processing cores and 48 GB of RAM memory.

V. Results and Analysis

A. Results of DF With Various Screeners

In this section, we first investigate the performance of the adopted feature screeners to handling high and ultrahigh dimensionality issue of original DF in RS image classification. To make a comparison among the different sample size and multiscreening ratios, we report the OA and the time costs in seconds from model training by using different sample sets presented in Figs. 3 and 4, respectively.

From the graphs in Fig. 3, various OA values can be observed for considered feature screeners with different sample size, screening ratios, and datasets setting. In small sample (10 samples per class) setting scenario that shown by the graphs in the first row of Fig. 3, HOLP and SIS are better than others on the Pavia University data, BcorSIS, Kfilter, MVSIS, and HOLP are better than others on the DFC2013 data, and SMLE is better



(a)

(b)

(c)

Fig. 4. Model training time in seconds versus screening ratio from DF with various screeners ROSIS Pavia (a), DFC2013 Houston (b), and AirSAR Flevoland (c) datasets using all training samples.

than others on the AirSAR Flevoland test data, but the SMLE screener almost shows the worst results on DFC2013 data, as shown in Fig. 3(b). Moreover, rapid decreasing trend of OA values from the most screeners are obtained as the screening ratio is smaller than 40% for all three datasets. In using all sample training scenarios that shown by the graphs in the second row of Fig. 3, OA values of SMLE, MVSIS, HOLP, and Kfilter are always higher than OA values from FC, SIS, RRCS, BcorSIS, CSIS, MDCSIS, and SRIS screeners on all considered datasets. Additionally, the rapid decreasing trend of OA values from SMLE, MVSIS, HOLP, and Kfilter screener are obtained as the screening ratio is smaller than 30% for all three datasets. And in the most cases, decreasing ratios of MDCSIS, CSIS, SIRS are much faster than others along with increasing values of screening ratios, see the learning curves in cyan lines with left triangles, green lines with upper triangles, and black lines with cross, respectively.

According to the bar charts shown in Fig. 4, we can easily see that first the lowest computational efficiency is shown by FC and MVSIS screeners on Pavia University data, by MVSIS screener on DFC2013 Houston data, and by BcorSIS and MVSIS screeners on AirSAR Flevoland data. While the secondary low computational efficiencies are shown by MDCSIS, RRCS, and SMLE screeners and PmRMRE feature selector, the highest computational efficiencies are shown by screeners including SIS, Kfilter, HOLP, CSIS, and SIRS in the most cases. Moreover, screening ratio set does not have obvious influences on the computational efficiency of the considered feature screeners. Summing-up with the findings from the Fig. 3 in previous paragraph, it can be concluded that HOLP screener is the optimum one to handling the high and ultrahigh dimensionality issue of original DF in RS image classification task, from both high OA values and highly efficient point of view in contrast with the other considered the screeners.

B. *Parameters Analysis for HOLP-DF*

According to the methodology details presented in previous Section III, and the working mechanism of conventional DF, subspace feature propagation ratio, feature screening ratio, sample size, and model depth are the critical parameters could influence the performance of the proposed HOLP-DF. Hence, we show the OA and model training time values versus the propagation ratio and screening ratio of HOLP-DF using all samples from considered datasets in Fig. 5 with the depth of 50 layers.

From the results shown in Fig. 5, it can be clearly seen that both propagation ratio and screening ratio values have considerable influences on both classification accuracy and model training efficiency. In further, influence from the propagation ratio on OA values is more critical than influence from the screening ratio. For example, there not exits obvious changes in OA values by increasing screening ratio values after the propagation ratio is higher than 40%, especially on the Pavia University and AirSAR Flevoland datasets, as shown in Fig. 5(a) and (c). On the contrary, influence from the propagation ratio and screening ratio is almost the same on computational complexity, and an optimum choice of screening ratio is less than 50% for efficiently model training. Summing-up with the previous findings, combination of 40% of propagation ratio with 40% of screening ratio is recommended for the proposed HOLP-DF from high classification accuracy, high model training efficiency, and small size of subspace data volume (only occupy 40% × 40% = 16% of original data volume) points of view.

