International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

Enhancing soil salinity prediction in semi-arid regions using machine learning models technology

Ahmed Elsayed Amin Abd Elaziz, Khaled Goda Soliman, Mohamed Said Abu-Hashim, Enas Mohamed Wagdi Abdel Hamed and Mohamed Said Metwally*

Department of Soil Sciences, Faculty of Agriculture, Zagazig University, Egypt

Abstract

This study addresses the significant challenge posed by soil salinization in the fertile Nile Delta region, which threatens agricultural productivity and food security. Conventional methods for soil salinity assessment often lack the requisite speed for timely decision-making to mitigate salinity in these lands, necessitating the exploration of advanced techniques. Leveraging the capabilities of machine learning algorithms, this research develops robust predictive models for soil salinity in the Nile Delta. Three state-of-the-art machine learning algorithms: Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Random Forest (RF), were rigorously evaluated using a comprehensive dataset derived from 120 soil samples collected across the region. The models underwent meticulous training and validation processes, incorporating cross-validation techniques and stringent performance evaluation metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Squared Error (RMSE), and R^2. The results unequivocally demonstrated the superior performance of SVM, achieving remarkable values of 0.006 dS/m for MSE, 0.079 dS/m for RMSE, 0.007 dS/m for MAPE, 0.062 dS/m for MAE and 1.0 for R^2 during the training phase, further corroborated by an 0.008 dS/m for MSE, 0.089 dS/m for RMSE, 0.012 dS/m for MAPE, 0.071 dS/m for MAE and 0.99 for R^2 during the validation stage. This study elucidates the immense potential of machine learning techniques in accurately predicting soil salinity, paving the way for proactive management strategies and sustainable crop production practices in the pivotal Nile Delta region, thus enhancing sustainable crop production and agricultural management.

Keywords: Soil salinity, Machine learning, Support Vector Machine, Smart farming, Agri-environmental informatics.

Full length article **Corresponding Author*, e-mail: ahmdamyn19p@gmail.com

1. Introduction

Soil salinity poses a critical challenge to agricultural productivity, with the Nile Delta region experiencing its detrimental effects. This vast and agriculturally significant area faces a complex interplay of factors contributing to soil salinity, affecting crop yields and food security [1]. Conventional methods often fall short of providing accurate and timely assessments, prompting the exploration of machine learning (ML) as potential solutions to this pressing issue. The Nile Delta, an agricultural heartland, grapples with soil salinity due to a combination of human activities and environmental factors [2]. Excessive use of irrigation practices, compounded by the intricate dynamics of the delta's ecosystems, results in elevated salinity levels. This not only hinders crop growth but also jeopardizes the delicate balance required for sustainable agriculture in the region. The socio-economic impact of soil salinity extends beyond the farm gate, affecting livelihoods and the overall food supply chain. The integration of AI and ML technologies emerges as a promising avenue to revolutionize soil salinity prediction and management [3]. Algorithms such as Random Forest, XGBoost, and Support Vector Machines offer the ability to process vast datasets, identifying subtle patterns that traditional methods might overlook. By analyzing the intricate interactions among soil composition, climate, and agricultural practices, AI models hold the potential to enhance our understanding of soil salinity dynamics and contribute to effective mitigation strategies. Objectives of the Study, this study seeks to harness the power of ML models to accurately predict soil salinity in the Nile Delta region. The overarching goal is to develop robust models that provide real-time insights, enabling proactive management of soil salinity. The study further aims to compare the performance of various ML models, analyze critical input variables, and offer recommendations for the adoption of the most effective models to address soil salinity challenges in the region. In the following sections, we will delve into the specifics of the study area, providing a comprehensive description of the Nile Delta's geography, climate, soil types, and cropping patterns. Subsequently, we will detail the data collection and

preprocessing methodologies employed in this research, shedding light on the critical steps taken to ensure the accuracy and reliability of the data used. The methods section will then explore the ML algorithms evaluated, the input variables considered for each model, and the intricacies of the model training and validation processes. The performance evaluation metrics employed will be discussed, providing transparency into the criteria used to assess the efficacy of the models. The results and discussion section will be dedicated to presenting and analyzing the prediction outcomes of different models for key soil salinity parameters. A comprehensive comparison of model performances based on evaluation metrics will be undertaken, accompanied by an in-depth analysis of important input variables across various models. This section will also explore the reasons behind differences in model performances, providing valuable insights into the nuances of predicting soil salinity using AI and ML.

2. Materials and methods

2.1. Study Area and soil sampling

The study focuses on the extension of the Nile Delta region, a crucial agricultural zone with distinctive characteristics. Located in the northeast Nile Delta Egypt, the Nile Delta is formed by the Nile River's intricate network of distributaries as they meet the Mediterranean Sea. On 20th March 2022, 120 soil samples were collected at a depth of 30cm from the study area (Fig. 1). The region's climate is predominantly Mediterranean, characterized by hot, dry summers and mild, wet winters. The delta's soil types vary, encompassing alluvial soils enriched by the river's sediment deposits. Common crops include rice, wheat, and various fruits and vegetables. However, the region faces challenges associated with soil salinity. The delta's proximity to the Mediterranean, coupled with intensive irrigation practices, contributes to the accumulation of salts in the soil. This salinization poses a significant threat to agricultural productivity, affecting crop growth and soil fertility.

2.2. Soil chemical properties analysis

The concentrations of soluble anions like HCO_3^-, Cl^-, SO_4^{2-} and cations like $Na^+, K^+, Ca^{2+}, Mg^{2+}$ were assessed from the extract of soil paste also ECe, utilizing the established methods outlined by [4]. pH was measured in a 1:2.5 soil-water suspension using pH meter. $CaCO_3$ content was measured using calcimeter method [5].

2.3. Description of data collected

Comprehensive data were collected to develop accurate models for predicting soil salinity. Soil samples were strategically collected from various locations across the Nile Delta, considering different soil types and land-use patterns. Analysis of these soil samples included key parameters: electric conductivity (EC_e), calcium (Ca^{2+}), magnesium (Mg^{2+}), potassium (K^+), sodium (Na^+), chloride (Cl^-), sulfate (SO_4^{2-}), and bicarbonate (HCO_3^-). These parameters were chosen to capture the diverse chemical composition influencing soil salinity.

2.4. Data preprocessing

To ensure the reliability of the dataset, several preprocessing steps were undertaken. Missing data were addressed through imputation methods, such as mean substitution, to maintain dataset integrity. Outliers, identified through robust statistical techniques, were either corrected or removed to prevent their undue influence on model training. Normalization and standardization techniques were applied to bring all variables to a consistent scale, facilitating the effective training of machine learning (ML) models. A correlation analysis identified and addressed multicollinearity among input variables, ensuring that redundant information did not compromise the models' performance. Spatial autocorrelation, a common issue in geospatial datasets, was mitigated through spatial smoothing techniques. This step aimed to reduce the impact of localized variations and enhance the generalizability of the models across the entire study area. The soil dataset was subsequently randomly divided into training and validation sets, with 20% of the data allocated for validation, equivalent to 24 soil samples. The remaining 80% of the data was designated for training, comprising 96 soil samples. This division facilitates the creation and evaluation of an effective model. The training set served to learn the machine learning models the underlying patterns in the data, while the validation set provided an independent dataset to assess the models' predictive performance.

