Received: 15 May 2019 | Revised: 2 September 2019 | Accepted: 9 September 2019

DOI: 10.1111/ejss.12893

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

Synthetic resampling strategies and machine learning for digital soil mapping in Iran

Ruhollah Taghizadeh-Mehrjardi[1,2] | **Karsten Schmidt**[1,3] | **Kamran Eftekhari**[4] | **Thorsten Behrens**[1] | **Mohammad Jamshidi**[4] | **Naser Davatgar**[4] | **Norair Toomanian**[5] | **Thomas Scholten**[1,3]

[1]Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Tübingen, Germany

[2]Department of Nature Engineering, Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran

[3]Collaborative Research Center (CRC) 1070 Resource Cultures, University of Tübingen, Tübingen, Germany

[4]Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran

[5]Soil and Water Research Department, Isfahan Agricultural and Natural Resources Research and Education Center, AREEO, Isfahan, Iran

Correspondence

Ruhollah Taghizadeh-Mehrjardi and Karsten Schmidt, University of Tübingen, Department of Geosciences, Soil Science and Geomorphology, Rümelinstraße 19-23, D-72070 Tübingen, Germany.
Email: ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de (R. T.) and
Email: karsten.schmidt@uni-tuebingen.de (K. S.)

Kamran Eftekhari, Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran.
Email: keftekhari@swri.ir

Funding information

Alexander von Humboldt-Stiftung, Grant/Award Number: Ref 3.4 - 1164573 - IRN - GFHERMES-P; German Research Foundation, Grant/Award Number: B02; Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran

Abstract

Most common machine learning (ML) algorithms usually work well on balanced training sets, that is, datasets in which all classes are approximately represented equally. Otherwise, the accuracy estimates may be unreliable and classes with only a few values are often misclassified or neglected. This is known as a class imbalance problem in machine learning and datasets that do not meet this criterion are referred to as imbalanced data. Most datasets of soil classes are, therefore, imbalanced data. One of our main objectives is to compare eight resampling strategies that have been developed to counteract the imbalanced data problem. We compared the performance of five of the most common ML algorithms with the resampling approaches. The highest increase in prediction accuracy was achieved with SMOTE (the synthetic minority oversampling technique). In comparison to the baseline prediction on the original dataset, we achieved an increase of about 10, 20 and 10% in the overall accuracy, kappa index and F-score, respectively. Regarding the ML approaches, random forest (RF) showed the best performance with an overall accuracy, kappa index and F-score of 66, 60 and 57%, respectively. Moreover, the combination of RF and SMOTE improved the accuracy of the individual soil classes, compared to RF trained on the original dataset and allowed better prediction of soil classes with a low number of samples in the corresponding soil profile database, in our case for Chernozems. Our results show that balancing existing soil legacy data using synthetic sampling strategies can significantly improve the prediction accuracy in digital soil mapping (DSM).

Highlights

- Spatial distribution of soil classes in Iran can be predicted using machine learning (ML) algorithms.
- The synthetic minority oversampling technique overcomes the drawback of imbalanced and highly biased soil legacy data.
- When combining a random forest model with synthetic sampling strategies the prediction accuracy of the soil model improves significantly.

- The resulting new soil map of Iran has a much higher spatial resolution compared to existing maps and displays new soil classes that have not yet been mapped in Iran.

KEYWORDS

covariates, imbalanced data, machine learning, random forest, soil legacy data

1 | INTRODUCTION

Soils are one of the most valuable natural resources in many ways. For countries like Iran, the role of a soil in providing a sustainable resource for food production and water management is most important and is an essential link to, for example, nature conservation and biodiversity (Emadodin, Narita, & Bork, 2012). Therefore, spatial soil information is essential to reduce risks in environmental and agricultural planning (Mesgaran, Madani, Hashemi, & Azadi, 2017). In Iran, legacy soil maps and the related soil profile information are the main sources of soil information. However, many regions of Iran have not been mapped on a scale fine enough to evaluate, monitor, understand and maintain important soil functions, such as water holding capacity or carbon storage (Roozitalab, Siadat, & Farshad, 2018). Furthermore, the existing coarse-scale (1:1 M) soil class map (Banaei, 2000), although helpful, obviously shows imbalanced data on soil classes and is impractical for land use planning and agricultural practices, due to its insufficient information content (SWRI, 2015). Hence, it is necessary to develop higher resolution soil maps to be able to provide decision makers with maps that give detailed spatial soil information, which can be used for improving land management and crop guidelines. This holds true not only for Iran but for many other countries and regions worldwide (Sanchez et al., 2009).

Digital soil mapping (DSM) has been successfully applied for mapping of soil classes at a large range of scales, including regional and local scales (Brungard, Boettinger, Duniway, Wills, & Edwards, 2015; Hounkpatin et al., 2018; Schmidt, Behrens, & Scholten, 2008), the national scale (Adhikari, Minasny, Greve, & Greve, 2014; Ramcharan et al., 2018) and the continental scale (Hengl et al., 2017; Teng, Rossel, Shi, & Behrens, 2018). Digital soil mapping can integrate soil point observations (e.g., classified soils) with various sources of grid-based geospatial covariates (e.g., satellite imagery, digital elevation models and climate data). This is enabled by using machine learning (ML) algorithms that relate the covariates to any soil information (McBratney, Santos, & Minasny, 2003).

In DSM, particularly at the national scale, the soil data mainly consist of legacy soil profiles (Stumpf et al., 2016). Although they provide valuable local information on soil classes or properties, the use of legacy soil data in machine learning is challenging due to a number of problems related to the nature of these data (Hounkpatin et al., 2018). Most of the common ML algorithms consider balanced training sets, that is, datasets where all classes are approximately represented equally. Because these algorithms treat all misclassifications equally, they have a bias towards classes with many instances, which often results in false accuracy estimates (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) and the misclassification or neglect of classes with only a few instances (Batista, Prati, & Monard, 2004). This is known as the class imbalance problem in ML (He & Garcia, 2008). In this respect, datasets that do not follow this criterion are called imbalanced data. Most soil class datasets are therefore imbalanced data (Hounkpatin et al., 2018). This stems from the spatial distribution pattern of soils, which is usually not equal. It can additionally be influenced by the sampling strategy.

Several approaches have been developed in the ML community to handle imbalanced data. One is the design of new models which can handle imbalanced data directly, for example, by applying cost functions that penalize wrong classification (Chawla et al., 2002). Another approach is to apply different evaluation metrics instead of the overall accuracy, such as precision and recall (He & Garcia, 2008). A third approach is to resample the data (Piri, Delen, & Liu, 2018). In this study. we focus on resampling methods but also test different evaluation metrics on the original as well as the resampled datasets.

Several resampling approaches have been proposed which can be separated into two groups: (a) data-driven and (b) algorithm-driven methods (He & Garcia, 2008). Most researchers have employed data-driven methods (Piri et al., 2018) which use resampling techniques to adjust the ratio between the classes in the training set (Chawla et al., 2002). In their simplest forms, random oversampling (ROS) increases the minority class data by the random replication of their occurrence, and random undersampling (RUS) decreases the number of majority class data by randomly removing data from the original dataset. This consequently allows ML algorithms to be learned from the balanced data without bias (He & Garcia, 2008). However, these classical random resampling techniques could increase the chances of overfitting or potentially discard useful observations (Piri et al., 2018). To account for such shortcomings, more sophisticated resampling techniques for speech and image recognition have been proposed, such as the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002), one-sided selection (Batista et al., 2004) and the adaptive synthetic sampling approach (ADASYN) (Branco, Torgo, & Ribeiro, 2016).

Although many of the resampling techniques have been proposed and successfully applied to cope with problems of imbalanced data in mathematics and informatics (Tkachenko, Doroshenko, Izonin, Tsymbal, & Havrysh, 2018), to the best of our knowledge, these techniques have not been widely tested in DSM studies for large areas and at a national scale. There are only a few studies related to different balanced sampling techniques for soil science, mostly limited in sampling size and techniques. For instance, Heung et al. (2016) tested one method (ROS) for generating balanced training data from a conventional soil survey. They indicated that the use of ROS resulted in computational limitations that might have been present because the training data were derived from soil survey data. Sharififar, Sarmadian, Malone, and Minasny (2019) evaluated two resampling techniques (ROS and RUS) to cope with the issue of imbalanced soil data with 452 profiles observations in an area covering about 12,000 ha. Therefore, our paper investigates how to solve the problem of imbalanced soil data using different resampling techniques and legacy soil information for a large area (1,648,195 km^2) and at the national scale.

We tested five ML algorithms on eight resampled balanced datasets generated from the original imbalanced dataset to compare the influence on different algorithms. We analysed and discussed the influence of the resampled balanced datasets as well as the prediction accuracies. Our main objective is to apply and test ML algorithms and resampling techniques for imbalanced legacy soil information in DSM.