In the Fig. 6, we present the OA values versus the number of samples for each class and number of layers (depth) of the considered DFs with 40% of propagation ratio with 40% of screening ratio set as recommended earlier. To comprehensively evaluate the performance of the proposed HOLP-DF, the original DF and DF with patch-based pooling [DF(PP)], morphological

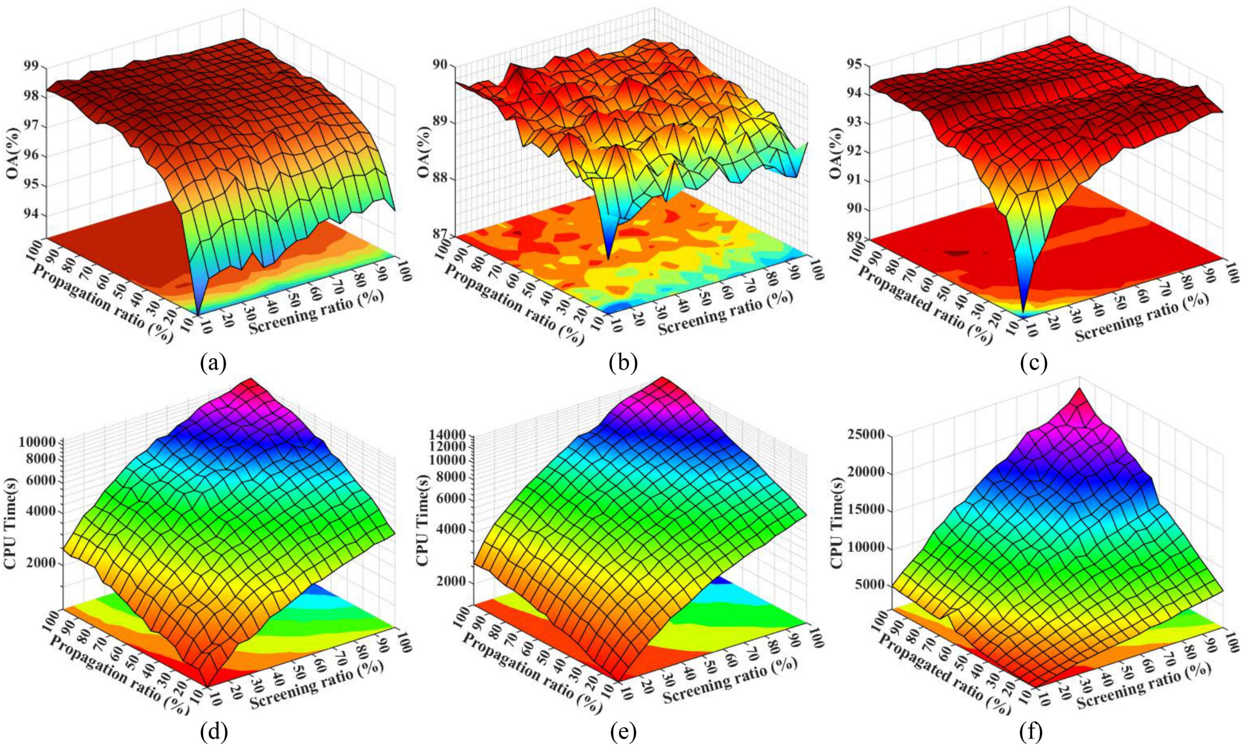


Fig. 5. OA (row 1) and CPU-Time (row 2) values versus screening and propagation ratios from HOLP-DF on ROSIS Pavia (a), (d), DFC2013 Houston (b), (e), and AirSAR Flevoland (c), (f) datasets.