2.5. Machine learning Algorithms

The study employed two advanced ML algorithms: Random Forest (RF), XGBoost (XGB) and support vector machine (SVM).

2.5.1. Random Forest (RF)

Random Forest is a powerful machine learning technique that belongs to the ensemble learning family of algorithms [6]. It has proven to be an effective approach for predicting soil salinity parameters due to its capability to model intricate relationships between input variables and target data. The algorithm constructs an ensemble of multiple decision trees, leveraging their collective strength to enhance the overall accuracy and robustness of the model [7]. In the realm of soil salinity prediction, Random Forest can be employed to explore the complex interplay between various environmental factors, such as meteorological conditions, soil properties, subsurface characteristics, and the resulting soil salinity levels. The model utilizes these environmental variables as input features, while the target variable represents the soil salinity value to be predicted. The algorithm is trained and validated using a dataset collected from a specific region of interest, such as the Manas River Basin in China's Xinjiang Uygur Autonomous Region [6]. One of the notable advantages of Random Forest is its ability to handle both categorical and numerical data seamlessly, making it suitable for a wide range of soil salinity prediction studies. Additionally, the algorithm is robust to missing data and outliers, which are common challenges encountered in environmental datasets. By leveraging its ensemble approach, Random Forest can effectively capture the intricate patterns and relationships present in the data, leading to improved predictive performance compared to individual decision trees [7].

2.5.2. Extreme Gradient Boosting (XGB)

XGBoost (Extreme Gradient Boosting) is an open-source machine learning library that provides a scalable,

distributed gradient boosting framework for various programming languages, including C++, Java, Python, R, and others [8] . This study utilized the XGBoost from pip library in Python to train and validate a prediction model using soil data. It is designed to be highly efficient, flexible, and portable, implementing machine learning algorithms widely used for regression, classification, and ranking problems. XGBoost is built on the principles of supervised machine learning, decision trees, ensemble learning, and gradient boosting [9]. XGBoost is renowned for its ability to achieve high accuracy in predictive modelling tasks, often outperforming other machine learning algorithms. It is particularly effective in handling large datasets with numerous features, and it includes regularization techniques to prevent overfitting [8]. The model is trained using a gradient boosting approach, where each iteration builds a new decision tree that focuses on the residual errors of the previous tree. Some key features of XGBoost include parallel processing, built-in cross-validation, and the ability to handle non-linear data patterns. It can also be integrated with distributed processing frameworks like Apache Hadoop, Apache Spark, and Dask for scalability [10].

2.5.3. Support Vector Machine (SVM)

SVM is a powerful type of supervised learning algorithm in machine learning, known for its effectiveness in solving classification and regression problems. They are particularly well-suited for binary classification tasks, where the objective is to classify the elements of a dataset into two distinct groups. The fundamental aim of an SVM algorithm is to find the optimal decision boundary, often referred to as a hyperplane, that separates the data points of different classes. This hyperplane is especially useful when working in high-dimensional feature spaces. The key principle behind SVMs is to maximize the margin, which is the distance between the hyperplane and the closest data points of each category, thereby enhancing the ability to discriminate between different classes with high accuracy [11]. These Ensemble learning techniques have demonstrated effectiveness in managing intricate connections within datasets and are particularly suitable for forecasting soil salinity.

2.6. Input variables for each model

For RF, SVM and XGB, the input variables comprised the comprehensive set of soil and environmental parameters collected during the data collection phase. This included EC_e, Na^+, K^+, Ca^{2+}, Mg^{2+}, $HCO_3^-, Cl^-, SO_4{}^{2-}$,along with pH and $CaCO_3$.

2.7. Model training and validation process

The models underwent a rigorous training process using the designated training dataset. Hyperparameter tuning was performed to optimize model performance. Following training, the models were validated using the independent validation dataset to assess their ability to generalize to new, unseen data. In the process of training and testing the models, a k-fold cross-validation approach (k = 5) was employed to prevent overfitting of the models [12].

2.8. Performance evaluation metrics

The evaluation of the accuracy and stability of machine learning models in predicting soil salinity parameters relied on five statistical measures, as proposed by [13]. These measures included the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(A_i - \widehat{A}_i\right)^2}{n}} \qquad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \widehat{A}_i\right)^2 \qquad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \widehat{A}_i\right| \qquad (4)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \widehat{A}_i}{A_i}\right| x\ 100 \qquad (5)$$

Where A_i is the predicated values, $\widehat{A}_i$ is the observed values and n is the number of soil samples.

3. Results

3.1. Statistical analysis for the soil dataset

The statistical descriptions of the dataset are presented in Table 1, which includes the count, mean, standard deviation, minimum, first quartile, median, third quartile, and maximum for soil variables. The dataset includes 120 samples, with pH ranging from 6.9 to 8.3, electrical conductivity (EC) ranging from 3.39 to 18.15 dS/m, sodium Na^+ ($mmol_c$/L) ranging from 10.37 to 30.84, potassium K^+ ($mmol_c$/L) ranging from 0.5 to 2.01, calcium Ca^{2+} ($mmol_c$/L) ranging from 3 to 55, magnesium Mg^{2+} ($mmol_c$/L) ranging from 8 to 175, bicarbonate HCO_3^- ($mmol_c$/L) ranging from 3 to 20, chloride Cl^- ($mmol_c$/L) ranging from 2 to 88, sulfate SO_4^{2-} ($mmol_c$/L) ranging from 9.22 to 211.81, and percentage of $CaCO_3$ ranging from 0.34 to 6.71. From the statistical analysis, we find that the variation between the input values leads to an increase in the accuracy of the prediction of the models used**.**

3.2. Correlation between soil parameters and EC

The Pearson correlation matrix analysis shows that there is a significant correlation between the different features and soil salinity ($p<0.05$) see Fig. 2, revealing that there were low and moderate negative correlations between EC and Na^+ concentration (-0.044), EC and HCO_3^- concentration (-0.52), and positive correlations between EC and Mg^{2+} concentration (0.97), EC and SO_4 concentration (0.95), and EC and $CaCO_3$ concentration (0.23), EC and Cl^- concentration (0.18), EC and Ca^{2+} concentration (0.63), EC and K^+ concentration (0.56). Additionally, there was a strong positive correlation between EC and Mg^{2+} and SO_4^{-2} concentration (0.97), (0.95) respectively.

