AJG

Article

Accurate Land Cover Classification via Training and Validation Data Refinement Using DNRS

Yoshie Ishii *[1], Junichi Susaki[2], Tsuguki Kinoshita[3] and Koki Iwao[4]

[1]Graduate School of Engineering, Kyoto University, Kyotodaigaku Katsura, Nishikyoku, Kyoto 615-8540, Japan. Email: <ishii.yoshie.4k@kyoto-u.ac.jp>

[2]Professor, Graduate School of Engineering, Kyoto University, Kyotodaigaku Katsura, Nishikyoku, Kyoto 615-8540, Japan. Email:<susaki.junichi.3r@kyoto-u.ac.jp>

[3]College of Agriculture, Ibaraki University, 3-21-1 Chuo, Ami 300-0393, Japan. Email: <tsuguki.kinoshita.00@vc.ibaraki.ac.jp>

[4]National Institute of Advanced Industrial Science and Technology (AIST), Geological Survey of Japan, Tsukuba Central 7, Higashi 1-1-1, Tsukuba 305-8567, Japan. Email: <iwao.koki@aist.go.jp>

*Corresponding author: Y. Ishii, Email: <ishii.yoshie.4k@kyoto-u.ac.jp>

Received: December 30, 2024; Accepted: April 24, 2025; Published: May 10, 2025

ABSTRACT

Land cover maps provide critical insights for a variety of applications, including monitoring natural disasters and hazards, assessing climate change impacts, and forecasting future environmental conditions. The accuracy of these maps significantly depends on the quality of the training data used in their generation. In addition, generating validation data is essential for creating land cover classification maps to demonstrate the accuracy of the maps. However, acquiring high-quality training and validation data is time and labor intensive as well as presents potential for error in data collection, interpretation, and ground surveys. This study introduces a novel training and validation data refinement method employing the directional neighborhood rough set approach to address these challenges. We applied this refinement method to training and validation data for land cover classification using Landsat-8 and FLC1 datasets. The results demonstrate that the proposed method effectively identifies reliable training and validation data, thereby enhancing the quality of land cover maps and providing the assessment method with several confidence levels depending on the purposes.

Keywords: Class Boundary, Directional Neighborhood Rough Set, Land Cover Classification, Training Data, Validation Data

1. INTRODUCTION

Land cover maps are utilized for a variety of applications, including urban and regional planning, evaluation of environmental vulnerability and impact, and monitoring of natural disasters and hazards (Talukdar et al., 2020). Land cover classification is predominantly performed through machine learning techniques, which are bifurcated into supervised and

unsupervised learning. Supervised learning boasts the capability to classify images with high precision, aligning with specific objectives. However, the creation of training data for supervised learning takes time and effort. Conversely, unsupervised learning, while being more economical and quicker because of unnecessary training data preparation, often struggles with achieving the desired classification accuracy, resulting in maps of relatively lower quality. As a result of the emphasis on accuracy over convenience, supervised classifiers have emerged as the preferred tools for land cover classification (Sheykhmousa et al., 2020).

Despite these advancements, the cost associated with producing training data is an issue that cannot be overlooked. Several strategies have been explored to reduce these costs. For example, polygon-based training data rather than pixel-based data has been proposed, although this approach suffers from a lack of spatial and feature-specific representativity (Stehman, 2009). Alternatively, utilizing existing maps or results from unsupervised classifications as training data has been suggested as a means to significantly cut down the costs involved in training data production. However, this method could propagate errors from the original or unsupervised maps into the newly generated map. Notably, Foody and Arora have demonstrated that the choice of training data exerts a more significant impact on the classification results compared to the differences among classifiers (Foody and Arora, 1997).

Recent advancements have introduced various methodologies to filter out erroneous training data or to identify and utilize reliable training data, aiming to enhance the accuracy of land cover maps. Kavzoglu utilized visual analyses to examine the distribution of training data across subspaces and histograms for each class, demonstrating how the exclusion of outlier samples based on these analyses could significantly improve the performance of neural network classifications (Kavzoglu, 2009). Radoux et al. proposed two automated methods for extracting training data from existing land cover maps focusing on global land cover map: the multiclass border reduction filter, which identifies class boundaries on the map, and a spectral filtering technique commonly employed in change detection to remove outliers from the spectral signature distribution (Radoux et al., 2014). Paris and Bruzzone introduced a method that utilizes Gaussian distributions to extract reliable training data from thematic products (Paris and Bruzzone, 2021). Bratic et al. suggested extracting dependable training data through the intersection of multiple maps (Bratic et al., 2023).

In addition, the creation of validation data is essential for creating land cover classification maps to demonstrate the accuracy of the maps. The creation of validation data is as costly as the creation of training data. In some cases, existing land cover classification maps can be used as a substitute for the validation data. Therefore, quality assurance is required for both the training and validation data.

This study proposes an alternative perspective to the conventional methodologies for extracting reliable training and validation data. It explores the application of rough set theory, which includes the concept of certainty approximation sets deemed apt for identifying dependable training and validation data. Originally, the

classical rough set theory was introduced by Pawlak in 1982, aiming to derive features and rules from data that may be ambiguous or incomplete (Pawlak, 1982). Rough set theory has since been extended and adapted across various domains, demonstrating its applicability to a wide range of fields. Several studies have employed extended rough set theories as land cover classifiers (Pan et al., 2010; Ishii et al., 2021). Among these classifiers, the grade-added rough set (GRS) possesses datasets featuring numerical explanatory variables and categorical objective variables, offering the advantage of minimizing information loss in comparison to methods requiring discretization (Ishii et al., 2021). Subsequently, the directional neighborhood rough set (DNRS) was introduced as a generalization of GRS, providing a more rigorous mathematical framework and addressing limitations inherent to GRS (Ishii et al., 2022). However, the applicability and effectiveness in the field of remote sensing remain to be validated.

In this study, an innovative method for refining training and validation data is proposed to enhance the accuracy of land cover maps and to assess the accuracy with precision, utilizing the lower approximation concept within the DNRS framework. DNRS exhibits a limitation with increasing dimensionality, often resulting in a conservative lower approximation. To address this issue, the "degree of certainty" within the lower approximation set is refined, applying this comprehensive DNRS approach to identify and extract the most effective training data for land cover classification and to assess the accuracy using reliable validation data.

2. EXTENSION OF CURRENT APPROACH

2.1 DNRS Approach

The basic DNRS approach (Ishii et al., 2022) is described in this section as a foundation for the new concept in DNRS introduced in section 2.2.

