Erciyes University Journal of Institue Of Science and Technology

Research Article
10.65520/erciyesfen.1999099

Imprint:
Volume: 42(3)
Year: 2026
Page: 297-321

Selver GÜNGÖR[a]
Berra Nur TUNÇ[b*]
Ümit Haluk ATASEVER[c]

[a] Res. Asst., Erciyes University, selvergungor@erciyes.edu.tr
[b] Res. Asst., Erciyes University, berranursen@erciyes.edu.tr
[c] Prof., Erciyes University, uhatasever@erciyes.edu.tr

* Corresponding Author

Received: 7/20/2026
Accepted: 9/14/2026

Citation:
Selver GÜNGÖR, Berra Nur TUNÇ, Ümit Haluk ATASEVER (2026). Probabilistic Stacked Ensemble and Automated Machine Learning Approach in Remote Sensing Data. *Erciyes University Journal of Institue Of Science and Technology*, 42(3), 297-321. https://doi.org/10.65520/erciyesfen.*1999099*

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

Probabilistic Stacked Ensemble and Automated Machine Learning Approach in Remote Sensing Data

Abstract

Background/Aim: In the classification of remote sensing data, Machine Learning (ML) models when used alone fall short in capturing spatial and spectral features. These models, especially in the classification of hyperspectral images with similar spectral characteristics, produce high generalization errors. In this study, a probability-based Stacked Ensemble and Automated Machine Learning (AutoML) approach that carries the confidence information of the base learners into the combination stage is proposed to overcome these issues.

Methods: The Stacked model is an ensemble method that combines the outputs of multiple ML or deep learning models with a second meta model. Naive Bayes, Decision Trees, Artificial Neural Networks, K-Nearest Neighbors and Support Vector Machines were used as base learners. The class probability scores produced by these models were presented as input to a meta-learner whose hyperparameters were optimized by AutoML, instead of being used directly as output.

Results: The proposed framework achieved the highest Macro-F1 score on both datasets (0.9297 on Indian Pines and 0.9573 on Pavia University), significantly outperforming every individual base learner as well as the tree-based ensembles, the voting strategies and the deep learning architectures included in the comparison.

Conclusion: The findings show that by combining the strengths of different models, more accurate and more class-balanced predictions can be made compared to single models.

Keywords: Remote Sensing, Stacked Ensemble, Machine Learning

Uzaktan Algılama Verilerinde Olasılık Tabanlı Yığılmış Topluluk ve Otomatik Makine Öğrenmesi Yaklaşımı

Öz

Uzaktan algılama verilerinin sınıflandırılmasında makine öğrenmesi modelleri tek başına kullanıldığında mekânsal ve spektral özellikleri yakalamada yetersiz kalmaktadır. Bu modeller, özellikle benzer spektral özelliklere sahip hiperspektral görüntülerin sınıflandırılmasında yüksek genelleme hataları üretmektedir. Bu sorunların üstesinden gelmek için çalışmada, sınıflandırma belirsizliğini en aza indiren olasılık tabanlı Yığılmış Topluluk (Stacked Ensemble) ve Otomatik Makine Öğrenmesi yaklaşımı önerilmiştir. Yığılmış model, birden fazla makine öğrenmesi veya derin öğrenme modelinin çıktılarını ikinci bir meta model ile birleştiren bir topluluk yöntemidir. Çalışmada temel öğreniciler olarak Naive Bayes, Karar Ağaçları, Yapay Sinir Ağları, K-En Yakın Komşu ve Destek Vektör Makineleri kullanılmıştır. Bu modellerin ürettiği sınıf olasılık skorları, doğrudan çıktı olarak kullanılmak yerine, hiperparametreleri Otomatik Makine Öğrenmesi ile optimize edilen bir

meta-öğreniciye girdi olarak sunulmuştur. Önerilen model, Indian Pines ve Pavia University veri setlerinde en yüksek Macro-F1 skorunu (sırasıyla 0.9297 ve 0.9573) elde ederek tüm bireysel temel öğrenicileri, ağaç tabanlı toplulukları, oylama yaklaşımlarını ve karşılaştırmaya dâhil edilen derin öğrenme mimarilerini anlamlı biçimde geride bırakmıştır. Elde edilen bulgular, farklı modellerin güçlü yönlerinin birleştirilmesiyle tekil modellere kıyasla daha doğru, daha az yanlı ve daha kararlı tahminler yapılabileceğini göstermiştir.

Anahtar kelimeler: Uzaktan Algılama, Yığılmış Topluluk, Makine Öğrenmesi

1.INTRODUCTION

Remote sensing technologies have become a critical data source in areas such as monitoring terrestrial resources, tracking environmental changes and smart city planning. Within this field, hyperspectral imagery (HSI) has gained ground as sensor technology has matured. A hyperspectral scene records hundreds of narrow and contiguous bands, so that every pixel carries a near-continuous reflectance signature instead of a few broad colour values and materials that are indistinguishable in multispectral data can be separated. This level of spectral detail is what makes HSI attractive for tasks requiring fine discrimination and classification is the operation through which that detail is turned into thematic information [1]. Although Machine Learning (ML) algorithms are widely used in processing these data, it is known that no single algorithm performs best across all data distributions and complex land cover classifications [2,3]. The limitations in the generalization ability of individual models and the risks of overfitting have led researchers to community learning methods that combine the predictive power of multiple models.

Recent studies in the literature demonstrate that the Stacking method outperforms individual models. Healey et al. [3] tested the Stacking method using a Random Forest (RF) based fusion rule in the mapping of forest changes. This approach has shown to significantly reduce error rates compared to individual change detection algorithms . Similarly Xu et al. [2] compared 19 different machine learning models in drought monitoring studies. They reported that a Stacking model using the CatBoost algorithm as a meta-learner provided the highest accuracy in predicting drought indices.

The success of the Stacking method has also been reported in areas with complex topographic features. Zandi et al. [4] reported that the stacked model they developed for high-resolution precipitation mapping which includes Multi-Layer Perceptron (MLP), Support Vector Machines (SVM) and RF reduced the error (RMSE) by 19% compared to the traditional locally weighted linear regression method . In land cover classification Cai et al. [5] proposed an object-based Stacking method for mapping wetlands with high heterogeneity. He found that this method provided 3.88% higher overall accuracy compared to individual classifiers.

Similar trends are observed in specific applications such as biomass and urban structure extraction. Zhang et al. [6] developed a Stacking algorithm that combines five different base learners to reduce systematic errors in forest aboveground biomass estimation. This method has been shown to reduce bias by over 22% compared to the best individual model. In the urban area Cao et al. [7] proposed a sparse autoencoder based Stacking model (SENet) that combines U-Net, FCN-8s and SegNet models for building extraction. This model has shown to provide 6.7% higher accuracy than individual deep learning models .

Most traditional Stacking applications only provide the final class predictions of base learners as input to the meta-learner. However ignoring the probability scores that indicate how much trust is placed in the models predictions leads to information loss especially in transition regions where spectral confusion is high. Studies in the literature generally prefer simple linear models such as Logistic Regression as meta-learners or algorithms that operate with default parameters. However manually tuning the hyperparameters and structure of the meta-learner is both time consuming and prone to human bias. The use of Automated Machine Learning (AutoML) techniques for optimization at the meta level is still limited in the literature [8].

CC: BY-NC-ND

The main objective of this study is to present the approach of using Probability based Stacked Ensemble and AutoML to fill these gaps in the literature. The study aims to enhance the model's decision confidence by presenting the calculated posterior probability scores for each class to the meta-learner as input instead of the class labels produced by the base learners (Naive Bayes (NB), Decision Trees (DT), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN) and SVM). Additionally the structure and hyperparameters of the meta-learner are optimized using AutoML techniques instead of manual trial and error methods. Thus a hybrid structure with the highest generalization capability independent of human intervention has been created. This approach aims to achieve more accurate and better class-balanced results in the classification of remote sensing data.

2. MATERIALS AND METHODS

In this study a two layered Stacked Generalization architecture, as shown in Figure 1, is proposed to improve classification performance by combining the complementary strengths of the base learners. The method has a hierarchical structure that combines the strengths of base learners based on different mathematical assumptions and processes the outputs of these models with a meta-learner optimized by AutoML. As mentioned in the literature, the stacking method reduces the variance of individual models and enhances generalization ability by combining the complementary strengths of different algorithms

The proposed framework was evaluated on two benchmark hyperspectral datasets, Indian Pines and Pavia University, which differ in sensor, spatial resolution, number of spectral bands and number of classes.

In order to measure the generalization ability of the dataset impartially 80% of the existing samples were randomly assigned as the training set and 20% as the test set. During the training phase the 5-Fold Cross-Validation method was applied to prevent the model from overfitting and to make hyperparameter optimization more stable. This ensured that the meta-model parameters determined in the AutoML process exhibited consistent performance across different layers of the dataset**.**

The stacking principle adopted here originates in the stacked generalization framework of Wolpert [9], in which the outputs of several first-level models are treated as the input of a second-level model rather than being combined by a fixed rule. Breiman [10] subsequently formalised the approach for regression and demonstrated that the second-level model must be trained on cross-validated predictions of the first level, otherwise the combination overfits; this requirement motivates the out-of-fold construction of the meta-features used in the present study. Comprehensive reviews of ensemble learning in remote sensing [11] show that these ideas have since become a standard component of land cover classification pipelines.

It should be emphasised that supplying class probabilities rather than hard labels to the meta-learner is not in itself novel; the practice has been documented in remote sensing since at least 2018 and is used, for example, in the deep ensemble scene classification framework of He et al. [12]. The contribution of the present work is therefore not the use of probabilistic meta-features as such, but the combination of posterior probability estimates with explicit ensemble diversity measures in a single meta-feature space, together with the automatic selection of the meta-learner instead of a predetermined choice.