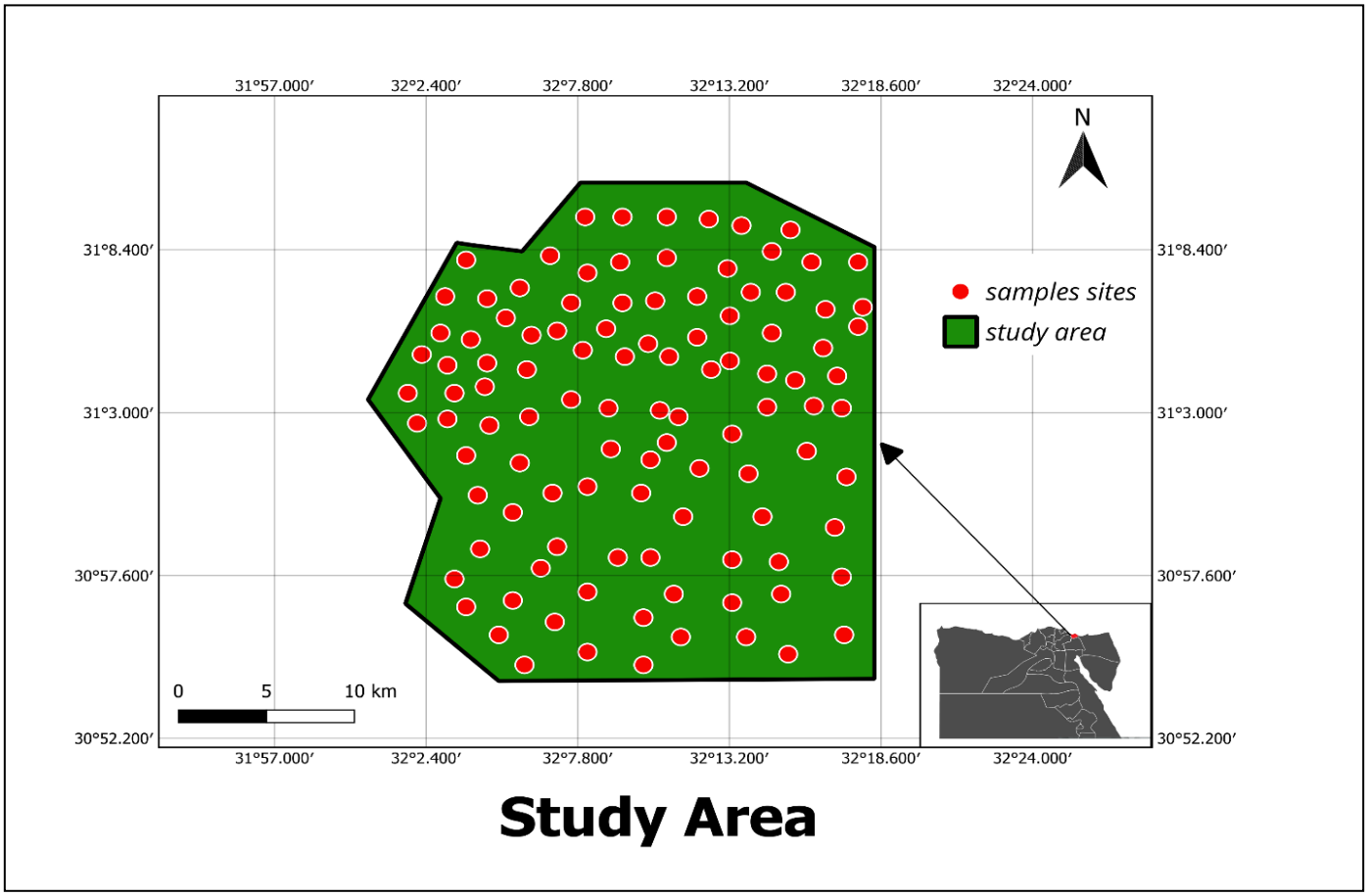


Figure 1: The study area and the locations of soil samples located in the northeast Nile Delta, Egypt.