2 | MATERIALS AND METHODS

2.1 | Study area

Iran, bounded to the north by the Caspian Sea and to the south by the Persian Gulf and Sea of Oman, is the second largest country in the Middle East and covers an area of about 1,648,195 km^2 (Figure 1a). Iran has a diverse topography, which is due to the existence of two major mountain systems, namely the Alborz and the Zagros ranges. The Alborz range starts from the north-west and extends like an arc on the south of the Caspian Sea to the east, separating humid climate sections in the north from arid central Iran. The Zagros range covers the western parts of Iran, extending to the south and separating the semi-arid western parts from central arid basins (Figure 1b). The average altitude is 1,200 m, the maximum elevation of the country is in the centre of the Alborz chain at 5,671 m above sea level and the minimum elevation is on the southern coast of the Caspian Sea at 28 m below sea level.

Most of the country has arid (65%) and semi-arid (20%) climate conditions. Temperature ranges from −20°C to

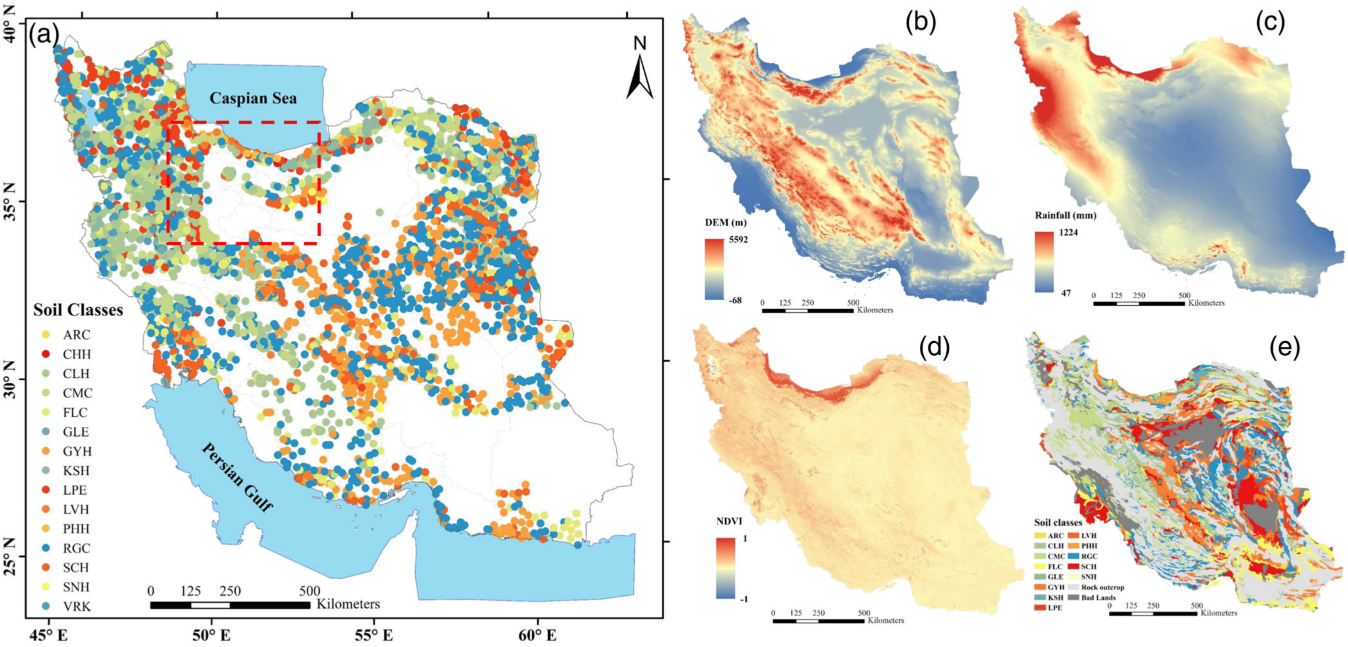


FIGURE 1 (a) The location of the study area (Iran) and spatial distribution of soil legacy dataset, (b) a digital elevation model (DEM), (c) mean annual rainfall (millimeter: mm), (d) normalized difference vegetation index (NDVI) and (e) a traditional soil map. ARC, Arenosols; CHH, Chernozems; CLH, Calcisols; CMC, Cambisols; FLC, Fluvisols; GLE, Gleysols; GYH, Gypsisols; KSH, Kastanozems; LPE, Leptosols; LVH, Luvisols; PHH, Phaeozems; RGC, Regosols; SCH, Solonchaks; SNH, Solonetz; VRK, Vertisols [Color figure can be viewed at wileyonlinelibrary.com] greater than 50°C throughout the country and during the year. The average annual rainfall is 250 mm, which is about one-third of the world's average precipitation, ranging from less than 100 mm in the central region of the country to 1,200 mm in the north (Mesgaran et al., 2017) (Figure 1c). It has been estimated that 70% of precipitation in Iran evaporates. Because of the climatic and topographic contrasts, together with different geological substrates, Iran shows a diversity of plant communities. The potential natural vegetation consists of oak, beech, linden and elm in the more humid sections of the north, and the thin cover of grasses and scattered shrubs in the semi-arid and arid regions. In addition, the variation in the density of plant cover throughout the country could be inferred from the analysis of the normalized difference vegetation index (NDVI), where the dominant trend indicated an increase from south to north, correlating with climate conditions (Figure 1d).

Diverse topography, climate, geology and vegetation cover have led to formation of a high variety of soils that cover about 58% of the Iranian landscapes (Roozitalab et al., 2018). The remaining landscapes are rocky mountains, outcrops, badlands, sand dunes, salinas (Dagh), lakes and others, covering about 69 million hectares or 42% of the total land area of the country. The conventional soil map provided by the Soil and Water Research Institute of Iran (Banaei, 2000) (Figure 1e) shows that Regosols (22%), Gypsisols (20%), Cambisols (17%), Solonchaks (15%) and Calcisols (12%) constitute about 87% of the total soil resources of the country. Soil groups including Kastanozems, Gleysols, Phaeozems and Luvisols, which have mainly developed in the Caspian Sea region, constitute less than 3% of the total soil cover. Gypsisols and Solonchaks are widespread in the central main plateau between the Alborz and Zagros mountain chains, where the climate is arid or super-arid and the environment is mostly desert. Calcisols and Cambisols are developed in semi-arid regions along the sloping landscapes of the two mountain chains and Regosols are developed on highlands with steep slopes. Kastanozems, Gleysols, Phaeozems and Luvisols are mostly developed under sub-humid to humid climates of the Caspian Sea (SWRI, 2015).

2.2 | Procedures

This work was conducted in three main steps (Figure 2): (a) preprocessing of soil datasets, (b) acquisition of covariates and (c) calibrating of ML algorithms.

2.3 | Preprocessing of soil datasets

2.3.1 | Original imbalanced dataset

In this study, we used the Soil Profile Database (SPDB) of Iran, which consists of 7,664 soil profiles (Banaei, 2000; SWRI, 2015). Sampling locations in the study area are presented in Figure 1a. The soil profile locations were selected by different sampling strategies, including stratified random sampling (~87%), grid sampling (~8%) and the conditioned latin hypercube sampling approach (~5%). However, the excavation and the description of soil profiles, laboratory analyses of soil samples and classification of soils were conducted using common methodologies (Sparks, 1998; WRB, 2006). The soils in SPDB are classified into 15 World Reference Base (WRB) groups: Arenosols, Chernozems, Calcisols, Cambisols, Fluvisols, Gleysols, Gypsisols, Kastanozems, Leptosols, Luvisols, Phaeozems, Regosols, Solonchaks, Solonetz, and Vertisols (Table 1).

2.3.2 | Resampled balanced datasets

We used and tested eight resampling techniques (random under- and over-sampling, synthetic minority oversampling, adaptive synthetic sampling, the introduction of Gaussian Noise, Tomek link, condensed nearest neighbours and one-sided selection method) using the “Utility Based Learning for Classification and Regression tasks” (UBL) package in R

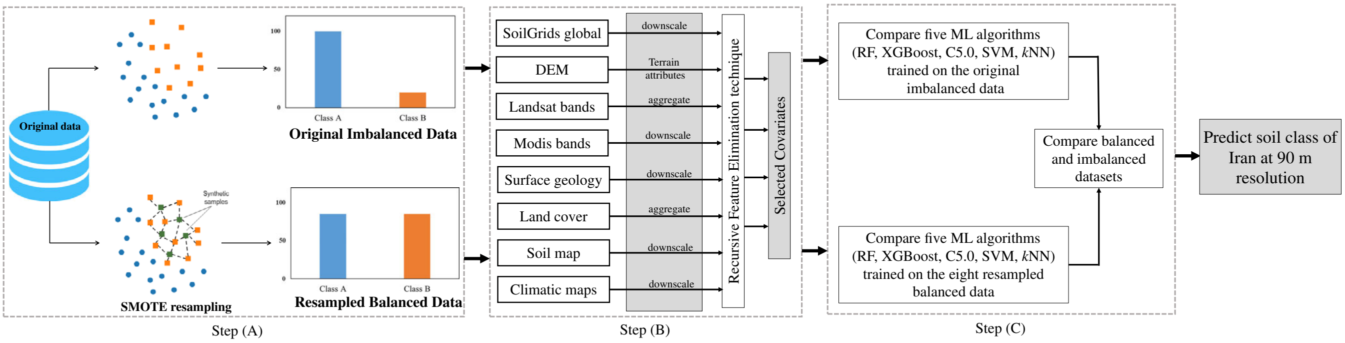

Original data
100
50
0
Class A
Class B
Original Imbalanced Data
Synthetic samples
SMOTE resampling
Resampled Balanced Data
Step (A)
SoilGrids global
DEM
Landsat bands
Modis bands
Surface geology
Land cover
Soil map
Climatic maps
downscale
Terrain attributes
aggregate
Recursive Feature Elimination technique
Selected Covariates
Step (B)
Compare five ML algorithms (RF, XGBoost, C5.0, SVM, kNN) trained on the original imbalanced data
Compare balanced and imbalanced datasets
Compare five ML algorithms (RF, XGBoost, C5.0, SVM, kNN) trained on the eight resampled balanced data
Predict soil class of Iran at 90 m resolution
Step (C)

