EFFICIENCY OF MACHINE LEARNING ALGORITHMS IN SOIL SALINITY DETECTION USING LANDSAT-8 OLI IMAGERY

S.Alamdar[1], F.Ghazban[1, *], A.Zarei[2]

[1] Dept. of Environment, University of Tehran, Enghelab Square, Ghods St, Tehran, Iran - (setare.alamdar,fghazban)@ut.ac.ir
[2] Dept. of Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran - arastou.zarei @ut.ac.ir

Commission IV, WG IV/3

KEY WORDS: Soil Salinity, Landsat 8-OLI, Machine Learning, Electrical Conductivity, Climate Change.

ABSTRACT:

Climate change is one of the biggest problems facing today's world. Rising temperatures and declining rainfall have had a profound effect on the planet, one of which is the destructive effects of soil salinity. Soil salinity phenomena commonly occur in arid and semi-arid regions. Maharloo Salt Lake, southeast of Shiraz, Iran, with an arid and semi-arid climate, has faced severe droughts in the past and is dealing with the soil salinity problem. One useful way to manage land and soil in such areas is regular monitoring of the soils and lands and keeping abreast of changes to prevent land degradation and erosion. With the advancement of technology, remote sensing techniques to monitor natural factors have become very popular. Landsat sensor images were used in this research, and several environmental indicators were extracted by combining satellite bands. Three machine learning algorithms, RF, GBM, and MLP, were used to evaluate methods for monitoring and mapping saline soils. The models were trained and then tested to compare the accuracy and performance of each model in predicting soil salinity. GBM algorithm showed the best performance with R2 = 0.89 and RMSE = 0.63 for testing the dataset after that RF model with R2 = 0.85 and RMSE = 0.71 and the worst performance was for MLP model with R2 = 0.75 and RMSE = 0.88. The figures mapped from the results of these algorithms for salinity distribution in this region showed that by choosing the appropriate algorithm and suitable in-situ data, it could be possible to estimate soil salinity to an excellent extent by satellite data.

1. INTRODUCTION

climate change has affected diverse area around the world and its impact will grow worldwide. soil is one of the most significant environmental parameters that has been changed directly and indirectly by climate change (Eswar et al., 2021). The rise of the sea level, infiltration, inappropriate and unprincipled irrigation and drainage, and the increase in surface temperature are among the factors that are aggravated by climate change and cause soil salinity (Singh et al., 2020). One of the main reasons for soil salinity due to climate change is the increase in evaporation and transpiration due to the increase in surface temperature and Also, it is one of the appropriate methods to investigate issues related to climate change and global warming, to identify and investigate the dynamics of changes in the earth's surface temperature on a local and regional scale and their impact on soil salinity (Jahan and Ur, 2021; Zarei et al., 2021b). Fertile soils are the most critical environment for food supply in the world; these days, soils are under much pressure due to natural and human factors and are abandoned due to their salinization (Gorji et al., 2017). Soil salinity causes many problems, including loss of soil fertility, damage to the ecosystem, impact on water quality, soil erosion, and land degradation (Khan et al., 2005). In recent years, the expansion of agricultural activities to provide more food, which is accompanied by severe climatic events such as reduced rainfall and increased evaporation, as well as the use of traditional irrigation systems in agriculture, has increased soil salinization (Jalali et al., 2021; Shrivastava and Kumar, 2015). According to Food and Agriculture Organization (FAO) estimate, 397 million hectares of land worldwide have saline soils, while it is estimated that these lands will expand by 2 million hectares annually (Koohafkan and Stewart, 2012; Peng et al., 2019). One of the severe problems of soil salinity is related to arid and semi-arid regions of the world, posing a severe threat to the environment and agricultural sustainability (Arnous and Green, 2015). For scientists and researchers trying to devise plans to reduce the effects of soil salinity, soil monitoring at the national, regional, and even local scales is very effective (Davis et al., 2019). Soil mapping is used to identify and determine soil distribution models so that they can be easily described and displayed to users. Despite the strengths of the traditional mapping in interpreting the primary process of soil formation, there are major shortcomings in describing the structure and dynamic properties of soil (Peng et al., 2019). Traditional soil monitoring is very time-consuming while requiring extensive mapping as well as laboratory activities. Due to the problems with traditional monitoring, the use of multi-spectral satellite imagery with high resolution for detecting and monitoring saline soils, is economically viable and highly recommended (Vermeulen and Van Niekerk, 2017). In various studies, salinity indices generated from simple or complex spectral compositions of remote sensing and satellite image data are used in order to detect the temporal-spatial distribution of saline soils (Elhag and Bahrawi, 2017). Since 1990 great variety of multi-spectral sensors such as SPOT, IKONOS, ASTER, Landsat series, IRS, and MODIS have made it possible to