The provided algorithmic implementation presents an advanced stacked generalization framework tailored for hyperspectral image classification, integrating automated machine learning and prediction diversity metrics. In align with the suggestions in recent geomatics literature regarding the potential of combining diverse ensemble learners for enhanced spatial analysis [13], this study implements a Probabilistic Stacked Ensemble approach in which the class-probability outputs of the base learners are carried into the combination stage. The training dataset is defined in Equation 1 as [14]:

$$D_{\text{train}} = \{x_i, y_i\}_{i=1}{}^{N}, \quad x_i \in \mathbb{R}^B, \quad y_i \in \{1,2,\dots,C\} \tag{1}$$

CC: BY-NC-ND

Where N is the number of labelled training samples, $x_i \in \mathbb{R}^B$ is the input pixel vector of the i-th sample defined over B spectral bands and $y_i \in \{1,2,\dots,C\}$ is the corresponding categorical class label among C total classes. As illustrated in Figure 1, the system initializes a diverse ensemble of $K = 5$ base learners, encompassing DT, KNN, NB, SVM and ANN. Through a 5-fold cross-validation strategy, the training data is partitioned to generate unbiased out-of-fold predictions. For each instance x_i, every base learner $L_k, k = 1, \dots, K$, outputs a categorical class prediction $\hat{y}_{ik}$ alongside a vector of continuous class probability scores defined in Equation 2 [14]:

$$P_k(x_i) = [P_{k1}(x_i), P_{k2}(x_i), \dots, P_{kC}(x_i)] \tag{2}$$

To capture the collective decision-making dynamics and potential uncertainties among the base models, the methodology computationally extracts four distinct diversity and confidence features for each spatial sample. The voting entropy quantifies the disagreement among the classifiers and is calculated as shown in Equation 3 [14]:

$$E(x_i) = -\sum_{c=1}^{C} p_c \log_2(p_c + \varepsilon) \tag{3}$$

Where p_c indicates the proportion of base learners predicting class c and ε serves as a minuscule constant to prevent undefined logarithms. The agreement rate measures consensus by determining the fraction of base learners that align with the majority vote, formulated as in Equation 4 [15]:

$$A(x_i) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{I}\left(\hat{y}_{ik} = \mathrm{mode}(\hat{y}_{i1}, \dots, \hat{y}_{iK})\right) \tag{4}$$

Here $\mathrm{I}(\cdot)$ denotes the indicator function. The prediction variance evaluates the statistical dispersion of the predicted categorical labels, expressed as Equation 5 [14,15]:

$$s^2(x_i) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{y}_{ik} - \overline{y}_i\right)^2 \tag{5}$$

Here $\overline{y}_i$ represents the mean of the K predictions obtained for sample i. Furthermore, the average confidence metric aggregates the maximum probability scores derived from each constituent model, computed as shown in Equation 6 [15]:

$$\overline{P}(x_i) = \frac{1}{K}\sum_{k=1}^{K} \max_c P_{kc}(x_i) \tag{6}$$

These engineered statistical features are subsequently concatenated with the raw out-of-fold class probability scores to construct a comprehensive meta-feature vector Equation 7 [15]:

$$z_i = \left[P_1(x_i), \dots, P_K(x_i), E(x_i), A(x_i), s^2(x_i), \overline{P}(x_i)\right] \tag{7}$$

This high-dimensional augmented representation functions as the input space for the Level-1 meta-learner. The framework employs an automated hyperparameter optimization routine to dynamically select and tune the most optimal meta-classifier, mathematically mapping the relationship as shown in Equation 8 [14,15]:

$$\hat{y}_i = f_{\mathrm{meta}}(z_i) \tag{8}$$

In computational instances where the automated search fails to converge optimally, the algorithm robustly defaults to a bootstrap aggregating ensemble method. Ultimately, this two-tier mathematical transformation is applied sequentially across the entire spatial domain of the hyperspectral image tensor to yield the definitive classification map.

CC: BY-NC-ND

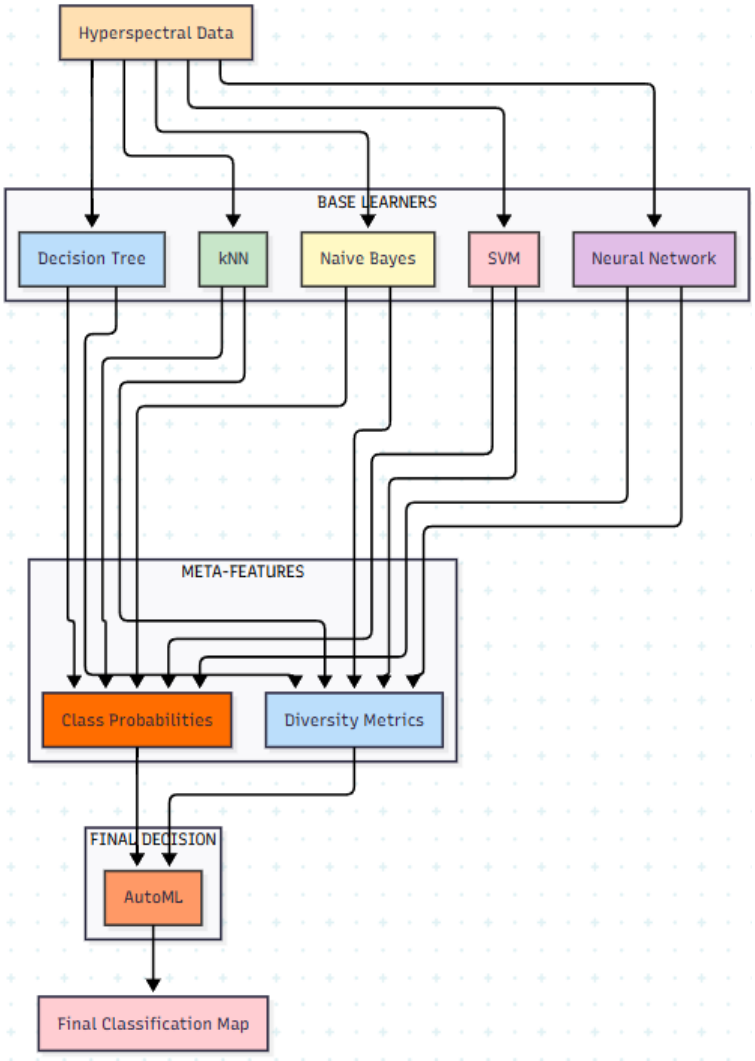

Figure 1. Experimental Methodology Flowchart

2.1. Basic Learners

To capture the diversity of spectral and spatial patterns in the dataset five different machine learning algorithms with proven success in the literature were selected as Level-0 models: SVM, DT, KNN, ANN and NB.

2.1.1. Decision Trees

DT is a non-parametric method that classifies data by hierarchically dividing the data space. The tree structure aims to make the best separation at each internal node, starting from the root node, based on a specific attribute. The Gini index or information gain is often used as the splitting criterion. The Gini index (G) is calculated as shown in Equation 9 [16]:

$$G = 1 - \sum_{c=1}^{C} {p_c}^2 \quad (9)$$

Here p_c represents the proportion of samples belonging to class c at a node and C represents the total number of classes. The algorithm performs splits that minimize the Gini value to increase homogeneity.

2.1.2. K-Nearest Neighbors

KNN operates as a memory based learning technique where the label of an unclassified sample

is determined through a majority consensus among its k most proximal neighbors within the feature space. To quantify this proximity the Euclidean distance defined in Equation 10 is frequently adopted as the standard metric for calculating the spectral separation between two distinct pixel vectors, u and v [17]:

$$d(u,v) = \sqrt{\sum_{j=1}^{B}(u_j - v_j)^2} \tag{10}$$

Here u and v represent two different pixel vectors and B represents the number of spectral bands. KNN is effective in modeling local data distributions but can be sensitive to noise [18].

2.1.3. Naive Bayes

Rooted in Bayes theorem, NB functions as a probabilistic classification technique under the premise that all input features exhibit conditional independence [19]. The algorithm operates by computing the posterior probability for every potential category, subsequently assigning the given data instance to the class that yields the maximum likelihood. For a class variable Y and pixel vector x this decision making process is formally defined by the rule expressed in Equation 11 [19,20].

$$\hat{y} = \arg\max_{c\in\{1,\ldots,C\}} \mathrm{P}(Y=c) \prod_{j=1}^{B} \mathrm{P}(x_j|Y=c) \tag{11}$$

Here $\mathrm{P}(Y=c)$ is the prior probability of class c and $\mathrm{P}(x_j|Y=c)$ is the conditional probability of the j-th spectral band given class c. NB is particularly known for its computational efficiency, especially with high dimensional data.

2.1.4. Support Vector Machines

SVM are statistical learning theory-based algorithms designed to identify the optimal hyperplane that maximizes the margin between classes. For datasets that lack linear separability, kernel functions are employed to project the data into a higher dimensional feature space. The corresponding optimization framework is formulated in Equation 12 [21].

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \lambda \sum_{i=1}^{N} \xi_i \tag{12}$$

Here w is the weight vector, b is the bias value, ξ_i are the slack variables and λ is the penalty parameter. The Radial Basis Function (RBF) kernel has been preferred to model complex non-linear relationships in remote sensing data [22].

2.1.5. Artificial Neural Networks

ANN are hierarchical architectures designed to emulate biological neural systems, typically comprising an input layer, one or more hidden layers and an output layer [23]. The MLP model is specifically favored for addressing non-linear modeling challenges. Within this structure neurons transform the weighted sum of input features via an activation function to generate a result. Following Equation 13, the output (y) represents this transformation process for a single neuron [24,25]:

$$y = \sigma\left(\sum_{j=1}^{B} w_j x_j + b\right) \tag{13}$$

Here w_j represents the weights, x_j the inputs, b the bias term and σ the activation function. The network updates the weights to minimize the error function using the backpropagation algorithm.

2.2. Competing Methods

The proposed framework was compared with nine methods drawn from three families. The first comprises tree-based ensembles. Random Forest grows a large number of trees on bootstrap samples, considers a random subset of bands at each split and aggregates the trees by voting [26]. AdaBoost instead builds trees sequentially, raising the weight of the samples misclassified at the previous stage so that later trees concentrate on the difficult cases [27].

CC: BY-NC-ND

The second family consists of three gradient boosting implementations that differ mainly in how trees are grown and how splits are located. XGBoost adds an explicit regularisation term to the objective and uses a sparsity-aware split-finding procedure [28]. LightGBM grows trees leaf-wise and lowers the cost of the split search through gradient-based sampling and feature bundling [29]. CatBoost applies ordered boosting to avoid the prediction shift that target leakage introduces [30]. Boosting of this kind has already been adopted as a meta-learner in remote sensing applications [2].

