Earth Science Informatics (2025) 18:7
https://doi.org/10.1007/s12145-024-01555-5

RESEARCH

An assessment of training data for agricultural land cover classification: a case study of Bafra, Türkiye

Mustafa Ustuner[1] · **Fatih Fehmi Simsek**[2]

Received: 20 August 2024 / Accepted: 19 October 2024

Abstract
The training data plays a pivotal role in the accuracy of a machine learning (ML) model in remote sensing. In this case, the set size and purity of the training data have a large influence in classification accuracy. The purpose of this experimental research is to investigate the impact of the different training set size on supervised machine learning classifiers for the agricultural land cover classification in remote sensing. The training set size for each class was incrementally increased at the following intervals: 1%, 5%, 10%, 20%, 30%, 40%, and 50% in our experiment. The remaining 50% of the full ground truth data was used for evaluating the model's accuracy. The test site is situated in Bafra Plain, Samsun, Turkey and the agricultural land cover classification was held using multispectral Sentinel-2 imagery with four ML models, namely Support Vector Machines (SVM), Random Forest (RF), Light Gradient Boosting Machines (LightGBM), and Kernel Extreme Learning Machines (KELM). The experimental results demonstrated that the highest classification accuracy was achieved by LightGBM (89.93%), and followed by RF (86.49%), KELM (78.38%) and SVM (72.49%). The classification accuracies of tree-based methods (RF and LightGBM) increased as the training set size grew, however, kernel-based methods (KELM and SVM) exhibited unstable results as the size of the training dataset varied. Furthermore, our findings highlight that each machine learning model demonstrates different sensitivity to variations in training set size with respect to agricultural land cover classification.

Keywords Training sample size · Agricultural land cover classification · Machine learning · LightGBM · KELM

Introduction

Ongoing advancements in sensor technology and artificial intelligence have fundamentally transformed Earth observation practices and catalyzed significant breakthroughs in Earth observation. The satellite images have been used for many environmental and socio-economic applications ranging from agricultural monitoring to land cover/use mapping, in which the extraction of the useful information from the satellite imagery is critical and needs some expertise (Aplin 2006; Kavzoglu 2009; Khatami et al. 2016; Wang et al. 2023). The classification of the satellite imagery is the most practical and widely-used way for extracting such information. Supervised classification is the most typical way among the other methods in classification techniques however there are many factors impacting the classification accuracy (Foody and Mathur 2004a, b; Lu and Weng 2007; Kavzoglu and Colkesen 2009; Ramezan et al. 2021; Moraes et al. 2024).

In recent years, a number of advanced machine learning models such as kernel extreme learning machines (KELM), light gradient boosting machines (LightGBM) and extreme gradient boosting (XGBoost) are properly applied for the remote sensing image classification however there are still several uncertainties and major challenges in supervised learning process, that needs to be diligently investigated, for acquiring the accurate classification (Foody and Mathur 2006; Lu and Weng 2007; Kavzoglu 2009; Paheding et al.

Communicated by Hassan Babaie.

✉ Mustafa Ustuner
mustuner@artvin.edu.tr

Fatih Fehmi Simsek
fehmi.simsek@tubitak.gov.tr

[1] Department of Geomatics Engineering, Artvin Coruh University, Artvin 08000, Türkiye

[2] TÜBİTAK Space Technologies Research Institute, Ankara 06800, Türkiye

Published online: 05 December 2024

2024). This is essentially due to the darkness in how supervised learning algorithm reacts to input features and the corresponding training data representing the ground labels to make a correct decision. In such cases, the quality (purity, size and composition) of the training data and its harmonization to a machine learning model directly influence the accuracy of the final classification (Foody and Mathur 2006; Ramezan et al. 2021; Weitkamp and Karimi 2023; Moraes et al. 2024). The resolution and dimension of the input imagery, intra-class spectral heterogeneity and the design of the training data (set size, purity and selection strategy) can potentially change the way of how learning algorithm defines the decision boundary or surface that separates the classes in the feature space. Among these uncertainties, the set size of the training data has the highest influence in classification accuracy (Ramezan et al. 2021; Hermosilla et al. 2022; Moraes et al. 2024). As a rule of thumb (a heuristic approach), some of the previous literature in remote sensing defines the sample size in training set as a function of the data dimension and recommends the usage of 10-30p cases for a single class in training dataset, where p denotes the feature numbers (Foody and Mathur 2004a, b; Foody et al. 2006; Kavzoglu 2009; Moraes et al. 2024).

Typically, the larger training dataset could derive higher classification performance compared to those obtained from the smaller training set, but nevertheless it's highly dependent upon the characteristic of the algorithm and the way of how an adopted model define the optimal decision (separating) boundary (Foody and Mathur 2004a, b, 2006; Maxwell et al. 2018; Ramezan et al. 2021). The collecting the ground truth sample is time-consuming, costly and needs some expertise hence researchers try to develop more practical machine learning models that practically require the less number of training samples for more accurate classification. Therefore, the larger training dataset might yield lower classification performance compared to smaller training set in some particular cases, contrary to what is conventionally and practically expected (Foody and Mathur 2006; Kavzoglu 2009; Ramezan et al. 2021; Amarsaikhan et al. 2024; Moraes et al. 2024). For instance, support vector machines (SVM) has been used many times in hyperspectral remote sensing over the last three decades because of having a good generalization capability and tackling the high dimensionality problem, which leads to achieve a desired level of performance (Mountrakis et al. 2011; Belgiu and Drăguţ 2016; Maxwell et al. 2018; Colkesen and Ozturk 2022; Amarsaikhan et al. 2024). However, some other classifiers such as artificial neural networks (ANNs) require a large number of samples to train the neural networks and determine the class boundaries. For ANNs, a limited number of training sample is inadequate to characterize the all classes (Kavzoglu 2009; Foody et al. 2016; Song et al. 2019; Gütter et al. 2022; Paheding et al. 2024). When the classification model reached the optimal number in training set to achieve the desired level of accuracy, the abundant data could cause the overfitting problem in learning process, which can lead to poor performance (low accuracy) of the classification model (Kavzoglu 2009; Ouma et al. 2023; Amarsaikhan et al. 2024; Moraes et al. 2024). Under these circumstances, how much training dataset that algorithm practically needs are still unclear to correctly represent each class. In fact, the training stage (quality and set size of training data) is of crucial on the classification accuracy regardless of a classification method itself (Foody and Mathur 2006; Amarsaikhan et al. 2024; Paheding et al. 2024).

The current literature focusing on the relationship between varying training set size and its impact on classification accuracy provides only limited understanding of these matters. Most of the previous researches questioned that point with conventional machine learning methods such as SVM, RF and ANN for remote sensing image classification.

For example, Song et al. (2012) analyzed the effects of training set size for land cover classification using SVM and two different ANN algorithms. They concluded that SVM is more robust to a small training set size while this small set size is insufficient for ANN learning in terms of classification performance. Li et al. (2014) compared the classification performances of the two unsupervised and 13 supervised classification algorithms. They performed classification based upon different size of training sets and concluded that Classification and Regression Trees (CART) based algorithms except for random forests (RF) and Adaboost are highly affected by the training sample size. Millard and Richardson (2015) evaluated the sampling strategy and sample size of the training data for random forest classification. Their results demonstrated that RF is eminently sensitive to the training sample size and larger sample size should lead to higher accuracy. Shang et al. (2018) investigated the impacts of training sample size on the classification accuracies of the four ML models in an urban fringe area. The classification accuracy of all models except the naive Bayes (NB) increases when the training sample size is increased in their experimental study. Ramezan et al. (2021) assessed the classification accuracy of six machine learning models in response to different size of training samples ranging from a large to a very small sample size. Their results suggest that the classification accuracy of different models varies depending on the training set size. Phinzi et al. (2023) explored the effects of the varying sizes of training sample on land use/cover classification using RF. They reached to a conclusion that to increase the size of training sample leads to a considerable increase on overall accuracy.

It is obviously seen from the aforementioned papers that increasing the size of training samples might lead to an increase on overall accuracy in some cases, especially for tree-based methods, and the training data size has an influence on overall accuracy of supervised classifiers. The relationship between the set size of training data and its influence on the accuracy was assessed by conventional and commonly used ML methods such as SVM, RF, ANN and MLC in above papers. Recently the gradient boosting algorithms such as XGBoost, LightGBM and kernel based methods such as KELM have been extensively used for the classification of remote sensing images however their response and sensitivity to the varying set size of training data in terms of classification accuracy have not been fully explored and comparatively investigated yet. Hence, such analysis and experiments have to be explored with the more recently-introduced methods such as KELM and LightGBM to understand their characteristics and their sensitivity to the training sample size. As far as we are aware, there are not any studies yet that have assessed the relationship between training set size and its influence on classification accuracy by using KELM or LightGBM models for agricultural land cover classification.

Therefore, this paper aims to investigate the impact of the different training set size on supervised machine learning classifiers for the agricultural land cover classification in remote sensing. For this purpose, seven different scenarios on training set size were employed: 1%, 5%, 10%, 20%, 30%, 40%, and 50%. The remaining 50% of the full ground truth data was used for assessing the model accuracy. In our experimental research, four machine learning models (KELM, LightGBM, RF and SVM) were tested for the agricultural land cover classification in Bafra, Türkiye. Specifically, the particular focus of this research is to assess and track how supervised machine learning classifiers react up to different scenarios of varying training set sizes in the context of agricultural land cover classification.

