JEE Journal of Ecological Engineering

Journal of Ecological Engineering 2022, 23(10), 61–67
https://doi.org/10.12911/22998993/152281
ISSN 2299–8993, License CC-BY 4.0

Received: 2022.07.12
Accepted: 2022.08.12
Published: 2022.09.01

Soil Salinity Classification Using Machine Learning Algorithms and Radar Data in the Case from the South of Kazakhstan

Timur Merembayev[1*], Yedilkhan Amirgaliyev[1], Sultan Saurov[2], Waldemar Wójcik[3]

[1] Institute of Information and Computational Technologies CS MES RK, 050010, 28 Shevchenko Str., Almaty, Kazakhstan

[2] S. Seifullin Kazakh Agro Technical University, 010011, 62 Zhenis Ave., Nur-Sultan, Kazakhstan

[3] Lublin University of Technology, 20-618, 38D Nadbystrzycka Str., Lublin, Poland

* Corresponding author's e-mail: timur.merembayev@gmail.com

ABSTRACT

Soil salinity is one of the major impact factors on agriculture in the South of Kazakhstan. Prediction and estimation of soil salinity before planting a season usually helps to plan for the leaching of the salt. In the paper, satellite data such as radar data and machine learning algorithms, were used to classify soil salinity. Numerical results were presented for the Turkestan region, which contains more than 102 points. The machine learning algorithms, including Gaussian Process, Decision Tree, and Random Forest, were compared. The evaluation of the model score was realized by using metrics, such as accuracy, Recall, and f1. In addition, the influence of the dataset features on the classification was investigated using machine learning algorithms. The research results showed that the Gaussian Process model has the best score among considered algorithms. In addition, the results are consistent with the outcome of the Shapley Additive exPlanations (SHAP) framework.

Keywords: environmental correlation, soil salinity, machine learning, remote sensing.

INTRODUCTION

Currently, due to anthropogenic impacts or natural phenomena, agricultural lands are increasingly subject to changes, including soil degradation. The studies on the problem of salinity and the struggle against salinity in Central Asia are important tasks. In different states and areas, various methods and technologies are used to identify and measure soil salinity. Along with groundwater, the drying up of the Aral Sea, the salinization of nearby areas occurs. As a result of salinization, agricultural lands are degraded. Salinization of agricultural land directly affects the productivity and quality of crops grown, and these factors affect the economy of the selected region. Especially in the example of Kazakhstan, the problem of salinity is mostly presented in the South of Kazakhstan regions. New effective methods need to be developed to estimate the degree of salinity of agricultural land and restore the degraded arable land. There are various methods and algorithms for determining saline lands and technologies for restoring arable land to increase productivity.

For assessing soil salinity, remote sensing methods have been actively used in the last decade (Pankova et al., 1978). The scientific papers provided the methods for estimating the salinity of irrigated arable land based on satellite data and GIS technologies (Fernandez-Buces et al., 2006; Gabdullin et al., 2015; Laiskhanov et al., 2016). The authors used satellite data and described the salinization of some tracts of irrigated arable land in Kazakhstan. Recently, mapping saline agricultural lands using satellite data and GIS technologies is the most common way to monitor the salinization of arable land. On the basis of the research of existing solutions, remote sensing and GIS technologies can successfully compete with technically complex and expensive methods of ground-based monitoring of arable land salinity. At the same time, data is collected from the spectral channels of the spectrum's visible and near infrared range. On the basis of channel combinations, various salinity indices are built, combined with vegetation indices, making it possible to build a logical system of communication between the actual salinity level (ground data) of arable land with the spectral characteristics of the underlying surface.

There are several examples of solving problems by restoring salinity parameters (salinity maps) of irrigated arable land based on relatively small ground data and current satellite imagery (Fernandez-Buces et al., 2006; Masoud et al., 2006; Asfaw et al., 2018). However, this approach is not always suitable for other arable lands, since they are characterized by varying conditions, especially in different seasons and weather conditions. Automatically using this approach does not always give good results. Accounting for seasonal characteristics, the composition of cultivated crops, the pace of the onset of spring, and cloudless satellite imagery calendar dates require additional satellite surveys each time.