The third family covers combination rules and deep architectures. Majority voting assigns the class predicted by the largest number of base learners, whereas soft voting averages their posterior probabilities and selects the maximum; both are standard reference points for ensemble combination [11]. The two deep models operate on the spectral vector alone: a one-dimensional convolutional neural network (1D-CNN) extracts local patterns along the spectral axis through successive convolution blocks, in the manner commonly adopted for remote sensing imagery [31] and a Transformer classifier replaces convolution with multi-head self-attention, so that every band can be related to every other band [32].

2.3. Meta-Learner and Optimization

A meta-learner is used in the second layer (Level-1) to combine the outputs of the base learners; the literature notes that its quality directly governs ensemble performance [6]. To remove the subjectivity of manual tuning, the meta-learner was not fixed in advance but selected automatically by an automated machine learning (AutoML) model-selection procedure under a fixed budget of 100 objective-function evaluations. The search strategy is size dependent: Bayesian optimisation is used for training sets below 10,000 observations and Asynchronous Successive Halving (ASHA) above that threshold, so Bayesian optimisation was applied to Indian Pines (8,199 samples) and ASHA to Pavia University (34,221 samples). The optimisation criterion is the five-fold cross-validation misclassification rate and the search space spans Decision Trees, discriminant analysis, Naive Bayes, support vector machines, nearest-neighbour models, ensembles and neural networks together with their hyperparameters.

The selected meta-learner differed between the two datasets: an error-correcting output codes (ECOC) model was chosen for Indian Pines and an ensemble model for Pavia University. This difference is itself informative, since it indicates that no single fixed meta-learner would have been adequate for both scenes and supports the use of automated selection rather than a predetermined choice.

Additionally in traditional stacking methods only the predicted class labels of the base learners are given to the meta-learner whereas in this study, posterior probability scores which are a richer source of information have been used. This method also known in the literature as Class Probabilities, allows the meta-learner to take into account the confidence levels of the base models when making decisions [3,33]. Each base learner produces not only the most probable class for an input x but also the probability of that example belonging to each class. The generated probability vectors are combined and presented as input to the meta-learner. This method allows the model to consider not only which class it belongs to but also the confidence level in its decision. The intention is to give the meta-learner more information in spectral transition regions, where the base learners disagree most.

The posterior probabilities are not taken directly from the raw decision scores of the base learners. Four of the five base learners are multiclass models built from binary sub-problems and their scores are mapped onto the probability scale by score-to-posterior transformation functions fitted on the training partition through internal cross-validation; the neural network produces normalised class posteriors directly at its output layer. No further post-hoc calibration, such as isotonic regression, was applied and calibration quality was not assessed separately, so these values are used as comparable confidence estimates rather than as verified calibrated probabilities.

The meta-feature vector supplied to the meta-learner is formed from two components. The first is the set of posterior class probabilities produced by each of the L = 5 base learners, which contributes L x C entries for a problem with C classes. The second is a set of four ensemble diversity measures computed across the base learners for every sample: voting entropy, agreement rate, prediction

CC: BY-NC-ND

variance and average confidence. The total dimensionality is therefore L x C + 4, which gives 5 x 16 + 4 = 84 features for Indian Pines and 5 x 9 + 4 = 49 features for Pavia University.

Voting entropy measures the disorder of the hard predictions of the base learners for a given sample; agreement rate is the proportion of base learners that agree with the majority prediction; prediction variance is the variance of the predicted class indices; and average confidence is the mean of the maximum posterior probability reported by each base learner. These quantities describe how much the ensemble members disagree on a particular pixel, which is information not contained in the probability vectors themselves. All meta-features are generated by five-fold stratified cross-validation within the training partition, so that no test sample contributes to their construction and are standardised to zero mean and unit variance using statistics estimated from the training partition only.

2.4. Performance Metrics

In the process of classifying remote sensing images, accuracy analysis is mandatory to determine the quality and reliability of the resulting thematic maps [34,31]. This evaluation is generally performed using an error matrix (confusion matrix) that summarizes the discrepancies between the classification result and ground reference data (ground truth) [34,35]. In the study Overall Accuracy, Kappa Coefficient (κ) and F1-Score metrics were used to measure classification performance.

2.4.1. Overall Accuracy

To quantify the overall correctness of a classification, the overall accuracy is frequently employed as a standard benchmark [34]. It is determined by dividing the total count of successfully classified pixels by the overall reference count [35]. Mathematically for both binary and multi-class problems, it is defined as the sum of correctly predicted instances (TP and TN) normalized by the total number of samples across all categories (TP, TN, FP and FN) [31,36,37].

Mathematically the overall accuracy (OA) is expressed as shown in Equation 14:

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{14}$$

Here TP represents true positives, TN true negatives, FP false positives and FN false negatives [31,38].

2.4.2. Kappa Coefficient

The Kappa coefficient (κ) is a more robust statistical indicator that evaluates classification accuracy by taking into account agreement occurring by chance [35]. Unlike overall accuracy the Kappa statistic uses not only the diagonal values but all data in the error matrix. The Kappa value ranges from 0 to 1; values approaching 1 indicate perfect agreement while 0 indicates a complete mismatch between the classification and ground data [26].

The calculation of the Kappa coefficient is performed using the Equation 15:

$$\kappa = \frac{M\sum_{i=1}^{C} x_{ii} - \sum_{i=1}^{C}(x_{i+}\cdot x_{+i})}{M^2 - \sum_{i=1}^{C}(x_{i+}\cdot x_{+i})} \tag{15}$$

In this equation C represents the number of classes in the matrix, x_{ii} represents the number of observations on the diagonal (correctly classified), x_{i+} and x_{+i} represent the marginal totals of the i-th row and column respectively and M represents the total number of observations [31,39].

2.4.3. F1-Score

The F1-Score serves as a pivotal metric for assessing model performance, particularly in scenarios characterized by class imbalance within the dataset [38]. It is derived by computing the harmonic mean of Precision and Recall [36]. Accordingly, the definitions for Precision and Recall are formulated as shown in Equation 16:

CC: BY-NC-ND

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{16}$$

The F1-Score is obtained by combining these two values with the following formula Equation 17:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \tag{17}$$

A high F1-Score indicates that the classifier minimizes both false positives (FP) and false negatives (FN) [31,36]. As expressed in Equation 17, this metric provides a balanced assessment by penalizing extreme values in either Precision or Recall. Because both datasets are strongly imbalanced, the per-class F1 values are additionally summarised by their unweighted mean, the Macro-F1 score, which gives every class the same weight regardless of its size and is therefore reported alongside overall accuracy throughout the results.

2.5. Experimental Setup and Implementation Details

All experiments were carried out within a single computing environment. A fixed random seed of 42 was used throughout for the data partition, the cross-validation folds and every model that draws random numbers. Ensembles whose members are trained in parallel are an exception in that their individual fits are not reproducible bit for bit, because the random draws take place independently in the parallel workers. The data partition and the fold assignment are unaffected by this. The complete hyperparameter configuration of all methods is reported in Table 1 so that the experiments can be reproduced.

Differences between the proposed framework and each competing method were tested with McNemar's test on the paired predictions of the common test partition. The exact binomial version of the test was used rather than the asymptotic chi-square approximation, so that the p-values remain valid when the number of discordant pairs is small. Fourteen pairwise comparisons are performed per dataset and the p-values are reported as raw, uncorrected values. To confirm that this does not affect the conclusions, the effect of a correction was checked separately. With fourteen comparisons, a Bonferroni correction tightens the significance threshold from 0.05 to 0.0036. Under this stricter threshold the outcome is unchanged on both datasets. Every comparison that is significant without correction remains significant and those that are not significant remain so.

Table 1. Hyperparameter configuration of all methods.

Model	Configuration
Decision Tree	MinLeafSize = 1; MaxNumSplits = 500
KNN	k = 5; Euclidean distance; standardised predictors
Naive Bayes	Kernel density estimation
SVM	RBF kernel; KernelScale = auto; BoxConstraint = 1; standardised; one-vs-one ECOC
Neural Network	Layers [50 20]; ReLU; Lambda = 1e-4; max. 1000 iterations; standardised
Random Forest	Bagging; 200 learning cycles; MinLeafSize = 1
AdaBoost	AdaBoostM2; 200 learning cycles; MinLeafSize = 1
XGBoost /LightGBM /CatBoost	500 rounds; learning rate 0.05; max. depth 6; row and column subsampling 0.8; seed 42
1D-CNN	3 convolutional blocks (32/64/128 filters); batch normalisation; global average pooling; dropout 0.3
Transformer	8 patches; 64-dim convolutional embedding; positional embedding; 2 self-attention layers (4 heads); dropout 0.3
Deep models (common)	Adam; 30 epochs; mini-batch size 128; learning rate 1e-3
Meta-learner	AutoML model selection; 100 evaluations; five-fold CV misclassification rate
Cross-validation	Five-fold stratified; meta-features from out-of-fold posteriors

CC: BY-NC-ND

To keep the mathematical description unambiguous, the symbols used throughout the paper are collected in Table 2.

Table 2. Notation used throughout the paper.

Symbol	Meaning
N	Number of training samples
M	Number of test samples
B	Number of spectral bands
C	Number of land cover classes
L	Number of base learners (L = 5)
K	Number of cross-validation folds (K = 5)
x_i	Spectral vector of the i-th sample, x_i in R^B
y_i	Class label of the i-th sample, y_i in {1, ..., C}
p^(l)(x)	Posterior probability vector produced by base learner l for sample x
z_i	Meta-feature vector of sample i, of dimension L x C + 4

2.6. Study Area and Datasets

The proposed framework was evaluated on two benchmark hyperspectral datasets that differ in sensor, spatial resolution, number of spectral bands, number of classes and degree of class imbalance. One scene represents an agricultural environment with severe class imbalance and the other a heterogeneous urban environment with a much finer spatial resolution, so that the generalisation capability of the framework can be assessed rather than its behaviour on a single type of scene. The main characteristics of the datasets and the corresponding training and test partitions are summarised in Table 3.

Table 3. Characteristics of the hyperspectral datasets and the training/test partitions.