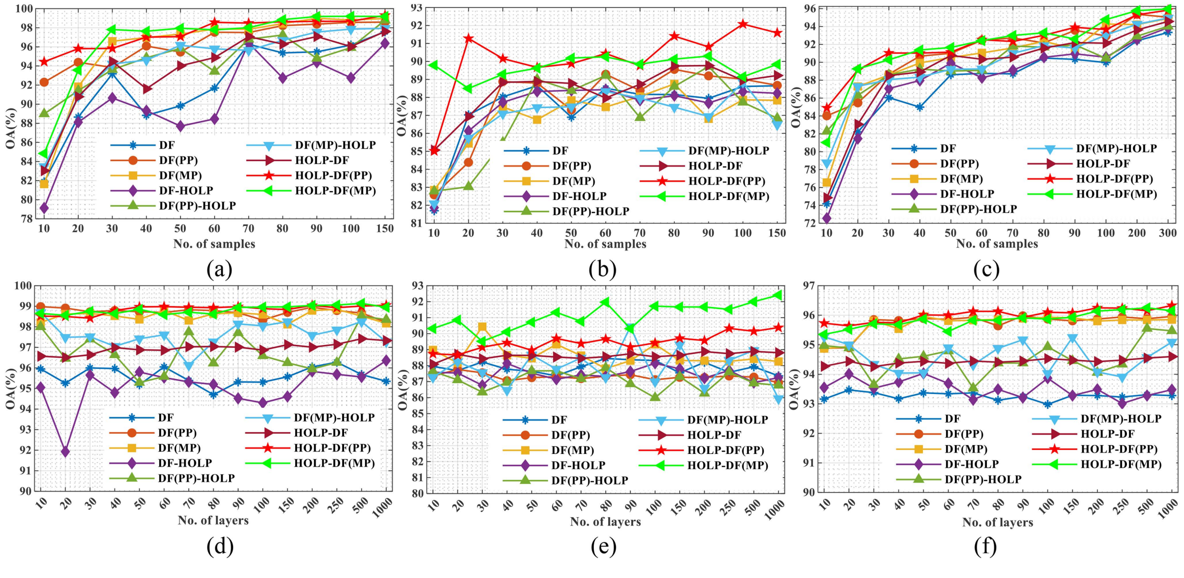


Fig. 6. OA curves versus no. of samples for per class (row 1) and layers (row 2) from considered methods on ROSIS Pavia (a), (d), DFC2013 Houston (b), (e), and AirSAR Flevoland (c), (f) datasets.

profiling [DF(MP)], and HOLP based screening with random subspace propagation (DF-HOLP) techniques are considered. To further investigate the performance of HOLP-DF using PP and MP technologies, experiments of HOLP-DF with PP [HOLP-DF(PP)] and MP [HOLP-DF(MP)] are also considered.

Based on the results shown in the first row of Fig. 6, it can be clearly seen that as follows.

1) OA curves from the proposed HOLP-DF is always stays upper place than OA curves from original DF.
2) OA curves from HOLP-DF with PP [HOLP-DF(PP)] and MP [HOLP-DF(MP)] are always stay at the highest place using fewer samples, see the green lines marked with left triangle and red lines marked with star.
3) OA curves from DFs with subspace propagation are always stay higher place than their counterparts of without using subspace propagation, see light blue lines marked with down triangle (DF(MP)-HOLP) versus green lines marked with left triangle [HOLP-DF(MP)], light green

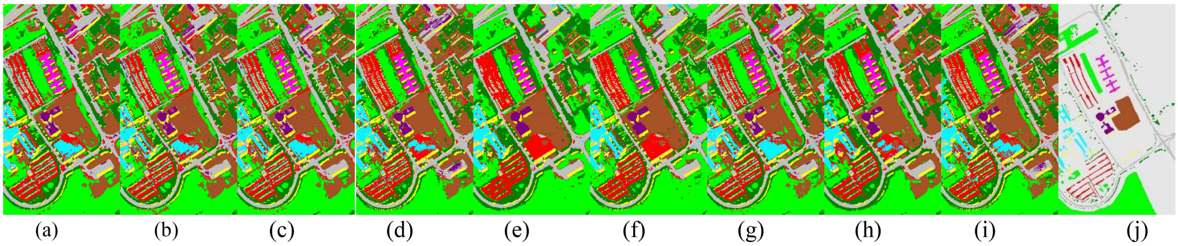


Fig. 7. Classification maps with OA values corresponding to the underlined numbers from Table II (a)–(i) and ground-truth map (j) for Pavia University data. (a) 98.40%. (b) 96.03%. (c) 97.77%. (d) 99.13%. (e) 99.78%. (f) 99.83%. (g) 98.01%. (h) 99.40%. (i) 99.23%.

TABLE II
OA, AA, and KAPPA Values for the Adoptee Classifiers on All Considered Datasets

Methods		DF	DF-CS	DF-FS	DF(PP)	DF(MP)	DF(PL)	DF(PP _PL)	DF(MP _PL)	CDCNN	DBDA	DBMA	FDSSCN	SSRN	RPNet	HOLP -DF	HOLP -DF(PP)	HOLP -DF(MP)
Pavia University	AA	98.71	97.74	98.60	99.33	99.29	99.31	99.81	99.86	95.99	99.23	99.25	99.61	99.58	97.36	98.53	99.41	99.20
	OA	98.40	96.03	97.77	99.32	99.36	99.13	99.78	99.83	96.49	99.46	99.50	99.68	99.59	96.30	98.01	99.40	99.23
	Ka	0.98	0.95	0.97	0.99	0.99	0.99	1.00	1.00	0.95	0.99	0.99	1.00	0.99	0.95	0.97	0.99	0.99
DFC2013 Houston	AA	87.53	88.28	88.47	88.93	88.85	91.35	91.21	92.15	91.76	91.2	91.69	91.95	89.93	84.27	90.92	92.22	91.01
	OA	86.57	85.80	86.10	86.17	86.06	89.33	89.66	90.58	90.26	88.01	90.24	90.91	88.06	81.63	89.14	90.44	89.03
	Ka	0.85	0.85	0.85	0.85	0.85	0.88	0.89	0.90	0.89	0.87	0.89	0.90	0.87	0.80	0.88	0.90	0.88
AirSAR Flevoland	AA	98.88	97.19	97.42	99.62	99.4	99.85	99.92	99.84	98.34	99.83	99.64	99.94	99.91	99.32	97.24	98.81	98.61
	OA	98.78	97.18	97.56	99.62	99.48	99.87	99.94	99.88	98.33	99.85	99.62	99.93	99.92	99.38	96.75	98.84	98.77
	Ka	0.99	0.97	0.97	1.00	0.99	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	0.99	0.97	0.99	0.99