FIGURE 2 Overview of employed methods. C5.0, decision tree; DEM, digital elevation model; *k*NN, *k*-nearest neighbour; ML, machine learning; RF, random forest; SMOTE, a dataset oversampled using the synthetic minority oversampling technique; SVM, support vector machine; XGBoost, extreme gradient boosting [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Number of training and validation data in each soil class

WRB groups	Abbreviation	Training data	Validation data	Total	In %
Arenosols	ARC	20	8	28	0.3
Chernozems	CHH	7	2	9	0.1
Calcisols	CLH	1,337	573	1,910	24.9
Cambisols	CMC	986	422	1,408	18.3
Fluvisols	FLC	316	135	451	5.9
Gleysols	GLE	45	19	64	0.8
Gypsisols	GYH	694	297	991	12.9
Kastanozems	KSH	168	71	239	3.1
Leptosols	LPE	234	100	334	4.3
Luvisols	LVH	64	27	91	1.2
Phaeozems	PHH	31	12	43	0.5
Regosols	RGC	1,001	428	1,429	18.6
Solonchaks	SCH	336	143	479	6.2
Solonetz	SNH	55	23	78	1.0
Vertisols	VRK	77	33	110	1.4

Abbreviation: In, intensity.

(Branco, Ribeiro, & Torgo, 2016) in order to balance the class distribution in the training dataset. In this paper, fully balanced soil datasets were generated; that is, datasets where all soil classes are represented by the same number of samples (Burez & Van den Poel, 2009; Estabrooks, Jo, & Japkowicz, 2004). For the sake of clarity, we plotted each of the eight resampling techniques in Figure 3 and then described them briefly in turn. In any case, a synthetic sample is a computer-generated new soil sample within a given covariate space based on different statistical assumptions and, therefore, counts as an artificial covariate space description of a given soil class. These synthetic samples are only valid in the tested environment. A more detailed description of the eight used resampling techniques can be found in Branco, Torgo, and Ribeiro (2016).

Random undersampling (RUS) is a simple strategy of randomly removing samples of the majority classes (e.g., Calcisols) according to the size of the minority class (e.g., Chernozems) to generate a balanced dataset. This is one of the earliest techniques used to alleviate imbalance in the dataset, however, it may increase the variance of the ML algorithms (He, Bai, Garcia, & Li, 2008).

Random oversampling (ROS) is a simple technique that increases the number of minority class data by random replication, whereby the replicate is content-identical to its source and no new contextual variation will be added. This technique builds a new set of representatives of the minority class (e.g., Chernozems) according to the size of the majority class (e.g., Calcisols) to balance the classes.

The synthetic minority oversampling technique (SMOTE) is a more sophisticated technique compared to ROS. It performs oversampling by creating synthetic examples, in which samples of a minority class (e.g., Chernozems) are interpolated in the covariate space to generate new examples of that specific class (Chawla et al., 2002). Thus, based on the *k*-nearest neighbours (*k*NN)of a minority class, linear functions between all adjacent neighbours and covariates are generated and one new synthetic sample is generated along this function. By repeating this methodological workflow, an equal number of samples for each class is generated.

The adaptive synthetic (ADASYN) sampling approach works similarly to SMOTE, as shown in Figure 3. However, it generates more synthetic examples for the minority class (e.g., Chernozems) along the linear function by weighting the distance. According to He and Garcia (2008) ADASYN focuses on the minority class examples according to their level of difficulty in learning. The essential idea of ADASYN is to use a weighted distribution for different minority classes as a criterion to decide the number of synthetic samples that need to be generated for each minority class (He et al., 2008). The weight is calculated according to:

$$w = \frac{\Delta}{K} \tag{1}$$

where Δ is the number of examples in the K nearest neighbours of a minority class that belong to the majority class and the value of w ranges from 0 to 1.

The introduction of Gaussian Noise (GN) was first proposed by Lee (1999) and adapted by Branco, Ribeiro, and

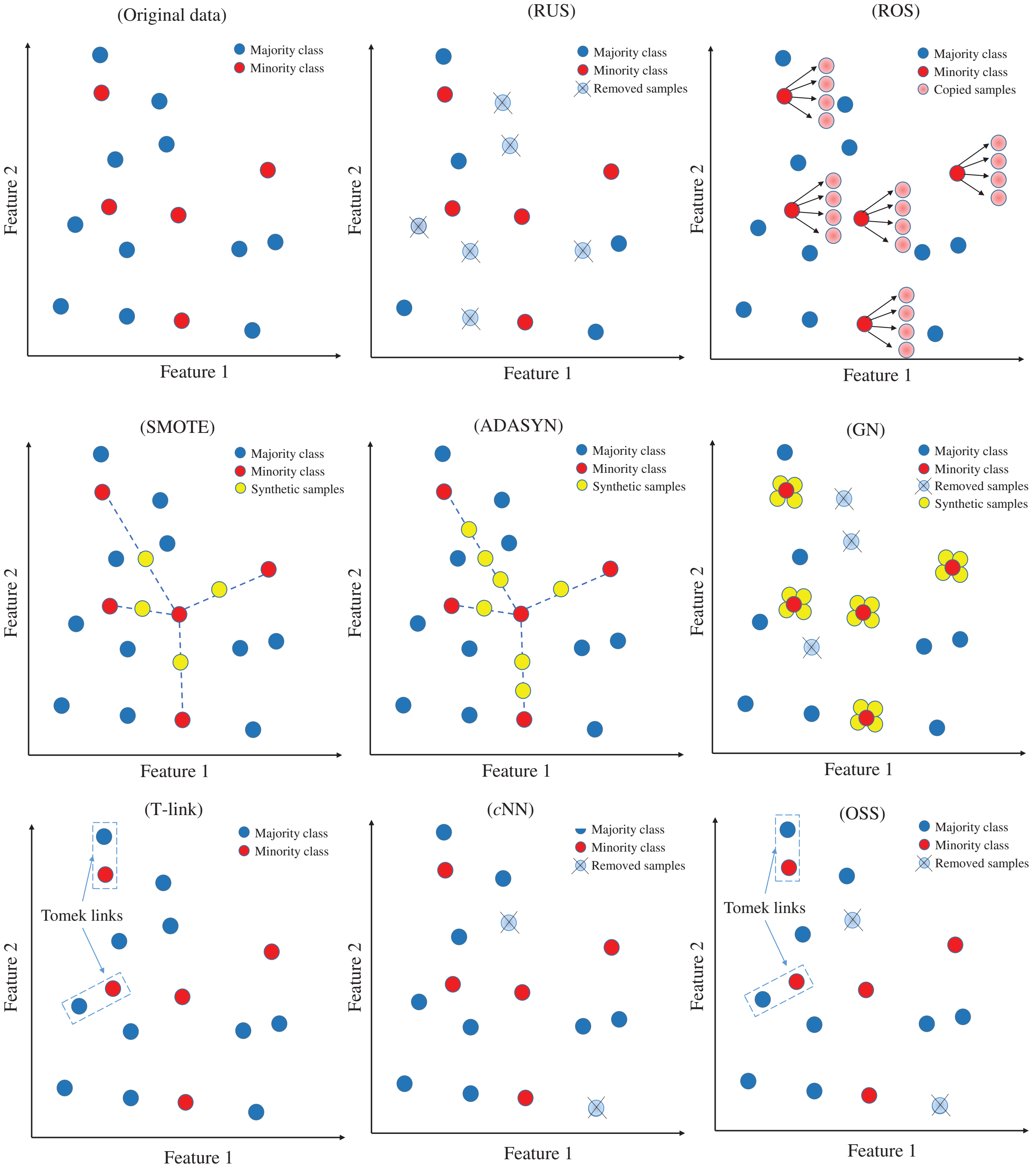


FIGURE 3 An illustrative overview of the eight resampling methods used in this study. ADASYN, adaptive synthetic sampling; *c*NN, condensed nearest neighbours; GN, the introduction of Gaussian noise; OSS, one-sided selection method; ROS, random oversampling; RUS, random undersampling; SMOTE, synthetic minority oversampling; T-link, Tomek link [Color figure can be viewed at wileyonlinelibrary.com]

Torgo (2016) to generate a balanced dataset using a combination of random under- and oversampling. The algorithm starts by applying random undersampling to the majority class (e.g., Calcisols). Regarding the oversampling method, a new synthetic example of the minority class (e.g., Chernozems) is obtained by introducing a small perturbation on existing samples through Gaussian noise, in which the noise depends on the standard deviation of each numeric covariate (evaluated on the examples of minority class). This means that each covariate value (i) of the new synthetic example new_i is built as follows:

$$new_i = ex_i + rnorm(0, sd(i) \times pert) \tag{2}$$

where ex_i represents the original example value for covariate i, $sd(i)$ represents the evaluated standard deviation for covariate i in the class under consideration (e.g., Chernozems) and $pert$ is a number indicating the level of perturbation to introduce when generating synthetic examples ($pert = 0.1$).