$DT =< U, A, V, \rho >$ is a decision table, where U is a nonempty finite set of m objects $\{x_1, x_2, \cdots, x_m\}$; $A = \{a_1, a_2, \cdots, a_n\}$ is a nonempty finite set of n attributes; and $V = \cup_{a \in A} V_a$, where V_a is the domain of attribute $a \in A$. $A = C \cup D$ consists of a condition attribute set C and a decision attribute set D. $\rho: U \times A \to V$ is an information function that allocates attribute value $\rho(x, a) \in V$ to an object $x \in U$ and an attribute $a \in A$. The datasets addressed in this study have numerical values for condition attributes and categorical values for the decision attribute. The condition attributes $\rho(x_i, a)$ is the degree of association an object x_i has with attribute a, defined as follows:

$$g(x_i, a) = \frac{\rho(x_i,a) - \min_{x \in U}\{\rho(x,a)\}}{\max_{x \in U}\{\rho(x,a)\} - \min_{x \in U}\{\rho(x,a)\}}, \quad (1)$$

where $g(x_i, a) \in [0,1]$. The difference of grade is as follows:

$$diff(x, y, a) = g(x, a) - g(y, a). \quad (2)$$

Furthermore, the n-dimensional vectors and difference of grade are expressed as follows:

$$\boldsymbol{g}(x) \equiv \big(g(x, a_1), g(x, a_2), \cdots, g(x, a_n)\big). \quad (3)$$

$$\boldsymbol{diff}(x, y) \equiv \big(diff(x, y, a_1), diff(x, y, a_2), \cdots, diff(x, y, a_n)\big). \quad (4)$$

To define a fundamental set (granule information), half-space and neighborhood sets

are introduced. Initially, the i-th hyperplane in real *n*-space $\mathbb{R}$ is defined using $x, y \in \mathbb{R}^n$ as follows:

$$H_i(\boldsymbol{\varepsilon}_i, \boldsymbol{g}(x)) = \{\boldsymbol{g}(y) \in \mathbb{R}^n | \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(y) = \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(x)\}, \quad (5)$$

where we assume that the *j*-th component of vector $\boldsymbol{\varepsilon}_i$ on the real *n*-dimensional space satisfies the following definition:

$$\varepsilon_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j' \end{cases}$$

where *i* and *j* are the subscripts indicating the number of dimensions. The real *n*-dimensional space can be divided into 2^n regions using the *n* hyper plane defined by (5). To define such regions, the upper and lower half-spaces are defined as follows:

$$H_i^+(\boldsymbol{\varepsilon}_i, \boldsymbol{g}(x)) = \{\boldsymbol{g}(y) \in \mathbb{R}^n | \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(y) \geq \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(x)\} = \{\boldsymbol{g}(y) \in \mathbb{R}^n | \boldsymbol{\varepsilon}_i \cdot \boldsymbol{diff}(y, x) \geq 0\}, \quad (6)$$

$$H_i^-(\boldsymbol{\varepsilon}_i, \boldsymbol{g}(x)) = \{\boldsymbol{g}(y) \in \mathbb{R}^n | \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(y) \leq \boldsymbol{\varepsilon}_i \cdot \boldsymbol{g}(x)\} = \{\boldsymbol{g}(y) \in \mathbb{R}^n | \boldsymbol{\varepsilon}_i \cdot \boldsymbol{diff}(y, x) \leq 0\}. \quad (7)$$

One of 2^n quadrants created by n hyper planes about object x in (6) and (7) can be expressed as

$$Q_B^l(x) = \{y \in U | y \in \cap_{i \in B} H_i^*(\boldsymbol{\varepsilon}_i, \boldsymbol{g}(x))\}, \quad (8)$$

where *l* is the subscript denoting the number of quadrants and $H_i^*(\varepsilon_i, g(x))$ denotes the half space. The sign of half space *H i*s decided using the following condition:

$$H_i^* = \begin{cases} H_i^+ \text{ if the } i-\text{th digit of } l \text{ in binary is } 0 \\ H_i^- \text{ if the } i-\text{th digit of } l \text{ in binary is } 1 \end{cases}.$$

Subsequently, the neighborhood set is defined. Neighborhood set $N_B^\delta(x)$ of object *x* on a partial set $B \subseteq C$ is as follows:

$$N_B^\delta(x) = \{y \in U | \Delta^B(x, y) \leq \delta\}, \quad (9)$$

where δ represents a neighborhood parameter. $\Delta^B(x, y)$ denotes the distance function and can be expressed as:

$$\Delta^B(x, y) = \left(\sum_{a \in B} |diff(x, y, a)|^P\right)^{\frac{1}{P}}, \quad (10)$$

where $P = \infty$, which corresponds to the Chebyshev distance.

Given a partial set $B \subseteq C$ and objects x, y, the fundamental set $R_B^{l\delta}(x)$ can be defined using (8) and (9) as follows:

$$R_B^{l\delta}(x) = \{y \in U \mid y \in Q_B^l(x) \cap N_B^\delta(x)\} \quad (11)$$

where *t* represents the number of objects and takes values ranging from 2 to 20. Card means cardinality of the set. Using this fundamental set, the directional neighborhood (DN)-lower approximation set and DN-upper approximation set of an arbitrary set *X* can be defined as follows:

$$\underline{R_B}(X) = \{x | R_B^{l\delta}(x) \subseteq X, \exists l\}, \quad (13)$$

$$\overline{R}_B(X) = \{x | R_B^{l\delta}(x) \cap X \neq \emptyset, \forall l\}. \quad (14)$$

In the case of land cover classification, the *X* means a set of training data whose elements belong to an arbitrary class for land cover classification.

2.2 Subdivision of DN-Lower Approximation Set

In this section, a new concept is induced in the DNRS approach. Equation (15) is a special case of (13).

$$\underline{R_B}(X) = \{x | R_B^{l\delta}(x) \subseteq X, \forall l\}. \quad (15)$$

This development acknowledges a variance in the degree of certainty between (13) and (15), a distinction not originally accounted for in DNRS. To elucidate this difference, an illustrative

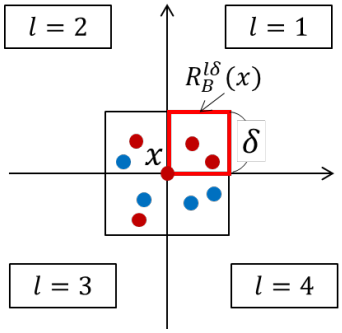
example is provided in Figure 1, where the target object is positioned at the axis intersection in both scenarios (a) and (b). To determine whether the object at the axis intersection qualifies as DN-lower, it should be evaluated in the context of surrounding objects. Both target objects in (a) and (b) meet the criteria for inclusion in the lower approximation set as per (13). However, the target object in (b) additionally fulfills the conditions of (15). The key distinction between (a) and (b) in Figure 1 lies in the proximity of different class objects to the target object in (a), in contrast with the exclusive presence of same-class objects near the target object in (b), underscoring the rationale for differentiating the degree of certainty between these scenarios. Thus, the degree of certainty is proposed to be directly proportional to the fraction of fundamental sets satisfying the lower approximation conditions across 2^n quadrants, expressed as follows:

$$\mu(x) = \frac{\mathrm{Card}(\{l \mid R_B^{l\delta}(x) \subseteq X\})}{2^n}. \qquad (16)$$

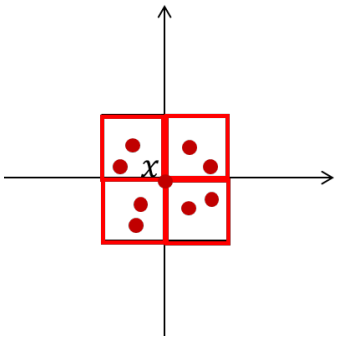
where Card means cardinality of the set. Extraction of DN-lower samples meeting a specified threshold α is then facilitated through the following equation:

$$\underline{R_B^{\alpha}}(X) = \{x | \mu(x) \geq \alpha\}. \qquad (17)$$

The selection of a higher α correlates with increased certainty. Notably, α represents relative certainty within a dataset. Typically, the occurrence of multiple classes within the same fundamental set is indicative of class boundaries in the feature space. Consequently, employing the DN-lower approximation set as a methodological tool serves to eliminate superfluous or uncertain training data situated at class boundaries, thereby enhancing the precision of training data selection for improved land cover classification.