lines marked with up triangle (DF(PP)-HOLP) versus red lines marked with star [HOLP-DF(PP)], and magenta lines marked with diamond (DF-HOLP) versus maroon lines marked with right triangle (HOLP-DF).

4) Only using HOLP based screening could limit and even degrade the classification accuracy of original DF by overfitting from the low ensemble diversity, see the magenta lines marked with diamond in Fig. 7(a), and also in Fig. 6(d) and (f).

In contrast with influence on OA values from the number of samples for training, influence from the depth of considered DFs is much smaller. Particularly see the maroon lines marked with right triangle, green lines marked with left triangle and red lines marked with star as almost horizontally presented in Fig. 6(d), (e) and (f) for HOLP-DF, HOLP-DF(MP), and HOLP-DF(PP), respectively. Moreover, OA curves from HOLP-DF(MP) and HOLP-DF(PP) are always stay at the highest place compared with the OA curves from on other considered DFs on all three datasets, and there are not obvious changes OA curves after the number of layers is larger than 30 in the experiments from Pavia University and AirSAR Flevoland datasets, as shown in Fig. 6(d) and (f).

And similarly, OA curves with the lowest values are always shown either by the original DF or by the DF-HOLP without using random subspace propagation strategy, as shown by sky blue line marked with asterisk and magenta lines marked with diamond. This proofs again that random subspace propagation can benefits the classification accuracy of HOLP-DF by increasing the diversity of ensemble.

C. Classification Results Comparison

To show the performance of the proposed HOLP-DF, we show OA, AA, and Ka values from the original gcForest (represented as DF for short), gcForest with confidence screening (DF-CS), and feature screening (DF-FS), and the proposed versions of DF with patch-based pooling [DF(PP)], morphological profiling (MP), and PL, which were proposed in our previous work [8] in Table II. Also results from the DL classifiers including double-branch, multiattention mechanism network (DBMA), double-branch dual attention mechanism network (DBDA), contextual deeper convolutional neural network (CDCNN), fast dense spectral-spatial convolutional network (FDSSCN), spectral-spatial residual network (SSRN), and random patches network (RPNet) are considered. Classification maps with OA values corresponding to the underlined numbers in Table II are presented in Figs. 7, 8, and 9 for the considered datasets. Notably, critical parameters including the number of layers, propagation ratio and screening ratio are set by 30, 40%, and 40%, respectively, as recommended by the previous results, whereas the parameters for other considered classifiers are set by default values as recommended in previous work [16].

Based on the results from Table II, again it is clear that compatible and even better classification results can be obtained by the proposed HOLP-DF by only using 16% of features from considered datasets compared with original DF. For example, while the original DF reached an OA value of 86.57% on DFC2013 Houston data, an OA value of 89.14% is reached by HOLP-DF. And in contrast with the classification results from DF-CS and DF-FS, AA, OA, and KAPPA values from the HOLP-DF are universally higher on Pavia University and DFC2013 Houston hyperspectral datasets. Take the DFC2013 Houston data as an example, the area at the lower part, which is covered by dense cloud shadows is more precisely classified by HOLP-DF [see Fig. 9(g)] compared with the maps from DF-CS and DF-FS, as shown in Fig. 9(b) and (c). Furthermore, classification performance of the proposed HOLP-DF can be further improved by extra utilizing of PP and MP features over original patch-based features. For instance, see that the area in the bottom right part of Pavia University and the area in the lower part of DFC2013 Houston data, are more correctly classified by HOLP-DF(PP) and HOLP-DF(MP), as shown in Figs. 7(h), (i), and 8(h), (i). By looking at the AA, OA, and KAPPA values from more sophisticated DL methods shown in Table II, the