Dataset	Sensor	Image size	Bands	Classes	Labelled px	Training / Test	Split
Indian Pines	AVIRIS	145 × 145	200	16	10,249	8,199 / 2,050	80 / 20
Pavia University	ROSIS	610 × 340	103	9	42,776	34,221 / 8,555	80 / 20

CC: BY-NC-ND

Table 4. Class distribution and the stratified training/test split for both datasets.

Dataset	Class	Labelled	Training	Test
Indian Pines	Alfalfa	46	39	7
	Corn-notill	1,428	1,142	286
	Corn-mintill	830	655	175
	Corn	237	189	48
	Grass-pasture	483	388	95
	Grass-trees	730	575	155
	Grass-pasture-mowed	28	20	8
	Hay-windrowed	478	385	93
	Oats	20	16	4
	Soybean-notill	972	786	186
	Soybean-mintill	2,455	1,972	483
	Soybean-clean	593	464	129
	Wheat	205	162	43
	Woods	1,265	1,018	247
	Buildings-Grass-Trees-Drives	386	314	72
	Stone-Steel-Towers	93	74	19
	Total	10,249	8,199	2,050
Pavia University	Asphalt	6,631	5,290	1,341
	Meadows	18,649	14,926	3,723
	Gravel	2,099	1,654	445
	Trees	3,064	2,454	610
	Painted metal sheets	1,345	1,050	295
	Bare soil	5,029	4,040	989
	Bitumen	1,330	1,061	269
	Self-blocking bricks	3,682	2,985	697
	Shadows	947	761	186
	Total	42,776	34,221	8,555

Table 4 makes the imbalance explicit. On Indian Pines the class sizes span two orders of magnitude, from 20 pixels for Oats to 2,455 for Soybean-mintill, so four classes contribute fewer than 100 labelled pixels each. On Pavia University the imbalance is milder but still pronounced, with Meadows accounting for 18,649 of the 42,776 labelled pixels. Because a stratified split was used, these proportions are preserved in both the training and the test partition.

2.6.1. Indian Pines

The Indian Pines scene was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over an agricultural test site in north-western Indiana, USA. The image covers 145 × 145 pixels at a spatial resolution of approximately 20 m. Of the 224 originally recorded spectral bands, 200 are retained after the removal of water absorption bands affected by atmospheric effects and a low signal-to-noise ratio. The reference data define 16 land cover classes over 10,249 labelled pixels.

The scene is dominated by agricultural crops observed at different growth stages, which results in highly similar spectral signatures among several classes. The dataset is also severely imbalanced:

CC: BY-NC-ND

the smallest class contains only 20 pixels whereas the largest contains 2,455. This combination of spectral similarity and pronounced class imbalance makes the scene particularly suitable for assessing class-balanced performance measures rather than overall accuracy alone.

2.6.2. Pavia University

The Pavia University scene was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) over the campus of the University of Pavia, Italy. The image covers 610 × 340 pixels at a spatial resolution of 1.3 m and comprises 103 spectral bands after the removal of noisy bands. The reference data define 9 urban land cover classes over 42,776 labelled pixels.

In contrast to the agricultural scene, this dataset represents a heterogeneous urban environment containing both natural and man-made surfaces and provides a considerably higher spatial resolution together with a substantially larger number of labelled samples. It therefore offers a complementary test of the proposed framework under conditions of low class imbalance and abundant training data.

2.6.3. Pre-processing and Experimental Partitioning

Classification is performed pixel-wise on the full spectral vector; no spectral or spatial dimensionality reduction was applied, so that the comparison between methods is not confounded by a feature extraction stage. Base learners that are sensitive to feature scaling were standardised to zero mean and unit variance, with the standardisation parameters estimated exclusively from the training partition and subsequently applied unchanged to the test partition and to the full image.

For both datasets the labelled pixels were partitioned into training and test subsets using an 80/20 ratio in a stratified manner, so that the class proportions of the reference data are preserved in both subsets. The partition was generated once with a fixed random seed of 42, stored and reloaded for every method, so that the base learners, the proposed framework and all nine competing methods were trained and evaluated on exactly the same training and test pixels. The same seed also fixes the assignment of the cross-validation folds and the identical fold vector was shared by every method that requires internal cross-validation, so that the reported differences originate from the classifiers themselves rather than from sampling variability. Meta-features were generated by five-fold stratified cross-validation within the training partition, so that no test sample contributes to the construction of the meta-feature space.

2.7. Ethical Approval

Ethical Approval: This study does not involve human participants or animal subjects; therefore, ethical approval was not required.

3. RESULTS AND DISCUSSION

The proposed probability-based stacked ensemble was compared with its five base learners (DT, KNN, NB, SVM, ANN) and with nine competing methods on the two datasets described in Section 2.6. The competing methods comprise two tree-based ensembles (Random Forest, AdaBoost), three gradient boosting implementations (XGBoost, LightGBM, CatBoost), two established combination strategies (majority voting and soft voting) and two deep learning architectures (a one-dimensional CNN and a Transformer classifier). The gradient boosting methods were trained with an identical budget (500 rounds, learning rate 0.05, maximum depth 6, row and column subsampling 0.8) so that the comparison reflects the method rather than the tuning effort. All models were trained and evaluated on identical partitions and all differences were assessed for statistical significance using McNemar's test.

3.1. Indian Pines

Table 5 reports the performance of the individual base learners and of the proposed framework. Among the base learners the ANN achieved the highest overall accuracy (0.9220), followed by the SVM (0.8961), whereas NB performed poorly (0.5566), which is consistent with its conditional

CC: BY-NC-ND

independence assumption being strongly violated by adjacent and highly correlated spectral bands.

Table 5. Performance of the base learners and the proposed framework on Indian Pines.

Model	Accuracy	Kappa	Macro-F1
DT	0.7551	0.7202	0.6863
KNN	0.8244	0.7987	0.8045
NB	0.5566	0.4974	0.5653
SVM	0.8961	0.8814	0.8902
ANN	0.9220	0.9110	0.8817
Proposed Stacked Ensemble	**0.9278**	**0.9176**	**0.9297**

The results obtained with the nine competing methods on the same partition are given in Table 6. The strongest competitor is LightGBM (0.9244), followed by XGBoost (0.9171), while AdaBoost (0.6439) and the one-dimensional CNN (0.6902) perform poorly on this scene. The weak result of the deep models is attributable to the limited number of training samples available for several of the sixteen classes.

Table 6. Performance of the competing methods on Indian Pines.

Method	Accuracy	Kappa	Macro-F1
Random Forest	0.8741	0.8559	0.8354
AdaBoost	0.6439	0.5786	0.5054
XGBoost	0.9171	0.9053	0.8305
LightGBM	0.9244	0.9137	0.8541
CatBoost	0.8571	0.8361	0.7749
Majority voting	0.8902	0.8745	0.8791
Soft voting	0.8888	0.8728	0.8723
1D-CNN	0.6902	0.6406	0.5368
Transformer	0.8639	0.8445	0.7981
Proposed Stacked Ensemble	**0.9278**	**0.9176**	**0.9297**

A comparison of Tables 5 and 6 reveals the characteristic behaviour of this dataset. The proposed framework attains the highest value of all three measures, but the margin in overall accuracy over the gradient boosting methods is small (0.9278 against 0.9244 for LightGBM). The difference in Macro-F1 is of another order: 0.9297 against 0.8541 for LightGBM and 0.8305 for XGBoost. A gap of up to 0.10 in Macro-F1 at an essentially equal overall accuracy indicates that the competing methods achieve their accuracy predominantly on the dominant classes, whereas the proposed framework distributes its performance more evenly across the sixteen classes. In a scene in which the smallest class comprises 20 pixels, this difference is of direct practical relevance. The comparison of the base learners with the proposed framework is shown in Figure 2.

CC: BY-NC-ND

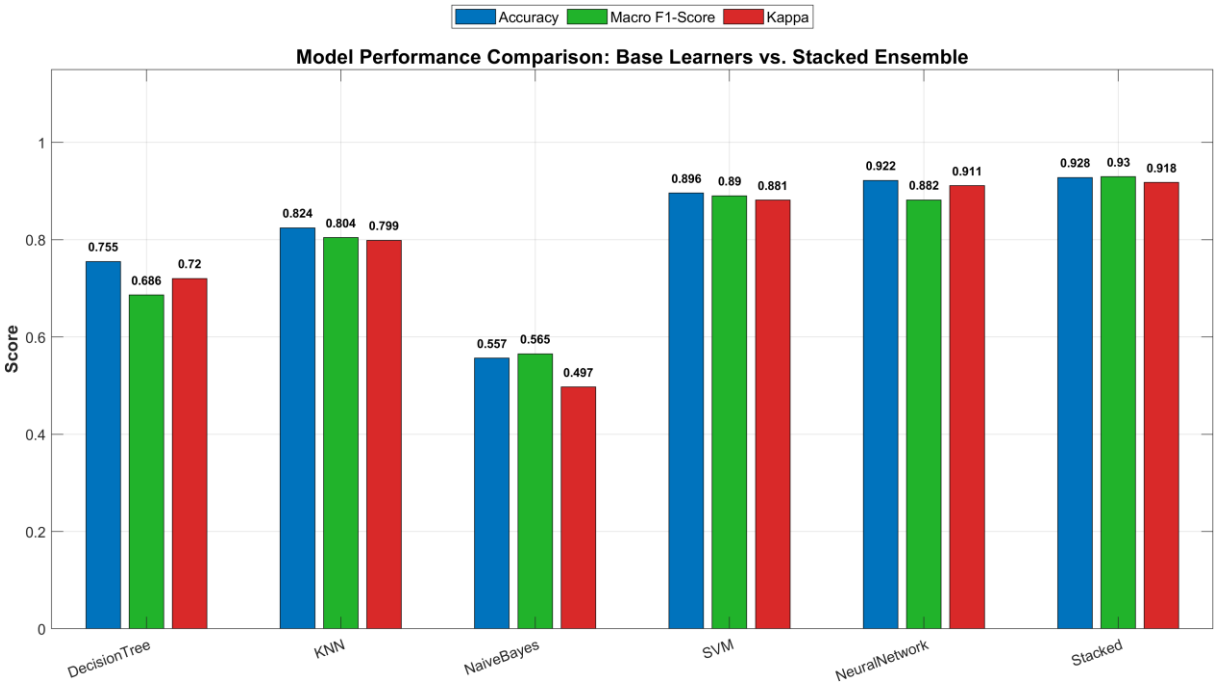

Figure 2. Indian Pines: comparison of the base learners and the proposed stacked ensemble.