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W1-2022-49-2023 | © Author(s) 2023. CC BY 4.0 License.

prepare thermal, low-cost, and high-speed monitoring of soil salinity (Dwivedi, 2001; Allbed and Kumar, 2013). Landsat-8 OLI satellite images were used to map soil salinity in Turkey (Gorji et al., 2017). In this study due to evaluate the efficiency of machine learning algorithm in identifying soil salinity in an arid and semi-arid region, three soil salinity indices were derived from Landsat-8 OLI and, Linear regression performed on indices with EC measurement, produced a soil salinity map with one of the indices with the most accuracy. In another study in Eshtehard Salt River located in Alborz, Iran (Zarei et al., 2021a) sentinel-2 satellite data were used to derive several indices to compare machine learning algorithms for soil salinity estimation. The result showed that the XGBoost method outperformed other models. However, the other two models, which were RF and GBM, also performed well.

In another research by (Aksoy et al.,2022) they assessed the performance of machine learning algorithms for soil salinity mapping in the Google Earth Engine platform. They showed that despite the CART algorithm providing a better prediction of soil salinity, Random Forest model estimated more reliable salinity levels in salt crusts, agricultural lands, drainage areas, and swamps. In a study by (Hoa et al., 2019) in the Mekong River Delta of Vietnam Sentinel, five machine learning models were used to prepare the salinity map. Multilayer Perceptron Neural Networks (MLP-NN), Radial Basis Function Neural Networks (RBF-NN), Gaussian Processes (GP), Support Vector Regression (SVR), and Random Forests (RF) were those algorithms whose performance was compared in estimating salinity. The results showed that the GP model showed the best performance in predicting salinity in this area (Rostami et al., 2022)

In the present study, soil salinity is monitored by measuring sample points and using Landsat-8 OLI data. The main goals of this research are : 1) To investigate salinity characteristics of the soils in Maharloo Salt Lake, 2) To determine the variables and indicators which are suitable for predicting soil salinity, 3) To evaluate the efficiency of machine learning algorithms in predicting soil salinity and validation with collected grand data, and finally 4) To investigate the efficiency of machine learning models to identify drier and saltier lake boundaries that have dried up in recent years due to climate change.

2. MATERIAL AND METHODS

2.1 Study Area

Maharloo Salt Lake is located in Tasht-Bakhtegan-Maharloo Basin in southwest Iran (Figure 1). It is a seasonal Salt Lake in Shiraz highland area. The seasonal river of Rudkhane-ye-Khoshk is flowing into the lake. The study area is located in Maharloo Lake with an area of about 10x35 km2 (latitude 29 31 21 N to 29 17 21 N and longitude of 52 43 57.51 E52 55 30 E) (Figure 2).

2.2 Datasets

2.2.1 In-Situ Data: The data set used in this study is based on the EC ground measurements, satellite images collected at the same interval with soil sampling, and several remote sensing indices for measurements of soil samples. TDR-350 device as utilized, soil moisture, and salinity meter, which provide an accurate measurement of EC, moisture, and temperature of soil (Figure 2). 177 points around Maharloo Lake were collected in Google Earth Pro software, and the ground data were measured (Figure 3). Based on the soil salinity classes Durand introduces, we assume five salinity classes shown in Table 1 (Durand, 1983). TDR-350 device stores the location for each point measured; thus, X and Y coordinate based on the UTM system and are used for locating the sampling points on SNAP software and the value of each environmental indices corresponding to each point was calculated by SNAP software. Finally, ground data and their corresponded environmental features were randomly divided into 70% for training data and 30 % for testing the dataset.