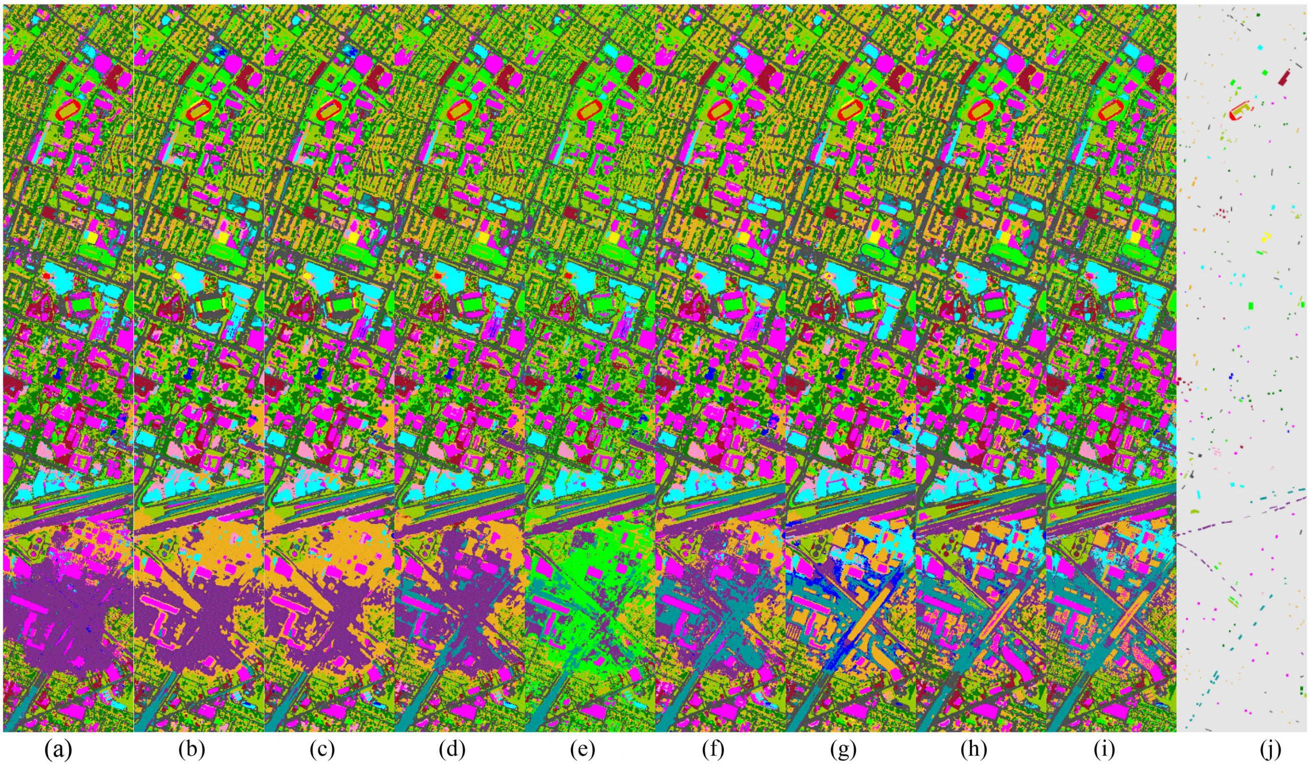


Fig. 8. Classification maps with OA values corresponding to the numbers from Table II (a)–(i) and ground-truth map (j) for DFC2013 Houston data. (a) 86.57%. (b) 85.80%. (c) 86.10%. (d) 89.33%. (e) 89.66%. (f) 90.58%. (g) 89.14%. (h) 90.44%. (i) 89.03%.