The Tomek link (T-link) method is the first of the so-called distance-based resampling methods. It builds on the idea that if two samples belonging to different classes are each other's nearest neighbour based on the covariate space, it will negatively influence the performance of the ML algorithms (Tomek, 1976). Therefore, the T-link method removes one of those two samples (i.e., the example of the majority class (e.g., Calcisols)), which increases the distance within the covariate space between the two classes and improves the learning.

The condensed nearest neighbours (*c*NN) method was proposed by Kubat and Matwin (1997) and adapted by Branco, Ribeiro, and Torgo (2016). In this undersampling strategy, a subset of samples that is consistent with the original data is generated. A subset of examples is said to be a consistent subset of the original data if, and only if, for every point in the original data its nearest neighbour in the subset has the same class. Beside the reduction of redundancies within the training set, we also strengthen the focus on a specific covariate range of a given class. This should allow us to better characterize the minority class (e.g., Chernozems) alongside minor changes in the classification accuracy of the majority classes.

The one-sided selection (OSS) method uses both the T-link undersampling strategy and the (*c*NN technique (Batista et al., 2004).

2.3.3 | Grid-based geospatial covariates

We collected and calculated 110 covariates, which were representatives of *scorpan* factors (McBratney et al., 2003) from five sources: (a) remote sensing images, (b) digital elevation model, (c) climatic maps, (d) available soil property maps and (e) digitized choropleth maps. All covariates were aggregated (average resampling) or disaggregated (bilinear resampling) to a common grid of 90 × 90 m spatial resolution.

Organisms (*o*) and parent material (*p*) factors were characterized by remote sensing images, using median values of six spectral bands (i.e., bands of 2, 3, 4, 5, 6 and 7) of Landsat 8 (Wulder et al., 2016) cloud-free images taken during 2016 with 30 × 30 m spatial resolution. Furthermore, remote sensing indices were also computed (e.g., salinity index and carbonate index). We used MODIS products with 250 × 250 m spatial resolution, including median values of two surface spectral reflectance of MODIS images, the enhanced vegetation index, the normalized difference vegetation index and daytime and night-time land surface temperature (Mira et al., 2015).

Furthermore, we derived 30 terrain attributes (e.g., mid-slope position and wetness index) from a preprocessed Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) with 90 × 90 m resolution using SAGA GIS (Conrad et al., 2015) to represent the terrain factor (*r*). The climate factor (*c*) has the potential to explain large parts of the variation of soil classes in Iran because of its high spatial variability and range over the country area. Therefore, 29 maps of climatic surfaces (e.g., annual mean precipitation and annual mean temperature) obtained by WorldClim (Hijmans et al., 2005) were used.

Additionally, we used 10 soil property maps (e.g., soil organic carbon and coarse fragments) as continuous covariates. These soil maps were provided by ISRIC (the International Soil Reference and Information Centre) (Hengl et al., 2017). As categorical predictor variables, we used five choropleth maps, which were compiled at different cartographic scales (e.g., soil map and land-use map) (Banaei, 2000). Four examples of covariates are shown in Figure 1.

In the next step, we used recursive feature elimination (RFE) to identify the most suitable covariates for modelling (Kuhn & Johnson, 2013) and to reduce the dimensionality. The RFE algorithm first ranks all covariates according to their importance, given by coefficients. Then, those predictor variables having the least influence on the prediction performance of models are eliminated. That procedure was recursively repeated until all the covariates were eliminated. Features were then ranked according to when they were eliminated.

2.3.4 | Calibration of ML algorithms

Five ML algorithms (i.e., C5, random forest, extreme gradient boosting, support vector machine and *k*-nearest neighbour) were evaluated using the caret package (Kuhn & Johnson, 2013) to build the relationship between soil classes and covariates.

Decision tree analysis, a commonly used technique in DSM, was implemented using the C5.0 algorithm (Quinlan, 1992). This method uses a series of binary rules (if–then statements) in the form of an algorithm having the structure of a tree consisting of nodes and leaves to classify the soil classes (Breiman, 1996). Considering small within-node variance and high between-node variance rules, the predictor variables are used to split the training dataset into two subsets. Partitioning is stopped if a minimum tolerated number of samples (pixels) is reached in a node of the tree. This threshold influences the size of the tree in terms of end nodes and is thus strongly related to overfitting and generalization. The terminal nodes, namely leaves, will present soil classes. Notably, for each terminal node the majority class label is assigned for the final classification results.

Two ensemble techniques, namely random forest (RF) and extreme gradient boosting tree (XGBoost) models, were also employed. Random forest develops a large number of independent decision trees using different subsets of the training data and a different combination of predictor variables (Breiman, 2001). The three well-known tuning parameters of RF models are mtry (the number of predictor variables), ntree (the number of trees) and sampsize (the size of sample to be used in each tree), and they are optimized by caret (Behrens, Zhu, Schmidt, & Scholten, 2010; Kuhn & Johnson, 2013). The final prediction is the average of all single trees. In addition, by calculating the mean decrease accuracy, the RF algorithm ranks the importance of each covariate.

Rather than building independent trees by RF, XGBoost models (Chen, He, Benesty, & Khotilovich, 2019) generate a number of trees sequentially, in which each new tree tries to improve the classification error of the previously constructed trees. In the first iteration, the XGBoost algorithm gives more weight to the samples that are badly predicted and the new trees are forced to focus on those difficult to learn samples.

Support Vector Machines (SVMs) developed by Cortes and Vapnik (1995) separate the dataset into different classes by constructing hyperplanes in a multidimensional space. To find an optimal hyperplane with the greatest possible margin between the hyperplane and any point within the training set, SVM with radial basis function (RBF) kernels uses an iterative training algorithm in order to minimize an error function.

The *k*NN is an instance-based learner used for classification and regression that simply tries to classify a new sample in a dataset according to a combination of the classes of the *k* instance(s) located the closest in covariate space distance to it in a training dataset (Hastie, Tibshirani, & Friedman, 2009) and, therefore, can be seen as a supervised classifier. Here, standard Euclidean distance was implemented to quantify a distance between the new samples and the training samples.

2.3.5 | Validation of ML algorithms

In order to test the accuracy of predictions of all ML algorithms, the soil dataset was divided randomly into two sets. The larger set was used for training (70% = 5,371 soil samples) and the smaller set was set aside for validation (30% = 2,293 soil samples). Five different accuracy metrics were used, including overall accuracy (OA), kappa index (K-index), recall, precision and F- score, which is in accordance with general recommendations for describing the quality of ML algorithms for imbalanced datasets (Chawla et al., 2002). All the accuracy metrics are functions of the confusion matrix as shown in Table 2. In brief, OA is a metric calculating the classifier overall accuracy; the K-index measures interrater agreement for instances; recall is the proportion of those instances that are correctly classified; precision is the proportion of those predicted instances that are correctly classified; and the F-score is the harmonic mean of precision and recall (Equations 3–7). Importantly, 10 iterations of training and validation were applied to provide more reliable accuracy metrics. Then the average values of the accuracy metrics and their standard deviations were calculated.

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$K - index = 1 - \frac{1 - OA}{1 - P_e} \tag{7}$$

where *TP*, *TN*, *FP* and *FN* are true positive, true negative, false positive and false negative, respectively (Table 2). P_e is the hypothetical probability of chance agreement.