Using satellite data can positively affect the assessment of the salinity of agricultural lands. In this approach, there is a need to use various MODIS products, which have been developing significantly in recent years. The technologies for processing time series of primary satellite imagery currently make it possible to obtain comparable long-term data series with an update period of 7–10 days. These products are usually based on MODIS data and have a medium spatial resolution, for example 250 m for vegetation indices.

There are various remote sensing data to identify and monitor saline areas, including aerial images, video images, infrared thermography, and multispectral images (Ondrasek et al., 2021). Multispectral scanning technology for the study of natural resources is a promising direction for solving the problem of monitoring and predicting soil salinity, an example of multispectral remote sensing satellites: Landsat MSS / TM (Multispectral scanning / Thematic Mapper), Sentinel, Astel, and Spot. The type and variety of images depend on electronic scanners that register reflected radiation in separate ranges. Landsat offers a much wider range of ranges than Spot and allows the detection of various elements on the surface.

When analyzing the data from Landsat MSS channels 3, 4, and 5 presented in the research (Abuelgasim et al., 2019), the authors recommended using them to identify soil salinity. Landsat TM data in the ranges 1 to 5 and 7 are good indicators for determining salt minerals, at least when they are the dominant component of the soil. Examples of more complex techniques include various clustering and classification algorithms for raw satellite data or machine learning (Hoa et al., 2019; Akramkhanov et al., 2012).

In the paper (Hoa et al., 2019) authors considered the Mekong Delta and proposed an approach for solving the problem of salinity mapping using radar images and machine learning algorithms. Using radar images, field data, and soil electrical conductivity measurements, the authors solved the regression problem. They provided a comparative analysis of several machine learning models of salinity results.

The authors of the paper (Akramkhanov et al., 2012) studied soil salinity in the Aral Sea basin. In the research, they tried to estimate the spatial distribution of soil salinity based on easily available environmental parameters (relief indices, remote sensing data, distance to ditches and long-term groundwater observations) using a neural network model. The environmental attributes and soil salinity ratios have been used to reach a score of soil salinity almost similar to mean soil salinity values (0.94 vs. 1.04 dS m-1 estimate). The author reached a score 70–90% for the test dataset.

The aim of the study was to evaluate the influence of generated features on a classification model of soil salinity. The features were obtained using radar data, a digital elevation model, temperature, and texture analysis of the area of interest. The potential features that should help to improve the score of a classification model were identified.

METHODS

The authors have followed the general workflow of a machine learning classifier which is showed in Figure 1. The process of the classifier model contains of the following steps:

- Obtaining salinity classification based on Landsat-5 spectral indices and expert assessment;
- Texture analysis of Sentinel 1 (radar data) by the Grey Level Co-occurrence Matrix (GLCM) method (Haralick et al., 1973);
- Augmentation of the dataset by digital elevation model (SRTM) and temperature (MODIS);

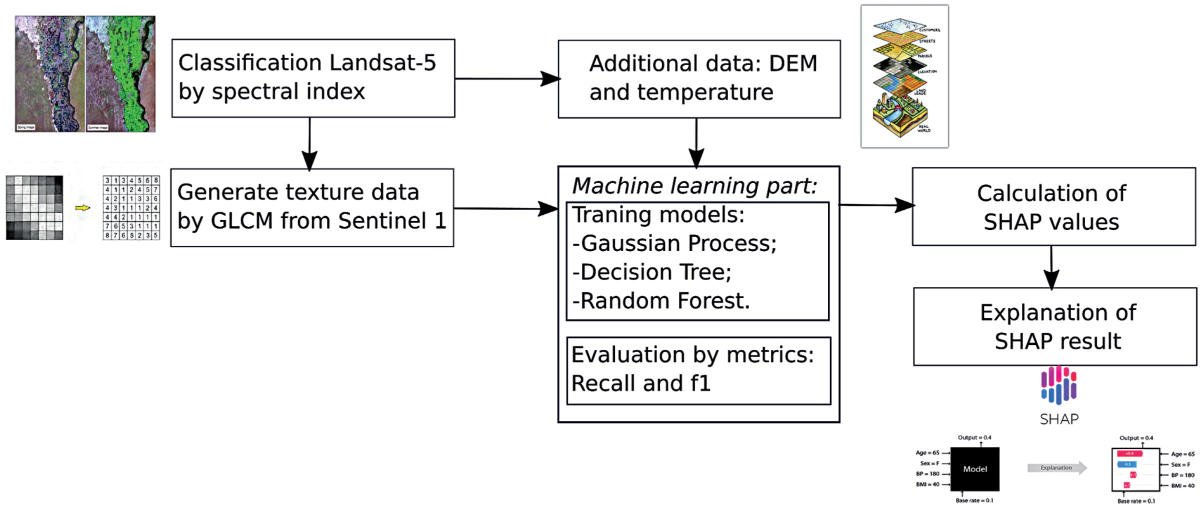
Classification Landsat-5 by spectral index
Additional data: DEM and temperature
Generate texture data by GLCM from Sentinel 1
Machine learning part:
Traning models:
-Gaussian Process;
-Decision Tree;
-Random Forest.
Evaluation by metrics: Recall and f1
Calculation of SHAP values
Explanation of SHAP result
SHAP
Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Explanation
Base rate = 0.1