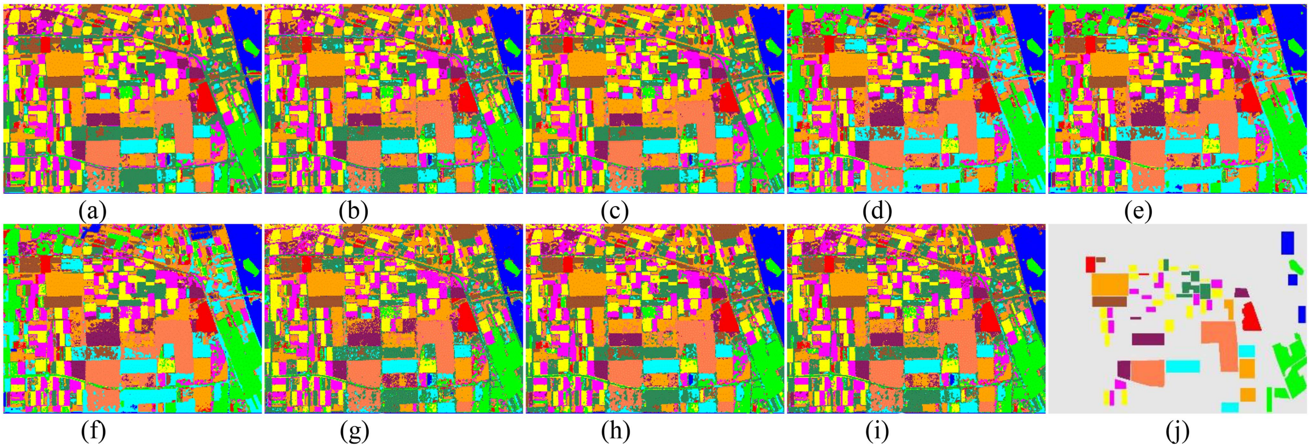


Fig. 9. Classification maps with OA values corresponding to the underlined numbers from Table II (a)–(i) and ground-truth map (j) AirSAR Flevoland data. (a) 98.78% (b) 97.18%. (c) 97.56%. (d) 99.87%. (e) 99.94%. (f) 99.88%. (g) 96.75%. (h) 98.84%. (i) 98.77%.

proposed HOLP-DF, HOLP-DF(PP), and HOLP-DF(MP) reached higher values than CDCNN (OA = 96.49%) and RPNet (OA = 96.30%) on Pavia University data, reached higher values than CDCNN (OA = 90.26%), DBDA(OA = 88.01%), DBMA(OA = 90.24%), SSRN (OA = 88.06%), and RPNet (OA = 81.63%) on DFC2013 Houston data, and reached only higher values than CDCNN (OA = 98.33%) on AirSAR Flevoland data.

VI. Conclusion

In this article, model-based HOLP feature screening method is introduced to overcome the intramodel high data dimensionality drawback of the conventional DF model (the gcForest) in hyperspectral and PolSAR image classification, where the random subspace propagation and REP techniques are also adopted to further boost the classification performance by increasing the ensemble diversity and decreasing the forests ensemble redundancy. To comparatively evaluate the performance of the proposed method, popular feature screeners, and state-of-the-art DL methods are selected in the experiments. According to the results from three widely used hyperspectral and PolSAR image classification benchmarks, the following results are concluded.

1) HOLP is an optimum solution to reduce the high and ultrahigh dimensionality in contrast with feature screening

algorithms like FC, SIS, RRCS, SMLE, BcorSIS, CSIS, Kfilter, MDCSIS, MVSIS, and SIRS, from both highly accurate and efficient execution points of view.

2) HOLP-DF capable of obtaining better classification results than original DF and its modified versions of DF-CS and DF-FS, and the optimum sets of model depth, propagation ratio and screening ratio parameters are 30, 40%, and 40, respectively.

3) Classification performance of HOLP-DF can be further boosted by extra using PP and MP techniques.

Although the proposed HOLP-DF algorithm show advanced performance in terms of classification accuracy, computational efficiency, and intramodel feature reduction effectiveness, the classification accuracy of HOLP-DF is still limited in some scenarios in contrast with the DF, which fused usage of PP and MP with PL techniques and in contrast with the state-of-the-art DL algorithms. Additionally, the HOLP feature screening algorithm may not be the best choice for those originally lower dimensional data such as PolSAR and even multispectral imageries. Therefore, we will focus on the study of other advanced feature screening algorithms for originally low dimensional but highly redundant intramodel dimensional RS image classification cases.

Acknowledgment

The authors would like to thank Prof. P. Gamba and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for freely providing the Pavia University and DFC2013 Houston hyperspectral datasets used in this article.

References

[1] A. M. Lechner, G. M. Foody, and D. S. Boyd, "Applications in remote sensing to forest ecology and management," *One Earth*, vol. 2, no. 5, pp. 405–412, 2020.

[2] A. El Jazouli, A. Barakat, R. Khellouk, J. Rais, and M. El Baghdadi, "Remote sensing and GIS techniques for prediction of land use land cover change effects on soil erosion in the high basin of the Oum Er Rbia River (Morocco)," *Remote Sens. Appl., Soc. Environ.*, vol. 13, pp. 361–374, 2019.

[3] X. Li et al., "Understanding land use/Land cover dynamics and impacts of human activities in the Mekong delta over the last 40 years," *Glob. Ecol. Conservation*, vol. 22, 2020, Art. no. e00991.

[4] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017.

[5] J. Song, S. Gao, Y. Zhu, and C. Ma, "A survey of remote sensing image classification based on CNNs," *Big Earth Data*, vol. 3, no. 3, pp. 232–254, 2019.

[6] Z. Lv, G. Li, Z. Jin, J. A. Benediktsson, and G. M. Foody, "Iterative training sample expansion to increase and balance the accuracy of land classification from VHR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 139–150, Jan. 2021.

[7] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discov.*, vol. 8, no. 6, 2018, Art. no. e1264.