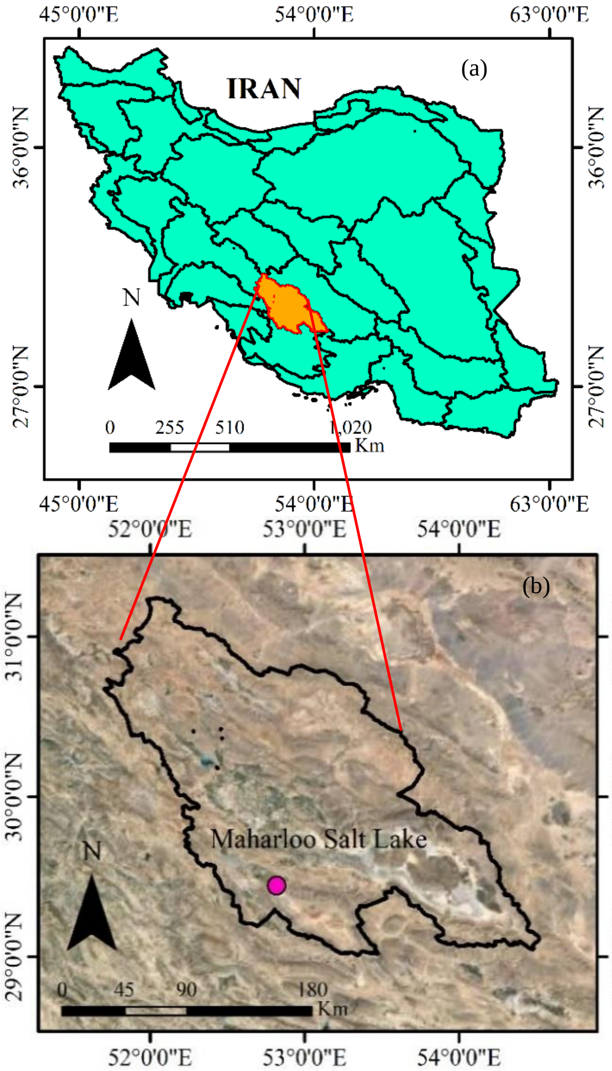


Figure 1. Location of a) Tasht-Bakhtegan-Maharloo Basin in southwest Iran, and b) Maharloo Salt Lake in its hydrologic Basin.

2.2.2 Satellite Imagery: One Landsat-8 images dated 24 September 2021 were selected as satellite images with 0 % cloud cover. Due to the date of ground points taken, which were on September 23th and 24th, the time of ground measuring data and acquisition of satellite image were corresponded. Recording to NASA, Landsat-8 is a satellite image with 30-meter spatial resolution and 12-meter locational accuracy (NASA, 2020).

Landsat-8 satellite imagery did not require atmospheric or geometric corrections because a used surface reflection images provided in a ready-made format at Google Earth Engine was used. Google Earth Engine is a cloud computing platform for preprocessing and processing satellite imagery and other geospatial data. This platform provides access to a wide range of satellite images and the ability to analyze these images. Therefore, the necessary preprocessors have already been done on the satellite images that are called on the GEE platform.

Figure 2. Top photo, TDR 350 device for In-situ data collection, Bottom photo, a general view of the study area.

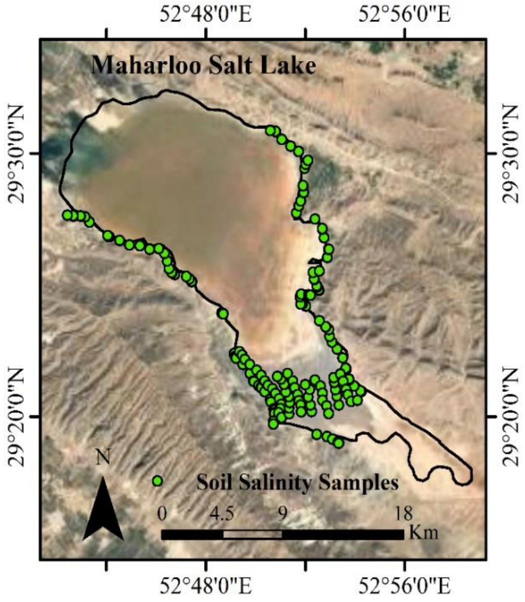


Figure 3. The study area of Maharloo Salt Lake with the ground data locations.