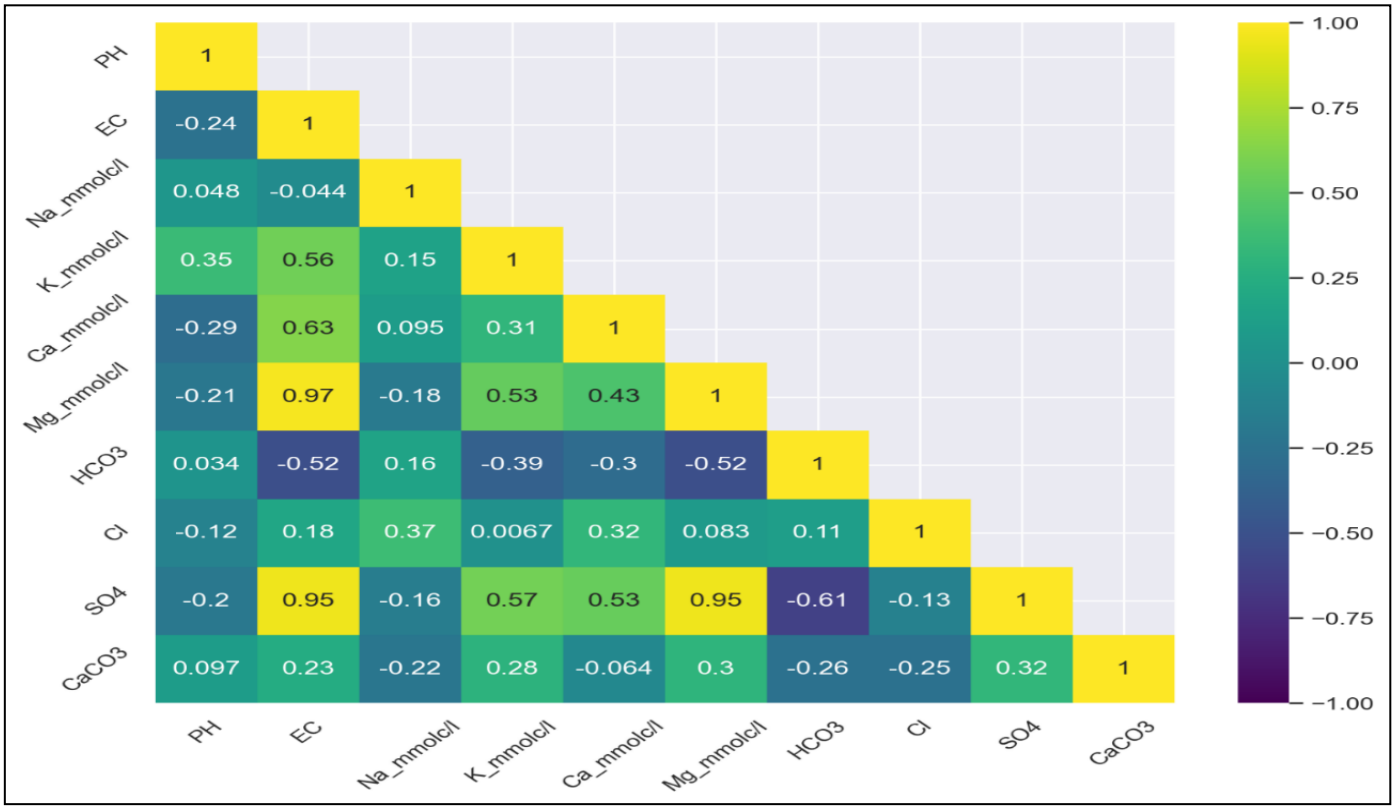


Figure 2: Analysis **of** Pearson Correlation Matrix for Soil Features

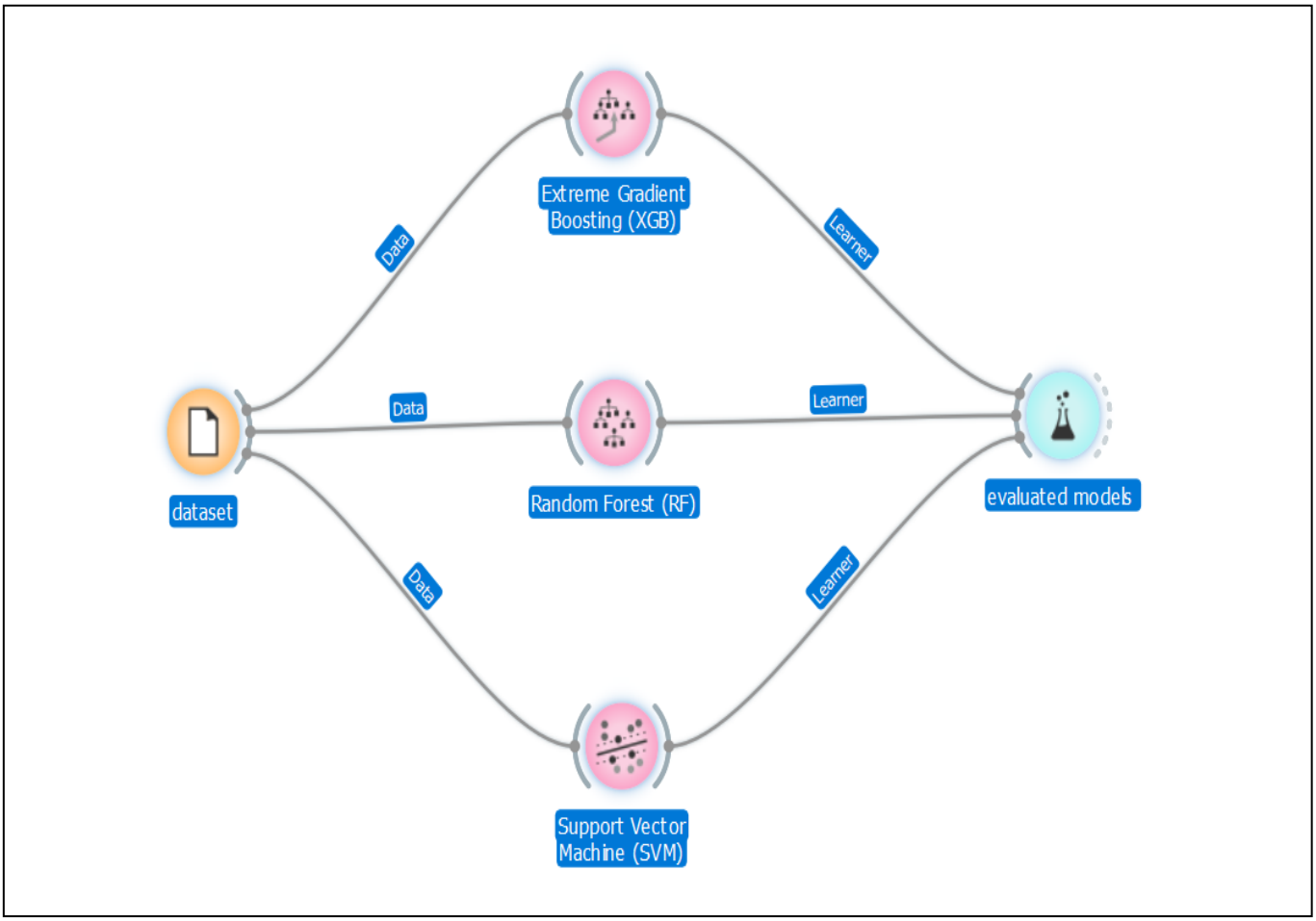
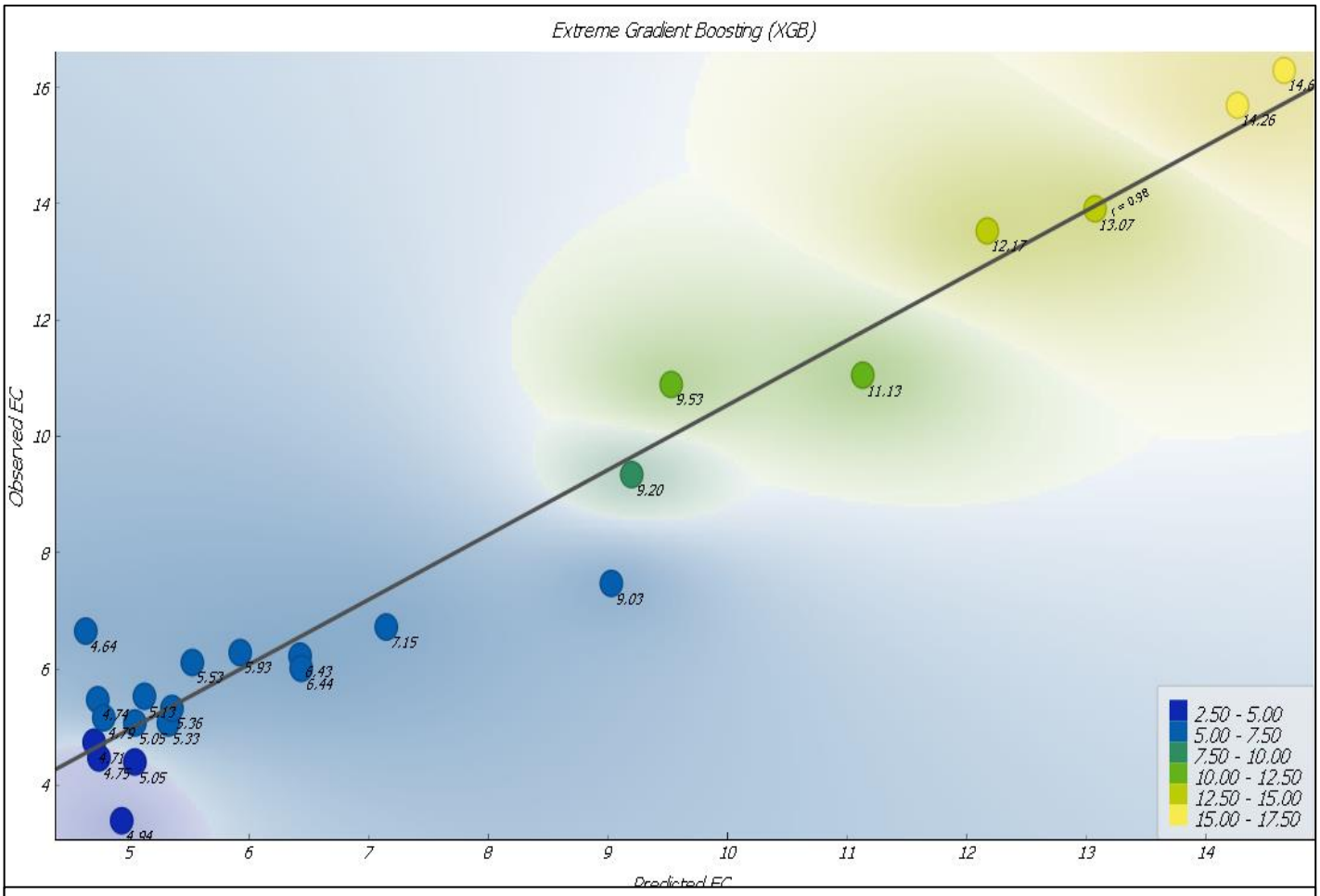