[8] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[9] X. Pan, C. Zhang, J. Xu, and J. Zhao, "Simplified object-based deep neural network for very high resolution remote sensing image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 218–237, 2021.

[10] G. Cheng, X. Xie, J. Han, L. Guo, and G. S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020.

[11] Z. H. Zhou and J. Feng, "Deep forest," *Nat. Sci. Rev.*, vol. 6, no. 1, pp. 74–86, 2019.

[12] Z.-H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proc. Int. Joint Conf. Artifical Intell.*, 2017, pp. 3553–3559.

[13] M. Pang, K. M. Ting, P. Zhao, and Z. H. Zhou, "Improving deep forest by confidence screening," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 1194–1199.

[14] M. Pang, K. M. Ting, P. Zhao, and Z. H. Zhou, "Improving deep forest by screening," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4298–4312, Sep. 2022, doi: 10.1109/TKDE.2020.3038799.

[15] S. Ni and H. Y. Kao, "PSForest: Improving deep forest via feature pooling and error screening," in *Proc. Asian Conf. Mach. Learn.*, 2020, pp. 769–781.

[16] A. Samat, E. Li, P. Du, S. Liu, and Z. Miao, "Improving deep forest via patch-based pooling, morphological profiling, and pseudo labeling for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9334–9349, Sep. 2021.

[17] J. Fan and R. Li, "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," in *Proc. 25th Int. Congr. Mathematicians*, Madrid, Spain, Aug. 2006, pp. 595–622.

[18] X. Wang and C. Leng, "High dimensional ordinary least squares projection for screening variables," *J. Roy. Statist. Soc., Ser. B, Statist. Methodol.*, vol. 78, no. 3, pp. 589–611, 2016.

[19] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, 2009.

[20] J. Xue and F. Liang, "A robust model-free feature screening method for ultrahigh-dimensional data," *J. Comput. Graphical Statist.*, vol. 26, no. 4, pp. 803–813, 2017.

[21] Q. Mai and H. Zou, "The fused kolmogorov filter: A nonparametric model-free screening method," *Ann. Statist.*, vol. 43, no. 4, pp. 1471–1497, 2015.

[22] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.

[23] G. Li, H. Peng, J. Zhang, and L. Zhu, "Robust rank correlation based screening," *Ann. Statist.*, vol. 40, no. 3, pp. 1846–1877, 2012.

[24] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 70, no. 5, pp. 849–911, 2008.

[25] W. Pan, X. Wang, W. Xiao, and H. Zhu, "A generic sure independence screening procedure," *J. Amer. Statist. Assoc.*, vol. 114, no. 526, pp. 928–937, 2018.

[26] R. Li, W. Zhong, and L. Zhu, "Feature screening via distance correlation learning," *J. Amer. Statist. Assoc.*, vol. 107, no. 499, pp. 1129–1139, 2012.

[27] L. P. Zhu, L. Li, R. Li, and L. X. Zhu, "Model-free feature screening for ultrahigh-dimensional data," *J. Amer. Statist. Assoc.*, vol. 106, no. 496, pp. 1464–1475, 2011.

[28] H. Cui, R. Li, and W. Zhong, "Model-free feature screening for ultrahigh dimensional discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 110, no. 510, pp. 630–641, 2015.

[29] X. Shao and J. Zhang, "Martingale difference correlation and its use in high-dimensional variable screening," *J. Amer. Statist. Assoc.*, vol. 109, no. 507, pp. 1302–1318, 2014.

[30] C. Xu and J. Chen, "The sparse mle for ultrahigh-dimensional feature screening," *J. Amer. Statist. Assoc.*, vol. 109, no. 507, pp. 1257–1269, 2014.

[31] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[32] Y. Guo, S. Liu, Z. Li, and X. Shang, "BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," *BMC Bioinf.*, vol. 19, no. 5, pp. 1–13, 2018.

[33] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: A review," *Sensors*, vol. 12, no. 4, pp. 4764–4792, 2012.

[34] J. M. Moyano, E. L. Gibaja, K. J. Cios, and S. Ventura, "Review of ensembles of multi-label classifiers: Models, experimental study and prospects," *Inf. Fusion*, vol. 44, pp. 33–45, 2018.

[35] R. Blaser and P. Fryzlewicz, "Random rotation ensembles," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 126–151, 2016.

[36] R. Blaser and P. Fryzlewicz, "Regularizing axis-aligned ensembles via data rotations that favor simpler learners," *Statist. Comput.*, vol. 31, no. 2, pp. 1–12, 2021.

[37] T. I. Cannings and R. J. Samworth, "Random-projection ensemble classification," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 79, no. 4, pp. 959–1035, 2017.