TABLE 2 An example of a confusion matrix

	Actual positive class	Actual negative class
Predicted positive class	TP (true positive)	FP (false positive)
Predicted negative class	FN (false negative)	TN (true negative)

TABLE 3 The accuracy of ML algorithms trained on the original imbalanced dataset

Models	Overall accuracy (%)	Kappa (%)	F-score (%)
RF	58.4 ± 2.1	49.8 ± 1.5	49.6 ± 2.2
XGBoost	58.1 ± 1.0	49.4 ± 1.2	49.1 ± 2.1
C5.0	56.4 ± 2.1	47.5 ± 2.4	48.4 ± 1.4
SVM	51.0 ± 3.1	41.5 ± 2.3	44.5 ± 1.3
*k*NN	48.0 ± 1.2	36.7 ± 1.5	37.0 ± 2.4

Abbreviations: C5.0, decision tree analysis; *k*NN, *k*-nearest neighbour; ML, machine learning; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

3 | RESULTS AND DISCUSSION

3.1 | The original imbalanced dataset

Table 3 summarizes the validation results of ML algorithms trained on the original imbalanced dataset. As can be seen, the highest accuracy was achieved by the RF model, in which OA, K-index and F-score were 58.4, 49.8 and 49.6%, respectively. Other models tested were XGBoost (OA of 58.1%, K-index of 49.4% and F-score of 49.1%) and C5.0 (OA of 56.4%, K-index of 47.5% and F-score of 48.4%). The *k*NN had the lowest performance, with OA, K-index and F-score of 48.0, 36.7 and 37.0%, respectively. Similar to *k*NN, the SVM model showed poor prediction power, with OA of 51.0%, K-index of 41.5% and F-score of 44.5%. This might be attributable to the fact that *k*NN and SVM classifiers are more sensitive to the imbalanced class distribution in comparison to the tree-based ML algorithms, and thus cannot handle multiclass imbalanced problems (Yang, Zhou, Zhu, Ma, & Ji, 2016). This is in line with the results of Piri et al. (2018), who pointed out the performance of SVM, which deteriorates dramatically when applied to imbalanced datasets. Because of mathematical characteristics, the SVM decision boundary is closer toward the minority class region compared to the ideal classification decision boundary in an imbalanced dataset, as in our case.

Generally speaking, the performances of ML algorithms trained on the original imbalanced dataset indicated that the two ensemble-based models (RF and XGBoost) show high and fairly similar accuracy (Table 3). Our findings are in line with results of several DSM literature reviews, which all confirmed the power of ensemble-based models compared to the other common ML algorithms (Brungard et al., 2015; Hounkpatin et al., 2018). However, a closer inspection of the calculated recall values of individual classes revealed that a considerable number of minority classes will be misclassified as majority classes, such as Chernozems, Phaeozems and Solonetz (Table 4). In other words, the minority classes are overpredicted. We observed that the highest recall values were obtained by the Calcisols (~70.8%, 1,910 samples) and the lowest recall values were attributed to the Chernozems (0%, nine samples). These results were expected for an imbalanced dataset, in which classes with lower sampling frequencies (e.g., Chernozems, Phaeozems and Solonetz) were mostly modelled less accurately, compared to the majority classes (e.g., Calcisols, Regosols and Cambisols). Consequently, the soil class map has bias toward the majority class. Similar problems have been reported by other DSM researchers who found that there is a positive relationship between the sample size and the accuracy of individual soil classes (Brungard et al., 2015; Hengl, Toomanian, Reuter, & Malakouti, 2007). As an example, Jafari, Finke, Van de Wauw, Ayoubi, and Khademi (2012) found a relatively poor prediction of some soil groups that had only a few pedon observations in relation to the area. Taken together, obtained findings on the imbalanced soil data need to be interpreted with caution.

3.2 | The resampled balanced datasets

Figure 4 shows the results obtained with the different resampling approaches. Similar to the original imbalanced dataset, RF outperforms all other models for all resampled balanced datasets. The results indicate that all ML algorithms trained on the SMOTE resampled balanced data achieved the overall best performance, which is in accordance with the studies from other fields (Chen et al., 2018; Zarinabad et al., 2017). The general trend reveals ADASYN and GN as the second and third most effective balancing approaches, respectively. The poorest performance was obtained with the RUS resampled data. This clearly indicates that discarding the samples randomly from the original dataset decreased the power of ML algorithms (Lauron & Pabico, 2016). However, the other three undersampling techniques (*c*NN, T-link and OSS) performed better compared to RUS. Nevertheless, they did not achieve superior performance in comparison to the ML algorithms trained on the three oversampling techniques (SMOTE, ADASYN and GN).

In general, we conclude that oversampled data perform better compared to undersampled data (Estabrooks et al., 2004). This can be explained by the fact that undersampling techniques ignore useful information embedded in the instances of the majority class and hence degrade classifier performance (Zarinabad et al., 2017). It is worth noting that the ML algorithms trained on the ROS resampled data achieved a poor performance in comparison to those trained on the other three oversampled data (SMOTE, ADASYN and GN). This indicates that in a dataset with a huge difference between minority (e.g., Chernozems: 0.12%) and majority (e.g., Calcisols: 24%) classes, replication (i.e., copied samples) of the minority group randomly at a

TABLE 4 Confusion matrix of RF trained on the original imbalanced dataset

	ARC	CHH	CLH	CMC	FLC	GLE	GYH	KSH	LPE	LVH	PHH	RGC	SCH	SNH	VRK	Precision (%)
ARC	**3.8**	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	**79.1**
CHH	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
CLH	1.6	0.0	**405.8**	98.8	29.4	2.0	25.2	13.6	21.6	1.4	0.2	87.0	7.2	3.0	4.8	**57.8**
CMC	0.2	0.0	41.2	**266.2**	20.0	5.2	4.4	8.4	6.4	7.4	3.0	25.8	3.2	0.0	7.0	**66.8**
FLC	0.2	0.0	7.4	4.0	**32.0**	0.2	2.4	0.2	0.2	1.2	0.2	8.8	2.0	0.0	0.4	**54.0**
GLE	0.0	0.0	2.0	1.2	0.4	**7.4**	0.0	0.0	0.2	0.6	0.2	0.0	0.2	0.0	0.2	**59.7**
GYH	0.0	0.0	24.6	5.8	13.0	0.0	**172.6**	0.0	1.6	0.0	0.0	48.6	19.4	10.2	0.0	**58.3**
KSH	0.0	0.4	6.0	6.6	2.0	0.6	0.0	**43.0**	2.8	2.0	2.0	2.4	0.0	0.0	2.2	**61.4**
LPE	0.0	0.2	4.2	3.2	3.4	0.0	2.2	0.6	**44.6**	0.4	0.8	6.0	0.0	0.0	0.8	**67.1**
LVH	0.0	0.0	1.2	3.6	0.0	0.0	0.2	1.8	0.8	**10.6**	2.0	0.4	0.0	0.0	0.0	**51.4**
PHH	0.0	0.0	0.0	0.2	0.2	1.6	0.0	1.0	0.0	0.0	**3.2**	0.0	0.4	0.0	0.0	**48.5**
RGC	2.2	1.4	69.6	19.8	28.8	0.8	70.8	1.6	20.6	3.2	0.4	**238.4**	16.2	0.8	1.0	**50.1**
SCH	0.0	0.0	8.0	6.6	5.4	1.0	19.2	0.0	0.4	0.2	0.0	9.4	**94.4**	1.0	0.0	**64.5**
SNH	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**8.0**	0.0	**97.5**
VRK	0.0	0.0	2.0	6.0	0.4	0.2	0.0	0.8	0.8	0.0	0.0	1.0	0.0	0.0	**16.6**	**59.7**
Recall (%)	**47.5**	**0.0**	**70.8**	**63.1**	**23.7**	**38.9**	**58.1**	**60.5**	**44.6**	**39.2**	**26.6**	**55.7**	**66.0**	**34.7**	**50.3**	

Abbreviations: ARC, Arenosols; CHH, Chernozems; CLH, Calsisols; CMC, Cambisols; FLC, Fluvisols; GLE, Gleysols; GYH, Gypsisols; KSH, Kastanozems; LPE, Leptosols; LVH, Luvisols; PHH, Phaeozems; RF, random forest; RGC, Regosols; SCH, Solonchaks; SNH, Solonetz; VRK, Vertisols.