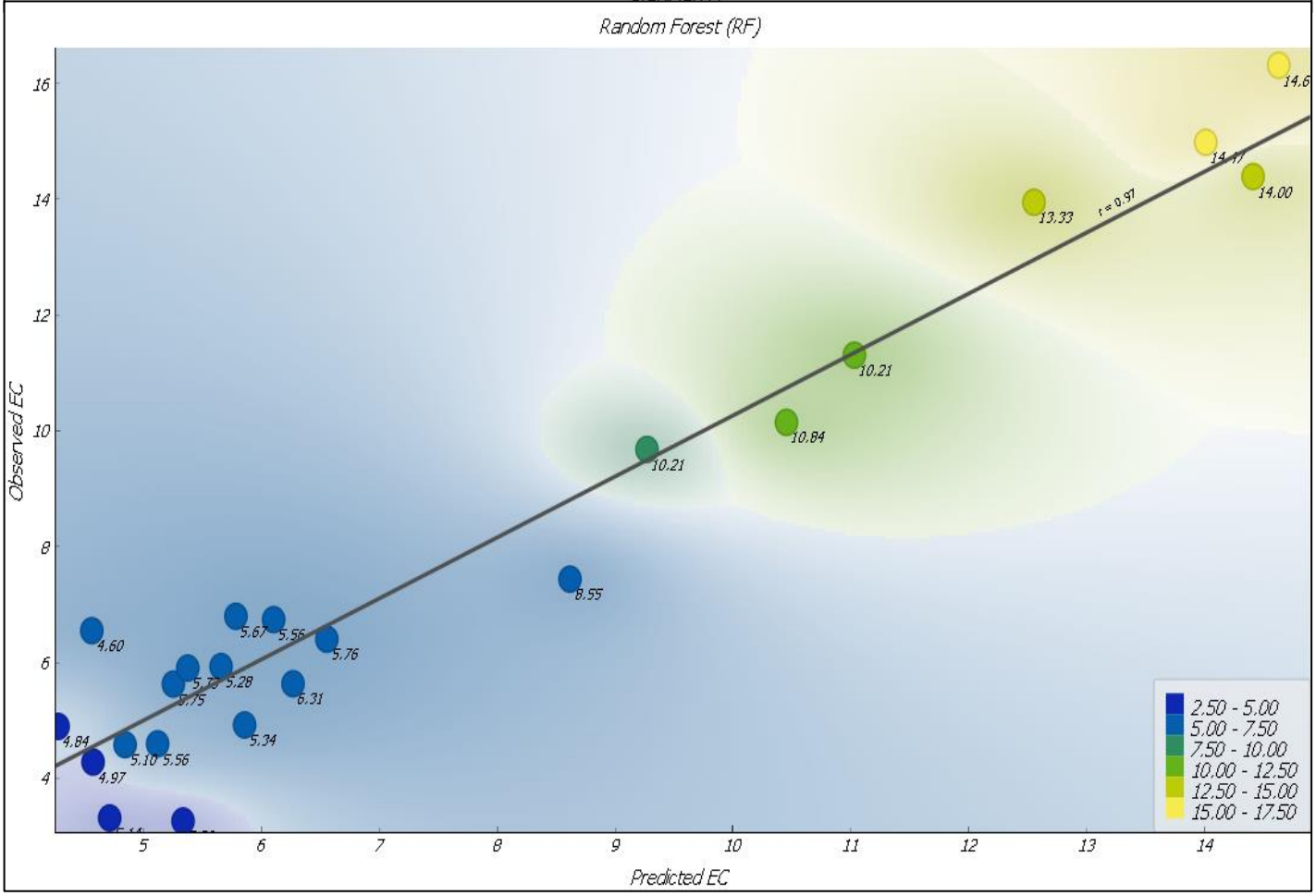
Figure 3: Framework for Machine Learning Models: SVM, RF, XGBoost

Table 1: Statistical descriptions of values of chemical soil properties

Features	count	mean	std	min	25%	50%	75%	max
pH	120	7.61	0.37	6.9	7.3	7.7	7.9	8.3
EC dS/m	120	9.43	3.79	3.39	6.08	9.1	12.96	18.15
Na _ mmol$_c$/l	120	17.42	4.38	10.37	13.66	16.73	20.98	30.84
K _ mmol$_c$/l	120	1.1	0.38	0.5	0.75	1.08	1.37	2.01
Ca _ mmol$_c$/l	120	17.3	11.03	3	10	14	22.25	55
Mg _ mmol$_c$/l	120	82.01	41.79	8	44	77.5	114.75	175
HCO_3^-	120	10.83	3.95	3	8	10	14	20
Cl^-	120	19.74	14.35	2	9	16.5	25	88
SO_4^{-2}	120	87.25	49.31	9.22	42.92	74.74	134.06	211.81
%$CaCO_3$	120	3.55	1.61	0.34	2.14	3.5	5	6.71

Notes: 25% — Q1; 50% — Q2; 75% — Q3; Std — Standard Deviation


Extreme Gradient Boosting (XGB)
Observed EC
Predicted EC
r = 0.98
2.50 - 5.00
5.00 - 7.50
7.50 - 10.00
10.00 - 12.50
12.50 - 15.00
15.00 - 17.50


Random Forest (RF)
Observed EC
Predicted EC
r = 0.97
2.50 - 5.00
5.00 - 7.50
7.50 - 10.00
10.00 - 12.50
12.50 - 15.00
15.00 - 17.50