Figure 1. Flowchart of workflow for salinity classification

- Application of machine learning algorithms and evaluation of the quality of the trained model;
- Calculation of importance values for classification model by SHAP;
- Explanation of obtained results.

In the conducted research, the approach from the research that processed radar data was considered (Hoa et al., 2019). The machine learning algorithms such as Gaussian process (GP), decision tree, and random forest were adopted for comparison analysis of model scores. In the performed research, the task was a binary classification of soil salinity.

In the paper (Nickisch et al., 2008), the authors described the theoretical and practical implications of GP to a binary classification problem. The authors have studied various approaches to improve the scores of the GP algorithm for binary classification. GP is a stochastic process that is given by a function of the mean $m(x) = E[f(x)]$ and a covariance function of $k(x, x') = V[f(x), f'(x)]$, where x_i is one soil field point from a set of soil data *X*, and each point is

Table 1. Abbreviation and description of features

Name	Description
point_class	Binary class
gamma_vv	Plarization VV
dissimilarity_vv	Dissimility of gray level co-occurrence matrix for polarization VV
contrast_vv	Contrast of gray level co-occurrence matrix for polarization VV
homogeneity_vv	Homogenety of gray level co-occurrence matrix for polarization VV
ASM_vv	ASM (homogeneity of an image) of gray level co-occurrence matrix for polarization VV
energy_vv	Energy of gray level co-occurrence matrix for polarization VV
correlation_vv	Correlation of gray level co-occurrence matrix for polarization VV
entropy_vv	Entropy of gray level co-occurrence matrix for polarization VV
gamma_vh	Plarization VH
dissimilarity_vh	Dissimility of gray level co-occurrence matrix for polarization VH
contrast_vh	Contrast of gray level co-occurrence matrix for polarization VH
homogeneity_vh	Homogenety of gray level co-occurrence matrix for polarization VH
ASM_vh	ASM (homogeneity of an image) of gray level co-occurrence matrix for polarization VH
energy_vh	Energy of gray level co-occurrence matrix for polarization VH
correlation_vh	Correlation of gray level co-occurrence matrix for polarization VH
entropy_vh	Entropy of gray level co-occurrence matrix for polarization VH
elv	SRTM of ground elevation model
slope	Calculated slope from DEM
temp	MODIS land surface temperature

assigned salinity or lack of salinity. $y_i \in \{0,1\}$ The prediction is achieved using the function f to be found, to achieve a binary classification per function f, the sigmoid function $sig: R \rightarrow [0,1]$ is used. Thus, the probability that the function f will predict the event y_i will be represented as:

$$P(y \vee x) = sig\big(y \cdot f(x)\big) \qquad (1)$$

The decision tree method is the data mining method, and it is successfully used to solve different classification problems. Morgan and Sonquist created and used an algorithm for the determinants of social conditions (Rokach et al., 2005). The advantages of decision trees are that they are computationally fast and work with multidimensional data. In addition, one decision tree can process the data, and the algorithm will be greedy; therefore, it continues to grow deeper into the tree.

The random forest was presented by Breiman as a training tree ensemble classifier (Breiman et al., 2001). The main idea of the algorithm is to get random vector values from the aggregated bootstrap sample (training dataset) and then train a lot of decision trees. However, a trained tree contains many trees, so it needs more computational resources.