EC (ds/m)	Salinity Category
EC < 0.6	None saline
0.6 < EC < 1	Slightly saline
1 < EC < 2	Moderately saline
2 < EC < 4	Very saline
EC > 4	Extremely saline

Table 1. Salinity classes were introduced by Durand (Durand, 1983).

3. METHODOLOGY

The flowchart of the methodology is presented in Figure 4**.** In preparation of the in-situ and satellite data, we used 177 points in the study are for measuring EC of the soil, 70 % of the selected points (n = 123) were used for training the machine learning models, and the remaining 30% points (n = 54) used to test the accuracy of the model. Furthermore, the purpose of the study is to test and evaluate the sustainability of three machine learning algorithms: RF, MLP, and GBM. These algorithms are used to model the relationship between the soil salinity indices derived from satellite images and the EC measured over Maharloo Salt Lak. image processing and machine learning, is a facilitating way to convert a set of measured data with features to a dataset containing lots of practical information (Richards, 2013). In this study, we use 18 features derived from Landsat-8 OLI images. 7 spectral bands were derived from the Landsat-8 OLI image, and 10 vegetation and salinity features were extracted from Landsat-8 OLI data (Table 2). These indices have shown the best performance in modeling the salinity prediction and determination in previous studies (Allbed and Kumar, 2013; Scudiero et al., 2014; Wang et al., 2019).

3.1 Machine Learning Algorithms

Machine learning algorithms can be divided into two main categories; Regression and Classification. In this study, we use three Regression models to estimate soil salinity. Regression models predict a relationship between a dependent and some independent variables. we use regression models, RF (Random Forest), GBM (Gradient Boosting Machine) and, MLP (Multi-layer Perceptron) to estimate the best algorithms for predicting soil salinity.

3.1.1 Random Forest (RF): Random Forest (RF) was proposed by Breiman (Breiman, 2001). This machine learning model is an algorithm that is produced based on the CART (Classification and Regression Trees) algorithm. This model contains several groups of decision trees. These groups do not correlate; each tree produces its evaluation, and the RF evaluation process is based on the average of those multiple decision trees. Moreover, a set of all subsets of the decision tree is formed, and the final Random Forest model is derived. The Random Forest model has shown excellent efficiency in soil assessment, such as mapping soil properties by.

3.1.2 Gradient Boosting Machine (GBM): Boosting is a method used to turn weak learners into strong learners. Boosting is done so that each new tree fits the modified version of the original data set. This algorithm uses a series of sets to eliminate noise and variance. Also, some boosting methods allow newer models to use the error of previous models, which is implemented regularly. Like the RF algorithm in GBM, two critical parameters (n estimator and max feature) are used in the grid search method to optimize the model (Friedman, 2001).

3.1.3 Multi-layer Perceptron (MLP): A Multi-layer Perceptron (MLP) Regression System is a multi-layered neural network training system that uses a multi-layer perceptron regression algorithm to solve regression problems. This model usually has three layers; input, hidden, and output. The number of input neurons in the model equals the number of input variables. In contrast, the number of neurons in the hidden layers must be calculated, and the search network optimization method was used to select the best number of hidden layer neurons. The number of output neurons is one, which in this article is the same as EC .An MLP algorithm consists of several layers of input nodes connected as a directed graph between the input and output layers (Haykin, 1998).