Study area and data

Study area

The study (160km^2) site is situated in Bafra Plain, Samsun, Turkey (35° 47'-35° 58' E, 41° 39' 41° 37' N) (Fig. 1). It is placed in the largest deltas of Turkey. It has a rich

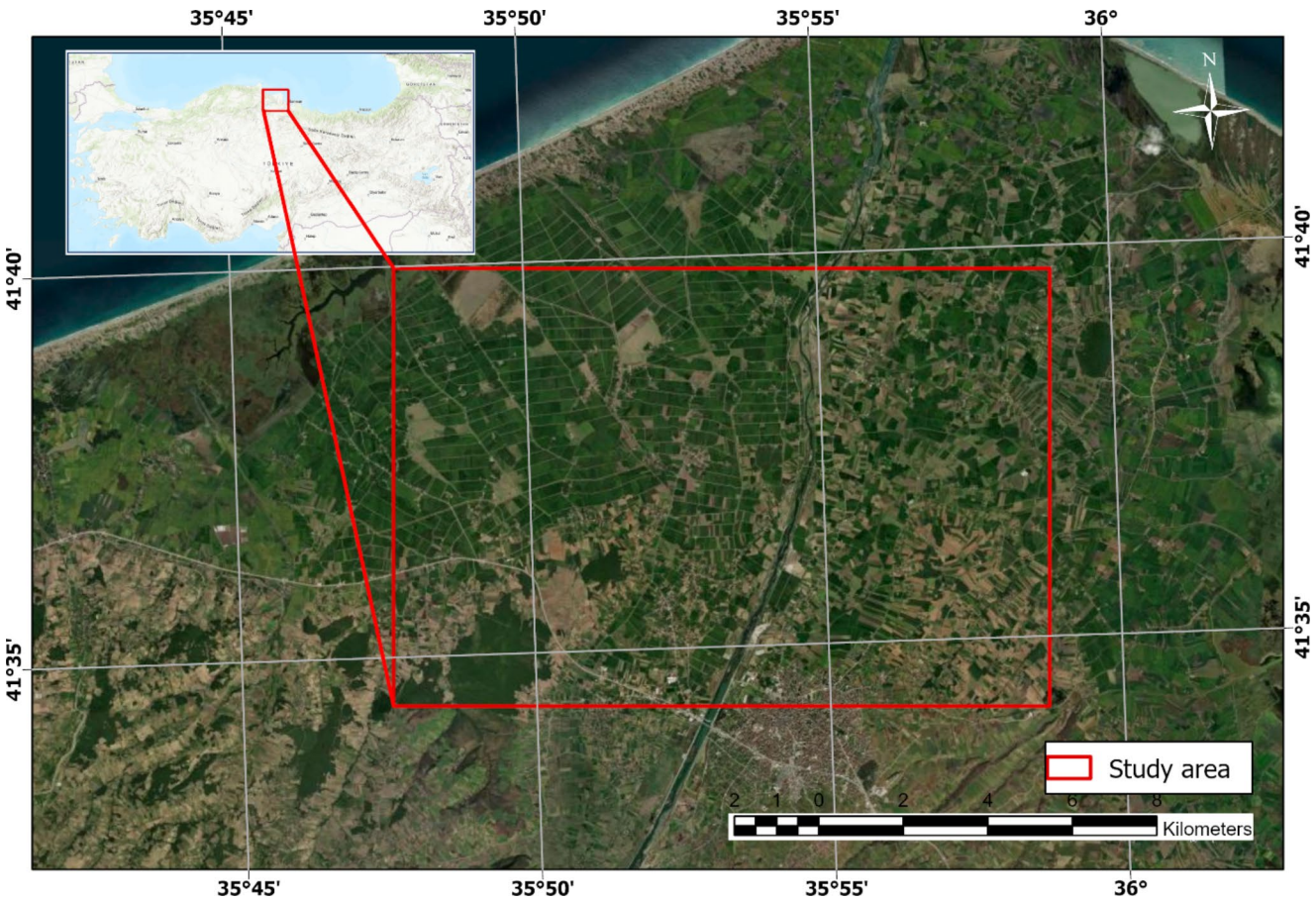

Fig. 1 Study area, Bafra, Samsun, Turkey

biodiversity due to the favorable climate, and sea, lake, marsh, swamp, grassland, pasture and, agricultural areas. The Bafra Plain, where the Kızılırmak flows into the sea, is located in the delta plain and the wetlands formed by the alluvium it carries. As the Bafra Plain has a very large wetland ecosystem and fertile soils, rice is grown in large areas of the region. Approximately the 10% of the rice production in Turkey comes from this region. (Akay et al. 2017)

Satellite data

Sentinel-2 satellites are passive remote sensing satellites launched by the European Space Agency (ESA). Each image has a number of thirteen spectral bands of varying spatial resolutions (10, 20, and 60 m) and a 5-days temporal resolution. The details for the spectral channels of the Sentinel-2 imagery are provided in Table 1.

The Sentinel-2 imagery was acquired as a Top-of-Atmosphere (TOA) reflectance Level-1 C product and processed via SEN2COR (2.12 version used, released on September 23 of 2024) for generating Sentinel-2 Level 2 A product. Level-2 A products provide atmospherically and terrain corrected surface reflectance (Bottom of Atmosphere (BOA)) data which can be used for vegetation analysis and land cover classification. All bands except B1-B9-B10 were incorporated for the classification process as these three bands are mainly used for atmospheric corrections and cloud screening.

The acquisition date of the Sentinel-2 imagery is 10th of September, 2022.

Land parcel identification system (LPIS) and physical block concept

The Integrated Administration and Control System (IACS) is a system by the European Union (EU) to manage and control agricultural subsidies under the Common Agricultural Policy (CAP). The purpose of this system is to use scientific methods to determine the classes of land assets, to define the exact boundaries in accordance with these classes, and to pay agricultural subsidies accurately through an information system with cross-checks and field checks (Inan et al. 2010). IACS benefits from the latest technology such as remote sensing, geographical information systems (GIS), and geo-databases to monitor agricultural land use and verify eligibility for the subsidies. In this context, "Land Parcel Identification System (LPIS)" is used to register agricultural lands (JRC-2001). LPIS is one of the vital components of the IACS and is a geographical database containing reference parcels where agricultural parcels are declared for area-based support (Harvolk et al. 2014).

A number of different approaches based on the farmer's block, the cadastral parcel, and an agricultural parcel and the physical block have been analyzed in European Union countries. The LPIS in the European Union consists of various sorts of reference parcels such as agricultural parcel, farmer block, physical block or the combination there of chosen by each Member State (national or regional) (Şimşek 2023) (See Fig. 2).

In Turkey, because of the country's large scale and heterogeneity of agricultural landscapes, the physical block was preferred as a reference parcel. The physical block refers to a 'production block' and it was prepared by the national administration. In this manner, the physical block refers to the smallest spatial unit of an agricultural activity (Şimşek and Durduran 2023). In the establishment of physical block boundaries; continuous vegetation pattern and continuous land cover (arable land, grassland, continuous vegetation) are accepted as valid boundaries. Physical blocks were divided into 3 according to their degree of permanence. Permanent boundaries; with high degree of permanence; transport networks, artificial surface, water, forest, trees, permanent land boundaries. Relatively permanent boundaries; farm roads, in-parcel roads, grass strips. The boundaries most likely to change; boundaries between agricultural parcels belonging to different and the same crop types.

Table 1 Sentinel-2 spectral channels

Band number	Band description	Central wavelength (μm)	Resolution (m)
B1	Coastol aerosol	0.443	60
B2	Blue	0.490	10
B3	Green	0.560	10
B4	Red	0.665	10
B5	Vegetation Red Edge	0.705	20
B6	Vegetation Red Edge	0.740	20
B7	Vegetation Red Edge	0.783	20
B8	Near Infrared	0.842	10
B8a	Vegetation Red Edge	0.865	20
B9	Water vapour	0.945	60
B10	SWIR - Cirrus	1.375	60
B11	SWIR	1.610	20
B12	SWIR	2.190	20

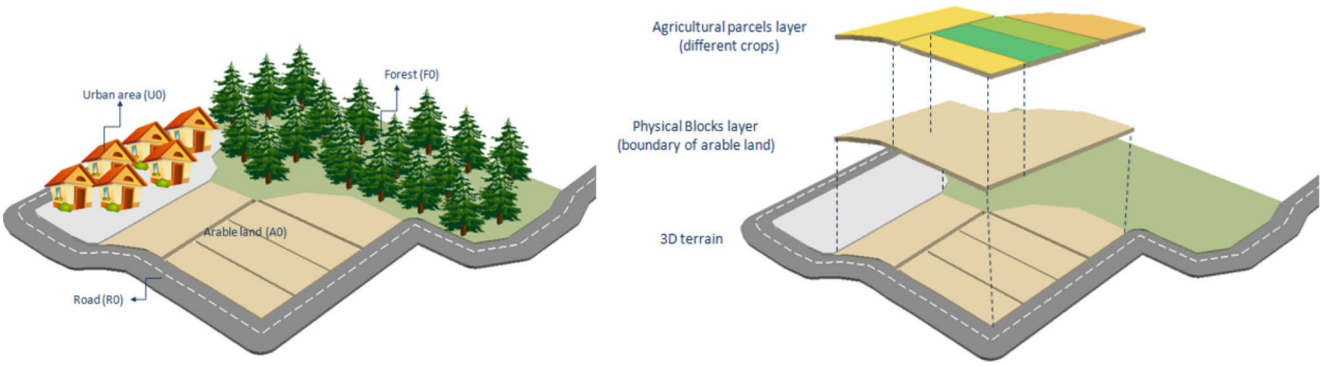

Fig. 2 Summarize the concept of land parcel identification system

Within the scope of LPIS, Photointerpretation and Digitization Guideline were created without digitizing the physical blocks. As part of the Ministry of Agriculture and Forestry's IACS project, physical blocks were drawn, land cover classes were determined and LPIS was created on a national scale using the Digitization Guideline as a reference on 1/5000 scaled aerial photographs in 30 m spatial resolution and satellite images from 2016. LPIS classes are split into two categories which are agricultural and non-agricultural lands and those cover twenty-four subcategories (Fig. 3). In Fig. 3, blue color represents agricultural land, light blue color represents farmer's block (arable lands, permanent tree crops, mixed agricultural areas, grasslands etc.) and yellow represent physical blocks (arable land, rice fields, greenhouses, tea, hazelnuts, vineyards, olives etc.), based on the LPIS, Photointerpretation and Digitization Guideline.

