JEE Journal of Ecological Engineering

Journal of Ecological Engineering, 2025, 26(11), 242–257
https://doi.org/10.12911/22998993/208093
ISSN 2299–8993, License CC-BY 4.0

Received: 2025.06.20
Accepted: 2025.08.01
Published: 2025.08.15

Assessing water quality in the context of climate change in the Red River Delta using the hybrid machine learning

Huu Duy Nguyen[1]

[1] Faculty of Geography, VNU University of Science, Vietnam National University, Ha Noi, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Vietnam
E-mail: nguyenhuuduy@hus.edu.vn

ABSTRACT

Salinity intrusion is considered to be one of the important environmental degradations, causing negative effects on agricultural development, which is increasingly serious due to climate change and sea level rise. The Red River Delta, considered the rice granary of Vietnam, is among the regions most vulnerable to salinity intrusion. Although this region is important for the country's food security, however, less study has been conducted to predict salinity intrusion. The objective of this study is to evaluate the salinity intrusion based on hybrid machine-learning alogrithms, namely Xgboost (XGB), random forest (RF), LightGBM, Xgboost-Decision tree (XGB-DT), random forest-Decision tree (RF-DT), LightGBM-Decision tree (LightGBM-DT), Xgboost-Linear regression (XGB-LR), random forest-linear regression (RF-LR), LightGBM-Linear regression (LightGBM-LR) in the Red River Delta in Vietnam. Hourly water level, precipitation, and temperature dataset from 2014 to 2023 were used to predict salinity intrusion. The results showed that all proposed models were effective in predicting salinity intrusion in a Red River Delta, with the value of $R^2 > 0.8$. Among them, the Xgboost-DT model was more accurate than other models with an R^2 score of 0.86. The salinity at the Ba Lat station reached its highest peak of 23–24 g/l, and all proposed models captured the trend of salinity in the study area, in which the LightGBM model and its hybrid model provided the highest precision in terms of time and intensity of salinity. The outcomes of this study play an important role in supporting decision-makers or farmers in establishing effective measurements and optimizing water resource management to reduce the effects of salinity intrusion on agriculture.

Keywords: salinity intrusion, red river delta, machine learning.

INTRODUCTION

Coastal regions generally have high population density and agricultural yields due to their rich ecosystems (Khosravi et al., 2024, Raheli et al., 2024). However, these regions face important changes mainly related to climate change and sea level rise, as well as human activities such as deforestation and changes in land use (Kraiem et al., 2024, Yu et al., 2024). These changes can affect the stability of coastal areas by altering carbon cycles, soil degradation, and reducing biological diversity. Additionally, these changes also affect natural resources through salinity intrusion (Nguyen et al., 2019). According to FAO estimates, salinity intrusion affects about 20% of the world's agricultural land, and this figure is expected to increase in the future. Among the regions affected by salinity intrusion, the Red River Delta is considered one of the most affected by this problem, causing significant damage to agricultural activities and food security (Yuen et al., 2021). Although this region plays an important role, it is characterised by a high population density (1.450 people/km^2) and contains approximately 1,079,407 hectares of agricultural land (accounting for 51.2% of the region's area), with an annual rice yield of 6.2 million tons – representing about 15% of the national production. To reduce the effects of salinity intrusion, several structural and nonstructural measures have been explored and studied in recent years, including groundwater exploitation reduction, dike systems development, etc. (Tran et al., 2021). Which water quality prediction is considered one of the

most effective and recommended approaches to reduce the effects of salinity intrusion, especially in coastal zone.. However, the effectiveness of these measures depends on several factors, such as precipitation, geological conditions, and hydrological regimes (Gül et al., 2010, Zhang et al., 2017).

Historically, numerical simulation models in hydrology have been used to explore complex flow dynamics (Simons et al., 1996, Geng and Boufadel, 2015, Wang et al., 2017). Although these models have been proven to be effective, these traditional models require diverse data sets, which are rarely available in several regions of the world. Additionally, the hydrological model requires the expertise of the modeler, which makes it less usable to capture complex salinity dynamics. In recent years, remote sensing techniques have been used to observe water quality in regions around the world (Chong et al., 2014, Elhag, 2016). The available satellite images with various resolutions allow to observe of coastal and the estuary regions. The success of satellite imagery in monitoring salinity intrusion has been justified in different studies using Landsat (Mishra et al., 2023), and Sentinel (Sakai et al., 2021). However, due to the complex phenomenon of salinity intrusion, the remote sensing approach has been limited by spatial and temporal resolution. Moreover, with the development of the volume and quality of satellite images, it is necessary to integrate satellite image data into power models to explore useful information from these data.