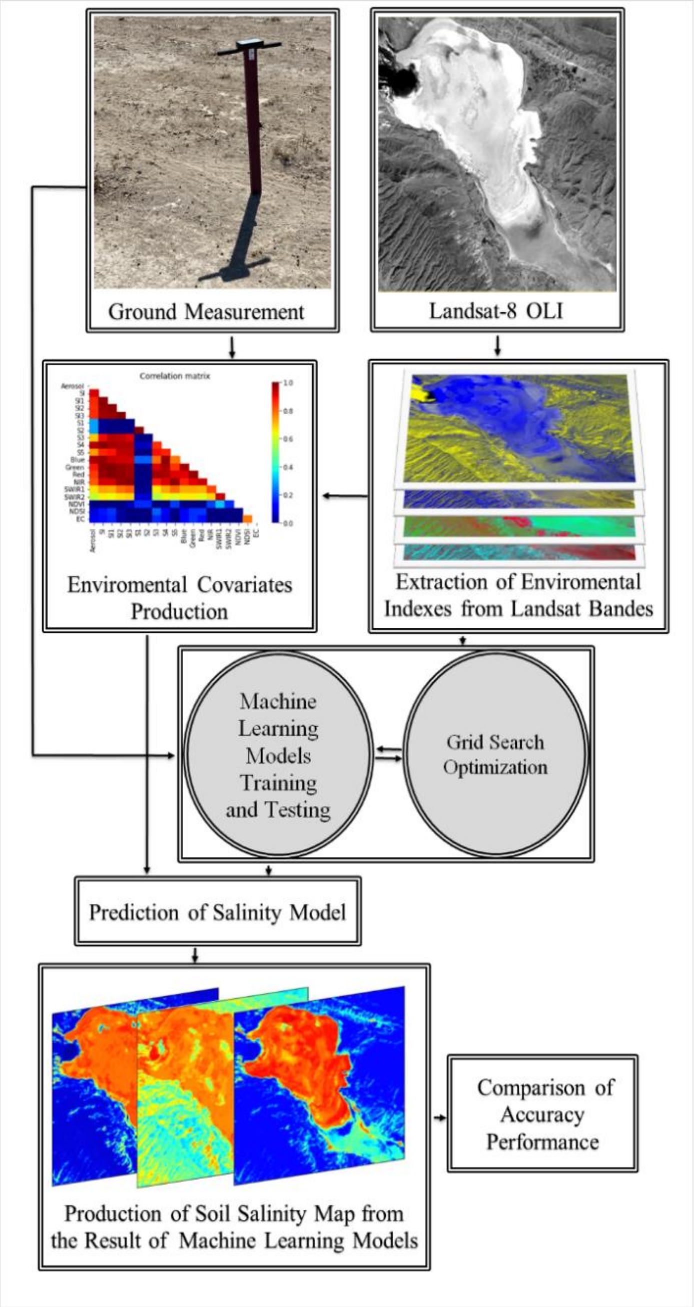

Ground Measurement
Landsat-8 OLI
Correlation matrix
Enviromental Covariates Production
Extraction of Enviromental Indexes from Landsat Bandes
Machine Learning Models Training and Testing
Grid Search Optimization
Prediction of Salinity Model
Comparison of Accuracy Performance
Production of Soil Salinity Map from the Result of Machine Learning Models

Figure 4. Flowchart of the methodology used in Maharloo Salt Lake to estimate soil salinity.

3.2 Performance Assessment

To measure the performance of soil salinity models, we used two statistical metrics. These statistical metrics are, RMSE (Root Mean Square Error) and R2 (coefficient of determination):

$$\text{RMSE}= \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)\ ;\ R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}\ , \quad (1)$$

Where n = sample size
y_i = the vector of observed values of the Variable being predicted
$\hat{y}_i$ = a vector of predicted dependent variables with n data points
$\bar{y}_i$ = the mean of the observed dependent variable

a higher coefficient of determination R^2 values and lower RMSE values Demonstrate better model performance (Wang et al., 2019).