The confusion matrices in Figure 3 show that the 148 residual errors of the proposed framework are strongly concentrated in the corn and soybean classes. The single largest confusion is Corn-notill predicted as Soybean-mintill (22 samples, 14.9% of all errors), followed by Soybean-clean predicted as Soybean-mintill (11), the reciprocal confusion between Soybean-mintill and Soybean-notill (11 in each direction) and Corn-mintill predicted as Soybean-mintill (9). These six pairs alone account for 48.6% of all errors. They are not arbitrary: they correspond to crop classes that differ mainly in tillage residue cover rather than in species, so their canopy spectra are nearly identical. The only substantial non-crop error is Buildings-Grass-Trees-Drives predicted as Woods (7 samples), which is expected because the former is a mixed class that itself contains trees. The link with spectral similarity is direct: the mean spectral angle between the mean signatures of the five most frequently confused pairs is only 1.20 degrees, against an average of 10.38 degrees over all 120 class pairs of this scene. The class-wise ROC curves in Figure 4 confirm that the proposed framework attains the highest or joint-highest area under the curve in the majority of classes.

CC: BY-NC-ND

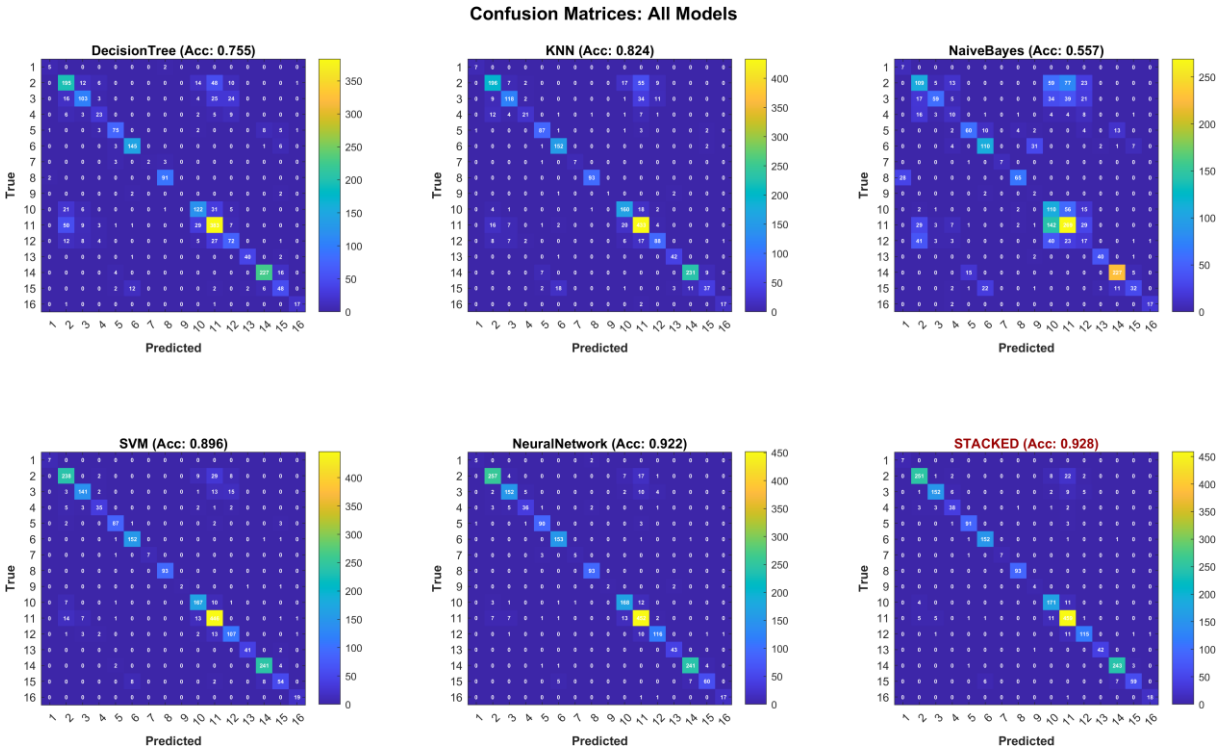

Figure 3. Indian Pines: confusion matrices of all models.

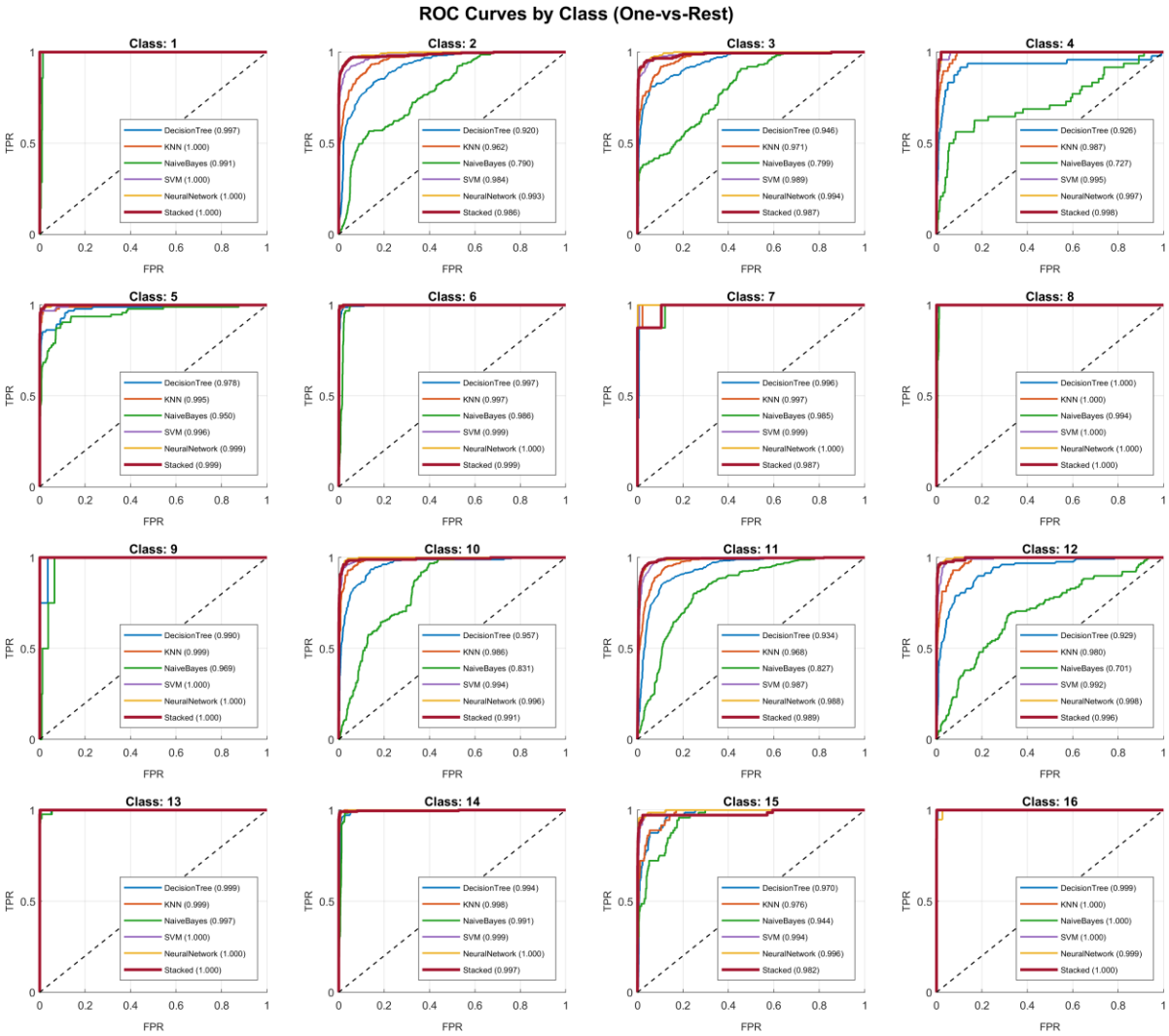

Figure 4. Indian Pines: class-wise ROC curves.

Because differences of a few tenths of a percentage point on a finite test set may arise by chance, every pairwise comparison was assessed with McNemar's test on the same test samples. The improvement over the Decision Tree, KNN, Naive Bayes, SVM, Random Forest, AdaBoost, CatBoost, both voting strategies and both deep learning architectures is significant in every case ($p < 0.001$). The differences with respect to the ANN base learner ($p = 0.188$), XGBoost ($p = 0.061$) and LightGBM ($p = 0.576$) are not significant. On this scene the framework should therefore be understood as matching the strongest competitors in overall accuracy while improving Macro-F1 by up to 0.076, rather than as improving accuracy itself.

The classification maps in Figure 5 provide a spatial view of the same behaviour. The map produced by the proposed framework preserves field boundaries more consistently and exhibits markedly less salt-and-pepper noise within homogeneous parcels than the maps of the individual base learners. The maps are used here as a qualitative complement to the quantitative measures; the quantitative performance of the framework is that reported for the independent test partition in Table 5.