Recall and F1 metrics were used to assess the quality of the model. These metrics were considered in order to avoid the problems with a possible imbalance in the sample, so there will be fewer points with salinity than points where there is no salinity.

DATA

Generated from radar data

The Turkestan region, Shardara district, was chosen for providing tests with machine learning algorithms. Figure 2 shows the area of interest, and the area is 858.84 km^2.

The Landsat 5 multispectral images of this territory were used for classification and obtained in June 2021, the classification was performed using spectral indices and with expert validation. Figure 2 shows the territory with the classification performed according to Landsat 5 data. 9 classes were identified, and 5 of them relate to soil salinity: strong, extreme, medium, initial, and no salinity.

On the basis of the classification data, the territory was labeled, 102 points randomly located and defined whether this location has salinity or not. Figure 2 showed the randomly selected points in the area of interest.

The dataset consists of binary soil values (salinity or no salinity) and 16 features that were generated using a radar image and the GLCM method. Table 1 contains abbreviations and descriptions of the dataset.

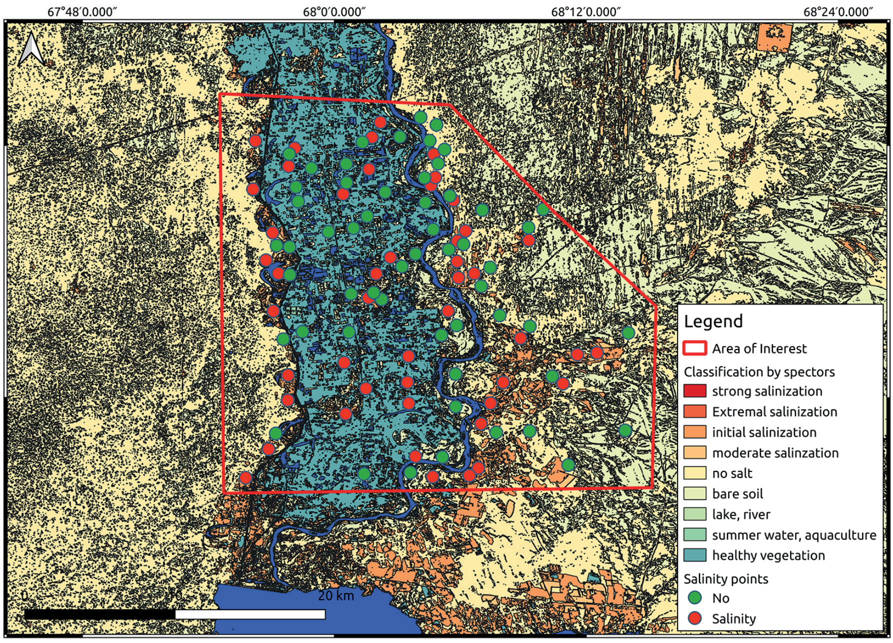

Figure 2. Randomly labeled 102 points in the area of interest.

Additional data

For radar data augmentation, using a digital elevation model and temperature data that may affect the salinity classification model were proposed. USGS MODIS Earth Surface Temperature or land surface temperature (LST) and ground elevation model (ELV) data were used to extend the data set. The obtained images have a resolution of 1 km for LST and 30 meters resolution for ELV. From the ELV data, two features were obtained, i.e. height above sea level and angle of land (slope of relief). Combining additional data with radar data, a dataset of 102 points and 19 features was obtained.

The data was split into training and test datasets, split at 85% and 15%, respectively. The distributions of training and test datasets had an equal proportion of classes in the two samples. The total dataset is 102 points, the training dataset contains 81 points, and the test dataset contains 15 points.

RESULTS

The comparison of the chosen algorithms was carried out on 16 features (radar data) and 3 additional features obtained from temperature and elevation points, a total dataset is 19 features. Table 2 shows the comparison of model scores on the test dataset for the Recall and f1 metrics. The GP has the highest score in the test set for class 1 (soil salinity) Recall = 0.60 and f1 = 0.67.

The augmentation with new features improved the accuracy for the three models and for all metrics. The improvement especially impacted for class 0 (no salinity): Recall = 0.89 and f1 = 0.84.