(a) Example of lower approximation object satisfying only (13)



(b) Example of lower approximation object satisfying both (13) and (15)

Figure 1: Illustration of variations in DN-lower approximation definitions within a two-dimensional feature space: x represents the target object.

2.3 Comparison with existing methods

In order to clarify the novelty and features of the proposed method, we compare the proposed method with existing methods. Table 1 shows the comparison of the refinement methods for training data. While almost all of the existing methods use existing land cover maps to create refined training data, Taskin (2009) and the proposed method use original training data. Creating refined training data from the existing land cover map is less laborious than creating that from training data. However, the accuracy of training data from the existing land cover maps

more or less depends on that of the existing land cover maps. In addition, the method is limited in the case that land cover maps exist. On the other hand, creating refined training data from original training data needs to prepare the original training data, and it is time-consuming. In addition, the proposed method needs hyperparameter settings. Therefore, it is relatively time-consuming compared to existing methods. Instead, there is no worry about the limitation depending on the accuracy of land cover maps. In addition, it is different from the existing research to discuss not only training data but also validation data. As shown in Table 1, some existing methods focus on class boundaries. However, the definition of class boundaries is different among them. The class boundaries in Taskin (2009) are based on visual analysis. This is a concern about the results being different depending on the person who interprets. The class boundaries defined by Radoux et al. (2014) mean on the map, not feature space. The most important feature of the proposed method is to decide the class boundary theoretically based on set theory and therefore non-parametrics. The same results are derived from the proposed method, not depending on the person. The distribution of class in feature space is not limited to parametric distribution such as Gaussian distribution.

Table 1: Comparison of the refinement methods for training data.

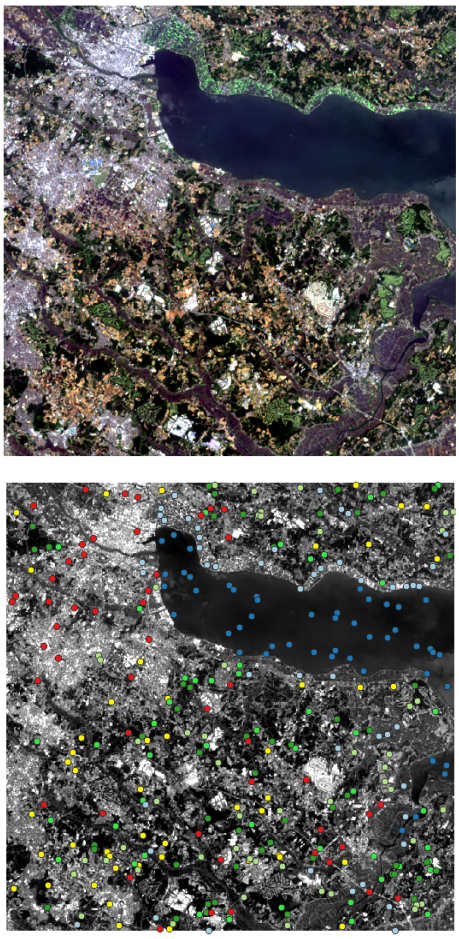
Article	Data to create refinement training	Method	Focus on	Hyperparameters
Taskin (2009)	Training data	Step1: Visual analysis of the training pixels for depicting the decision boundaries in two and tree dimensional graphs Step2: Mixed and atypical pixels were detected and eliminated using visual histogram analysis	Class boundary in feature space and spectral outlier from histgram	None
Radoux et al. (2014)	Existing global land cover map	Step1: Local training Step2: Trimming training data using MBRF (multiclass border reduction filter) or Spectral filtering	Class boundary on map and spectral outlier based on Mahalanobis distance	Width of the central tile and of the corresponding training area (Step1) and MBRF or window size for trimming (Step2)
Paris and Bruzzone (2021)	Existing land cover map	Step1: Understand the source domain propoerties Step2: Decompose the source domain Step3: Training data extraction using Gaussian-distribution	Mixed pixel decomposition and outlier form Gaussian-distribution	None
Bratic et al. (2023)	Existing global land cover map	Majority vote of multiple land cover maps	Reliability based on majority	None
Proposed method	Training data	Using DNRS lower approximation	Class boundary in feature space	t and alpha

3. DATASETS

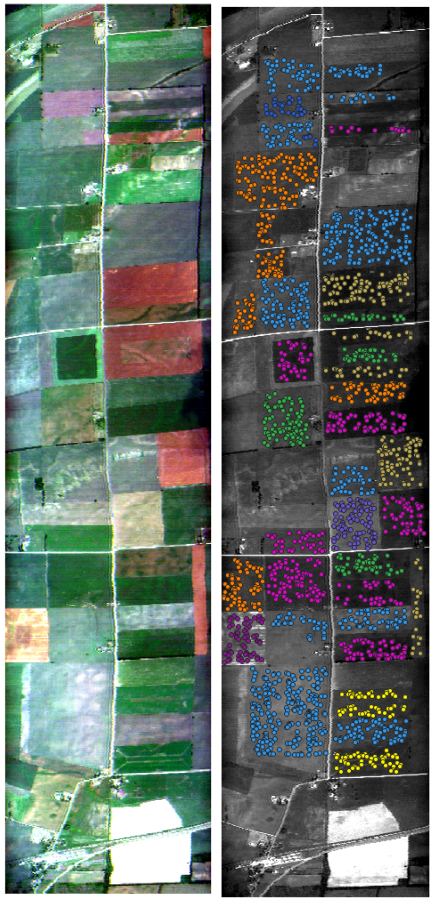
This study utilizes two distinct remote sensing datasets to evaluate the effectiveness of the DNRS approach in extracting reliable training and validation data. The first dataset originates from the Landsat-8 operational land imager (OLI), covering the southern part of Ibaraki, Japan. The second dataset, referred to as flightline C1 (FLC1), encompasses the agricultural land in the southern region of Tippecanoe County, Indiana (Landgrebe, 1994). For the purposes of this study, these datasets are henceforth designated as Landsat-8 and FLC1, respectively.