Creating ground truth data from LPIS

IACS and LPIS data enables us to study in a nation-wide scale for the land use change in the spatial and temporal domain (Tomlinson et al. 2018). In this experimental research, the physical blocks created by LPIS were utilized as a reference data. As rice fields cover the majority areas in the test site, they were generated as a separate class from agricultural land. Due to there are no shrub land, tree crops and greenhouses classes within the study area, they are not included in the classification process (Table 2).

Physical blocks are in polygon type and vector format, and polygons with an area of less than 5000 m^2 were not incorporated into classification. As an initial processing, the vector data was converted to a raster format and 10 m negative buffer was performed for each class in raster data. Through that way, the noise and border pixels are primarily removed (Fig. 4).

Table 3 shows the number of polygons and pixels for each class included in the classification study. The pixels corresponding to 1-5-10-20-30-40-50% of the reference data were selected as training data (Table 3) and the pixels corresponding to 50% of the reference data were selected as test data and used in the classification study.

Image classification

In this section, basic principles of the machine learning models (SVM, RF, LightGBM and KELM) performed in this research study are provided with the parameter setting. The classification process is performed in Python environment.

Random forest

The Random Forest (RF) has been commonly preferred for the classification and regression problems. RF, which is an ensemble classifier, uses multiple decision trees and is the improved version of bagging (bootstrap aggregating) which splits per node by using the best split among the variables (Pal 2005; Gislason et al. 2006). However, the input variables (i.e., a subset of predictors) are randomly chosen for building the trees in RF. After the training process, the decision for assigning each unknown sample to a specific class is made based upon the majority voting of the ensemble (Rodriguez-Galiano et al. 2012; Akar and Güngör 2015).

The RF algorithm utilizes CART to generate the decision trees, where uses a criterion (e.g. Gini index) to perform the splitting tree nodes. The Gini index can be written as

$$\sum \sum_{j \neq i} \left(f\left(C_i, T\right) / |T|\right)\left(f\left(C_j, T\right) / |T|\right),$$

where T denotes the training set, C_i represent the class for a random pixel and $f\left(C_i, T\right) / |T|$ is the probability for class C_i (Pal 2005; Akar and Güngör 2015).

For the utilization of the RF model, two model parameters were needed to be tuned by the user. These parameters are the number of trees to grow (N) and number of variables (mtry) for splitting at each node (Pal 2005;

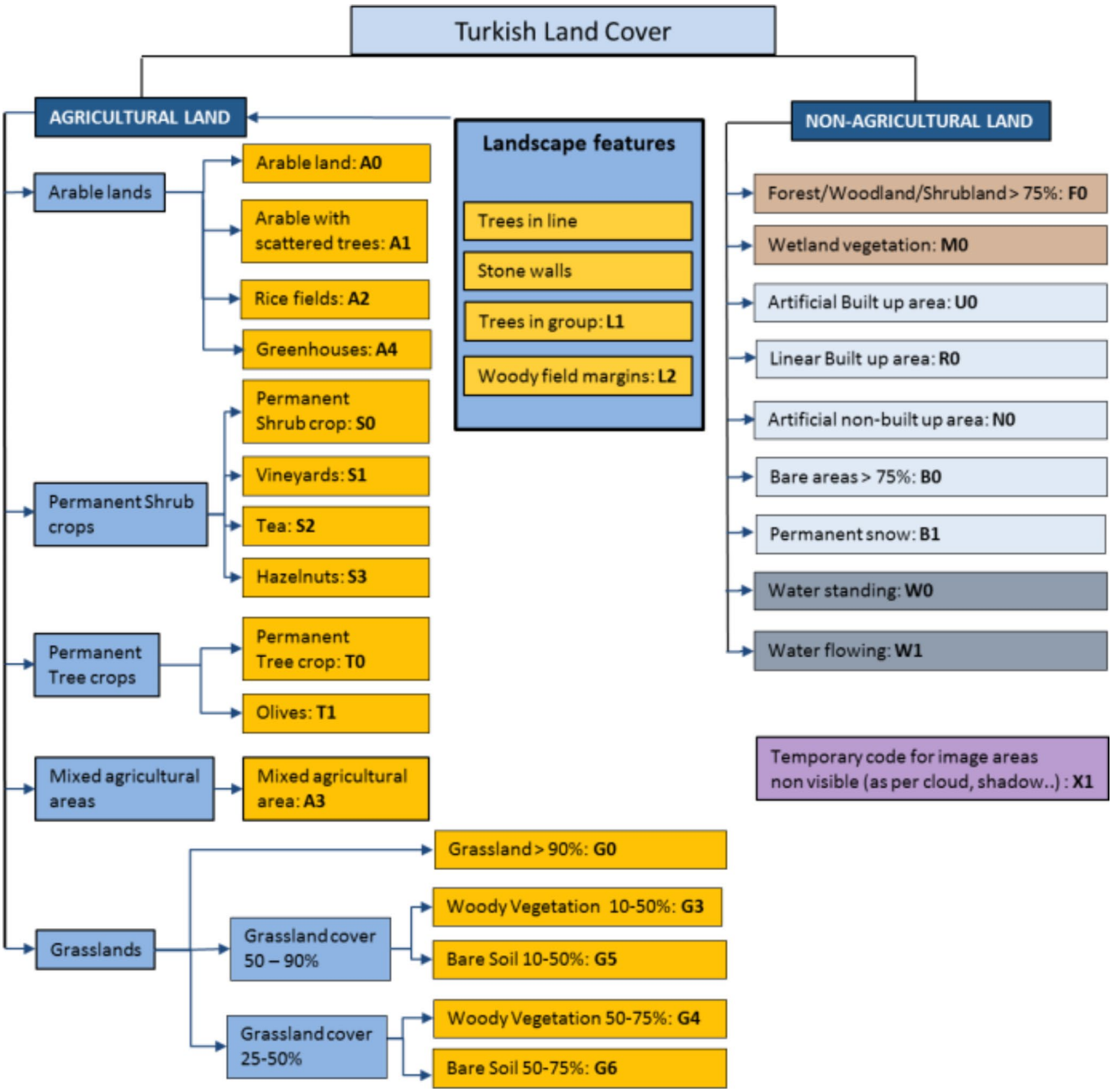

Fig. 3 LPIS classes in Turkey

Millard and Richardson 2015). In our experiment, the user-defined parameters were set to 800 and 5 for N and mtry, respectively.

Light gradient boosting machines

The LightGBM algorithm is the enhanced version of the Gradient Boosting Decision Tree algorithm (GBDT), and getting popular due to offering superior performance with respect to other ensemble learning algorithms. What makes LightGBM special and distinct from other tree-based learning algorithms is about the tree growth strategy (Ustuner and Balik Sanli 2019; McCarty et al. 2020; Jafarzadeh et al. 2021). LightGBM grows trees leaf-wise (i.e., vertically), while other ensemble learning algorithms which grow trees in level-wise (i.e., horizontally) (Ke et al. 2017; Liu et al. 2017). This is the main factor of the LightGBM's superb performance in reducing the computation time and memory utilization (Ke et al. 2017; McCarty et al. 2020; Chen et al. 2021).

The GBDT is commonly used in machine learning and can achieve high prediction accuracy; however, it's

Table 2 LPIS class grouping for land cover classification

Proposed LC classes	Original LC class	Code and name of land cover
Arable land	Land cover type A0	A0-arable land
	Land cover type A1	A1-arable with scaltered trees
	Land cover type A3	A3-mixed agricultural area
	Land cover type G0	G0-grassland > %90
Rice	Land cover type A2	A2-rice field
Grassland	Land cover type B0	B0-bare areas > %75
	Land cover type G5	G5-grassland cover %50–90 Bare soil %10–50
	Land cover type G6	G6-grassland cover %25–50 Bare soil %50–75
Forest	Land cover type G3	G3-grassland cover %50–90 Woody veg. %10–50
	Land cover type G4	G4-grassland cover %25–50 Woody veg. %50–75
	Land cover type F0	F0-forest/woodland/shrubland > %75
Artificial surface	Land cover type N0	N0-artificial non-built up areas
	Land cover type R0	R0-linear built up areas
	Land cover type U0	U0-artificial built up areas
Water	Land cover type M0	M0-wetland vegetation
	Land cover type W0	W0-water standing
	Land cover type W1	W1-water flowing

time-consuming and computationally complex. LightGBM could address the issues encountered by GBDT with its two new functionalities: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (Ke et al. 2017; Candido et al. 2021; Chen et al. 2021). GOSS is a new sampling approach that down samples the data instances based on the gradients of the loss function. It keeps instances of large gradients and then remove the instances with small gradients by random. GOSS uses a subset of the data based on the magnitude of these gradients. And EFB is primarily designed for reducing the feature dimension (i.e., number of features) and bundles the continuous (e.g. exclusive) features into a single feature (Ke et al. 2017; Ustuner and Balik Sanli 2019; Chen et al. 2021). This step reduces the computational complexity and hence accelerates the training process of the model. Hyper parameters of LightGBM are provided in Table 4.