To understand how the model was classified and what features influenced the model score, the SHAP package was used (Lundberg & Lee, 2017). The result of SHAP is presented for the Random Forest model since SHAP does not have the ability to use the Gaussian Process model. The use of the library has a wide range of applied tasks (Merembayev et al., 2021; Amirgaliyev et al., 2019; Muhamedyev et al., 2020) and is an effective tool for understanding a model. SHAP is a kind tool for explaining the various patterns, and it gives a significant value to each feature. SHAP creates an explanatory model for a one-row-forecast pair to explain the prediction result. The SHAP values are computed by means of the values across all possible features. Explanation models (tree and kernel) do not infer probabilities due to a limitation associated with non-linear transformations, but they do provide raw objective function limit values that fit the model.

The first interpretation that can be created with the explanation values is the summary plot shown in Figure 3, which presents the model's most important features and a visual representation of how they impacted it. Due to the nature of these patterns, each sample needs to be studied separately, so most of these plots are simply a composite of all of these samples.

Figure 3 presents the most important features of the model. The values of contrast_vv, correlation_vh can be good indicators that can be used for soil separation.

Figure 4 shows the dependencies of individual features (contrast_vv, correlation_vh). These graphs plot the value of the features in relation to their SHAP value, enabling to understand how they are related easily. SHAP also displays the value of the second variable, which is automatically selected depending on its interaction with the function in question. Thus, the library helps to find multidimensional dependencies in data. Moreover, in Figure 4, there was the dependency that higher variable values are associated with higher SHAP values. It can also be seen that this relationship is not linear but closer to a threshold of about 0.01 for the feature contrast_vv. For the correlation_vh

Table 2. The comparison of three models for test dataset

Models	Class	16 features (radar data)			19 features (temp+DEM)		
		Precision	Recall	F1	Precision	Recall	F1
Gaussian process	0	0.75	0.67	0.71	0.80	0.89	0.84
	1	0.50	0.60	0.55	0.75	0.60	0.67
Decision tree	0	0.67	0.44	0.53	0.55	0.67	0.60
	1	0.38	0.60	0.46	0.30	0.60	0.40
Random forest	0	0.67	0.67	0.67	0.75	0.67	0.71
	1	0.40	0.40	0.40	0.50	0.60	0.55

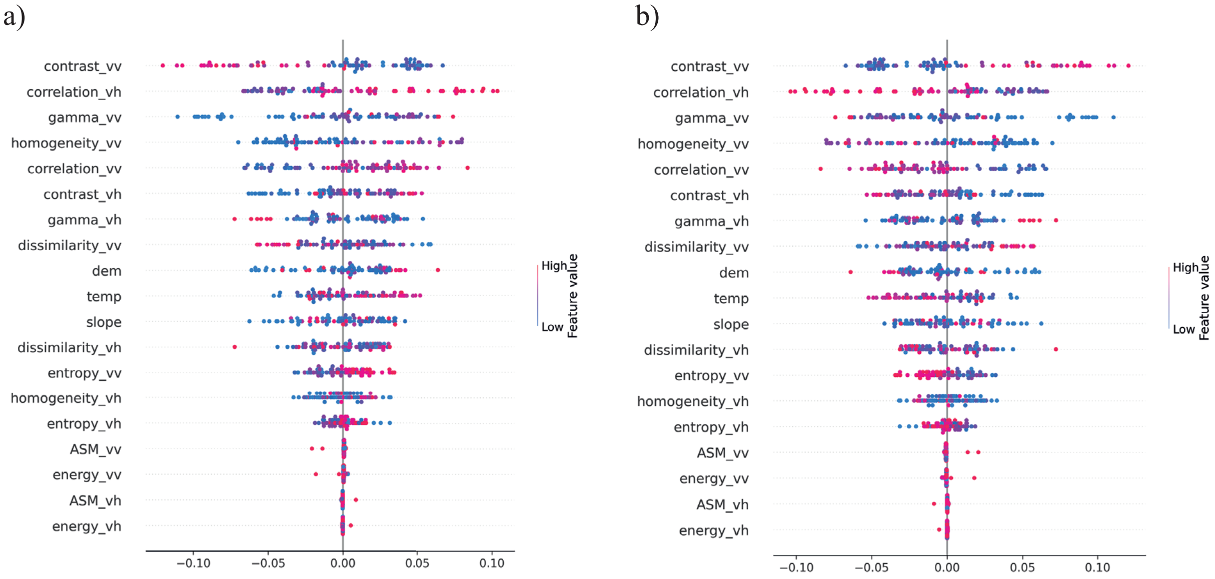

Figure 3. SHAP summary plot of the RF model; a) for contrast_vv feature; b) for correlation_vh feature