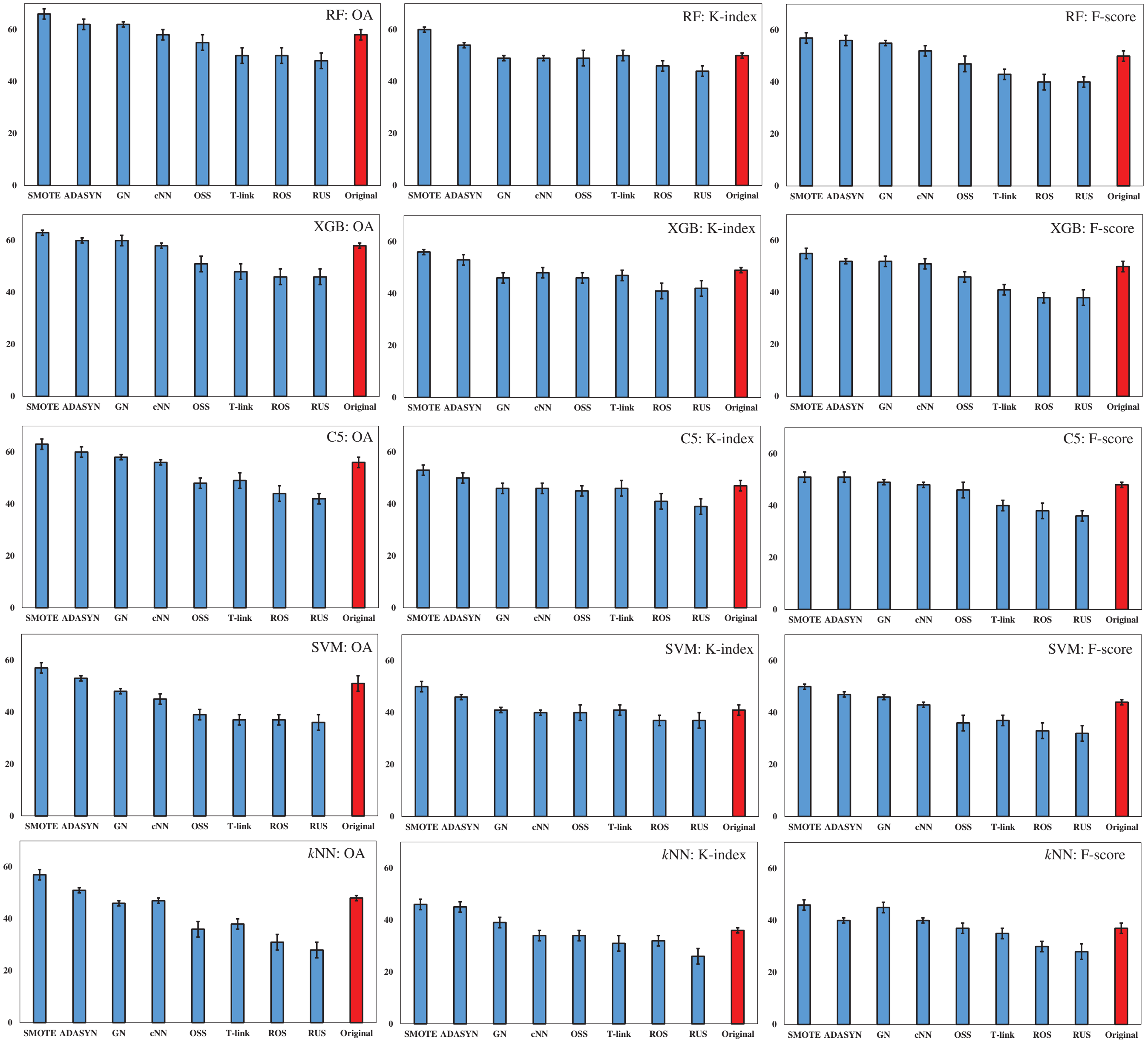


FIGURE 4 The accuracy metrics of five ML algorithms trained on the original imbalanced data (red) and eight resampled balanced datasets (blue). ADASYN, a dataset oversampled using the adaptive synthetic sampling approach; C5.0, decision tree analysis; *c*NN, a dataset undersampled using condensed nearest neighbours; GN, a dataset oversampled using the introduction of Gaussian noise; *k*NN, *k*-nearest neighbuor; ML, machine learning; OSS, a dataset undersampled using the one-sided selection method; RF, random forest; ROS, a dataset oversampled using random oversampling; RUS, a dataset undersampled using random undersampling; SMOTE, a dataset oversampled using the synthetic minority oversampling technique; SVM, support vector machine; T-link, a dataset undersampled using Tomek link; XGBoost, extreme gradient boosting [Color figure can be viewed at wileyonlinelibrary.com]

high rate is not necessarily the best option for solving the imbalanced learning problem. Several authors agree because ROS makes exact copies of the minority class examples; this might increase the likelihood of overfitting (Zarinabad et al., 2017). In addition, Chawla et al. (2002) stated that the simple replication of samples in ROS can make the decision region smaller and more specific for the minority samples. However, the synthetic examples (e.g., SMOTE) cause the ML algorithms to create larger and less specific decision regions (Amin et al., 2016; Haixiang et al., 2017). Therefore, the trained ML algorithms on the ROS resampled data could not precisely be generalized to the unseen data. This is one reason why we obtained a poor performance of ML algorithms using ROS resampled data, which is in line with the findings of Hounkpatin et al. (2018), who pointed out the poor power of generalization of RF models trained by ROS resampled data. Contrary to our findings, Sharififar, Sarmadian, Malone, and Minasny (2019) indicated a significant improvement in ML learning when they made balanced soil data using ROS. They also pointed out that balancing the soil classes led to a notable decrease in uncertainty of ML algorithms (Sharififar, Sarmadian, & Minasny, 2019).

Furthermore, balancing the dataset improved the prediction power of ML algorithms, compared to the original imbalanced data shown as baseline in Figure 4. This is particularly true when the data are preprocessed using the SMOTE resampling technique, which is better than any other resampling techniques and independent of the ML algorithm (Chawla et al., 2002; Tantithamthavorn, Hassan, & Matsumoto, 2018). For instance, RF trained on the SMOTE resampled dataset increased OA, K-index and F-Score by ~10, ~20 and ~10%, respectively, in comparison to the RF model trained on the original imbalanced dataset. This result confirmed that the SMOTE resampling technique could successfully handle the skewed class distribution, as also reported by Lauron and Pabico (2016). Like SMOTE, the ADASYN algorithm has achieved sufficiently higher performance. This is because ADASYN not only provided a balanced data distribution but also forced the ML algorithms to focus on complex minority class examples in the dataset (Amin et al., 2016). Our results generally indicate that ML algorithms could be better trained by synthetic resampling techniques (SMOTE, ADASYN and GN), which is in line with several published literature reviews (Piri et al., 2018; Zarinabad et al., 2017).

Contrary to our expectations, we found no prediction improvement when using *c*NN, T-link and OSS undersampling techniques in comparison to original data because useful information on the majority class might be lost through undersampling techniques. In other words, undersampling techniques can significantly improve the accuracy of classifiers if redundant information (e.g., majority class instances with nearly identical information or meaning) is present in the data space (Devi & Purkayastha, 2017). We do not see these effects within our prediction problem and therefore the effect of undersampling techniques is only minor or even negative, as important information gets lost. Tang, Krasser, Alperovitch, and Judge (2008) also observed that undersampling techniques might not provide highly accurate classification. Contrary to our findings, Sharififar, Sarmadian, Malone, and Minasny (2019) indicated a significant improvement in ML learning when they made balanced soil data using an undersampling technique. These results indicate that there is no universally good choice of how to resample the dataset (Haixiang et al., 2017); that is, the best resampling technique and ML algorithm for one dataset can be worse than no resampling for another (Rodriguez-Torres, Carrasco-Ochoa, & Martínez-Trinidad, 2019). Overall, the performance of resampling techniques depends heavily on the levels of imbalance, sizes of datasets and ML algorithms (Maldonado, López, & Vairetti, 2019).

We should consider that the main purpose of resampling techniques is not improving the overall accuracy of models but enhancing the accuracy of each soil class, particularly minority soil classes. To test this assumption, we compared the recall values of each soil class obtained by the RF model trained on the original imbalanced dataset with those obtained by the RF trained on the SMOTE resampled balanced dataset (Figure 5). From the plot depicted in Figure 5 it is possible to get a sense that the SMOTE resampling technique improves the accuracy of most of the soil classes, compared to the original dataset. For instance, the accuracy of Calcisols (1,910 soil samples) obtained by two RF models trained on the original imbalanced dataset and SMOTE resampled balanced dataset reached maximum values of 70.8

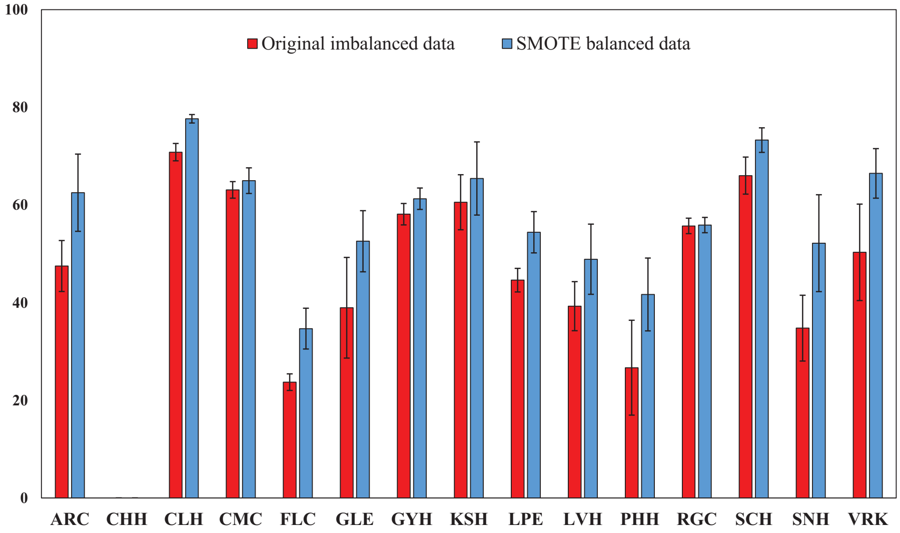


FIGURE 5 The recall values of each soil class obtained by the random forest model trained on the original imbalanced data (red) and SMOTE balanced data (blue). ARC, Arenosols; CHH, Chernozems; CLH, Calcisols; CMC, Cambisols; FLC, Fluvisols; GLE, Gleysols; GYH, Gypsisols; KSH, Kastanozems; LPE, Leptosols; LVH, Luvisols; PHH, Phaeozems; RGC, Regosols; SCH, Solonchaks; SNH, Solonetz; VRK, Vertisols [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Relative improvement (RI) of recall values in each soil class based on the RF trained on the original imbalanced data and SMOTE balanced data