Feature	Formula
Band 1-7 Landsat-8	
NDVI	$\frac{NIR - Re\,d}{NIR + Re\,d}$
NDSI	$\frac{Red - NIR}{NIR + Red}$
SI	$\sqrt{Blue * Red}$
SI1	$\sqrt{Green * Red}$
SI2	$\sqrt{Green^2 + Red^2 + NIR^2}$
SI3	$\sqrt{Green^2 + NIR^2}$
S1	$\frac{Blue}{Red}$
S2	$\frac{Blue - Re\,d}{Blue + Re\,d}$
S3	$\frac{Green * Red}{Blue}$
S4	$\frac{Blue * Red}{Green}$
S5	$\frac{NIR * Red}{Green}$

Table 2. Features extracted from Landsat-8, including seven bands of Landsat-8 OLI, vegetation, and salinity indices.

4. RESULT AND DISCUSSION

After preparing In-situ data and processing satellite images, the environmental variables and indicators were derived prior to producing the models. The correlation between the features such as different bands of Landsat-8 OLI, vegetation indexes, and salinity indexes, after that their relationship with the EC parameter was extracted as a matrix; therefore, to find the relationship between those indexes and efficiency in detecting salinity parameter, the correlation matrix was derived (Figure 5). According to the location of in-situ data, the required satellite images were subset, and their pixel size was estimated. A matrix was prepared to be given to machine learning algorithms containing 18 environmental indicators derived from satellite images for each soil point and salinity parameter corresponding to that point. The data in the matrix was divided into 70 % and 30 % for training and testing the data, respectively.

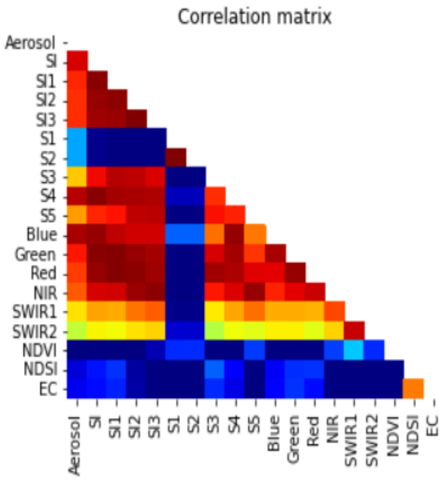
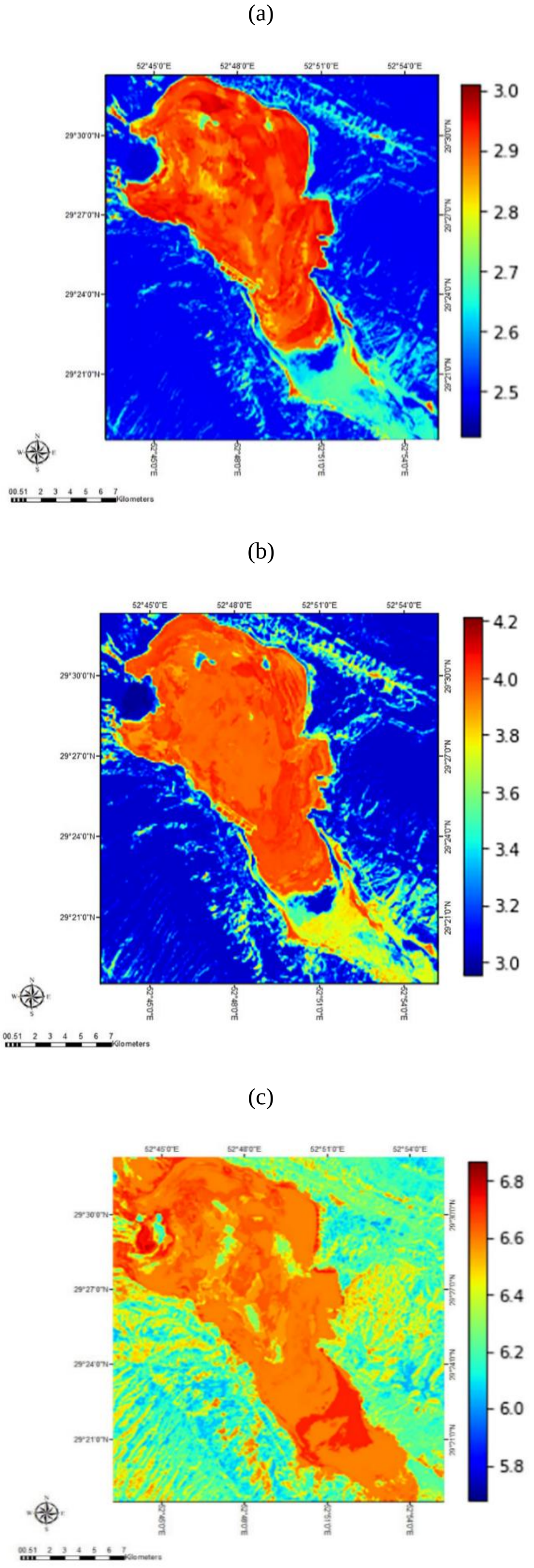