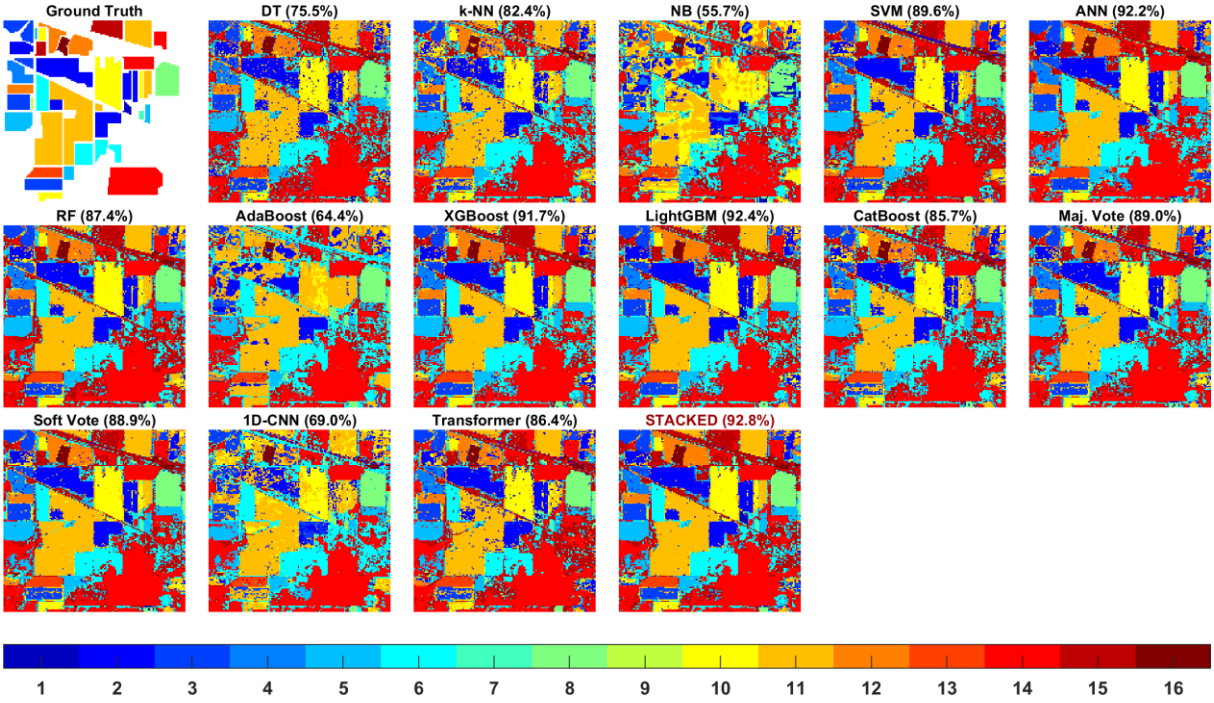

Figure 5. Indian Pines: ground truth and classification maps of all models.

3.2. Pavia University

Table 7 reports the performance of the base learners and of the proposed framework on the Pavia University dataset. The overall level of performance is higher than for Indian Pines, which reflects the finer spatial resolution, the smaller number of classes and the considerably larger training set of this scene.

Table 7. Performance of the base learners and the proposed framework on Pavia University.

Model	Accuracy	Kappa	Macro-F1
DT	0.8892	0.8533	0.8715
KNN	0.8988	0.8638	0.8907
NB	0.6943	0.6117	0.6887
SVM	0.9531	0.9378	0.9443
ANN	0.9655	0.9543	0.9536
Proposed Stacked Ensemble	**0.9674**	**0.9569**	**0.9573**

The corresponding results for the competing methods are given in Table 8. In contrast to Indian Pines, the deep learning architectures perform considerably better on this scene (0.9351 for the CNN and 0.9473 for the Transformer), which is consistent with the much larger number of training samples available here.

Table 8. Performance of the competing methods on Pavia University.

Method	Accuracy	Kappa	Macro-F1
Random Forest	0.9356	0.9141	0.9244
AdaBoost	0.8054	0.7330	0.7312
XGBoost	0.9529	0.9374	0.9435
LightGBM	0.9555	0.9409	0.9456
CatBoost	0.9288	0.9049	0.9138
Majority voting	0.9452	0.9269	0.9358
Soft voting	0.9441	0.9257	0.9344
1D-CNN	0.9351	0.9139	0.9203
Transformer	0.9473	0.9296	0.9373
Proposed Stacked Ensemble	**0.9674**	**0.9569**	**0.9573**

The proposed framework ranks first on all three measures, followed by LightGBM (0.9555) and XGBoost (0.9529). The margin over the strongest base learner is nevertheless small, which is expected: when a single base learner already approaches the performance attainable from the given feature space, the room for improvement available to any combination strategy is correspondingly limited. The comparison of the base learners with the proposed framework is shown in Figure 6.

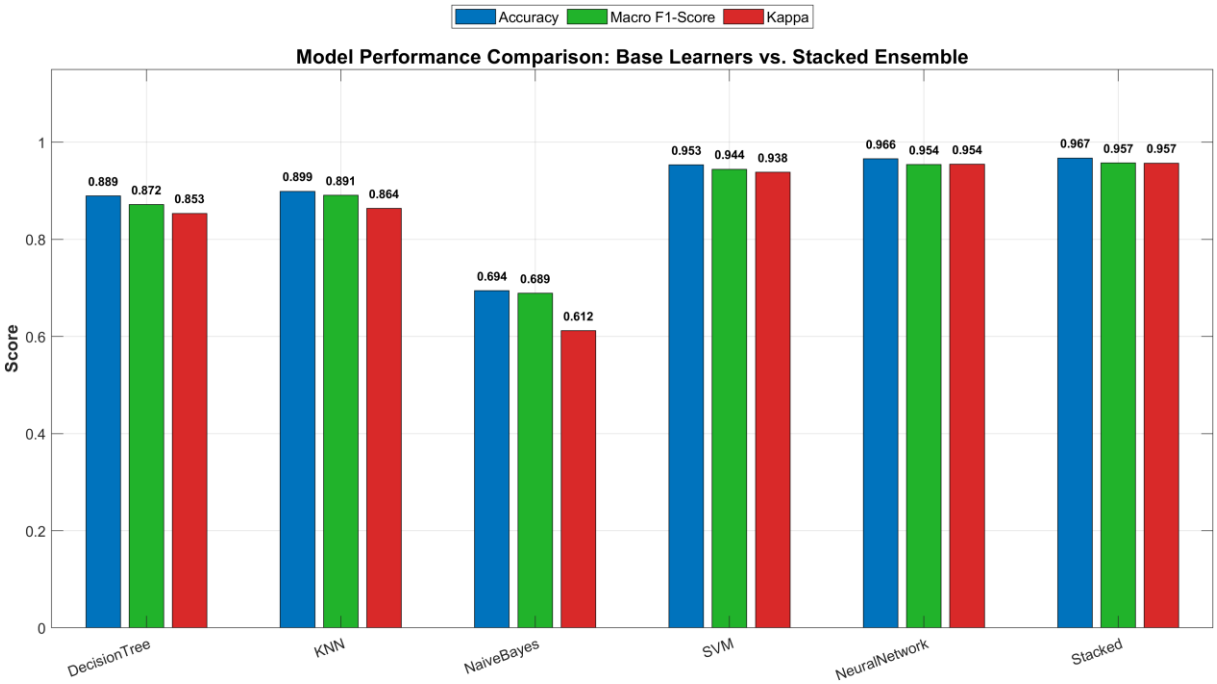

Figure 6. Pavia University: comparison of the base learners and the proposed stacked ensemble.

The confusion matrices in Figure 7 indicate that the 279 residual errors of the proposed framework in this urban scene arise from two distinct mechanisms. The first is spectral overlap between man-made mineral surfaces: Gravel and Self-blocking bricks are confused in both directions (43 and 35 samples, together 27.9% of all errors), Asphalt is predicted as Self-blocking bricks in 27 cases and Bitumen as Asphalt in 24 cases. The second is the reciprocal confusion between Meadows and Bare soil (41 and 28 samples), which is driven less by spectral identity than by mixed pixels at parcel boundaries and by sparse vegetation cover over soil. The spectral angle supports this distinction: Gravel and Self-blocking bricks differ by only 0.96 degrees, the smallest value among all 36 class pairs of this scene, whereas Meadows and Bare soil differ by 11.03 degrees and are therefore

CC: BY-NC-ND

separable in principle. The ROC curves in Figure 8 show consistently high area under the curve values across all nine classes.

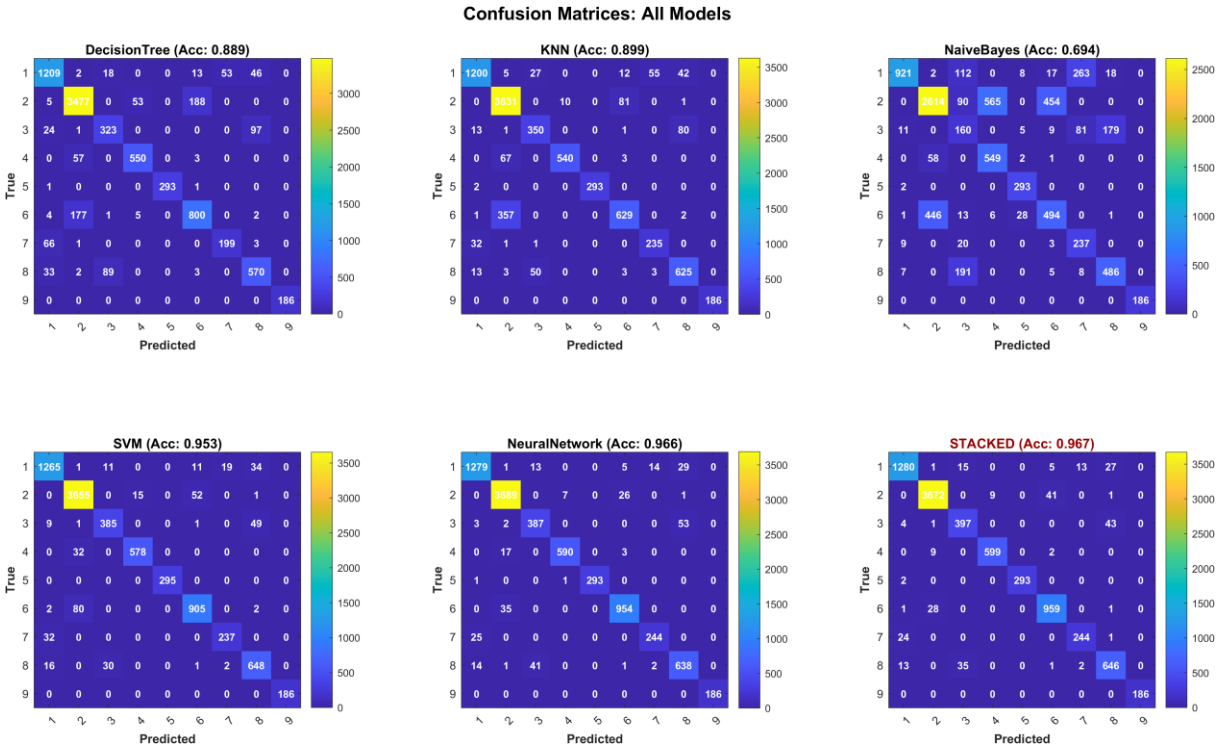

Figure 7. Pavia University: confusion matrices of all models.