WRB groups	In %	Recall obtained by RF trained on the original imbalanced data (%)	Recall obtained by RF trained on the SMOTE balanced data (%)	RI (%)	Area obtained by RF trained on the original imbalanced data (Km^2)	Area obtained by RF trained on the SMOTE balanced data (Km^2)
Arenosols	0.3	47.5	62.5	31.6	51.6	232.2
Chernozems	0.1	0.0	0.0	0.0	38.7	771.2
Calsisols	24.9	70.8	77.6	9.6	367,744.2	274,985.3
Cambisols	18.3	63.1	65.0	3.0	59,481.7	82,147.7
Fluvisols	5.9	23.7	34.6	46.3	26,356.6	20,539.2
Gleysols	0.8	38.9	52.6	35.2	304.8	714.8
Gypsisols	12.9	58.1	61.3	5.5	308,773.4	332,895.3
Kastanozems	3.1	60.5	65.4	8.1	15,645.3	15,453.8
Leptosols	4.3	44.6	54.4	22.0	46,116.2	42,159.5
Luvisols	1.2	39.2	48.9	24.7	9,499.0	9,828.3
Phaeozems	0.5	26.6	41.6	56.6	728.1	698.6
Regosols	18.6	55.7	55.9	0.3	660,505.4	715,567.5
Solonchaks	6.2	66.0	73.3	11.0	113,320.5	112,568.7
Solonetz	1.0	34.7	52.2	50.4	383.8	385.4
Vertisols	1.4	50.3	66.5	32.1	178.5	180.9

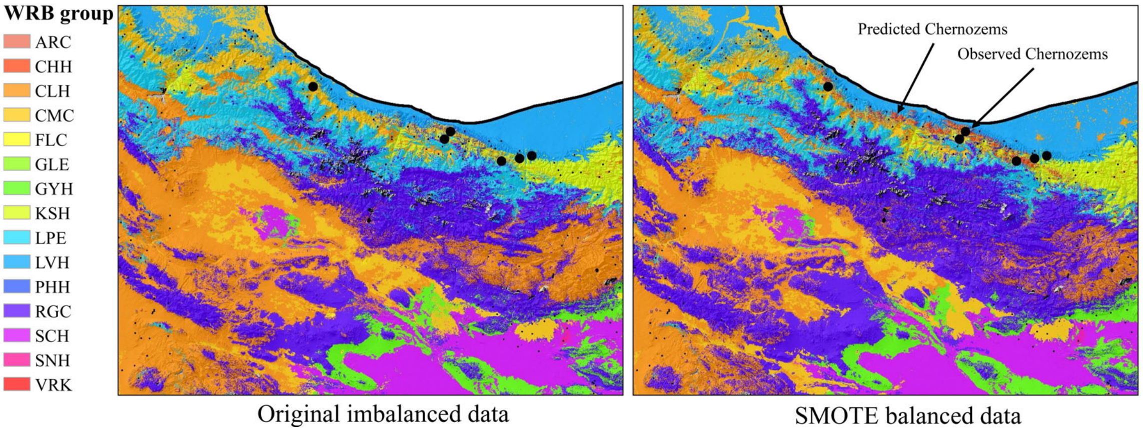
Abbreviations: In, intensity; RF, random forest; RI = ((Recall $_{\text{SMOTE Resampled data}}$ - Recall $_{\text{Original data}}$) / Recall $_{\text{Original data}}$) × 100; SMOTE, a dataset oversampled using the synthetic minority oversampling technique; WRB, world reference base system.

and 77.6%, respectively. This improvement can be seen very clearly for the prediction of Fluvisols (451 soil samples), in which the accuracy increased from 23.7 to 34.6%, when applying two RF models, respectively.

In addition, we calculated the relative improvement (RI) of the recall values obtained by RF models (original data and SMOTE) in order to compare the performances of the two models in detail (Table 5). It can be shown that the RF trained on the SMOTE resampled balanced dataset clearly improves the recall values of the 15 soil classes with an average of 22.4%, compared to RF trained on the original imbalanced dataset. For instance, the highest obtained relative improvement (RI) was 56.6% for Phaeozems (43 soil samples), followed by 50.4% for Solonetz (78 soil samples) and 46.3% for Fluvisols (451 soil samples). Apart from Chernozems, comparing the relative improvement and intensity of the observations for each class (Table 5) indicated a negative trend line (R^2 of −0.54), which demonstrated that the SMOTE resampling technique forced the RF to focus on the minority classes (Piri et al., 2018; Zarinabad et al., 2017).

Moreover, comparing the predicted area by RF models (original data and SMOTE) indicated that the minority soil classes are much better represented in the map resulting from SMOTE resampled data (Table 5). For instance, the RF model trained on the SMOTE resampled balanced dataset could increase the predicted area for two minority soil classes of Chernozems and Arenosols significantly from 38 to 771 km^2 and from 51 to 232 km^2, compared to the RF model trained on the original imbalanced data. This finding again suggests that balanced training datasets exhibit better classification rates in comparison to the imbalanced original dataset.

As an additional visual analysis, we showed a small section of the study area in the northern parts of the country (Figure 1a) after the SMOTE balancing approach has been applied (Figure 6), to explore the differences between two maps generated by RF models (original data and SMOTE). As seen, the general spatial distribution of soil classes is similar. This is particularly true for the majority soil classes (e.g., Calcisols, Regosols and Cambisols); however, there are some differences in the areas predicted as Chernozems (i.e., the minority soil classes). When the predicted soil classes compared with actual soil profiles was overlaid on the maps (Figure 6), we could conclude that the RF model trained on the SMOTE resampled data model was much more successful in DSM. These results are in line with the works of Sharififar, Sarmadian, Malone, and Minasny (2019) and Sharififar, Sarmadian, and Minasny (2019), who indicated balancing the soil dataset helped overcome the issue of modelling imbalanced soil data by improving the predictive models' results.



Original imbalanced data

SMOTE balanced data

FIGURE 6 The spatial distribution of soil classes obtained by the random forest model trained on the original imbalanced data (left) and SMOTE balanced data (right). The small section of the study area in the northern parts of country is depicted in Figure 1a. ARC, Arenosols; CHH, Chernozems; CLH, Calcisols; CMC, Cambisols; FLC, Fluvisols; GLE, Gleysols; GYH, Gypsisols; KSH, Kastanozems; LPE, Leptosols; LVH, Luvisols; PHH, Phaeozems; RGC, Regosols; SCH, Solonchaks; SNH, Solonetz; VRK, Vertisols [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | Comparison of digital and traditional maps

Under natural conditions, genesis of soil classes is typically a result of geological, topographic, climatic, hydrologic and geomorphologic factors interacting with the biosphere. Therefore, the understanding and interpretation of the spatial distribution of soils and its variability is of great concern in prediction methods and in interpretation of the terrestrial systems (Mesgaran et al., 2017). Accurate soil maps as well as soil models very much depend on how much the soil–landscape interrelationship is sampled and analysed in an unbiased way. The legacy soil databases, which were used in this study, do not have a geographically ideal distribution and intensity in terms of equal probability among all provinces and land uses in the country (SWRI, 2015). Although the recall values of models struggling on such datasets are a function of their soilscape representativeness, our results show that the ML algorithm, like RF trained on resampled SMOTE datasets, could reduce the effect of distribution bias significantly and can produce more accurate results compared to the traditional soil maps of Iran (Figure 1d). We used recall values to quantify this effect. Recall value describes the presence or absence of a soil class. Here, we use Chernozems soil distribution and compared it in our modelled map with the previously mapped units.

In chronological order, four soil maps were prepared for Iran by Kovda and Lebedev (1942), Dewan and Famouri (1964), Banaei (2000) and Hengl et al. (2007). The first was a schematic map (1/6,000,000 scale) describing global characteristics of structured zones of different environments and does not have soil classes like Chernozems. The second mentioned the existence of Chernozems intercalated with Chestnut soils (Dewan & Famouri, 1964) but did not delineate Chernozems on the map. Also, Banaei (2000) did not present Chernozems within his soil landscapes. Hengl et al. (2007) produced a map using fuzzy format but the distribution of Chernozems was separately mapped. Roozitalab et al. (2018) properly defined the presence and the distribution of Chernozems in Gillan and Mazandran provinces. The digital map of soil classes produced based on the combination of the SMOTE resampling technique and RF model has a much higher spatial resolution compared to existing maps in Iran and displays new soil classes such as Chernozems in the north parts of the country that have not yet been mapped in Iran. Therefore, the map generated in this study can be considered as an improved map of soil classes in Iran.

4 | CONCLUSION

In this paper, we tried to prepare a digital map of soil classes at the national scale with the resolution of 90 × 90 m in Iran. Herein, we tested five ML algorithms on nine datasets, including the original imbalanced soil dataset and eight resampled balanced soil datasets, to explore if resampling can enhance the prediction power of ML algorithms in DSM problems. The following conclusions can be drawn from this study.