Kernel extreme learning machines

The neural network classifiers have been extensively utilized in remote sensing since the 1990s. Among the different learning strategies, the back-propagation algorithm, proposed by Rumelhart et al. (1986), is the most preferred neural network model in remote sensing however there are crucial stages in model tuning and structural design of the backpropagation neural classifiers to maximize the performance (i.e., tuning of learning rate, momentum, number of nodes in each layer, and number of hidden layers etc.) (Benediktsson et al. 1990; Paola and Schowengerdt 1995; Kavzoglu and Mather 2003; Kavzoglu 2009). This method relies on iterative gradient descent training, wherein the error is back-propagated from output layer into input layer. There was also reported in many studies that network training parameters are of a great impact on the trained network performance and the learning algorithm. During the training stage of the backpropagation algorithm, it could encounter the local minima and might result in over-train. Another important drawback for back-propagation algorithm is its slow convergence rate (Pal 2009; Kavzoglu 2009). In order to overcome these challenges, Huang et al. (2006) introduced the single-hidden layer feed-forward neural networks (SLFN) model known as extreme learning machines (ELM), where input weights and biases are assigned at random. ELM generates the hidden node parameters randomly rather than iterative process, which makes the learning quite fast (Pal 2009; Kavzoglu 2009; Pal et al. 2013; Chen et al. 2014). In their research, Huang et al. (2006) highlighted that ELM outperformed the backpropagation neural network and does not need the tuning of the input weights, unlike conventional backpropagation algorithm. Within the last three decades, SVM and RF are the most preferred ML methods for the classification of remote sensing images. SVM, which is a kernel-based learning classifier, has been extensively preferred in hyperspectral image classification due to tackling the high dimensionality problem, which leads to achieve high classification performance (Foody and Mathur 2004a, b; Bazi and Melgani 2006; Chen et al. 2014).

The idea of utilizing the kernel functions attracted many researchers to develop the kernel-based approaches since kernels can map the data from input space into feature space (i.e. high dimensional space) (Pal et al. 2013). For this goal, Huang et al. (2012) introduced the kernel extreme learning machines (KELM) for classification. ELM exploits the smallest norm least squares for the adjustment of the input weights. In this step, the randomized assignment of the input weights and bias are held, which results in a significant change in the accuracy with the use of the same number of hidden nodes (Pal et al. 2013; Chen et al. 2014). To tackle this problem, Huang et al. (2012) modified the ELM hidden layer through a kernel function and proposed the use of kernels in ELM structure, and hence the need of randomness in assigning the weights are thereby eliminated. SVM needs to satisfy the Mercer's theorem however ELM does not need to meet such a condition. Thus, KELM can provide a cohesive solution for multiclass classification problems (Pal et al. 2013; Chen et al. 2014; Ergul and Bilgin 2020).

There are two different parameters to be tuned for KELM, which are kernel bandwidth γ and penalty parameter C.

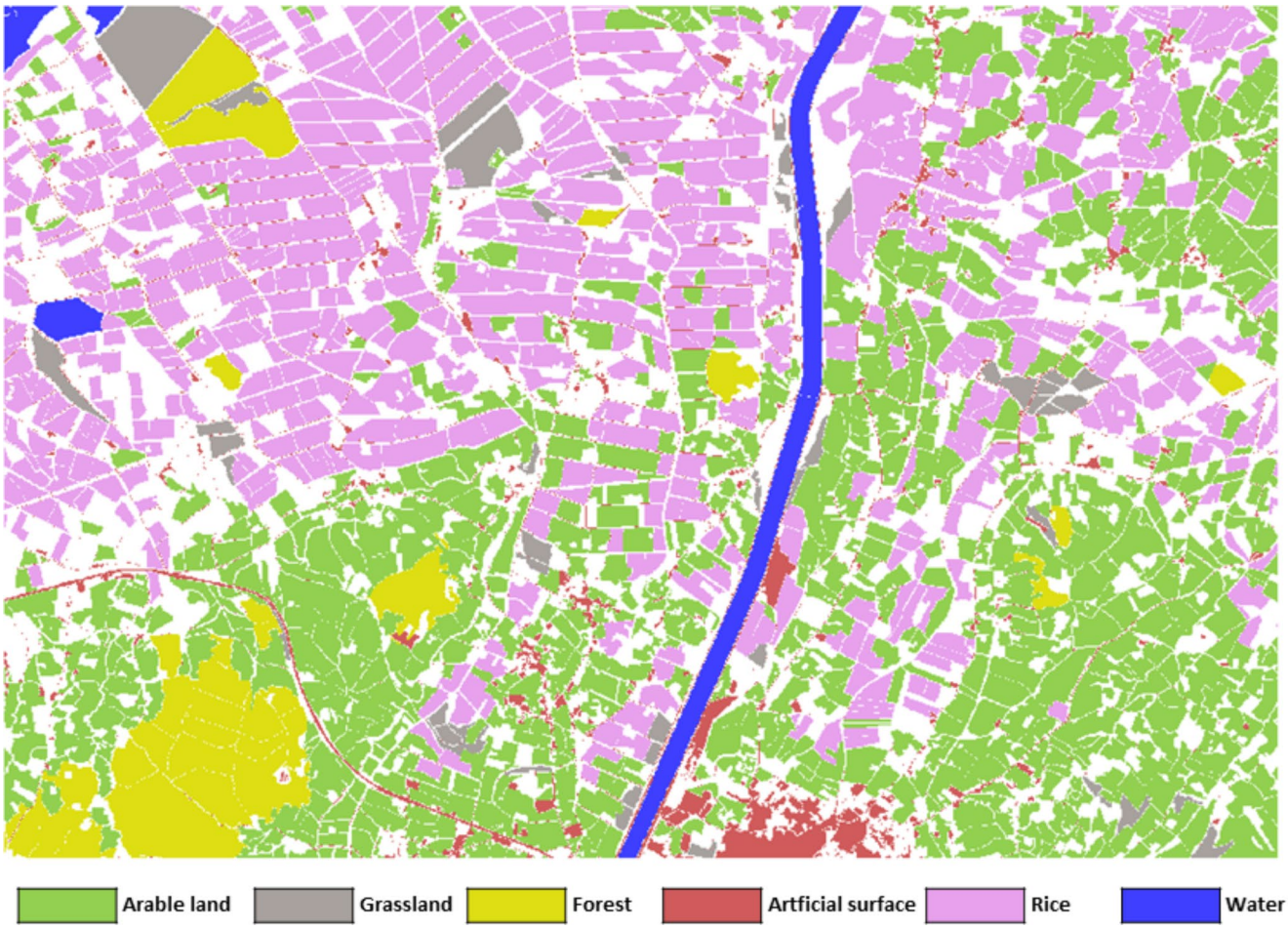

Fig. 4 Ground truth data for study area

Table 3 Number of polygons and pixels for each classes

Class name	Polygon number	Total pixel number	Train 1%	Train 5%	Train 10%	Train 20%	Train 30%	Train 40%	Train 50%	Testing data
Arable land	761	482,916	4809	24,043	48,374	96,523	144,471	193,047	241,588	241,328
Artificial surface	1676	48,668	505	2468	4906	9845	14,517	19,435	23,992	24,676
Forest	129	81,817	762	4181	8136	16,550	24,718	32,756	40,975	40,842
Grassland	131	41,403	422	2142	4188	8094	12,241	16,477	20,643	20,760
Rice	574	431,987	4355	21,423	43,077	86,408	130,044	173,017	216,130	215,857
Water	61	31,927	334	1678	3190	6323	9624	12,755	16,031	15,896
Total	**3332**	**1,118,718**	**11,187**	**55,935**	**111,871**	**223,743**	**335,615**	**447,487**	**559,359**	**559,359**

Table 4 LightGBM parameters

Boosting type	GOSS
Maximum depth	7
Number of leaves	800
Number of estimators	1500
Number of samples for constructing bins	350,000
Learning rate	0.2

These parameters were set to 0.1 and 500 for γ and C, respectively.

Support vector machines

Support Vector Machines (SVM), based on statistical learning theory developed by Cortes and Vapnik (1995), was primarily designed as a linear binary classifier in two-dimensional space by defining the optimum hyperplane with maximum margin (i.e. decision boundary). The SVM was commonly used for the classification and regression

problems in remote sensing and demonstrated outstanding performance in many applications, especially for the classification of hyperspectral images over the last three decades. In case of the two classes are not linearly separable in the input space, the SVM benefits the kernel trick by introducing the new variables (called slack variables) in SVM optimization. Through the kernels, SVM transform the data points into higher dimensional space and hence turns to non-linear boundaries to linear ones in the higher dimensional space (Melgani and Bruzzone 2004; Foody and Mathur 2006; Kavzoglu and Colkesen 2009; Mountrakis et al. 2011).

The popularity of SVM in remote sensing/pattern recognition comes from its great ability of handling small training dataset in producing higher classification accuracy compared to conventional methods in image classification. This is because SVM only uses the support vectors (i.e. a subset of points that lie on the margin) to place the hyperplane with maximum margin (Kavzoglu and Colkesen 2009; Mountrakis et al. 2011). In our experiment, the linear kernel was used as a kernel function in SVM classification. There is one parameter (penalty parameter C) to be tuned for SVM, which was set to 300.

Experimental results and discussion

In this section, the overall accuracies (OA) for the classification models are presented, based on varying training set size. Figure 5 displays the overall accuracies, which are derived from the error matrix (confusion matrix), for each classification method. The highest accuracy was received by the LightGBM (89.93%), followed by RF (86.49%), KELM (78.38%) and SVM (72.49%).

$$\mathrm{OA}=\frac{\text{Number of correctly classified instances}}{\text{Total Number of instances classified}} \quad (1)$$

In this experimental research, the classification accuracy is increased with the increasing training set size, with an exception of KELM and SVM which have unstable results in changing the set size of the training data. The highest classification accuracies for LightGBM and RF were obtained with the 50% training set size while KELM and SVM received the highest accuracy with 10% and 20% training set size, respectively. The lowest classification accuracies for all methods, except SVM, were received by using the 1% set size. The SVM produced the lowest classification accuracy by using the 40% set size (Fig. 5). The classified images are also depicted in Fig. 6.