The Landsat-8 dataset, captured on May 31st, 2014, incorporates features from the 1st to the 7th bands with a spatial resolution of 30 m, spanning an image size of 667 × 667 pixels. The target area is depicted in Figure 2 (a). The classification

includes seven land cover classes: Water (1), Cropland (2), Sparse Grass (3), Grass (4), Forest (5), Paddy (6), and Built-up (7). The numbers in the brackets correspond to the Class No. used in the Results and Discussion section (e.g. Figure 5 (a) and (b)). A stratified random selection process produced approximately 200 training data points per class, presenting in a total of 1412 training data points. We randomly selected 50 data points for each class for validation purposes, yielding a comprehensive validation dataset of 350 points. The validation points are presented in Figure 2 (a). The validation and training data points were interpreted using Google Earth imagery and supplemented by field surveys where feasible.



(a) Landsat-8 image (R=4, G=3, B=2) (top) and validation points (bottom)



(b) FLC1 (R=9, G=6, B=3) (left) and validation points (right)

Figure 2: Images and Validation points used for experiments.

Table 2: Features of datasets.

Landsat-8 dataset	FLC1 dataset
● Training and validation data are obtained on a pixel basis	● Training and validation data are obtained on a polygon basis
● Small size dataset	● Large size dataset
● Satellite-based dataset	● Airborne-based dataset
● 7 bands	● 12 bands
● Seven land cover classes: Water, Cropland, Sparse Grass, Grass, Forest, Paddy, and Built-up	● Nine vegetation classes: Alfalfa, Bare Soil, Corn, Oats, Red Clover, Rye, Soybeans, Wheat, and Wheat-2

The FLC1 dataset, provided by Purdue University (Landgrebe, 1994), features 12 spectral bands and was acquired in June 1966. The image size is 949 × 220 pixels. It delineates nine vegetation classes: Alfalfa (1), Bare Soil (2), Corn (3), Oats (4), Red Clover (5), Rye (6), Soybeans (7), Wheat (8), and Wheat-2 (9). The numbers in brackets correspond to the Class No. used in the Results and Discussion section (e.g. Figure 5 (c) and (d)). The original dataset, comprising 70,594 test data points as shown in Figure 7 in Appendix A, was partitioned into one validation dataset and thirty training datasets, each containing approximately 2,000 test data points. For the scope of this analysis, only five out of the thirty training datasets were utilized, deemed sufficient to discern the trends of the FLC1 dataset. Figure 2 (b) depicts the FLC-1 image and validation points.

The purpose of using these two datasets is to demonstrate the versatility of the method proposed. Therefore, we intentionally selected datasets whose features are different. Table 2 lists the features of these two datasets. Both the Landsat-8 and FLC1 datasets underwent normalization to fit within the [0.0, 1.0] range, ensuring uniformity in data processing and analysis.

4. EXPERIMENTS

Figure 3 depicts the four workflow patterns for land cover classification. (a) depicts general workflow as control. (b) depicts the workflow that tests the effectiveness of reliable training data extraction using DNRS. (c) depicts the workflow that tests the effectiveness of reliable validation data extraction using DNRS. (d) depicts the workflow that tests the effectiveness of both reliable training and validation data using DNRS. The DNRS technique is employed to procure assured training and/or validation data.

Detailed procedure of DNRS training/validation extraction is shown in Figure 4. In the step of "Search the optimal δ for each training/validation data (t=10)" in Figure 4, the optimal δ which satisfies Equation (12) is calculated for each dimension of each element of training/validation data. In this study, the parameter t, the number of elements that should be included in a fundamental set, was fixed $t = 10$. Although it takes time to adapt the parameter t, one can obtain more accurate results. However,

the parameter *t* is not the main theme in this study, so we used the fixed value. In the step of "Calculate the rate of lower approximation" in Figure 4, the rate of dimension that satisfies DN-lower approximation definition to all dimensions is calculated for each training/validation data. If the element of training/validation data satisfies the condition, it is recognized as certain training/validation data and used for machine learning and/or accuracy assessment. In this step, the threshold denoted as α, is meticulously surveyed from 0.000 to 1.000 in increments of 0.005 for both the Landsat-8 and FLC1 datasets. For a more detailed examination of trend nuances within the FLC1 dataset, the α threshold for training data was further investigated from 0.900 to 1.000 in finer intervals of 0.001.

The SVM (support vector machine) classifier was selected for land cover classification due to its widespread acceptance and proven efficacy in the field, particularly in handling class boundaries. The tuning of SVM hyperparameters was carried out across a specified range, with C values of 1.0, 10.0, 100.0, and 1,000.0, and γ values of 0.001, 0.01, 0.1, 1.0, and 10.0. In addition, a control experiment was conducted, wherein the original training data was directly subjected to SVM learning, without the intermediary step of DNRS-based refinement. Kappa coefficient was used for the accuracy assessments of land cover maps.