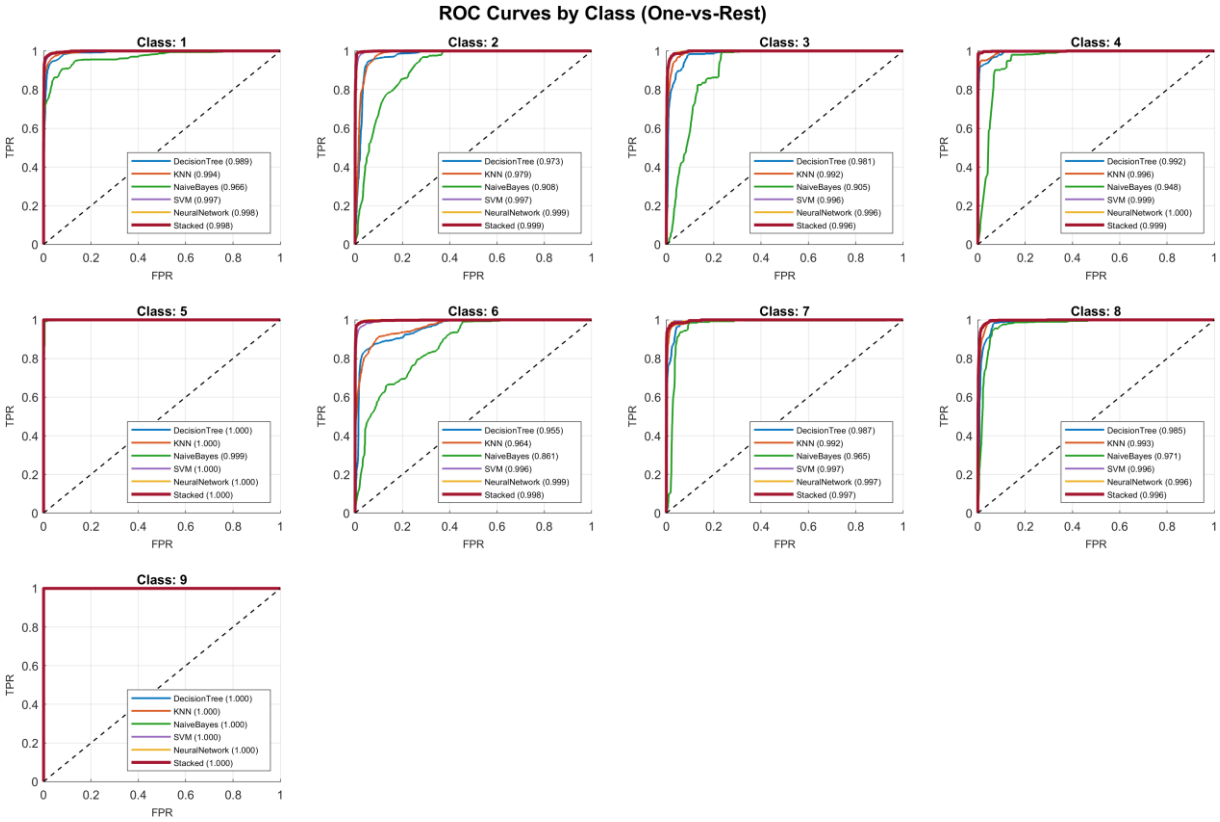

Figure 8. Pavia University: class-wise ROC curves.

Because every method was trained and evaluated on exactly the same stratified partition, the comparisons were assessed with McNemar's test rather than by reading the accuracy values against one another. For a given pair of methods the test considers only the discordant test pixels, that is, those classified correctly by one method and incorrectly by the other and tests the null hypothesis that the two discordant counts are equal; rejecting it means that the observed difference is larger than what the sampling of a finite test set would produce by chance. On this dataset the improvement of the proposed framework is significant against every competing method at $p < 0.001$, including the three gradient boosting implementations and both deep learning architectures. The single exception is the ANN base learner ($p = 0.125$), which is consistent with the small margin reported in Table 7 and with the observation that a strong individual learner leaves little room for any combination strategy on this scene.

The classification maps in Figure 9 show that the proposed framework yields spatially coherent delineations of buildings, roads and vegetated areas and reduces the isolated misclassified pixels that are visible in the maps of the weaker base learners.

These results can also be read against the wider literature. Breiman [10] showed that the benefit of a stacked combination depends on the first-level models being both accurate and mutually decorrelated and the reviews of ensemble methods in remote sensing [11] report the same requirement. The behaviour observed here is consistent with that principle: the margin of the proposed framework is largest on the scene in which the base learners differ most in their per-class behaviour and narrowest where a single base learner already approaches the performance attainable from the feature space. Constructing base learner pools that are strong and genuinely complementary, rather than merely numerous, therefore appears to be the decisive design choice for probabilistic stacking in hyperspectral classification.

In every experiment the meta-learner received the posterior probabilities and the four diversity measures as a single vector and the two were never varied independently. The results therefore establish the value of the combined representation rather than the share contributed by either part and the improvement should not be credited to the probabilities alone. Isolating the two would require an ablation in which each component is withheld in turn. This is left to future work.

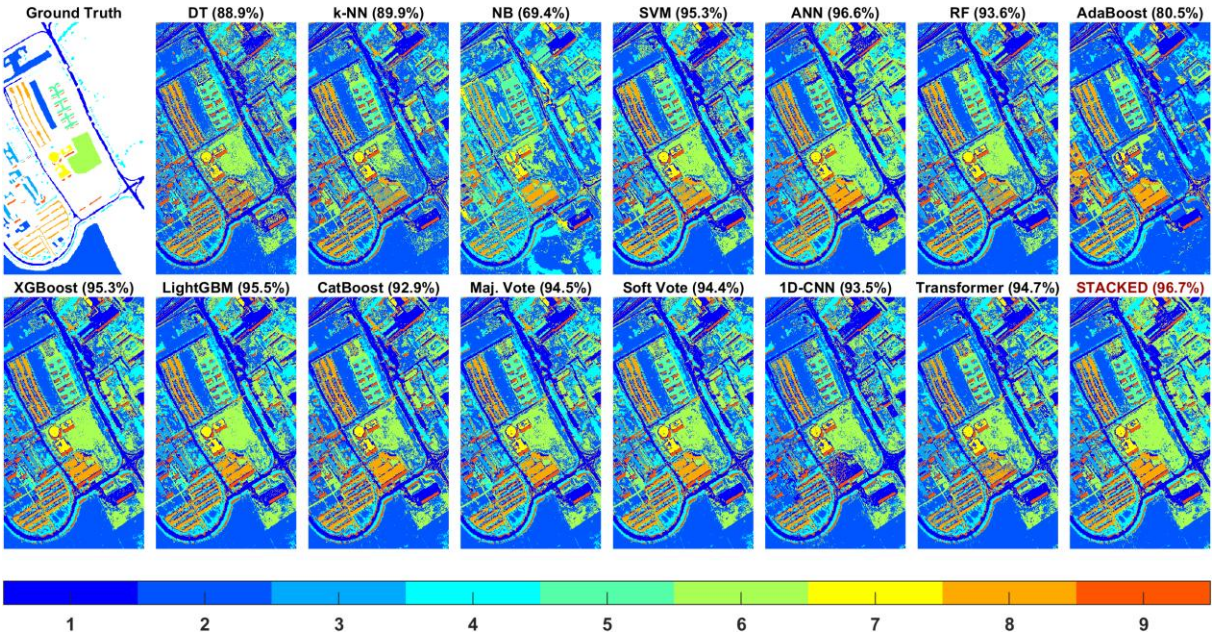

Figure 9. Pavia University: ground truth and classification maps of all models.

4. CONCLUSION

This study proposed a probability-based stacked ensemble framework integrated with automated meta-learner selection for the classification of hyperspectral remote sensing imagery. The framework was evaluated on two benchmark datasets with markedly different characteristics: Indian Pines, an agricultural scene with sixteen classes and severe class imbalance and Pavia University, a high-resolution urban scene with nine classes and abundant training data.

On both datasets the framework achieved the highest Macro-F1 score among the fifteen methods compared and significantly outperformed the classical individual classifiers, the tree-based ensembles, the voting strategies and both deep learning architectures.

In hyperspectral scenes the class distribution is typically far from uniform and overall accuracy is dominated by the majority classes. Where minority classes carry the information of interest, as is the case for sparse crop types or small urban features, a criterion such as Macro-F1 provides a more appropriate basis for model selection and the framework proposed here is advantageous precisely under that criterion. On Indian Pines it exceeded the best gradient boosting implementation by 0.076 in Macro-F1 while the two methods were statistically indistinguishable in overall accuracy. A further practical benefit is that automated meta-learner selection removes the manual trial and error normally required to configure the combination stage and the model selected differed between the two datasets, which indicates that no single fixed meta-learner would have been adequate for both scenes.

These properties map onto concrete application domains. In the agricultural setting the classes that the framework separates most reliably relative to its competitors are crop types that differ mainly in tillage residue rather than in species, which is the discrimination required for crop-type inventories, tillage-practice monitoring and the within-field management units used in precision agriculture. These are also the classes that occupy the smallest area and are therefore the first to be lost when overall accuracy drives model selection. In the urban setting the framework improves the separation of gravel, brick, asphalt and bituminous surfaces, which is the discrimination underlying impervious-surface mapping, roof and pavement material inventories and the surface-cover inputs of urban heat island studies. Because the framework operates on the spectral vector alone and adds only a lightweight second stage on top of existing classifiers, it can be placed on top of an operational pipeline without changing its data preparation.