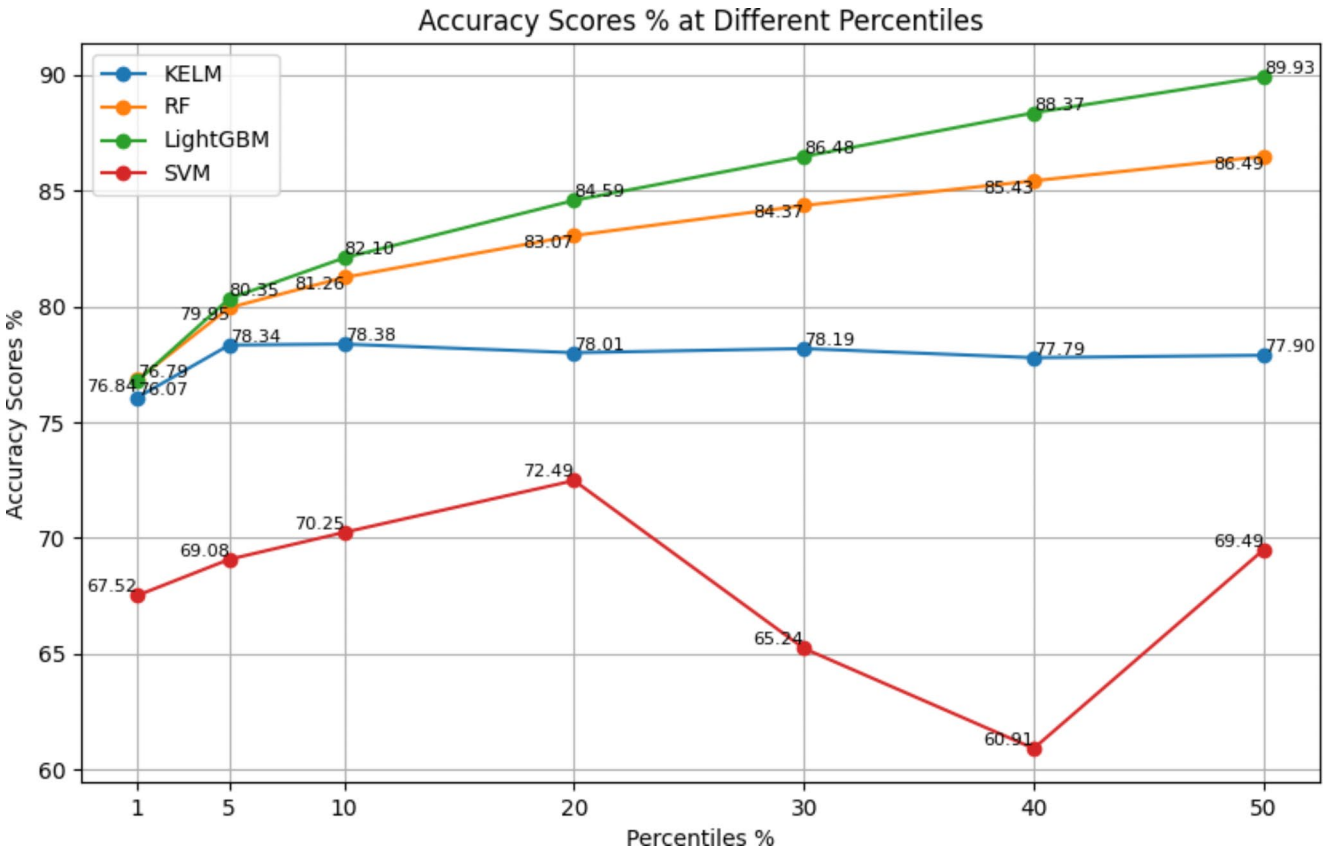

Fig. 5 Overall accuracies

In order to understand what classes led to increase or decrease in the overall accuracy, the class-based accuracies were assessed for each classification method. For this purpose, F1-score which is defined as the harmonic mean of the producer accuracy (PA) and user accuracy (UA) was calculated, in Eq. (2).

$$\text{F1-score}=2*\frac{(\text{PA}*\text{UA})}{(\text{PA}+\text{UA})} \quad (2)$$

Generally speaking, the F1-scores for all classes in LightGBM classification were increased with different rates by increasing the set size. Among the all classes, the water received the highest increase in F-1 score from 0.59 to 0.84, followed by grassland from 0.65 to 0.85 by changing the training set size from 1 to 50% (Fig. 7). Rice and arable land have almost same upward trend to the increasing training set size. The most stable (i.e. small change in F1-score) class among the others is forest, which is only increased from 0.91 to 0.97. The F-1 score for the forest class remains unchanged by increase the training set size from 20 to 30%. The lowest F-1 scores for all classes in LightGBM classification were received by the use of 1% training set size (Fig. 7). The forest and rice classes obtained the highest F-1 scores among other classes in every training set size in LightGBM classification. The gradually increase of the training set size has a different impact and sensitivity on the F-1 score for each class.

Similar to LightGBM, the F1-scores for all classes in RF classification were increased with different rate by increasing the set size. Among the all classes, the water received the highest increase in F-1 score from 0.56 to 0.77, followed by grassland from 0.65 to 0.80 and artificial surface from 0.58 to 0.73 by changing the training set size from 1 to 50% (Fig. 8). Rice and arable land have almost same upward trend to the increasing training set size, in a similar way of LightGBM. The most stable (i.e. small change in F1-score) class among the others is forest, which is only increased from 0.90 to 0.95. The F-1 score for the forest class remains unchanged by increase the training set size from 30 to 40%. The lowest F-1 scores for all classes in RF classification were received by the use of 1% training set size (Fig. 8). The forest and rice classes obtained the highest F-1 scores among other classes in every training set size in RF classification. Similar to LightGBM, the gradually increase of the training set size has a different impact and sensitivity on the F-1 score for each class in RF classification (Fig. 8).

For KELM, to change the training set size has different impacts on F1-scores for each class. While the F1-score for some particular classes such as artificial surface was increasing by increasing set size, it was decreasing or remain unchanged for some other classes such as forest or water. Even though the KELM in terms of overall classification accuracy is stable, the F-1 score for some particular class (water) demonstrates unsteady behavior. This led to the necessity for the careful selection of the training set size for agricultural land cover classification. Among the all classes, the artificial surface received the highest increase in F-1 score from 0.43 to 0.63. The F-1 score for forest class remained unchanged, which means there is no impact of changing set size on the classification of forest in our experimental research. However, the increase of the training set size might have led to a decline for some particular class. For instance, the water received a decline in F-1 score from 0.61 to 0.56 by changing the training set size from 10 to 50% (Fig. 9). The lowest F-1 scores for each class in KELM classification were received by the use of 1% training set size, with an exception of water class. Similar to LightGBM and RF, the forest and rice classes obtained the higher F-1 scores among other classes in KELM classification (Fig. 9).

For SVM, there are some particular classes such as artificial surface and grassland which are completely misclassified for the training set size 1% and 10% however demonstrated sharp increase on F1-score for the 5% training set size from 0 to 0.22 for artificial surface and from 0 to 0.42 for grassland (Fig. 10). Except the use of 5% training set size, the artificial surface is nearly misclassified (0%) in SVM classification and this led to receive low overall accuracy. The changing training set size has a remarkable impact on F1-score for all classes except rice. The F-1 scores for the rice class do not show any considerable change based on the training set size. The lowest F-1 scores for artificial surface, grassland and water were received by the use of 1% training set size in SVM classification. Besides, the lowest F1-scores for arable land and rice were received by the use of 40% training set size (Fig. 10). From these observations, it can be clearly seen that SVM is very sensitive to the training set size changes and each class of interest might react very differently from each other. The increase of the training set size from 30 to 40% might have led to a decline on F-1 score for some particular class such as grassland and arable land. However, the increase of the training set size from 40 to 50% might have led to an increase on F-1 score for some particular class such as water and arable land. Similar to other models employed, the forest and rice classes obtained the higher F-1 scores among other classes in SVM classification (Fig. 10).

Discussion

Based on the observations presented above, it can be inferred that training set size has varying impacts and sensitivities on both overall accuracy and F-1 scores for agricultural