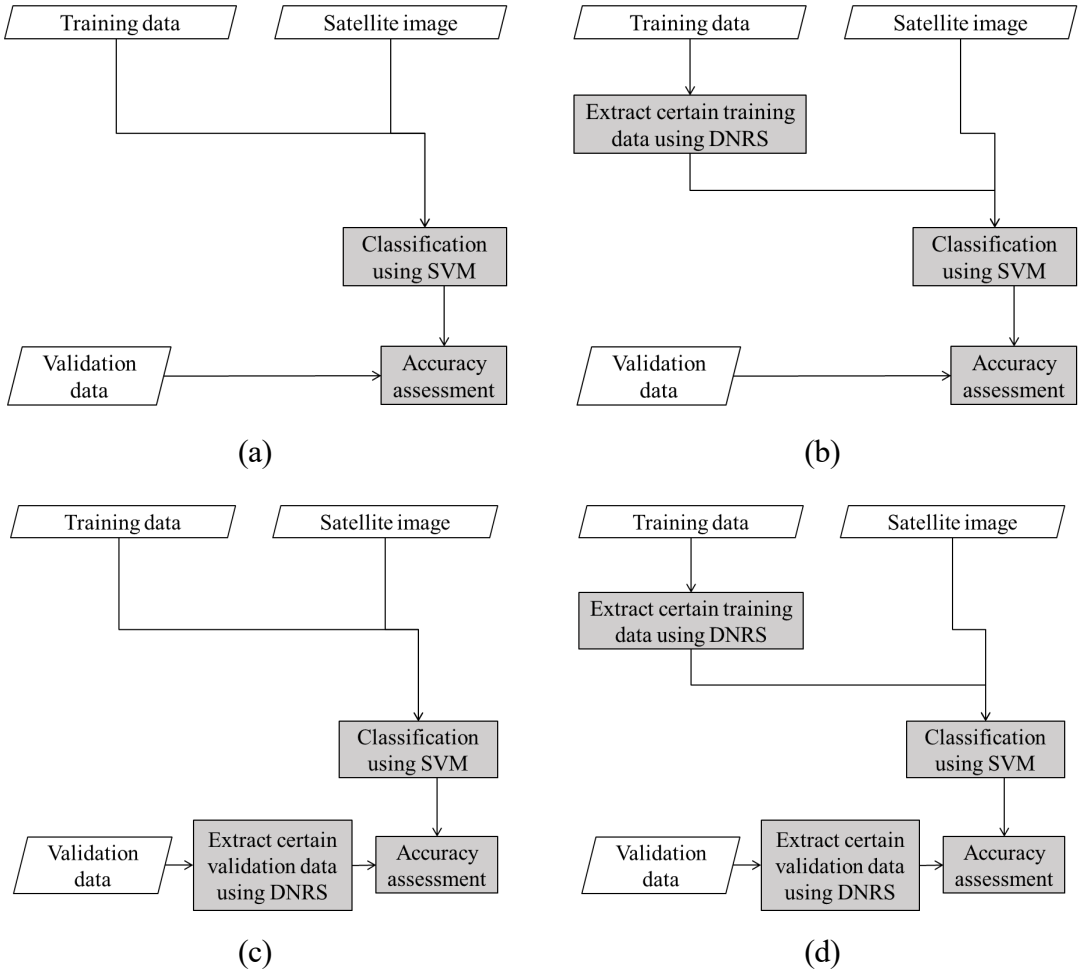


Figure 3: Workflows for land cover classification: (a) depicts the general workflow as control. (b) depicts the workflow that tests the effectiveness of reliable training data extraction using DNRS. (c)

depicts the workflow that tests the effectiveness of reliable validation data extraction using DNRS. (d) depicts the workflow that tests the effectiveness of both reliable training and validation data using DNRS.

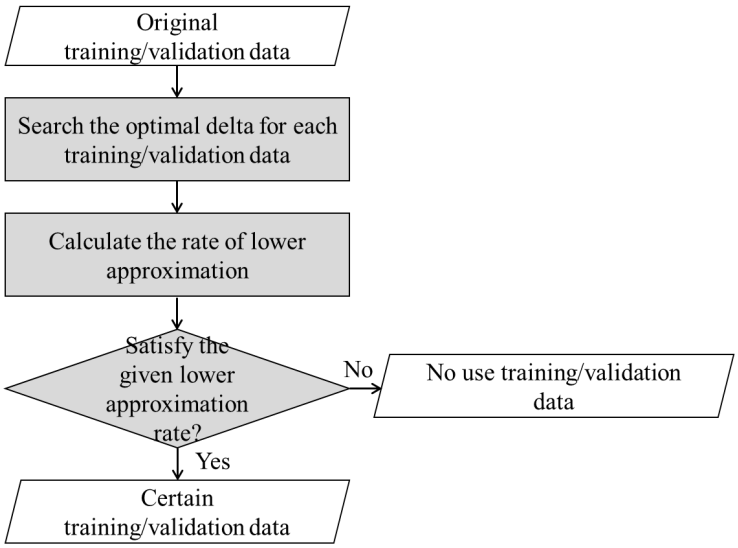


Figure 4: Flow of training/validation data extraction using DNRS.

5. RESULTS

The kappa coefficients of land cover maps, reflecting the conditions of SVM hyperparameters, are presented in Table 3 for the Landsat-8 dataset and Table 4 for the FLC1 dataset. In the case of FLC1, five datasets were randomly selected from the original pool of thirty to ascertain the prevailing trends. Table 4 presents the kappa coefficient for one representative dataset from these five. Four out of five datasets are presented in Appendix B since these results indicate similar trends. In these tables, the rows represent the proportion of training data corresponding to the original training dataset. The first to the third columns list the kappa coefficients, whereas the fourth and fifth columns list the optimal combination of hyperparameters obtained when tuning was performed using the validation dataset. The term “points” within the brackets in Tables 2 and 3 specifies the sample count, and α represents the threshold for DN-lower approximation, as defined in (17). The kappa coefficient in the first row corresponds to scenarios where the original training data was employed for model training. Tables 2 and 3 present results that are limited to instances where the proportion of assured training data ranges from 50% to 100%, in increments of approximately 10%. This non-uniform interval for the assured training data proportion is attributed to the challenge of precisely controlling the sample count reduction through DNRS, since the extent of reduction is contingent upon the sample distribution within the feature space. The kappa coefficient highlighted in bold in these tables represents the highest values among the training datasets when assessed with identical validation data, indicating the most favorable classification outcomes under given conditions.

Table 3: Kappa coefficients and hyperparameters for Landsat-8 dataset.

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (350 points, α=0.000)	78% (274 points, α=0.830)	50% (175 points, α=0.870)	C	Gamma
Rate of assured training data to original one	Original	100% (1412 points, α=0.000)	0.863	0.914	0.957	100.0	10.0
	Reduced by DNRS	89% (1263 points, α=0.720)	0.873	0.923	0.957	1000.0	0.1
		80% (1134 points, α=0.750)	**0.877**	**0.931**	0.957	100.0	10.0
		71% (1002 points, α=0.775)	0.863	0.923	0.957	1000.0	0.1
		61% (868 points, α=0.800)	0.853	0.923	**0.964**	100.0	10.0
		51% (718 points, α=0.825)	0.807	0.867	0.950	100.0	10.0

Table 4: Kappa coefficients and hyperparameters for FLC1 dataset (Dataset No. 1).

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (2318 points, α=0.000)	75% (1739 points, α=0.931)	50% (1152 points, α=0.953)	C	Gamma
Rate of assured training data to original one	Original	100% (2224 points, α=0.000)	**0.948**	**0.972**	0.975	100.0	10.0
	Reduced by DNRS	90% (1997 points, α=0.910)	0.941	0.969	0.974	100.0	10.0
		80% (1779 points, α=0.926)	0.924	0.969	**0.980**	1000.0	1.0
		70% (1552 points, α=0.937)	0.913	0.964	0.974	1000.0	10.0
		60% (1344 points, α=0.945)	0.896	0.953	0.971	1000.0	10.0
		50% (1116 points, α=0.952)	0.887	0.944	0.976	1000.0	1.0

6. DISCUSSION

6.1 Landsat-8 dataset

The kappa coefficient when the training data is reduced by DNRS is higher than that of the original training data in all the validation data cases (in three columns) presented in Table 3. The comparison in the first column in Table 3 corresponds to the comparison between (a) and (b) presented in Figure 3. This comparison demonstrates the effectiveness of training data refinement by DNRS in the condition of original validation data. The accuracy for the scenarios with reduced training data was found to improve when the proportion of assured training data was at 89% and 80% in the first column in Table 3. Conversely, this accuracy decreased at lower proportions of the assured training data, specifically at 61% and 51%. The initial improvements in accuracy can be attributed to the removal of uncertain training data, enhancing the performance of the classifier by focusing on more reliable samples. However, the subsequent drops in accuracy indicate the detrimental impact of information loss. Essentially, while a higher threshold of DN-lower approximation correlates with increased certainty in the training data, it paradoxically leads to diminished accuracy due to the insufficiency of the data required to

accurately represent the true distribution of classes. These results indicate that the most appropriate α value for obtaining a balance between the elimination of low reliable data and not occurring information loss exists. It is considered that this most appropriate α depends on the distribution of training data in the feature space. A comparison of the three columns corresponding to the kappa coefficients reveals similar trends. The comparison in the first row in Table 3 corresponds to the comparison between (a) and (c) in Figure 3, which represents the effectiveness of the validation data refinement presented by DNRS. The results of the first row indicate the validation data is reduced by DNRS, and the accuracy increases under the condition of the same original training data. However, it is crucial to carefully interpret the results. These results do not imply that simply reducing the validation data will improve accuracy. Instead, they indicate that accuracy assessment can be performed according to the purpose, e.g., when the land cover map is to be assessed only with highly reliable validation data or also with unreliable validation data.