[38] H. Ahn, H. Moon, M. J. Fazzari, N. Lim, J. J. Chen, and R. L. Kodell, "Classification by ensembles from random partitions of high-dimensional data," *Comput. Statist. Data Anal.*, vol. 51, no. 12, pp. 6166–6179, 2007.

[39] L. I. Kuncheva, J. J. Rodríguez, C. O. Plumpton, D. E. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 531–542, Feb. 2010.

[40] Y. Tian and Y. Feng, "RaSE: Random subspace ensemble classification," *J. Mach. Learn. Res.*, vol. 22, no. 45, pp. 1–93, 2021.

[41] G. Martinez-Munoz, D. Hernández-Lobato, and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 245–259, Feb. 2009.

[42] G. Tsoumakas, I. Partalas, and I. Vlahavas, "An ensemble pruning primer," in *Applications of Supervised and Unsupervised Ensemble Methods*. Berlin, Germany: Springer, 2009, pp. 1–13.

[43] H. Chen, P. Tiňo, and X. Yao, "Predictive ensemble pruning by expectation propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 7, pp. 999–1013, Jul. 2009.

[44] Z. H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, no. 1/2, pp. 239–263, 2002.

[45] X. Yao and Y. Liu, "Making use of population information in evolutionary artificial neural networks," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 28, no. 3, pp. 417–425, Jun. 1998.

[46] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proc. Int. Conf. Mach. Learn.*, 1997, vol. 97, pp. 211–218.

[47] R. Chakraborty and B. C. Vemuri, "Statistics on the stiefel manifold: Theory and applications," *Ann. Statist.*, vol. 47, no. 1, pp. 415–438, 2019.

[48] T. Banerjee, G. Mukherjee, and P. Radchenko, "Feature screening in large scale cluster analysis," *J. Multivariate Anal.*, vol. 161, pp. 191–212, 2017.

[49] E. Barut, J. Fan, and A. Verhasselt, "Conditional sure independence screening," *J. Amer. Stat. Assoc.*, vol. 111, no. 515, pp. 1266–1277, 2016.

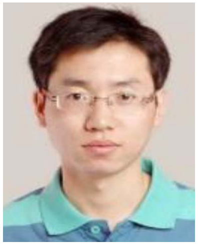
[50] Q. Mai and H. Zou, "The kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, vol. 100, no. 1, pp. 229–234, 2013.

[51] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "mRMRe: An R package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, 2013.

Alim Samat (Member, IEEE) received the B.S degree in geographic information system from Nanjing University, Nanjing, China, in 2009, the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2012, and the Ph.D. degree in cartography and geography information system from Nanjing University, Nanjing, China, in 2015.

He was a Visiting Ph.D. Student with the Department of Industrial and Information Engineering, University of Pavia, Pavia, Italy, in 2014. He is currently an Associate Professor with the State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Ürümqi, China. His main research interests include multisource remote sensing image processing and pattern recognition in arid land resource and environment applications.

Dr. Samat is a reviewer for several international remote sensing journals, including Remote Sensing Environment, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATION AND REMOTE SENSING, *IEEE Geoscience and Remote Sensing Letters*, *Remote Sensing*, and *International Journal of Remote Sensing*.



Erzhu Li received the Ph.D. degree in cartography and geographic information system from Nanjing University, Nanjing, China, 2017.

He is currently a Lecturer and Researcher with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in urban remote sensing applications.

Dr. Li is a Reviewer for several international remote sensing journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATION AND REMOTE SENSING, *IEEE Geoscience and Remote Sensing Letters*, and *Remote Sensing*.

Wei Wang received the B.S. degree in geographic information science from the School of Resources and Environment, Shandong Agricultural University, Taian, China, in 2017.

He is currently a joint Ph.D. student with the Department of Geography, Ghent University and Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences. He is interested in research on the application of the Google Earth Engine platform and machine learning. His research interests include sand and dust storm identification and risk mapping in arid zones of Central Asia.

Sicong Liu (Senior Member, IEEE) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Beijing, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, 2015.

He is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal data processing, change detection, and spectral signal analysis in multispectral/hyperspectral images.

Dr. Liu is a Reviewer for several international remote sensing journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATION AND REMOTE SENSING, *IEEE Geoscience and Remote Sensing Letters*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Remote Sensing*.

Ximing Liu received the B.S. degree in remote sensing science and technology from Central South University, Changsha, China, in 2020. She is currently working toward the M.E. degree in resource and environment remote sensing with the University of Chinese Academy of Sciences, Beijing, China.

Her research interests include remote sensing image processing, urban impervious surface area extraction, and machine learning.