Figure 5. Correlation between environmental features and the relation of environmental indices and electrical conductivity.

The performance of three machine learning regression algorithms, RF, GBM, and MLP, was investigated in order to analyse the soil salinity distribution by these algorithms. Various parameters were defined and used to optimize the model, this process for each model and their results were generated in Google Colab. After training and testing the models, the salinity for each image pixel was plotted using the results of regression models. After training and testing the models, the salinity for each image pixel was plotted using the results of regression models (Figure 6). The three models of the learning machine that salinity maps were drawn and displayed according to the salinity per pixel are compared based on statistical parameters R2 and RMSE. After optimizing the machine learning models, the best output of these models, in RF model R2 = 0.85, in GBM R2 = 0.89 and in MPL model R2 = 0.75.

As shown in Table 3, the statistical parameters for each model are fully shown. MPL model showed the worst performance in predicting salinity in the soil. GBM has the best performance between tree machine learning models and one of the reason for excellent performance of this model is that GBM model provides several hyper-parameter tuning options that make the function fit very flexible. Moreover, GBM performed better than the RF model because Gradian Boosting trees are more precise than the trees of the Random Forest model. Also, the trees of the GBMModel are trained in a way that corrects each other's mistakes, and these trees can receive and capture a more complex pattern in the input data than the trees in the RF model. The common problem with neural network models such as MLP is that they require much more data than other learning machine models so that the other models can perform better in low data volumes.


(a)

(b)

(c)

Figure 6. Salinity is predicted by the results of learning machine models. a) Random Forest, b) Gradient Boost Machine, and c) Multi-layer Perceptron.

5. CONCLUSION

In this investigation, 18 environmental indicators were extracted from Landsat sensor bands; the relationship between each of these indicators is calculated. The three machine learning models used to predict salinity were RF, GBM, and MLP for this study area. GBM algorithm showed the best performance with $R^2 = 0.89$ and RMSE = 0.63 for testing the dataset. There is an increasing drought in the study area due to climate change. The lake is drying up, and great amount of salt deposited in the lake especially in the shores and margins. In addition, the entire lake bed surface is completely covered by salt layers.

All three models demonstrate very clearly that this salinity boundary, which is higher on the lake's shores than elsewhere. It is also concluded that the GBM and RF models show the salinity distribution more enhanced and closer to reality. It is of importance to mention that, all three models showing noticeably that the lake environment have become drier and saltier due to recent climate change experienced in the region and the country entirely.

REFERENCES

Aksoy, S., Yildirim, A., Gorji, T., Hamzehpour, N., Tanik, A., Sertel, E., 2022. Assessing the performance of machine learning algorithms for soil salinity mapping in Google Earth Engine platform using Sentinel-2A and Landsat-8 OLI data. *Adv. Sp. Res*. 69, 1072–1086. https://doi.org/10.1016/j.asr.2021.10.024.

Allbed, A., Kumar, L., 2013. Soil Salinity Mapping and Monitoring in Arid and Semi-Arid Regions Using Remote Sensing Technology: A Review. *Adv. Remote Sens*. 02, 373–385. https://doi.org/10.4236/ars.2013.24040.