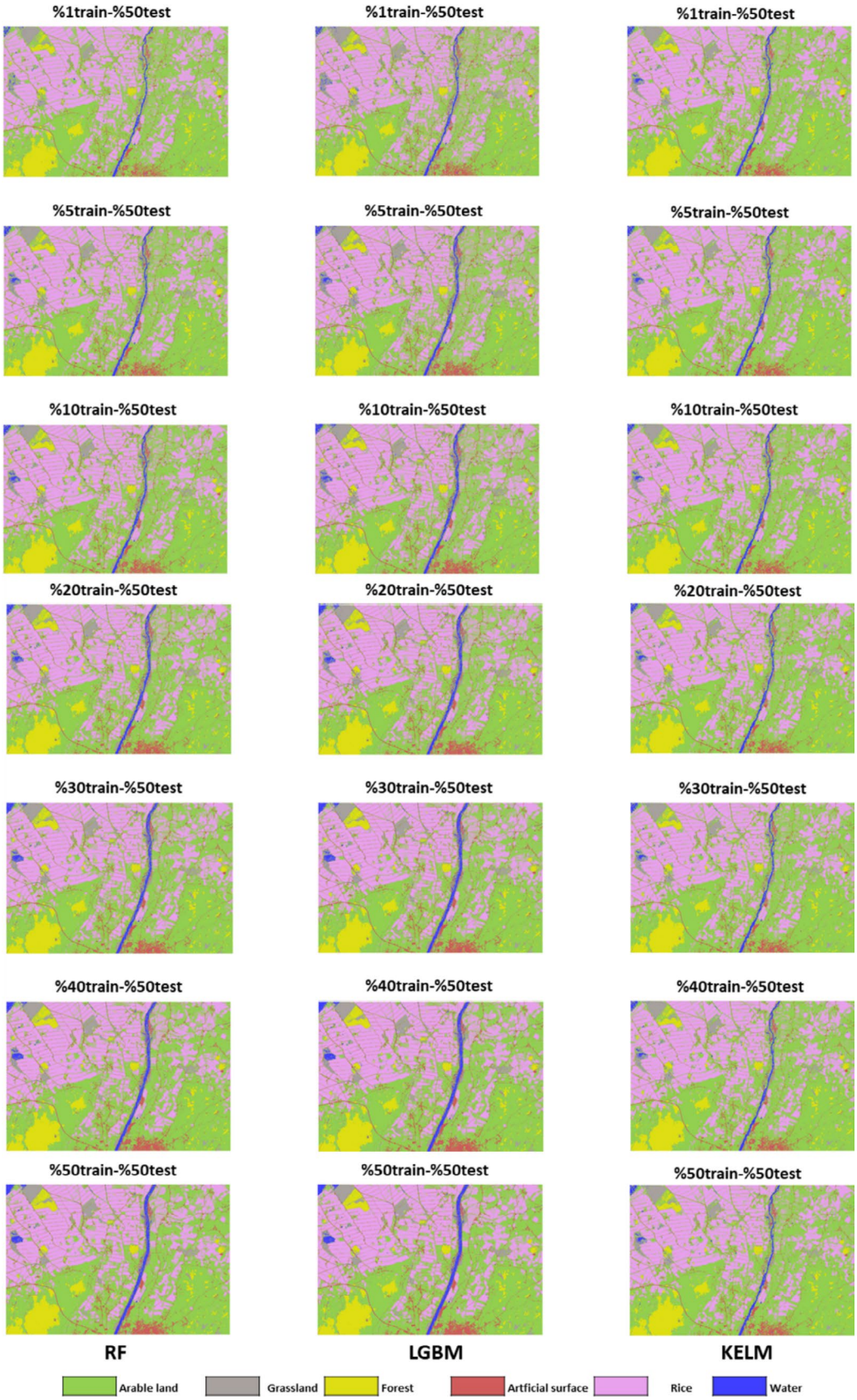

Fig. 6 Classified images

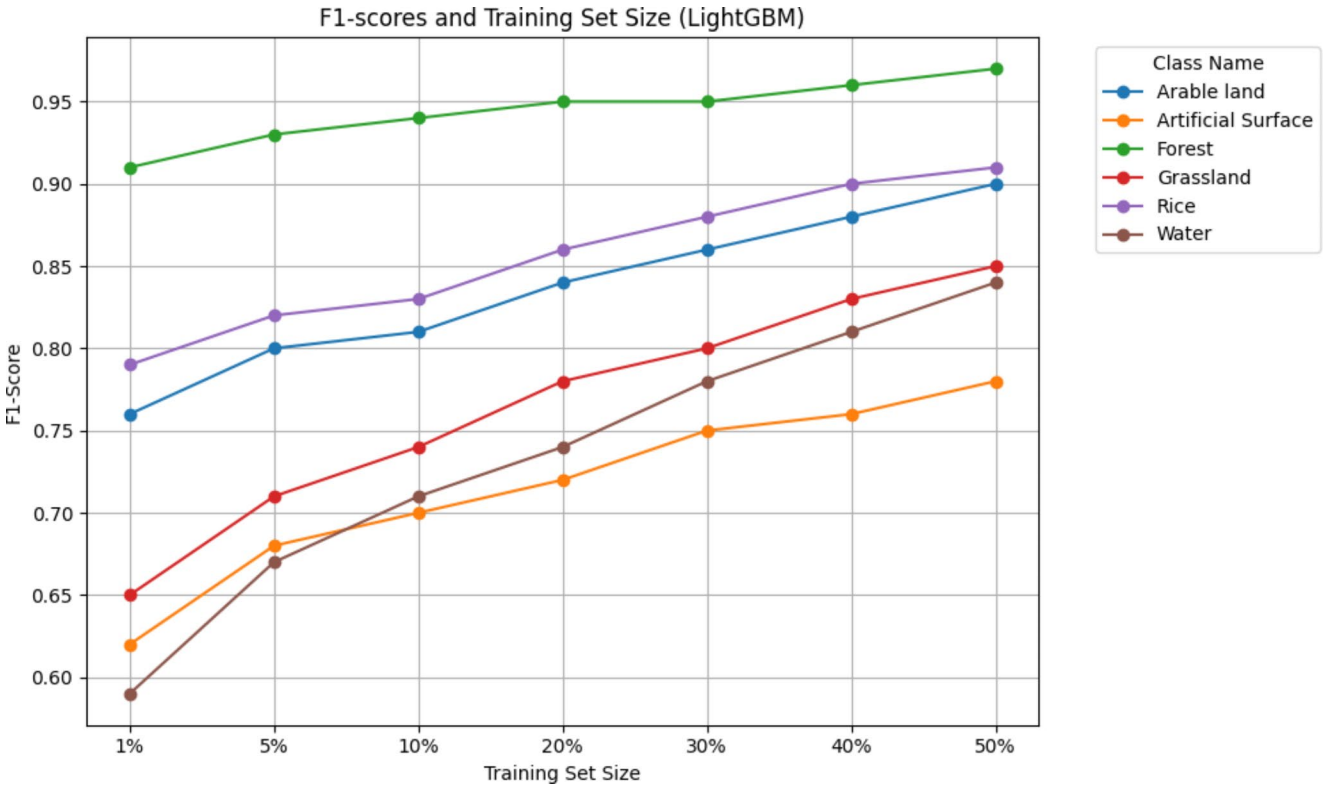

Fig. 7 F1-scores and training set size for LightGBM

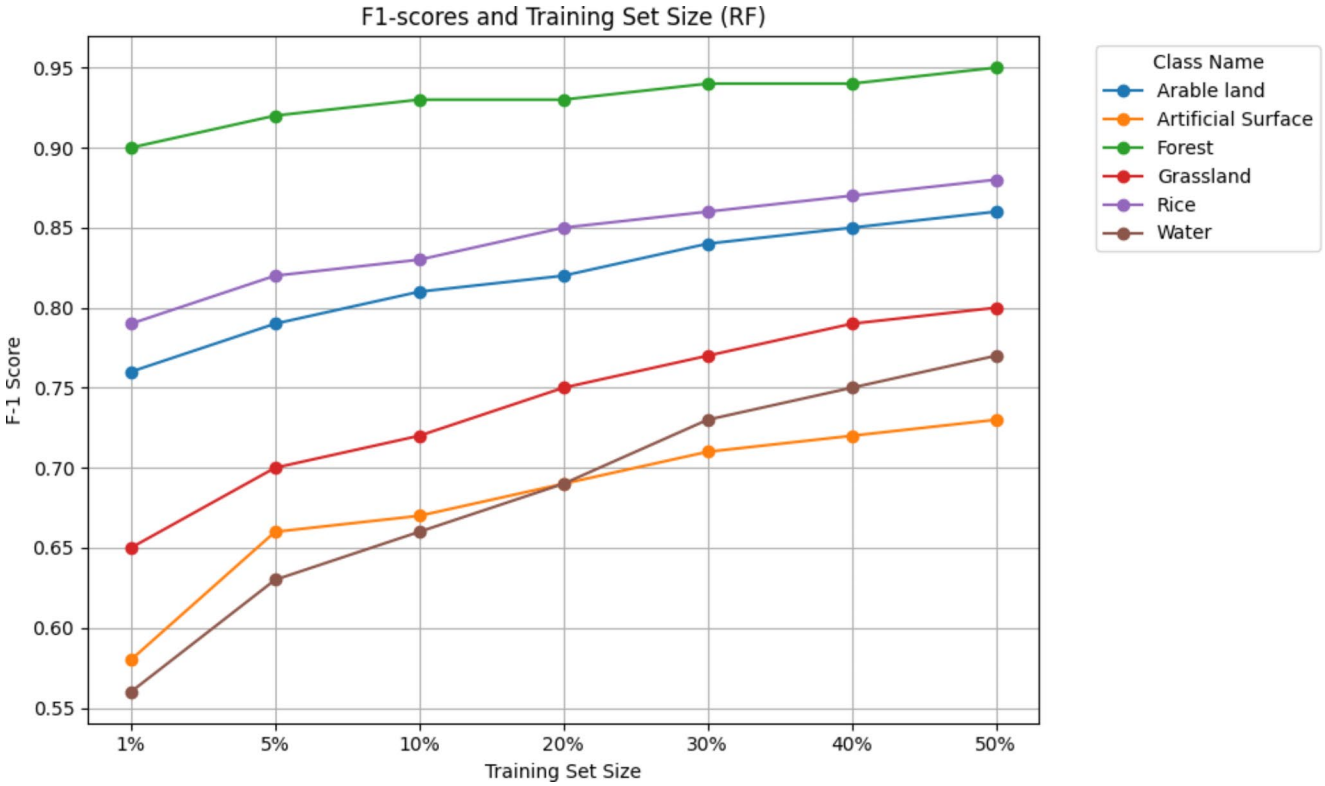

Fig. 8 F1-scores and training set size for RF

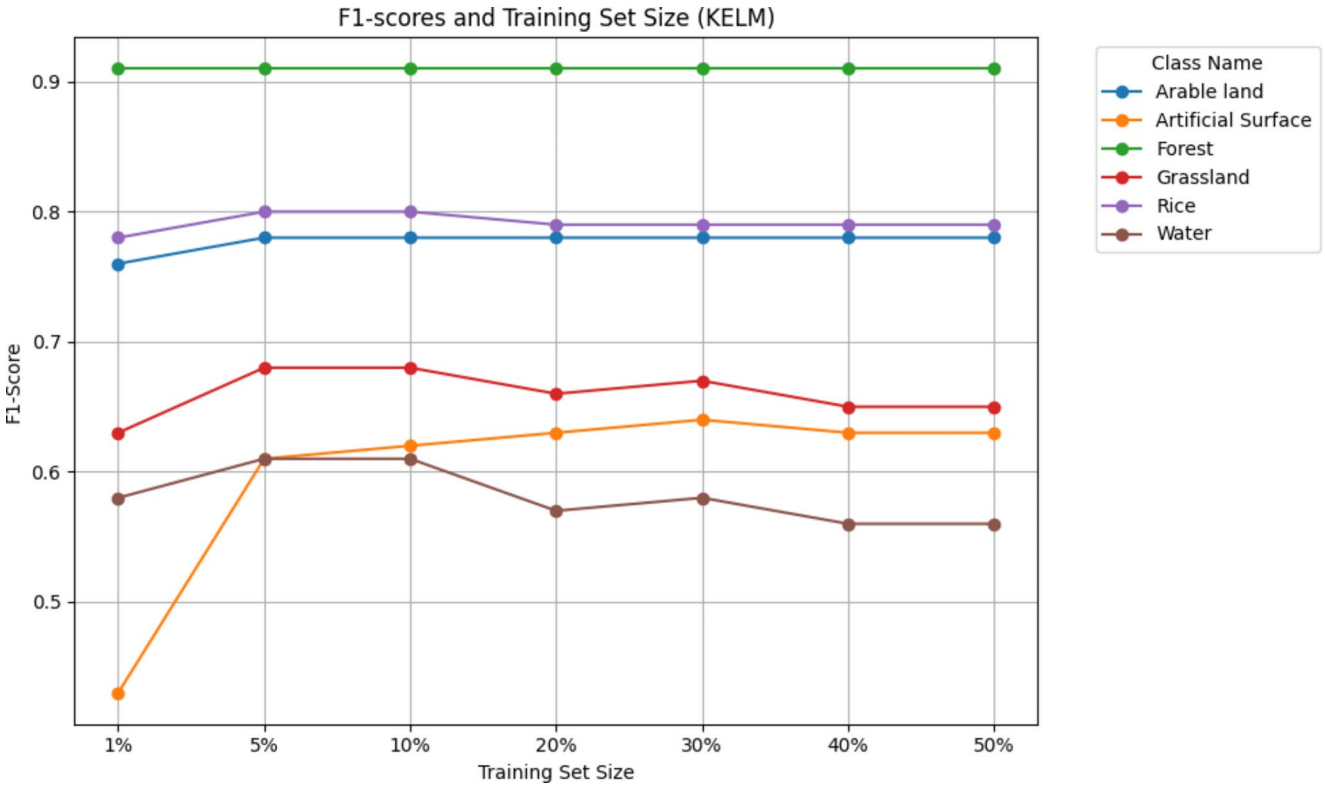

Fig. 9 F1-scores and training set size for KELM

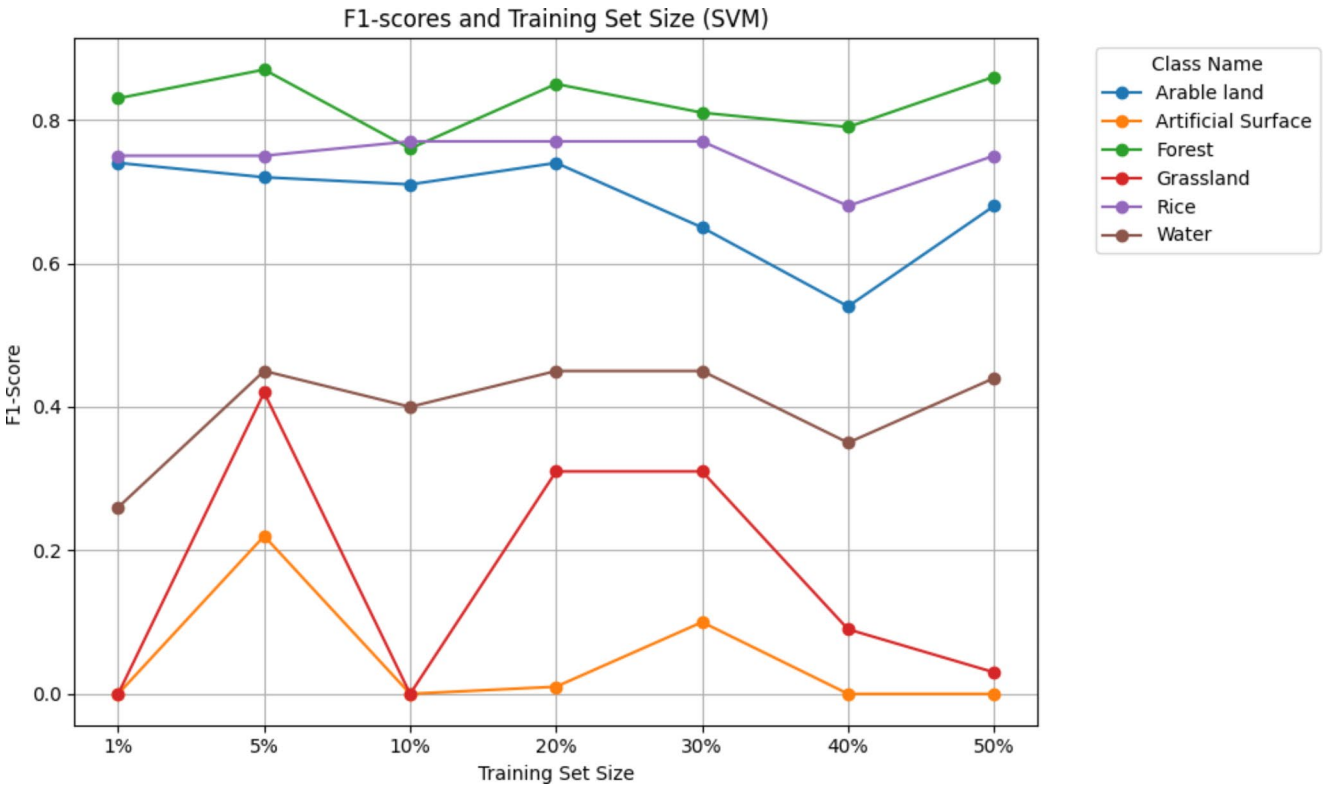

Fig. 10 F1-scores and training set size for SVM

land cover classification in our experimental research. Each machine learning model applies different decision rule to distinguish a particular class from other classes even though same dataset and training data were employed. The learning strategy for each classification model may vary due to the differences in decision-making process depending upon the training data purity, set size, landscape heterogeneity and class-confusion. For instance, arable land and rice classes are sensitive to varying training data size for RF, SVM and LightGBM, however are not much sensitive (i.e. stable) for KELM when we checked the Figs. 7, 8, 9 and 10. Furthermore, they are very sensitive and demonstrated unstable results in SVM even though SVM is also a kernel-based method like KELM. The forest and rice classes obtained comparatively higher F-1 scores from other classes in each ML model. This might be because these two classes simply spectrally-separable from the other classes. In such a case, a basic ML model (i.e. clustering) can easily distinguish these classes from others and changing the training set size do not have much impact on F-1 score. This might be also explained by the nature of real-world data of land cover, which has a dynamic structure. In most cases, the larger training dataset could derive higher classification performance compared to those obtained from the smaller training set such as ANN. However, this is not correct in every condition because of the complexity and uncertainty in the supervised learning for remote sensing classification. For instance, the F-1 score for water class was increased in LightGBM and RF classification however it was decreased in KELM when training set size was increased from 1 to 50%. Another interesting result is about the classification of artificial surface and grassland in SVM. There are some particular classes such as artificial surface and grassland which are completely misclassified in SVM for the training set size 1% and 10%, however demonstrated sharp increase on F1-score for the 5% training set size from 0 to 0.22 for artificial surface and from 0 to 0.42 for grassland (Fig. 10). Except the use of 5% training set size, the artificial surface is almost misclassified (0%) in SVM classification and this led to receive low overall accuracy.

F1-scores for all classes were increased in tree-based methods (LightGBM and RF) as the training set size grew, however this is not correct for KELM and SVM. This is because each classification method applies a different decision rule to separate the classes and each decision rule has a different sensitivity and response to any particular class due to their class purity and its adaptation to the learning algorithm. These particular results clearly address the necessity of the careful and appropriate design of the training set size (i.e. complete description of the spectral response for each class) for accurate agricultural land cover classification.