6.2 FLC1 dataset

Although the kappa coefficient when the training data is reduced by DNRS is higher than that of the original training data in 50% of the validation data cases (in the third column) presented in Table 4, the kappa coefficient of the original training data is higher than that when the training data is reduced by DNRS in original and 75% of the validation data (in the first and second column) presented in Table 4. This is because the effectiveness of the training data refinement achieved by DNRS is not observed in the original and 75% of the validation data is caused by the approach employed in obtaining the validation data. The validation data for the FLC1 dataset was taken polygon base in the cropland region, as shown in Table 2. That is, the noisy data may be included in the validation data. Consequently, even if the training data is refined by DNRS, the effectiveness of the training data refinement is not observed when the quality of validation data is insufficient. The effectiveness of the training data refinement is observed when the quality of validation data is also sufficiently ensured, as shown in the third column in Table 4.

6.3 Visual analysis in feature space

Figure 5 presents the results of visualizing the training data at $\alpha = 0.0$ (the original training data) and $\alpha = 0.75$ (the highest accuracy training data) for Landsat-8 dataset and the training data at $\alpha = 0.0$ (the original training data) and $\alpha = 0.926$ (the highest accuracy training data) for FLC1 dataset in the feature space to visually verify the effectiveness of the proposed method. Figures 4 (a) and (c) depict the distribution of the training data using the original training data for each dataset, and Figures 4 (b) and (d) depict the distribution of the training data extracted by DNRS lower approximation at $\alpha = 0.75$ for Landsat-8 and $\alpha = 0.926$ for FLC1. Seven bands of Landsat-8 images and 12 bands of FLC1 are used for the reliable training data extraction process; thus, the feature space in Figure 5 indicates the subspace. The major difference between Figures 4 (a) and (b), and (c) and (d) is the samples near the class boundaries (black circles in the scatter plots). While there is training data around the class boundaries in Figures 4 (a) and (c), the number of training data around the class boundaries is reduced in Figure 5 (b) when compared to that of Figure 5 (a), and

Figure 5 (d) when compared to Figure 5 (c). These results indicate that the proposed training data refinement method is characterized by the training data near the class boundaries being less reliable and excluded by lower approximation in DNRS. At the same time, the distribution of the reduced training data becomes closer to the true class distribution on the feature space, as shown in the accuracy improvement in Tables 2 and 3 under the conditions of the reliable validation datasets.

6.4 Number of data for each class

Figures 6 and 7 indicate the transformation of the number of validation data for each class. Figure 6 shows that the classes corresponding to vegetation (Cropland, Sparse grass, and Grass) are reduced by DN-lower approximation. These classes have many class boundaries on the feature space with each other. Similarly, Figure 7 shows that as the DNRS threshold α is increased, some classes decrease significantly while others do not.

Therefore, the number of validated data may be highly skewed by the class when reliable validated data is extracted using DNRS. Conversely, if the results are used effectively, non-separable classes can be reliably classified by merging them and then performing the land cover classification again.

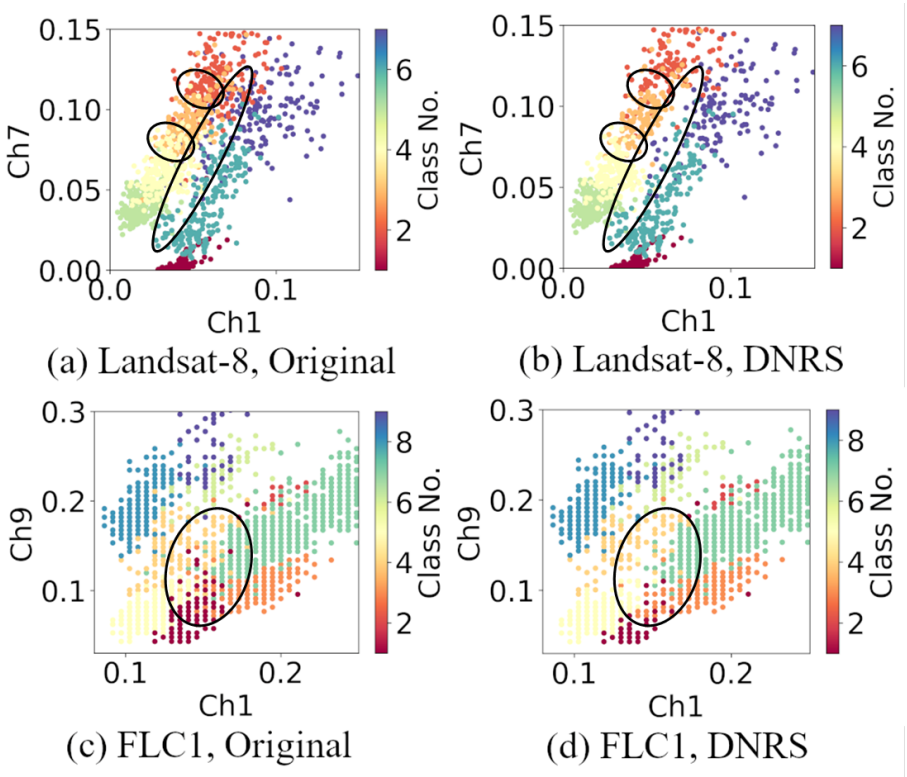


Figure 5: Representation of training data distribution across different classes in the feature subspace. (a) is the distribution of the original training dataset for Landsat-8. (b) is the distribution of training data extracted by DNRS lower approximation at α=0.75 for Landsat-8. (c) is the distribution of the original training dataset for FLC1. (d) is the distribution of training data extracted by DNRS lower approximation at α=0.926 for FLC1. The color bar identifies distinct classes, and "Ch" indicates the band number. Black circles in the scatter plots are the regions where the variation between (a) and (b), and between (c) and (d) are relatively large.