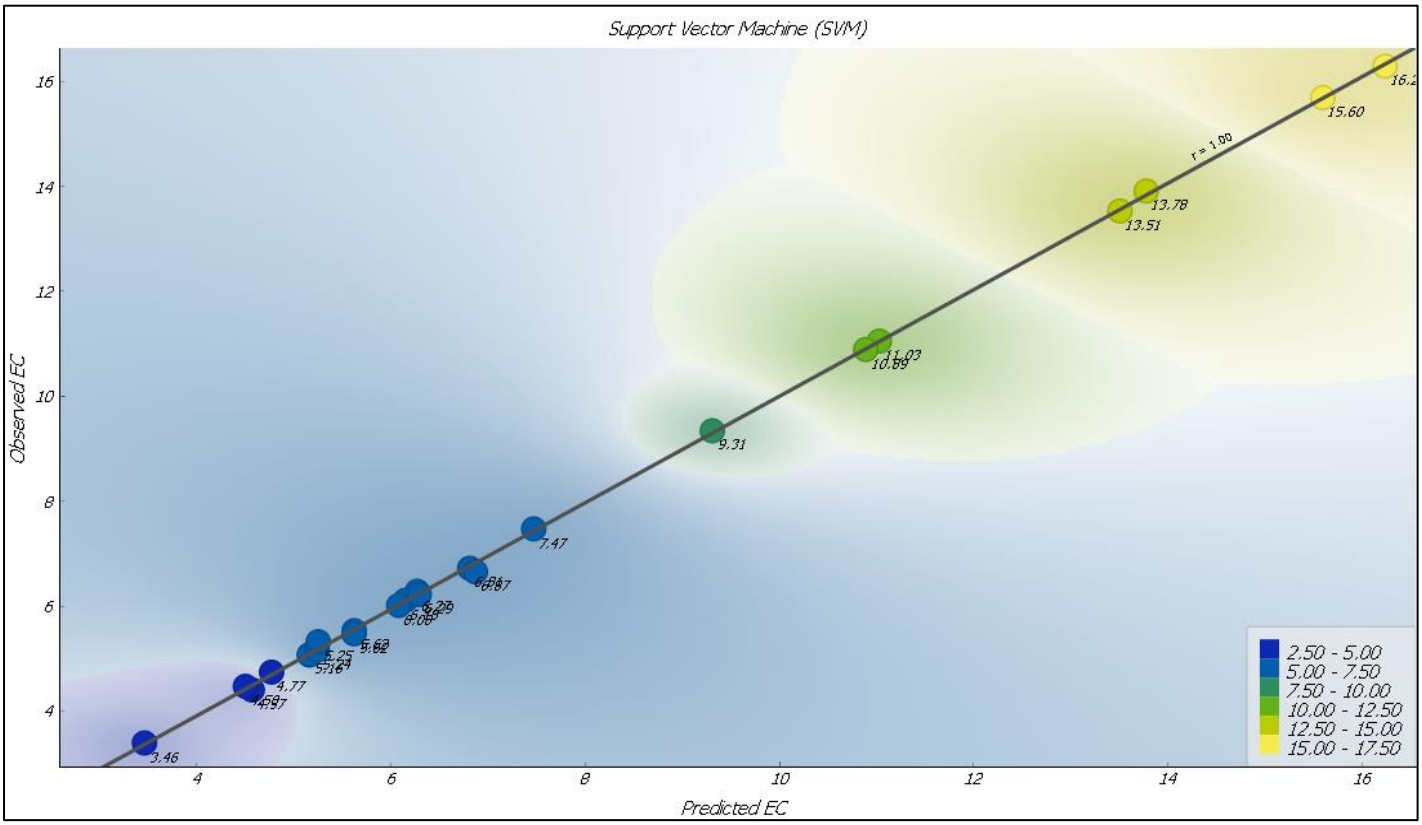


Figure 4: Scatter Plots: Observed vs. Predicted Values for Three ML Models.

Table 2: Statistical values of the three machine learning models during training–testing stage of input data

Model	Train time per second	Test time per second	MSE	RMSE	MAE	MAPE	R^2
XGBoost	0.469	0.012	0.547	0.739	0.478	0.049	0.959
SVM	0.218	0.045	0.006	0.076	0.062	0.007	1
Random Forest (RF)	0.272	0.033	0.532	0.73	0.507	0.055	0.96

Table 3: Statistical values of the three machine learning models during validation stage of input

Model	MSE	RMSE	MAE	MAPE	R^2
SVM	0.008	0.089	0.071	0.012	0.99
Random Forest (RF)	0.814	0.902	0.725	0.114	0.94
XGBoost	0.847	0.92	0.698	0.098	0.938

These findings suggest that there may be complex relationships between these variables that could impact the accuracy of soil salinity predictions.

3.3. Assessing the predictive accuracy of machine learning algorithms

Three machine learning models Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), and Random Forest (RF) were trained, tested, and evaluated utilizing the soil analysis dataset to create an accurate soil salinity prediction model Fig. 3. Table 2 shows statistical values which expresses of the three machine learning models performance during training–testing stage of dataset, there are significant differences in the results of the forecast and the accuracy of the models, and this is due to the type of model used and the input and output data. The performance of each model was assessed based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and R-squared value. SVM performed exceptionally well, yielding the lowest Mean Squared Error (MSE)= 0.006 dS/m, Root Mean Square Error (RMSE)= 0.076 dS/m, Mean Absolute Error (MAE)= 0.062 dS/m, Mean Absolute Percentage Error (MAPE)= 0.007 dS/m, and the highest Coefficient of Determination R^2 =1 compared to XGBoost and Random Forest. xGBoost had the longest training time per second (0.469 seconds), followed by RF (0.272 seconds) and SVM (0.218 seconds). However, SVM had the shortest test time per second (0.006 seconds), compared to xGBoost (0.547 seconds) and RF (0.532 seconds). Specifically, SVM achieved an MSE of 0.547 dS/m, RMSE of 0.739 dS/m, MAE of 0.478 dS/m, and R^2score of 0.959, indicating superior accuracy compared to RF and XGBoots.so SVM performed better than RF and XGBoost, respectively. Table 3 describes statistical values which expresses of the three machine learning models performance during validation stage of predicted dataset. During the validation phase, the performance of each algorithm varied slightly. Notably, SVM outperformed both RF and xGBoost in terms of MSE (0.008 dS/m vs. 0.814 dS/m and 0.847 dS/m, respectively), RMSE (0.089 dS/m vs. 0.902 dS/m and 0.921 dS/m, respectively), and MAE (0.071 dS/m vs. 0. 725 dS/m and 0.698 dS/m, respectively). Moreover, SVM also achieved a higher R^2 value (0.999) than either RF or XGBoost (0.941 and 0. 938, respectively). Taken together, these results indicate that SVM performed best overall in predicting soil salinity based on the input data.