Conclusion

This research investigates the impact of the different training set size on supervised machine learning classifiers for the agricultural land cover classification in remote sensing. The experimental results demonstrated that the highest classification accuracy was achieved by LightGBM (89.93%), and followed by RF (86.49%), KELM (78.38%) and SVM (72.49%). The classification accuracies of tree-based methods (RF and LightGBM) increased as the training set size grew, however, kernel-based methods (KELM and SVM) exhibited unstable results as the size of the training dataset varied. Typically, the main desire in training a supervised learning algorithm is to provide a fully representative samples for each class (i.e. a complete description for each class in feature space), which is practically achieved by a large training set. For instance, artificial neural networks require a large number of samples to train the neural networks and characterize the classes. When considered from this point of view, it was expected to have an upward trend in KELM classification accuracy similar to RF and LightGBM due to the fact that KELM is based on neural networks. However, the classification accuracy for KELM has not been increased when training set size grew. A likely explanation for this unforeseen outcome is that KELM, as a natural extension of ELM to kernel learning, benefits from the kernels to define the decision boundary separating the classes. The kernels in KELM possibly changed the way of how typical artificial neural networks models need a large number of training data samples.

Even though the changes in overall accuracy provide general overview of how the training set size impacts the image classification, it was necessary to understand what classes led to increase or decrease in the overall classification accuracy. Based on the observations presented above, it can be inferred that training set size has varying impacts and sensitivities on both overall accuracy and F-1 scores for agricultural land cover classification in our experimental research, especially for some particular classes such as water or rice. For instance; while the F-1 score for the water class was increased in LightGBM and RF classification, there was a decline in KELM and an instability in SVM classifications; hence, the appropriate selection and design of the training set size are essential for accurate land cover classification.

To the best of our knowledge, this research investigated the impact of the training set size, for the first time, for KELM classification as well as compared its performance with LightGBM, RF and SVM for agricultural land cover classification. Owing to the changing structures of agricultural land cover, it has been challenging to generalize the results. However, our experimental findings could provide a

general idea and recommend further directions for designing/planning the training set size and its sensitivity in KELM in terms of land cover classification.