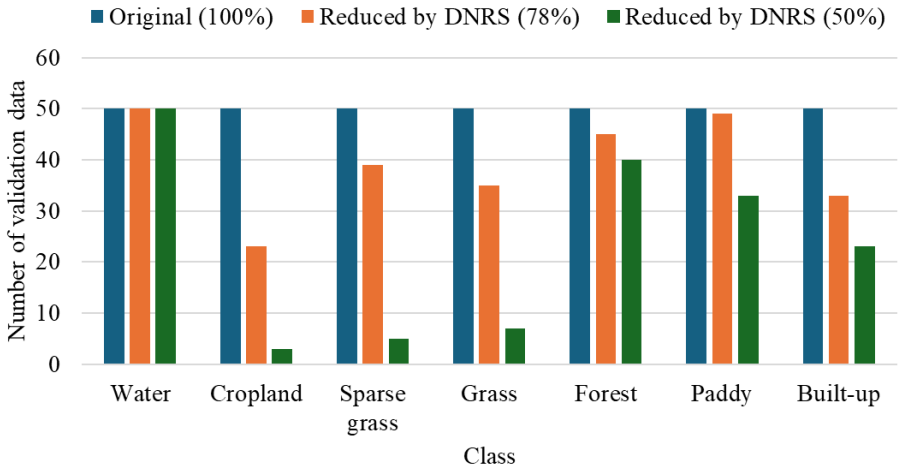


Figure 6: Number of validation data for each class for Landsat-8.

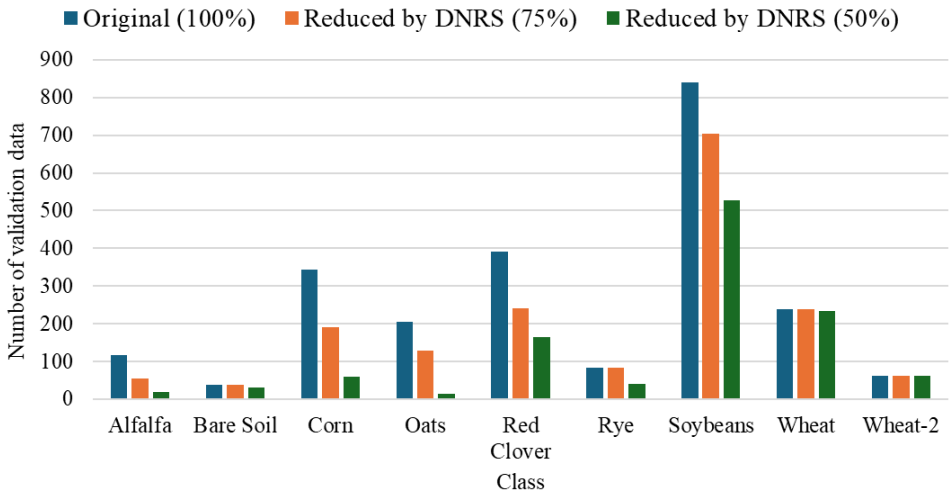


Figure 7: Number of validation data for each class for FLC1.

6.5 Optimal sample size

For training data, the optimal sample size is automatically determined based on the proposed method by selecting the parameter when the accuracy was the highest, using the same validation data. However, the selection of the optimal sample size for validation data is a little bit complicated. As shown in Tables 3 and 4, the more validation samples are decreased, the higher the accuracy is. While, the accuracy stability becomes less when the validation samples are reduced, based on the law of large numbers. If the purpose of accuracy assessment is to assess the accuracy using only reliable validation data, the optimal sample size of the validation data is the minimum number of the validation data by using proposed method, satisfying statistically adequate number. The definition of the statistically adequate number is discussed in a lot of existing studies (Cochran, 1977; Congalton, 1991; Foody, 2008) and depends on the situations of indices, sampling, etc. (e.g. overall accuracy, kappa coefficient, user's accuracy, producer's accuracy, simple random sampling, stratified sampling).

6.6 Summary of experiment results

These experiments demonstrate that the proposed training and validation refinement method using DN-lower approximation can effectively select training data and help improve the accuracy of land cover classification. It also provides the assessment method using validation data under several confidence levels depending on the purposes. Visual analysis of the training data demonstrated that the data determined to have low confidence levels by DNRS are those that lie on the boundaries of the classification class on the feature space. This property also applies to the validation data. Therefore, the higher the confidence level of the validation data determined by the DNRS, the more the validation data in the boundary areas are reduced. That is, the confidence level of the validation data corresponds to the amount of validation data included in the boundary regions on the feature space.

The limitations of the proposed method are as follows: This method needs to optimize two hyperparameters: one is t in Equation (12) and α in Equation (17), and the range of hyperparameters depends on the dataset as shown in Tables 2 and 3. Therefore, the cost of parameter setting is needed. Second, there is a possibility that the deviation of the amount of training and/or data among classes yields because of the characteristics of this method which reduces the elements of class boundary in the feature space. Figures 5 and 6 imply the trend of this limitation.

For future work, the robustness of the proposed method will be checked by assuming various satellite imagery, and land cover classification classes.

7. CONCLUSION

In this study, we present a novel method for refining the training and validation data by utilizing the DN-lower approximation concept thereby enhancing the accuracy of land cover classification and assess the accuracy with more precision. The proposed method is based on the rough set theory, which enables all decision rules to hold, and has few theoretical black box aspects, making it a suitable method for reliable training and validation data extraction.

The effectiveness of the proposed method was evaluated by performing a land cover classification into seven land cover classes using the Landsat-8 dataset and into nine vegetation classes using the FLC1 dataset. The results demonstrate that reducing the training data using the proposed method improves the land cover classification accuracy when compared to using the original training data. In addition, we provide the accuracy assessment method for land cover maps at several confidence levels using validation data extracted by DN-lower approximation. The confidence level of the validation data based on DNRS corresponds to how much validation data in the boundary regions on the feature space is included.

In future work, we aim to verify the robustness of the proposed method by assuming various satellite imagery and land cover classification classes.

REFERENCES

Bratic, G., Yordanov, V., and Brovelli, M. A. 2023. “High-resolution land cover classification: cost-effective approach for extraction of reliable training data from

existing land cover datasets." *Int. J. Digit. Earth.* 16 (1): 3618-3636. https://doi.org/10.1080/17538947.2023.2253784.

Cochran, W. G., 1997. "*Sampling techniques*" (3rd. ed.). John Wiley and Sons, New York.

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

Foody, G. M., and Arora, M. K. 1997. "An evaluation of some factors affecting the accuracy of classification by an artificial neural network." Int. J. Remote Sens. 18 (4): 799-810. https://doi.org/10.1080/014311697218764.

Foody, G. M. 2008, "Sample size determination for image classification accuracy assessment and comparison." *Proceedings of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences*. 154-162.

Ishii, Y., Hasi, B., Iwao, K., and Kinoshita, T. 2021. "A new land cover classification method using grade-added rough sets." *IEEE Geosci. Remote Sens. Lett.* 18 (1): 8-12. https://doi.org/10.1109/LGRS.2020.2965297

Ishii, Y., Iwao, K., and Kinoshita, T. 2022. "A new rough set classifier for numerical data based on reflexive and antisymmetric relations." *Mach. Learn. Knowl. Extr*. 4 (4): 1065-1087. https://doi.org/10.3390/make4040054.

Kavzoglu, T. 2009. "Increasing the accuracy of neural network classification using refined training data." *Environ. Model. Softw*. 24 (7): 850-858. https://doi.org/10.1016/j.envsoft.2008.11.012.