Several limitations should be acknowledged. The framework operates purely on spectral information and does not exploit spatial context; incorporating spatial-spectral features is expected to yield further improvement and constitutes the most promising direction for future work. The reported results are based on a single stratified partition per dataset, so repeated partitioning would provide a more robust estimate of variability. The individual contribution of the components of the meta-feature space was not isolated here, so a dedicated ablation of the probability and diversity components is left to further work. Finally, the deep learning methods included in the comparison are purely spectral models; a comparison against spatial-spectral architectures would provide a more complete assessment.

Peer-review: External, Independent.

Acknowledgements:
This work was supported by the Scientific Research Projects Coordination Unit of Erciyes University under project number FBA-2025-14681.

Declarations:

1. Statement of Originality:

This work is original. The manuscript has not been published elsewhere and is not under consideration by any other journal.

2. Author Contributions:

Concept: SG,BNT,ÜHA; ***Conceptualization:*** SG,BNT,ÜHA; ***Literature Search:*** SG,BNT,ÜHA; ***Data Collection:*** SG,BNT,ÜHA; ***Data Processing:*** SG,BNT,ÜHA; ***Analysis:*** SG,BNT,ÜHA; ***Writing – original draft:*** SG,BNT,ÜHA; ***Writing – review & editing:*** SG,BNT,ÜHA.

3. Ethics approval:

Not applicable.

4. Funding/Support:

This study was supported by the Scientific Research Projects Coordination Unit of Erciyes University under project number FBA-2025-14681.

5. Competing Interests:

The authors declare no competing interests.

6. GenAI Usage Statement:

During the preparation of this study, the Claude (Anthropic) tool was used in the code development stage. The operations performed were reviewed by the authors, and the final version was edited under the responsibility of Selver Güngör, Berra Nur Tunç and Ümit Haluk Atasever.

7. Sustainable Development Goals:

REFERENCES

[1] Günen, M. A., Atasever, U. H., & Beşdok, E. (2020). Analyzing the contribution of training algorithms on deep neural networks for hyperspectral image classification. Photogrammetric Engineering & Remote Sensing, 86(9), 581-588. https://doi.org/10.14358/PERS.86.9.581

[2] Xu, X., Chen, F., Wang, B., Harrison, M. T., Chen, Y., Liu, K., Zhang, C., Zhang, M., Zhang, X., Feng, P., Hu, K. (2024). Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach. Journal of Hydrology, 634, 131102. https://doi.org/10.1016/j.jhydrol.2024.131102

[3] Healey, S. P., Cohen, W. B., Yang, Z., Brewer, C. K., Brooks, E. B., Gorelick, N., Hernandez, A. J., Huang, C., Hughes, M. J., Kennedy, R. E., Loveland, T. R., Moisen, G. G., Schroeder, T. A., Stehman, S. V., Vogelmann, J. E., Woodcock, C. E., Yang, L., & Zhu, Z. (2018). Mapping forest change using stacked generalization: An ensemble approach. Remote Sensing of Environment, 204, 717–728. https://doi.org/10.1016/j.rse.2017.09.029

[4] Zandi, O., Zahraie, B., Nasseri, M., & Behrangi, A. (2022). Stacking machine learning models versus a locally weighted linear model to generate high-resolution monthly precipitation over a topographically complex area. Atmospheric Research, 272, 106159. https://doi.org/10.1016/j.atmosres.2022.106159

[5] Cai, Y., Li, X., Zhang, M., & Lin, H. (2020). Mapping wetland using the object-based stacked generalization method based on multi-temporal optical and SAR data. International Journal of Applied Earth Observation and Geoinformation, 92, 102164.

CC: BY-NC-ND

https://doi.org/10.1016/j.jag.2020.102164

[6] Zhang, Y., Ma, J., Liang, S., Li, X., & Liu, J. (2022). A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets. GIScience & Remote Sensing, 59(1), 234–249. https://doi.org/10.1080/15481603.2021.2023842

[7] Cao, D., Li, H., Xu, T., Ma, J., & Guo, H. (2021). A Stacking Ensemble Deep Learning Model for Building Extraction from Remote Sensing Images. Remote Sensing, 13(19), 3898. https://doi.org/10.3390/rs13193898

[8] Duman, H. S., & Başaraner, M. (2022). Şekil göstergeleri ve topluluk öğrenmesi sınıflandırma algoritmaları ile bina detaylarının şekil karmaşıklık analizi. Geomatik, 7(3), 197-208. https://doi.org/10.29128/geomatik.947334

[9] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[10] Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1), 49-64. https://doi.org/10.1007/BF00117832

[11] Zhang, Y., Liu, J., & Shen, W. (2022). A review of ensemble learning algorithms used in remote sensing applications. Applied Sciences, 12(17), 8654. https://doi.org/10.3390/app12178654

[12] He, Z., Li, G., Wang, Z., He, G., Yan, H., & Wang, R. (2024). Deep ensemble remote sensing scene classification via category distribution association. Remote Sensing, 16(21), 4084. https://doi.org/10.3390/rs16214084

[13] Barua, A., Raihan, M. M. S., & Akter, L. (2024). A Bayesian Optimization-Based Framework for Building Stacked Ensemble Models in Liver Disease Prediction. 2024 27th International Conference on Computer and Information Technology (ICCIT), 3086–3091. https://doi.org/10.1109/ICCIT64611.2024.11022610

[14] Li, M., & Yan, Y. (2024). Comparative Analysis of Machine-Learning Models for Soil Moisture Estimation Using High-Resolution Remote-Sensing Data. Land, 13(8), 1331. https://doi.org/10.3390/land13081331

[15] Naimi, A. I., & Balzer, L. B. (2018). Stacked Generalization: An Introduction to Super Learning. European Journal of Epidemiology, 33(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z

[16] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612–619.

[17] Garg, R., Kumar, A., Prateek, S., Kumar, S., & Kumar, R. (2022). Land cover classification of spaceborne multifrequency SAR and optical multispectral data using machine learning. Advances in Space Research, 69(4), 1726–1742. https://doi.org/10.1016/j.asr.2021.06.028

[18] Apaydın, C. & Abdikan, S. (2021). Fındık Bahçelerinin Sentinel-2 Verileri Kullanılarak Piksel Tabanlı Sınıflandırma Yöntemleriyle Belirlenmesi. Geomatik,6(2), 107-114, 10.29128/geomatik.705988.

[19] Giri, R. N., Kumar, C., Guha, A., Singh, G., Kumar, S., & Pande, K. (2024). A stacked ensemble learning-based framework for mineral mapping using AVIRIS-NG hyperspectral image. Journal of Earth System Science, 133, 107. https://doi.org/10.1007/s12040-024-02317-z

[20] Balaji, V. R., Suganthi, S. T., Rajadevi, R., Kumar, V. K., Saravanakumar, N., & Pandiyan, S. (2020). Skin disease detection and segmentation using dynamic graph cut algorithm and classification

through Naive Bayes classifier. Measurement, 163, 107922. https://doi.org/10.1016/j.measurement.2020.107922

[21] Awad, M., & Khanna, R. (2015). Support Vector Regression. In Efficient Learning Machines (pp. 67–80). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

[22] Dapke, P., Syed, A., Nagare, S., Bandal, S., & Baheti, M. (2025). A Comparative Analysis of MachineLearning Techniques for LULC Classification Using Landsat-8 Satellite Imagery. International Journal of Engineering and Geosciences, 10 (1), page numbers.

[23] Abdulwahid, M., Feizizadeh, B., & Yakar, M. (2026). Integrating Socio-Political Dynamics andGeospatial Machine Learning: An Integrated-Hybrid ANN approach for sustainable Urban Planningin Post-Conflict Baghdad. International Journal of Engineering and Geosciences, 11(2), 408-431.

[24] Tasdemir, S., & Ozkan, I. A. (2019). ANN approach for estimation of cow weight depending on photogrammetric body dimensions. International Journal of Engineering and Geosciences, 4(1), 36–44. https://doi.org/10.26833/ijeg.427531

[25] Kumar, C., Chatterjee, S., Oommen, T., & Guha, A. (2020). Automated lithological mapping by integrating spectral enhancement techniques and machine learning algorithms using AVIRIS-NG hyperspectral data. International Journal of Applied Earth Observation and Geoinformation, 86, 102006. https://doi.org/10.1016/j.jag.2019.102006

[26] Avcı, C. Budak, M. Yagmur, N., & Balcık, F. B. (2023). Comparison between random forest and support vector machine algorithms for LULC classification. International Journal of Engineering and Geosciences, 8(1), 01-10.

[27] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

[28] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

[29] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3149-3157.

[30] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6639-6649.

[31] Akar, Ö., Saralıoğlu, E., Güngör, O., & Bayata, H. F. (2024). Semantic segmentation of very-high spatial resolution satellite images: A comparative analysis of 3D-CNN and traditional machine learning algorithms for automatic vineyard detection. International Journal of Engineering and Geosciences, 9(1), 12-24. https://doi.org/10.26833/ijeg.1252298

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 6000-6010.

[33] Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. Journal of Artificial Intelligence Research, 10, 271–289. https://doi.org/10.1613/jair.594

[34] Huang, D., Xu, S., Sun, J., Liang, S., Song, W., & Wang, Z. (2017). Accuracy assessment model for classification result of remote sensing image based on spatial sampling. Journal of Applied

CC: BY-NC-ND

Remote Sensing, 11(4), 046023. https://doi.org/10.1117/1.JRS.11.046023

[35] Petrovska, I., & Dimov, L. (2020). Accuracy assessment of unsupervised land cover classification. Scientific Journal of Civil Engineering, 9(2), 83-88.

[36] Çelik, F., & Çelik, K. (2025). Optimizing Land Image Classification: A Deep Learning Approach with ResNet-101 on the EuroSAT Dataset. International Journal of Engineering and Geosciences, 10(2), 197-206. https://doi.org/10.26833/ijeg.1538708

[37] Ayalke, G. Z., & Şişman, A. (2024). Google Earth Engine kullanılarak makine öğrenmesi tabanlı iyileştirilmiş arazi örtüsü sınıflandırması: Atakum, Samsun örneği. Geomatik, 9 (3), 375-390.

[38] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1-13. https://doi.org/10.1186/s12864-019-6413-7

[39] Ozturk, O. (2026). Derin Öğrenme ile İHA Görüntüleri Tabanlı Yığma Duvar Segmentasyonu. Geomatik, 11(1), 120-128.

CC: BY-NC-ND