Arnous, M.O., Green, D.R., 2015. Monitoring and assessing waterlogged and salt-affected areas in the Eastern Nile Delta region, Egypt, using remotely sensed multi-temporal data and GIS. *J. Coast. Conserv*. 19, 369–391. https://doi.org/10.1007/s11852-015-0397-5.

Davis, E., Wang, C., Dow, K., 2019. Comparing Sentinel-2 MSI and Landsat 8 OLI in soil salinity detection: a case study of agricultural lands in coastal North Carolina. *Int. J. Remote Sens*. 40, 6134–6153. https://doi.org/10.1080/01431161.2019.1587205.

Dwivedi, R.S., 2001. Soil resources mapping: A remote sensing perspective. *Remote Sens. Rev*. 20, 89–122. https://doi.org/10.1080/02757250109532430.

Elhag, M., Bahrawi, J.A., 2017. Soil salinity mapping and hydrological drought indices assessment in arid environments based on remote sensing techniques. *Geosci. Instrumentation, Methods Data Syst*. 6, 149–158. https://doi.org/10.5194/gi-6-149-2017.

Eswar, D., Karuppusamy, R., Chellamuthu, S., 2021. Drivers of soil salinity and their correlation with climate change. *Curr. Opin. Environ. Sustain*. 50, 310–318. https://doi.org/10.1016/j.cosust.2020.10.015.

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using Remote Sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. PLoS One 12, 1–21. https://doi.org/10.1371/journal.pone.0170478.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Gorji, T., Sertel, E., Tanik, A., 2017. Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey. *Ecol. Indic*. 74, 384–391. https://doi.org/10.1016/j.ecolind.2016.11.043.

Gorji, T., Sertel, E., Tanik, A., 2017. Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey. *Ecol. Indic*. 74, 384–391.

Hoa, P.V., Giang, N.V., Binh, N.A., Hai, L.V.H., Pham, T.D., Hasanlou, M., Bui, D.T., 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 Sens*. 11, 1–21. https://doi.org/10.3390/rs11020128.

Jahan, E., Ur, T., 2021. Simulation of future land surface temperature under the scenario of climate change using remote sensing & GIS techniques of northwestern Rajshahi district , Bangladesh. *Environ. Challenges* 5, 100365. https://doi.org/10.1016/j.envc.2021.100365.

Khan, N.M., Rastoskuev, V. V., Sato, Y., Shiozawa, S., 2005. Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agric. Water Manag.* 77, 96–109. https://doi.org/10.1016/j.agwat.2004.09.038.

Peng, J., Biswas, A., Jiang, Q., Zhao, R., Hu, J., Hu, B., Shi, Z., 2019. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. *Geoderma* 337, 1309–1319. https://doi.org/10.1016/j.geoderma.2018.08.006.

Rostami, A., Shah-Hosseini, R., Asgari, S., Zarei, A., Aghdami-Nia, M., Homayouni, S., 2022. Active Fire Detection from Landsat-8 Imagery Using Deep Multiple Kernel Learning. *Remote Sens*. 14. https://doi.org/10.3390/rs14040992.

Singh, S., Ghosh, N.C., Gurjar, S., 2020. Index-based assessment of suitability of water quality for irrigation purpose under Indian conditions. *Environ Monit Assess* 190: 29. https://doi.org/10.1007/s10661-017-6407-3.

Vermeulen, D., Van Niekerk, A., 2017. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. *Geoderma* 299, 1–12. https://doi.org/10.1016/j.geoderma.2017.03.013.

Zarei, A., Hasanlou, M., Mahdianpari, M., 2021a. A comparison of machine learning models for soil salinity estimation using multi-spectral earth observation data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*. 5, 257–263. https://doi.org/10.5194/isprs-annals-V-3-2021-257-2021.

Zarei, A., Shah-hosseini, R., Ranjbar, S., Hasanlou, M., 2021b. ScienceDirect Validation of non-linear split window algorithm for land surface temperature estimation using Sentinel-3 satellite imagery : Case study ; Tehran Province , Iran. *Adv. Sp. Res*. 67, 3979–3993. https://doi.org/10.1016/j.asr.2021.02.01