Landgrebe, D. 1994. "Multispectral Data Analysis: A Moderate Dimension Example©," Purdue University, West Lafayette, Indiana, U.S.

Pan, X., Zhang, S., Zhang, H., Na, X., and Li, X. 2010. "A variable precision rough set approach to the remote sensing land use/cover classification." *Comput. Geosci*. 36 (12): 1466-1473. https://doi.org/10.1016/j.cageo.2009.11.010.

Paris, C., and Bruzzone, L. 2021. "A Novel Approach to the Unsupervised Extraction of Reliable Training Samples From Thematic Products." IEEE T. Geosci. and Remote Sens. 59 (3): 1930-1948. https://doi.org/10.1109/TGRS.2020.3001004.

Pawlak, Z. 1982. "Rough sets." *Int. J. Comput. Inf. Sci*. 11: 341-356. https://doi.org/10.1007/BF01001956.

Radoux, J., Lamarche, C., Bogaert, E. V., Bontemps, S., Brockmann, C., and Defourny, P. 2014. "Automated training sample extraction for global land cover mapping." *Remote Sens*. 6 (5): 3965-3987. https://doi.org/10.3390/rs6053965.

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., and Homayouni, S. 2020. "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review." *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* 13: 6308-6325. https://doi.org/10.1109/JSTARS.2020.3026724.

Stehman, S. V. 2009. “Sampling designs for accuracy assessment of land cover.” *Int. J. Remote Sens.* 30 (20): 5243-5272. https://doi.org/10.1080/01431160903131000.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., and Rahman, A. 2020. “Land-use land-cover classification by machine learning classifiers for satellite observations—a review.” *Remote Sens.* 12 (7): 1135. https://doi.org/10.3390/rs12071135.

APPENDIX A

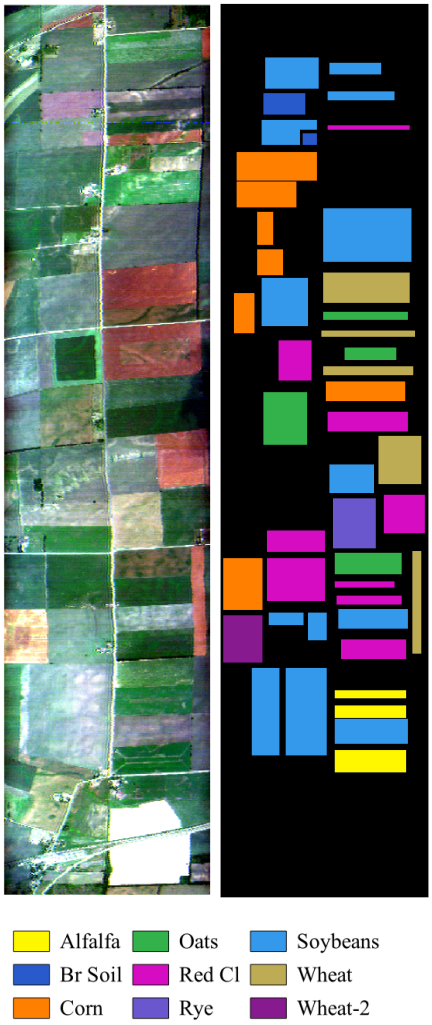


Figure 8: Distribution of original FLC1 test dataset.

APPENDIX B

The original FLC1 dataset was divided into 30 training datasets and one validation dataset. In this study, five out of 30 training datasets were used for experiments. In this appendix, the results of four out of five training datasets are listed below to show that all data sets are mostly similar in trend.

Table 5: Kappa coefficients and hyperparameters for FLC1 dataset (Dataset No. 2).

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (2318 points, α=0.000)	75% (1739 points, α=0.931)	50% (1152 points, α=0.953)	C	Gamma
Rate of assured training data to original one	Original	100% (2241 points, α=0.000)	**0.945**	0.972	0.982	1000.0	1.0
	Reduced by DNRS	90% (2014 points, α=0.910)	0.941	0.973	**0.986**	100.0	10.0
		80% (1790 points, α=0.926)	0.930	**0.975**	0.984	100.0	10.0
		70% (1570 points, α=0.936)	0.911	0.967	0.982	1000.0	10.0
		60% (1343 points, α=0.946)	0.877	0.943	0.983	100.0	10.0
		51% (1133 points, α=0.953)	0.867	0.944	0.979	100.0	10.0

Table 6: Kappa coefficients and hyperparameters for FLC1 dataset (Dataset No. 3).

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (2318 points, α=0.000)	75% (1739 points, α=0.931)	50% (1152 points, α=0.953)	C	Gamma
Rate of assured training data to original one	Original	100% (2241 points, α=0.000)	**0.947**	**0.971**	0.974	100.0	10.0
	Reduced by DNRS	90% (2009 points, α=0.909)	0.939	0.967	0.968	1000.0	10.0
		80% (1798 points, α=0.925)	0.933	0.967	0.970	1000.0	10.0
		70% (1578 points, α=0.936)	0.915	0.963	0.976	1000.0	10.0
		60% (1344 points, α=0.945)	0.901	0.955	**0.977**	1000.0	10.0
		50% (1116 points, α=0.952)	0.888	0.941	**0.977**	1000.0	10.0

Table 7: Kappa coefficients and hyperparameters for FLC1 dataset (Dataset No. 4).

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (2318 points, α=0.000)	75% (1739 points, α=0.931)	50% (1152 points, α=0.953)	C	Gamma
Rate of assured training data to original one	Original	100% (2284 points, α=0.000)	**0.952**	**0.974**	0.974	100.0	10.0
	Reduced by DNRS	90% (2051 points, α=0.912)	0.936	0.968	0.978	100.0	10.0
		80% (1827 points, α=0.927)	0.930	0.970	**0.981**	1000.0	1.0
		70% (1606 points, α=0.936)	0.909	0.956	0.980	100.0	10.0
		60% (1378 points, α=0.944)	0.898	0.951	0.978	100.0	10.0
		50% (1133 points, α=0.953)	0.871	0.932	0.977	100.0	10.0

Table 8: Kappa coefficients and hyperparameters for FLC1 dataset (Dataset No. 5).

			Rate of assured validation data to original one			SVM's Hyperparameters	
			Original	Reduced by DNRS			
			100% (2318 points, α=0.000)	75% (1739 points, α=0.931)	50% (1152 points, α=0.953)	C	Gamma
Rate of assured training data to original one	Original	100% (2306 points, α=0.000)	**0.945**	**0.970**	0.976	100.0	10.0
	Reduced by DNRS	90% (2077 points, α=0.911)	0.930	0.968	0.975	100.0	10.0
		80% (1846 points, α=0.927)	0.929	0.967	0.972	1000.0	10.0
		70% (1617 points, α=0.936)	0.906	0.956	0.977	1000.0	10.0
		60% (1390 points, α=0.944)	0.884	0.946	**0.978**	1000.0	1.0
		50% (1160 points, α=0.952)	0.853	0.925	0.976	100.0	10.0