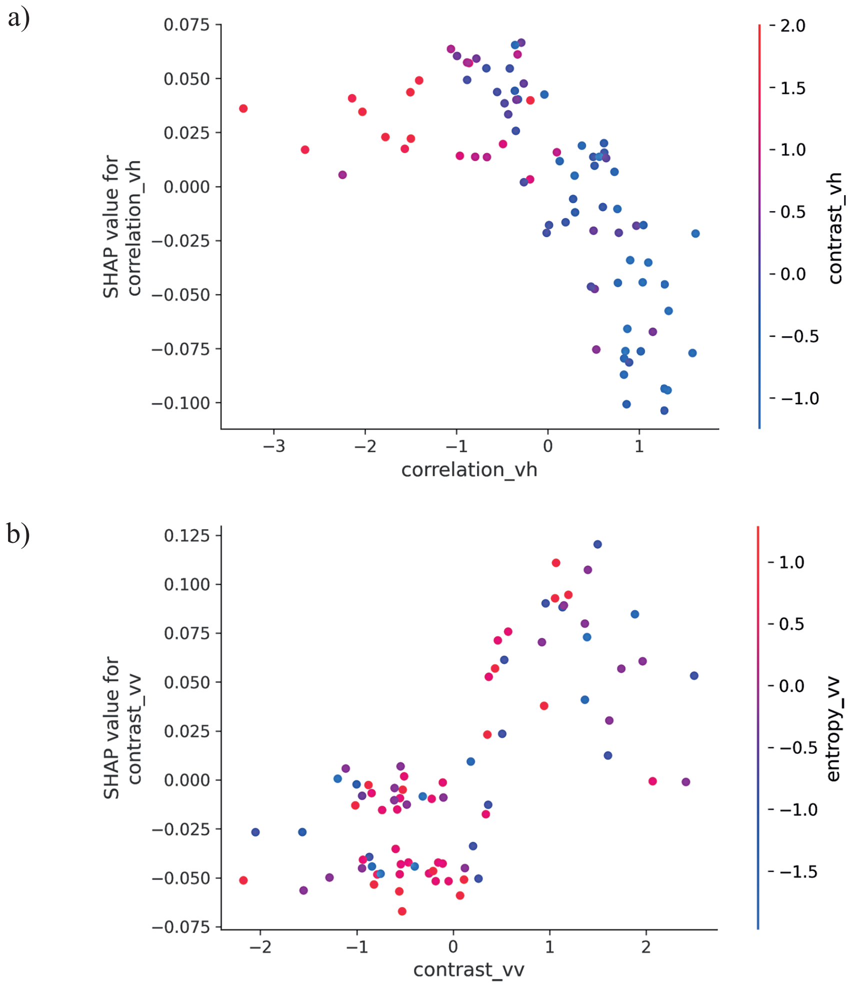

Figure 4. SHAP values for the dependency of features; a) for contrast_vv feature; b) for correlation_vh feature

feature, it can be seen how the color indicates that when correlation_vh is low, contrast_vh has a SHAP value lower than when it is higher.

CONCLUSIONS

This research analyzed the machine learning algorithms for salinity classification using satellite data (radar, temperature, and elevation). The obtained data indicate that the Gaussian Process model reached the best results with a comparison of the considered algorithms. The Gaussian Process model has an accuracy of 0.60 and 0.67 for the recall and f1 metrics, respectively, for class 1. The use of additional features such as elevation points and temperature improved the model's score.

The SHAP framework was used to examine the effect of features on the target classification and identify complex relationships between features. The SHAP result on the considered dataset revealed that the important features for class forecast were contrast_vv, correlation_vh, and gamma_vv.