1. Random forest was the best method to predict soil classes in all datasets. Therefore, RF can be recommended as the most reliable model to predict spatial distribution of soil classes of Iran.
2. Resampling the original datasets, particularly with the SMOTE technique, increased OA, K-index and F-Score in comparison to the original dataset. These results clearly indicate that standard ML algorithms could be better trained by the balanced SMOTE resampled dataset

than imbalanced legacy data from existing soil maps. This is vital in DSM studies because they mostly rely on such imbalanced soil legacy data, in which the application of ML algorithms can generate highly biased soil class maps.

3. The resulting new soil map of Iran produced based on the combination of the SMOTE resampling technique and random forest model has a much higher spatial resolution compared to four existing soil maps in Iran. Therefore, the map can be considered as the latest version of a soil map in Iran.

ACKNOWLEDGEMENTS

The work of the first author has been supported by the Alexander von Humboldt Foundation under grant number: Ref 3.4 - 1164573 - IRN - GFHERMES-P. The authors wish to acknowledge the contribution of the Soil and Water Research Institute of Iran for providing the Iranian soil dataset. We are particularly grateful to Dr Kambiz Bazargan (SWRI) and Dr Fereydoon Sarmadian (University of Tehran). Karsten Schmidt and Thomas Scholten thank the German Research Foundation (DFG) for supporting this research through the Collaborative Research Center (SFB 1070) "ResourceCultures" (subprojects Z, S and B02). Open access funding enabled and organized by Projekt DEAL. [Correction added on 26 October 2020, after first online publication: Projekt DEAL funding statement has been added.]

Data Availability Statement

The data that support the findings of this study are not publicly available due to privacy or ethical restrictions.

ORCID

Ruhollah Taghizadeh-Mehrjardi https://orcid.org/0000-0002-4620-6624
Karsten Schmidt https://orcid.org/0000-0003-0337-3024
Kamran Eftekhari https://orcid.org/0000-0001-9786-4566
Thorsten Behrens https://orcid.org/0000-0003-0632-2804
Norair Toomanian https://orcid.org/0000-0003-3347-7748
Thomas Scholten https://orcid.org/0000-0002-4875-2602

REFERENCES

Adhikari, K., Minasny, B., Greve, M. B., & Greve, M. H. (2014). Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma*, *214*, 101–113.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., … Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940–7957.

Banaei, M. (2000). Soil resources and use potentiality map of Iran, 1: 1000000. Soil and Water Research Institute, Tehran.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*, 20–29.

Behrens, T., Zhu, A. X., Schmidt, K., & Scholten, T. (2010). Multiscale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*, 175–185.

Branco, P., Ribeiro, R.P. & Torgo, L. (2016). UBL: an R package for utility-based learning. Retrieved from https://CRAN.R-project.org/package=UBL/ [accessedon 08 September 2018].

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, *49*, 1–50.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, *239*, 68–83.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626–4636.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., He, T., Benesty, M. & Khotilovich, V. (2019). 'xgboost'. R package version 0.90.0.2. Retrieved from https://CRAN.R-project.org/package=xgboost/

Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., & Yang, B. (2018). Machine learning based mobile malware detection using highly imbalanced network traffic. *Information Sciences*, *433*, 346–364.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., … Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, *Geoscientific Model Development*, *8*, 1991–2007.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Devi, D., & Purkayastha, B. (2017). Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. *Pattern Recognition Letters*, *93*, 3–12.

Dewan, M. L., & Famouri, J. (1964). *The soils of Iran*. Food and Agriculture Organization of the United Nations. Rome, Italy: FAO.

Emadodin, I., Narita, D., & Bork, H. R. (2012). Soil degradation and agricultural sustainability: An overview from Iran. *Environment, Development and Sustainability*, *14*, 611–625.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*, 18–36.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer, New York: Springer Series in Statistics.

He, H., Bai, Y., Garcia, E.A. and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (P.1322–1328). Hong Kong, China: IEEE Press.

He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, *21*, 1263–1284.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotic, A., … Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*, e0169748.

Hengl, T., Toomanian, N., Reuter, H. I., & Malakouti, M. J. (2007). Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, *140*, 417–427.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, *265*, 62–77.

Hijmans, R., Cameron, S., Parra, J., Jones, P., Jarvis, A., & Richardson, K. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *25*, 1965–1978.

Hounkpatin, K. O. L., Schmidt, K., Stumpf, F., Forkuor, G., Behrens, T., Scholten, T., … Welp, G. (2018). Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). *Scientific Reports*, *8*, 9959.

Jafari, A., Finke, P. A., Van de Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA- great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science*, *63*, 284–298.

Kovda, V.A. & Lebedev, Y.P. (1942). The soil map of Iran at scale of 1:6 million. Retrieved from http://isric.org/content/search-library-and-map-collection/ [accessed on 10 September 2017].

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 179–186). Nashville, TN: ICML.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.

Lauron, M.L.C. & Pabico, J.P. (2016). Improved sampling techniques for learning an imbalanced data set. *arXiv preprint arXiv: 1601.04756*.

Lee, S. S. (1999). Regularization in skewed binary classification. *Computational Statistics*, *14*, 277.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, *76*, 380–389.

McBratney, A. B., Santos, M. L. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52.

Mesgaran, M. B., Madani, K., Hashemi, H., & Azadi, P. (2017). Iran's land suitability for agriculture. *Scientific Reports*, *7*, 7670.

Mira, M., Weiss, M., Baret, F., Courault, D., Hagolle, O., Gallego-Elvira, B., & Olioso, A. (2015). The MODIS (collection V006) BRDF/albedo product MCD43D: Temporal course evaluated over agricultural landscape. *Remote Sensing of Environment*, *170*, 216–228.

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, *106*, 15–29.

Quinlan, J. R. (1992). Learning with continuous classes. In A. Adams, L. Sterling (Eds.), *5th Australian joint conference on artificial intelligence, Hobart, Tasmania* (pp. 343–348). Singapore: World Scientific.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., & Thompson, J. (2018). Soil property and class maps of the conterminous United States at 100-meter spatial resolution. *Soil Science Society of America Journal*, *82*, 186–201.

Rodriguez-Torres, F., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019). Deterministic oversampling methods based on SMOTE. *Journal of Intelligent & Fuzzy Systems*, *36*, 4945–4955.

Roozitalab, M. H., Siadat, H., & Farshad, A. (2018). *The soils of Iran*. Cham, Switzerland: Springer.

Sanchez, P. A., Ahamed, S., Carre, F., Hartemink, A. E., Hempel, J., Huising, J., … Zhang, G. L. (2009). Digital soil map of the world. *Science*, *325*, 680–681.

Schmidt, K., Behrens, T., & Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma*, *146*, 138–146.

Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019). Addressing the issue of digital mapping of soil classes with imbalanced class observations. *Geoderma*, *350*, 84–92.

Sharififar, A., Sarmadian, F., & Minasny, B. (2019). Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. *Computers and Electronics in Agriculture*, *159*, 110–118.

Sparks, D. L. (1998). *Soil physical chemistry*. Boca Raton, FL: CRC Press.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., … Scholten, T. (2016). Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *Journal of Plant Nutrition and Soil Science*, *179*, 499–509.

SWRI. (2015). *Soil and water research institute of Iran*.

Tang, Y., Krasser, S., Alperovitch, D., & Judge, P. (2008). Spam sender detection with classification modeling on highly imbalanced mail server behavior data. In B. Prasad, P. Sinha, A. Ram & E. K. Kerr (Eds.), *Proceedings of the International Conference on Artificial intelligence and pattern recognition* (pp. 174–180). Orlando, FL: Curran Associates.

Tantithamthavorn, C., Hassan, A.E. & Matsumoto, K. (2018). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. arXiv:1801.10269v1.

Teng, H. F., Rossel, R. A. V., Shi, Z., & Behrens, T. (2018). Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena*, *164*, 125–134.

Tkachenko, R., Doroshenko, A., Izonin, I., Tsymbal, Y., & Havrysh, B. (2018). Imbalance data classification via neural-like structures of geometric transformations model: Local and global approaches. In Z. Hu, S. Petoukhov, I. Dychka, M. He (Eds.), *International conference on computer science, engineering and education applications* (pp. 112–122). Cham, Switzerland: Springer.

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transactions on systems, Man, and Cybernetics*, *6*, 448–452.

WRB, I. W. G. (2006). World reference base for soil resources 2006. *World Soil Resources Reports*, *103*, 128.

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., … Roy, D. P. (2016). The global

Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, *185*, 271–283.

Yang, J., Zhou, J., Zhu, Z., Ma, X., & Ji, Z. (2016). Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. *Journal of Biological Research-Thessaloniki*, *23*, 13.

Zarinabad, N., Wilson, M., Gill, S. K., Manias, K. A., Davies, N. P., & Peet, A. C. (2017). Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy. *Magnetic Resonance in Medicine*, *77*, 2114–2124.

How to cite this article: Taghizadeh-Mehrjardi R, Schmidt K, Eftekhari K, et al. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. *Eur J Soil Sci*. 2020;71:352–368. https://doi.org/10.1111/ejss.12893