3.3. Model fitting and validation

Fig. 4 illustrates scatter plots depicting the fitting results depending on observed and predicted values during the validation stage for SVM, RF, and XGBoost. These scatter plots were generated to visually represent the relationships between the observed and predicted values. It is evident from the scatter plots that SVM yielded the best model-fitting results (R^2 = 0.99, MSE=0.008, RMSE= 0.089, MAE=0.071, MAPE=0.012), with most points aligning closely along the diagonal line. However, one point exhibited a slight deviation, where the observed value was 5.25 dS/m while the predicted value was 4.51 dS/m. RF also demonstrated a good model-fitting results (R^2 = 0.94, MSE=0.814, RMSE= 0.902, MAE=0.725, MAPE=0.114), although some points deviated slightly from the diagonal line. Notably, the model's accuracy in predicting observed values below 8.55 dS/m was lower compared to those above this threshold. On the other hand, XGBoost displayed a lesser degree of alignment, with several points noticeably deviating from the diagonal line. Moreover, the model's accuracy in predicting observed values below 9.07 dS/m was inferior to those above this threshold. Consequently, SVM and RF appear to be more effective in predicting soil salinity levels compared to XGBoost. These findings offer visual support for the efficacy of the proposed predictive models, particularly the one based on the SVM method, given its consistently high R^2 values throughout the validation process.

4. Discussion

Soil salinization has emerged as a major global environmental issue, threatening agricultural productivity and food security [14,15]. Therefore, accurate prediction of soil salinity is critical for effective management strategies aimed at mitigating its impacts. This study demonstrated the potential utility of machine learning techniques for developing robust predictive models of soil salinity based on a range of physiochemical parameters. Our findings showed that all three algorithms evaluated in this study, namely xGBoost, SVM, and RF, yielded promising results, although SVM performed best overall.

4.1. Exploration of soil dataset characteristics

The statistical analysis of the soil dataset sheds light on the diverse range of soil parameters crucial for predicting soil salinity levels [16]. The statistical analysis (Table 1) confirmed the variability within the soil properties, with pH ranging from 6.9 to 8.3, and electrical conductivity (EC) ranging from 3.39 to 18.15 dS/m. This variation aligns with previous studies conducted in [17], where a wide range of soil salinity levels were reported in agricultural regions. The presence of this variability is crucial for the development of robust prediction models, as it allows the models to learn from a diverse range of data points [18] Correlation analyses reveal intricate relationships between soil parameters and EC, a key indicator of soil salinity [19]. The Pearson correlation analysis (Fig. 2) identified significant correlations ($p<0.05$) between EC and several soil properties, including Mg^{2+} (0.97), SO_4^{2-} (0.95), $CaCO_3$ (0.23), Cl^- (0.18), and Ca^{2+} (0.63). These findings are consistent with established knowledge, as these elements are known to contribute to soil salinity [20]. The strong positive correlations between EC and Mg^{2+} and SO_4^{2-} particularly highlight their influence on overall soil salinity levels. Understanding these relationships is essential for developing targeted strategies to manage soil salinity in agricultural fields.

4.2. Comparative assessment of machine learning algorithms

The comparative evaluation of machine learning algorithms, including Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Random Forest (RF), provides valuable insights into their performance in predicting soil salinity levels [21] .While SVM exhibits superior performance in terms of predictive accuracy metrics such as Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R-squared value, it is essential to consider the strengths and limitations of each algorithm .Previous studies have also reported SVM's effectiveness in various predictive

tasks, attributed to its ability to handle nonlinear relationships and high-dimensional data [22]. However, RF and XGBoost may offer advantages in certain scenarios, such as interpretability and scalability. These findings contribute to the ongoing discourse on the selection and optimization of machine learning algorithms for soil salinity prediction [23].

4.3. Assessing the predictive accuracy of machine learning models

The evaluation of the three machine learning models (XGBoost, SVM, and Random Forest) revealed that SVM outperformed the other models in terms of prediction accuracy (Table 2, 3). SVM achieved the lowest MSE (0.006 dS/m), RMSE (0.076 dS/m), and MAE (0.062 dS/m), and the highest R^2 (1.0) during training and the highest R^2 (0.99) during validation stages. These results suggest that SVM effectively captured the underlying patterns within the soil dataset and produced highly accurate predictions of soil salinity. The superiority of SVM compared to XGBoost and Random Forest might be attributed to its ability to handle complex non-linear relationships between the input soil properties and the target variable (EC) [24] .

4.4. Insights from model fitting and validation

The model fitting and validation stage provide further insights into the performance of machine learning algorithms in predicting soil salinity levels [21]. While SVM demonstrates superior model fitting results, RF also exhibits satisfactory alignment between observed and predicted values. XG Boost, although displaying less alignment, may offer computational advantages in certain contexts . Comparative analyses of model performance highlight the importance of considering multiple metrics and trade-offs when selecting a predictive model [25]. Additionally, rigorous validation procedures ensure the reliability and generalizability of predictive models across diverse datasets and environmental conditions. These insights contribute to advancing the understanding of machine learning applications in soil salinity prediction, thereby enhancing agricultural management, and promoting sustainable crop production in the Nile Delta .The scatter plots (Fig. 3) provided a visual representation of the model-fitting performance. SVM exhibited the best alignment between the observed and predicted EC values (R^2 = 0.99), indicating a strong agreement between the model's predictions and the actual soil salinity measurements. Conversely, XGBoost displayed a weaker alignment, with several data points deviating from the diagonal line, suggesting a less accurate prediction of soil salinity, particularly for values below 9.07 dS/m. These findings highlight the importance of model selection for achieving reliable soil salinity predictions. Overall, the results demonstrate the potential of machine learning, particularly SVM, for providing accurate and efficient soil salinity assessments.

4.5. Agreement and divergence in research findings

Our findings align with previous research indicating SVM's superiority in soil salinity prediction tasks [26]. However, although xGBoost initially showed promising outcomes during the initial stages of development, ultimately, SVM emerged as the most accurate approach for predicting soil salinity levels. This could potentially be attributed to SVM's ability to handle complex datasets efficiently while avoiding overfitting issues commonly encountered when dealing with large feature spaces [27]. By capitalizing on SVM remarkable capabilities and exploring complementary approaches, researchers can unlock novel ways to confront pressing environmental challenges linked to soil degradation and contamination . Overall, this study highlights the potential of machine learning algorithms in predicting soil salinity and offers a promising avenue for future research. By improving the accuracy and reliability of soil salinity predictions, farmers and decision makers can make informed decisions regarding sustainable crop production, soil fertility, and enhanced agricultural management. The comprehensive analysis of machine learning algorithms provides valuable insights into their efficacy and suitability for soil salinity prediction tasks. While SVM emerges as a top performer, the choice of algorithm should consider various factors, including dataset characteristics, computational resources, and modeling objectives.