In future research, the authors will focus on building a machine learning model to quantify salinity (soil conductivity), use hyperspectral data, and apply deep learning algorithms such as convolution neural networks.

Acknowledgments

This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan BR10965172.

REFERENCES

1. Abuelgasim, A., Ammad, R. 2019. Mapping soil salinity in arid and semi-arid regions using Landsat 8 OLI satellite data, Remote Sensing Applications: Society and Environment, 13, 415–425.
2. Akramkhanov, A., Vlek, P.L. 2012. The assessment of spatial distribution of soil salinity risk using neural network. Environmental monitoring and assessment,184(4), 2475–2485.
3. Amirgaliyev, Y., Shamiluulu, S., Merembayev, T., Yedilkhan, D. 2019. Using machine learning algorithm for diagnosis of stomach disorders. In: International Conference on Mathematical Optimization Theory and Operations Research, 343–355.
4. Asfaw, E., Suryabhagavan, K., Argaw, M. 2018. Soil salinity modeling and mapping using remote sensing and GIS: The case of Wonji sugar cane irrigation farm, Ethiopia. Journal of the Saudi Society of Agricultural Sciences, 17, 250–258.
5. Breiman L. 2001. Random forests. Machine Learning, 45, 5–32.
6. Fernandez-Buces, N., Siebea, C., Cramb, S., Palacio, J.L. 2006. Mapping soil salinity using a combined spectral response index for bare soil and vegetation: A case study in the former lake Texaco, Mexico. J. Arid Environm, 65(4), 644–667.
7. Gabdullin, B., Zhogolov, A., Savin, I., Otarov, A., Ibrayeva, M., Golovanov, D. 2015. Application of multi-spectral satellite data for interpretation of soil salinization of the irrigated areas (case study of Southern Kazakhstan). Moscow University Geography Bulletin, 5, 34–41.
8. Haralick, R.M., Shanmugam, K., Dinstein, I.H. 1973. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics, 6, 610–621.
9. Hoa, P.V., Giang, N.V., Binh, N.A., Hai, L.V.H., Pham, T.D., Hasanlou, M., Tien Bui, D. 2019. Soil salinity mapping using SAR sentinel-1 data and advanced machine learning algorithms: A case study at Ben Tre Province of the Mekong River Delta (Vietnam). Remote Sensing, 11(2).
10. Laiskhanov, S.U., Otarov, A., Savin, I.Y., Tanirbergenov, S.I., Mamutov, Z.U., Duisekov, S.N., Zhogolev, A. 2016. Dynamics of soil salinity in irrigation areas in South Kazakhstan. Polish J. Env. Studies, 25, 2469–2475.
11. Lundberg, S.M., Lee, S.I. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
12. Masoud, A.A., Koike, K. 2006. Arid land salinization detected by remotely-sensed landcover changes: A case study in the Siwa region, NW Egypt. J. Arid Environm, 66(1), 151–167.
13. Merembayev, T., Kurmangaliyev, D., Bekbauov, B., Amanbek, Y. 2021. A Comparison of machine learning algorithms in predicting lithofacies: Case studies from Norway and Kazakhstan. Energies, 14(7), 1896.
14. Muhamedyev, R., Yakunin, K., Kuchin, Y.A., Symagulov, A., Buldybayev, T., Murzakhmetov, S., Abdurazakov, A. 2020. The use of machine learning "black boxes" explanation systems to improve the quality of school education. Cogent Engineering, 7(1), 1769349.
15. Nickisch, H., Rasmussen, C.E. 2008. Approximations for binary Gaussian process classification. Journal of Machine Learning Research, 9, 2035–2078.
16. Ondrasek, G., Rengel, Z. 2021. Environmental salinization processes: Detection, implications & solutions, Science of the Total Environment, 754.
17. Pankova, E.I., Mazikov, V.M., Isaev, V.A., Jamnova, I.A. 1978. Experience in the use of aerial photographs for the characteristics of soil salinity rainfed areas serozem area. Pochvovedenie, 3, 82–85. (in Russian)
18. Rokach, L., Maimon, O. 2005. Decision trees. In Data Mining and Knowledge Discovery Handbook. Springer: Berlin, Germany, 165–192.