Acknowledgements The authors would like to thank the General Directorate of Agricultural Reform, Ministry of Agriculture and Forestry in Turkey for providing the physical blocks used in this experimental research.

Author contributions All authors contributed to the study's conception and design. Material preparation and data collection were performed by Fatih Fehmi Simsek. The data processing and analysis were conducted by all authors. The first draft of the manuscript was written by Mustafa Ustuner, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Akar Ö, Güngör O (2015) Integrating multiple texture methods and NDVI to the Random Forest classification algorithm to detect tea and hazelnut plantation areas in northeast Turkey. Int J Remote Sens 36(2):442–464. https://doi.org/10.1080/01431161.2014.995276

Akay H, Sezer İ, Mut Z, Dengiz O (2017) Bafra Ovası Sol Sahilinde Yetiştirilen Bazı Çeltik Çeşitlerinin Verim Ve Kalite Performanslarının Belirlenmesi. KSÜ Doğa Bilimleri Dergisi 20:297–302. https://doi.org/10.18016/ksudobil.349264

Amarsaikhan E, Enkhjargal D, Jargaldalai E, Amarsaikhan D (2024) Comparison of machine learning and parametric methods for the discrimination of urban land cover types. Geocarto Int 39(1). https://doi.org/10.1080/10106049.2024.2380372

Aplin P (2006) On scales and dynamics in observing the environment. Int J Remote Sens 27(11):2123–2140

Bazi Y, Melgani F (2006) Toward an optimal SVM classification system for hyperspectral remote sensing images. IEEE Trans Geosci Remote Sens 44(11):3374–3385

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogrammetry Remote Sens 114:24–31

Benediktsson JA, Swain PH, Ersoy OK (1990) Neural network approaches Versus Statistical methods in classification of Multisource Remote Sensing Data. IEEE Trans Geosci Remote Sens 28(4):540–552

Candido C, Blanco AC, Medina J, Gubatanga E, Santos A, Ana RS, Reyes RB (2021) Improving the consistency of multi-temporal land cover mapping of Laguna lake watershed using light gradient boosting machine (LightGBM) approach, change detection analysis, and Markov chain, vol 23. Society and Environment, Remote Sensing Applications, p 100565

Chen C, Li W, Su H, Liu K (2014) Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine. Remote Sens 6(6):5795–5814

Chen C, Yan J, Wang L, Liang D, Zhang W (2021) Classification of Urban Functional areas from Remote sensing images and time-series user Behavior Data. IEEE J Sel Top Appl Earth Observations Remote Sens 14:1207–1221

Colkesen I, Ozturk MY (2022) A comparative evaluation of state-of-the-art ensemble learning algorithms for land cover classification using WorldView-2, Sentinel-2 and ROSIS imagery. Arab J Geosci 15(10):942

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Ergul U, Bilgin G (2020) MCK-ELM: multiple composite kernel extreme learning machine for hyperspectral images. Neural Comput Appl 32(11):6809–6819

Foody GM, Mathur A (2004a) A relative evaluation of multiclass image classification by support vector machines. IEEE Trans Geosci Remote Sens 42(6):1335–1343

Foody GM, Mathur A (2004b) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens Environ 93(1–2):107–117

Foody GM, Mathur A (2006) The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens Environ 103(2):179–189

Foody GM, Mathur A, Sanchez-Hernandez C, Boyd DS (2006) Training set size requirements for the classification of a specific class. Remote Sens Environ 104(1):1–14

Foody GM, Pal M, Rocchini D, Garzon-Lopez CX, Bastin L (2016) The sensitivity of mapping methods to reference data quality: training supervised image classifications with imperfect reference data. ISPRS Int J Geo-Information 5(11):199

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recognit Lett 27(4):294–300

Gütter J, Kruspe A, Zhu XX, Niebling J (2022) Impact of training set size on the ability of deep neural networks to deal with omission noise. Front Remote Sens 3:932431

Harvolk S, Kornatz P, Otte A, Simmering D (2014) Using existing landscape data to assess the ecological potential of miscanthus cultivation in a marginal landscape. GCB Bioenergy 6(3):227–241

Hermosilla T, Wulder MA, White JC, Coops NC (2022) Land cover classification in an era of big and open data: optimizing localized implementation and training data selection to improve mapping outcomes. Remote Sens Environ 268:112780

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501

Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybernetics Part B (Cybernetics) 42(2):513–529

Inan HB, Sagris V, Devos W, Milenov P, Oosterom PV, Zevenbergen J (2010) Data model for the collaboration between land administration systems and agricultural land parcel ıdentification systems. J Environ Manage 91(12):2440–2454

Jafarzadeh H, Mahdianpari M, Gill E, Mohammadimanesh F, Homayouni S (2021) Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: a comparative evaluation. Remote Sens 13(21):4405

JRC (European Commission Directorate General Joint Research Centre) (2001) Land parcel identification systems in the frame of regulation (EC) 1593/2000 Version 1.4 (Discussion Paper), 27s

Kavzoglu T (2009) Increasing the accuracy of neural network classification using refined training data. Environ Model Softw 24(7):850–858

Kavzoglu T, Colkesen I (2009) A kernel functions analysis for support vector machines for land cover classification. Int J Appl Earth Obs Geoinf 11(5):352–359

Kavzoglu T, Mather PM (2003) The use of backpropagating artificial neural networks in land cover classification. Int J Remote Sens 24(23):4907–4938

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Advances in neural information processing systems, p 30

Khatami R, Mountrakis G, Stehman SV (2016) A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. Remote Sens Environ 177:89–100

Li C, Wang J, Wang L, Hu L, Gong P (2014) Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. Remote Sens 6(2):964–983

Liu L, Ji M, Buchroithner M (2017) Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. Remote Sens 9(12):1299

Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int J Remote Sens 28(5):823–870

Maxwell AE, Warner TA, Fang F (2018) Implementation of machine-learning classification in remote sensing: an applied review. Int J Remote Sens 39(9):2784–2817

McCarty DA, Kim HW, Lee HK (2020) Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification. Environments 7(10):84

Melgani F, Bruzzone L (2004) Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans Geosci Remote Sens 42(8):1778–1790

Millard K, Richardson M (2015) On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. Remote Sens 7(7):8489–8515

Moraes D, Campagnolo ML, Caetano M (2024) Training data in satellite image classification for land cover mapping: a review. Eur J Remote Sens 57(1). https://doi.org/10.1080/22797254.2024.2341414

Mountrakis G, Im J, Ogole C (2011) Support vector machines in remote sensing: a review. ISPRS J Photogrammetry Remote Sens 66(3):247–259

Ouma YO, Keitsile A, Nkwae B, Odirile P, Moalafhi D, Qi J (2023) Urban land-use classification using machine learning classifiers: comparative evaluation and post-classification multi-feature fusion approach. Eur J Remote Sens 56(1):2173659

Paheding, S., Saleem, A., Siddiqui, M. F. H., Rawashdeh, N., Essa, A., & Reyes, A. A. (2024). Advancing horizons in remote sensing: a comprehensive survey of deep learning models and applications in image classification and beyond. Neural Computing and Applications, 36(27), 16727–16767. https://doi.org/10.1007/s00521-024-10165-7

Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222

Pal M (2009) Extreme-learning-machine-based land cover classification. Int J Remote Sens 30(14):3835–3841

Pal M, Maxwell AE, Warner TA (2013) Kernel-based extreme learning machine for remote-sensing image classification. Remote Sens Lett 4(9):853–862

Paola JD, Schowengerdt RA (1995) A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. Int J Remote Sens 16(16):3033–3058

Phinzi K, Ngetar NS, Pham QB, Chakilu GG, Szabó S (2023) Understanding the role of training sample size in the uncertainty of high-resolution LULC mapping using random forest. Earth Sci Inf 16(4):3667–3677

Ramezan CA, Warner TA, Maxwell AE, Price BS (2021) Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. Remote Sens 13(3):368

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogrammetry Remote Sens 67:93–104

Rumelhart DE, McClelland JL, PDP Research Group (1986) Parallel distributed processing, volume 1: explorations in the microstructure of cognition: foundations. MIT Press

Shang M, Wang SX, Zhou Y, Du C (2018) Effects of training samples and classifiers on classification of landsat-8 imagery. J Indian Soc Remote Sens 46:1333–1340

Şimşek FF, Durduran S (2023) Land cover classification using Land Parcel Identification System (LPIS) data and open source EO-Learn library. Geocarto Int Adv Online Publication. https://doi.org/10.1080/10106049.2022.2146760

Şimşek FF (2023) Arazi Parsel Tanımlama Sistemi Verileri Kullanılarak Ülkesel Ölçekte Arazi Örtüsü ve Arazi Kullanım Sınıflandırması. Turkish J Remote Sens GIS 4(2):276–288. https://doi.org/10.48123/rsgis.1268155

Song X, Duan Z, Jiang X (2012) Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image. Int J Remote Sens 33(10):3301–3320

Song J, Gao S, Zhu Y, Ma C (2019) A survey of remote sensing image classification based on CNNs. Big Earth Data 3(3):232–254. https://doi.org/10.1080/20964471.2019.1657720

Tomlinson SJ, Dragosits U, Levy PE, Thomson AM, Moxley J (2018) Quantifying gross vs. net agricultural land use change in Great Britain using the ıntegrated administration and control system. Sci Total Environ 628–629:1234–1248

Ustuner M, Balik Sanli F (2019) Polarimetric target decompositions and light gradient boosting machine for crop classification: a comparative evaluation. ISPRS Int J Geo-Information 8(2):97

Wang Y, Sun Y, Cao X, Wang Y, Zhang W, Cheng X (2023) A review of regional and global scale land Use/Land cover (LULC) mapping products generated from satellite remote sensing. ISPRS J Photogrammetry Remote Sens 206:311–334

Weitkamp T, Karimi P (2023) Evaluating the effect of training data size and composition on the accuracy of smallholder irrigated agriculture mapping in Mozambique using remote sensing and machine learning algorithms. Remote Sens 15(12):3017