5. Conclusions

This study demonstrated the power of machine learning techniques, particularly the Support Vector Machine (SVM) algorithm, in accurately predicting soil salinity levels in the agriculturally important Nile Delta region. SVM's outstanding performance, with a low MSE of 0.006 dS/m and RMSE of 0.076 dS/m during training, and an R^2 value of 1.0 during training and 0.99 during validation, affirms its ability to capture the complex relationships between soil parameters and salinity. The results recommend adopting SVM for predicting soil salinity in the region. The study's findings suggest that SVM is the most suitable model for predicting soil salinity in the Nile Delta region, offering accurate and reliable predictions crucial for informed decision-making in agricultural management. This study highlights SVM as the standout algorithm, while acknowledging the potential of alternative ML techniques like Random Forest and Extreme Gradient Boosting, both of which showed promise. Further exploration and optimization of these methods, tailored for specific scenarios or integrated with additional data sources, are warranted. Additionally, incorporating environmental and climatic factors could enhance predictive accuracy. The study sets the stage for future research in precision agriculture and sustainable crop production, leveraging ML, remote sensing, and real-time soil salinity monitoring. Future directions include combining algorithms, employing geospatial data and IoT monitoring, edge computing, developing decision support systems, exploring model transferability, and integrating economic and socio-cultural factors for sustainability.

References

[1] A. Singh. (2022). Soil salinity: A global threat to sustainable development. Soil Use and Management. 38. (1) 39-67.

[2] E.S. Mohamed, E.G. Morgun, S.M.G. Bothina. (2011). Assessment of soil salinity in the Eastern Nile Delta (Egypt) using geoinformation techniques. Moscow University Soil Science Bulletin. 66. (1) 11-14.

[3] J. Wang, J. Peng, H. Li, C. Yin, W. Liu, T. Wang, H. Zhang. (2021). Soil Salinity Mapping Using

Machine Learning Algorithms with the Sentinel-2 MSI in Arid Areas, China. Remote Sensing. 13. (2)

[4] L. Allison, L.A. Richards. (1954). Diagnosis and improvement of saline and alkali soils. Soil and Water Conservative Research Branch, Agricultural Research Service. pp. (166).

[5] A.L. Page. (1982). Methods of soil analysis. Part 2. Chemical and microbiological properties. American Society of Agronomy, Soil Science Society of America. pp. (1159).

[6] M.E. Fadl, M.E.M. Jalhoum, M.A.E. AbdelRahman, E.A. Ali, W.R. Zahra, A.S. Abuzaid, C. Fiorentino, P. D'Antonio, A.A. Belal, A. Scopa. (2023). Soil Salinity Assessing and Mapping Using Several Statistical and Distribution Techniques in Arid and Semi-Arid Ecosystems, Egypt. Agronomy. 13. (2) 583.

[7] P. An, W. Wang, X. Chen, Z. Zhuang, L. Cui. (2023). Machine learning brings new insights for reducing salinization disaster. Frontiers in Earth Science. 11.

[8] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2. 1. (4) 1-4.

[9] Z.A. Ali, Z.H. Abduljabbar, H.A. Taher, A.B. Sallow, S.M. Almufti. (2023). Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. Academic Journal of Nawroz University. 12. (2) 320-334.

[10] X. developers. (2022). XG Boost Documentation. https://xgboost.readthedocs.io/en/stable/.

[11] F. Tabsharani. (2023). support vector machine (SVM). https://www.techtarget.com/whatis/definition/support-vector-machine-SVM.

[12] A. El Bilali, A. Taleb, Y. Brouziyne. (2021). Groundwater quality forecasting using machine learning algorithms for irrigation purposes. Agricultural Water Management. 245. 106625.

[13] M. Despotovic, V. Nedic, D. Despotovic, S. Cvetanovic. (2015). Review and statistical analysis of different global solar radiation sunshine models. Renewable and Sustainable Energy Reviews. 52. 1869-1880.

[14] A. Allbed, L. Kumar, P. Sinha. (2014). Mapping and Modelling Spatial Variation in Soil Salinity in the Al Hassa Oasis Based on Remote Sensing Indicators and Regression Techniques. Remote Sensing. 6. (2) 1137-1157.

[15] M. Zaman, S.A. Shahid, L. Heng, S.A. Shahid, M. Zaman, L. Heng. (2018). Soil salinity: Historical perspectives and a world overview of the problem. Guideline for salinity assessment, mitigation and adaptation using nuclear and related techniques. 43-53.

[16] M. Bouaziz, M.Y. Chtourou, I. Triki, S. Mezner, S. Bouaziz. (2018). Prediction of soil salinity using multivariate statistical techniques and remote sensing tools. Advances in Remote Sensing. 7. (4) 313-326.

[17] K. Lal, R. Meena, S. Gupta, C. Saxena, G. Yadav, G. Singh. (2008). Diagnosis and Management of Poor Quality Water and Salt Affected Soils.

[18] S. Gu, S. Jiang, X. Li, N. Zheng, X. Xia. (2023). Soil salinity simulation based on electromagnetic induction and deep learning. Soil and Tillage Research. 230. 105706.

[19] U. Werban, K. Kuka, I. Merbach. (2009). Correlation of electrical resistivity, electrical conductivity and soil parameters at a long-term fertilization experiment. Near Surface Geophysics. 7. (1) 5-14.

[20] D.L. Corwin, K. Yemoto. (2019). Measurement of soil salinity: Electrical conductivity and total dissolved solids. Soil Science Society of America Journal. 83. (1) 1-2.

[21] H. Shi, O. Hellwich, G. Luo, C. Chen, H. He, F.U. Ochege, T. Van de Voorde, A. Kurban, P. De Maeyer. (2021). A global meta-analysis of soil salinity prediction integrating satellite remote sensing, soil sampling, and machine learning. IEEE Transactions on Geoscience and Remote Sensing. 60. 1-15.

[22] U. Thissen, R. Van Brakel, A. De Weijer, W. Melssen, L. Buydens. (2003). Using support vector machines for time series prediction. Chemometrics and intelligent laboratory systems. 69. (1-2) 35-49.

[23] C. Xiao, Q. Ji, J. Chen, F. Zhang, Y. Li, J. Fan, X. Hou, F. Yan, H. Wang. (2023). Prediction of soil salinity parameters using machine learning models in an arid region of northwest China. Computers and Electronics in Agriculture. 204. 107512.

[24] D. Andrade Foronda, G. Colinet. (2023). Prediction of Soil Salinity/Sodicity and Salt-Affected Soil Classes from Soluble Salt Ions Using Machine Learning Algorithms. Soil Systems. 7. (2) 47.

[25] C. Beverly, M. Bari, B. Christy, M. Hocking, K. Smettem. (2005). Predicted salinity impacts from land use change: comparison between rapid assessment approaches and a detailed modelling framework. Australian Journal of Experimental Agriculture. 45. (11) 1453-1469.

[26] X. Guan, S. Wang, Z. Gao, Y. Lv. (2013). Dynamic prediction of soil salinization in an irrigation district based on the support vector machine. Mathematical and Computer Modelling. 58. (3-4) 719-724.

[27] C.C. Chang, C.J. Lin. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2. (3) 1-27.