Machine learning models provide a promising solution for improving modelling through their ability to efficiently interpret linear and non-linear problems and process large datasets (Nasir et al., 2022, Rajeev et al., 2025). Unlike traditional models that simulate through physical parameters such as terrain and hydrometeorology, machine learning methods rely on available data and the relationship between water quality and dependent factors. This reduces the reliance on difficult-to-collect manual field measurements. Once trained, machine learning models can predict salinity quickly and efficiently with large datasets, allowing machine learning models to predict more scenarios than traditional models (Wang et al., 2022, Tran et al., 2025). Furthermore, machine learning models offer high adaptability, as they can be continuously updated with new data, ensuring that predictions remain directly linked and responsive to environmental changes. They can also support high-throughput analytics, allowing decision-makers and local authorities to assess multiple scenarios related to environmental changes and land use planning (Ireland et al., 2015). However, individual machine learning models are often affected by the overfitting and underfitting problems, so many scientists have used ensemble and optimisation methods to increase the accuracy and avoid above problems (Nguyen et al., 2025).

By applying some well-known methods, water quality prediction models, especially water salinity based on advanced machine learning models, can overcome the limitations of traditional models, leading to more accurate predictions and better interpretation, and can help policymakers and farmers take timely measures in water resource management for agricultural development. These models promise to be able to learn synthetic features, have high interpretability, and account for the variability of salinity intrusion phenomena (Khullar and Singh, 2022, Zhu et al., 2022, Yan et al., 2024). This research focuses on developing highly accurate and applicable synthetic models that can be applied in the real world that help managers, policymakers, and farmers effectively manage irrigation water resources and minimise the impact of salinity intrusion on agriculture, especially in the context of climate change.

The novel contribution of this paper to existing studies lies in the integration of unique machine learning models to create highly accurate ensemble models to accurately predict saltwater intrusion. The combined approach proposed in this paper can exploit the strengths of each distinct algorithm while minimising errors in the forecasting process through the ability of continuous learning and flexible correction. This study is particularly applied in the Red River Delta, which is highly affected by climate change and sea level rise, resulting in an increasingly severe saltwater intrusion and has not been investigated in previous studies. The results of this study can help policymakers and populations optimise irrigation water management, minimise the impact of saline intrusion on agriculture, and contribute to ensuring food security.

STUDY AREA AND MATERIAL

Study area

The Red River Delta is located in the north of Vietnam, extending from latitude 21°34 N to alluvial plains around 19°5 N, from 105°17 E to

107°7 E, with a natural area of 21,253 km^2, which represents 6.42% of the area (Figure 1). The Red River Delta is a low-lying area, mainly a plain, with a dense river system with main river systems such as the Red River, Duong River, Luoc River, and Thai Binh River. The water regime of the Red River Delta depends on the flow from the source of the Red River and is distributed seasonally. The water regime of the Red River Delta is divided into different seasons. The flood season lasts from June to October, accounting for 70–80% of the total annual water volume, while the dry season lasts from November to May next year, with low flow. The Red River Delta has rich and diverse land resources. In particular, alluvial soil accounts for most of the area, with 70%, suitable for growing many types of crops, especially rice and other crops. In addition, acidic and saline sulphate soils are concentrated in coastal areas, which are suitable for the development of aquaculture.

The Ba Lat estuary area, located between Thai Binh and Nam Dinh provinces, is considered an area heavily affected by salinity intrusion, especially during the dry season, when the water level of the Red River Delta upstream is reduced. Specifically, according to monitoring the salinity at the Ba Lat station can reach 15.1 at the peak of the tide and about 7 km from the mouth of the river. Therefore, monitoring and forecasting saline intrusion play an important role in supporting local authorities and farmers in establishing effective measures to minimise the impact of saline intrusion.

In recent years, salinity in the Red River Delta has caused significant damage to agricultural development. According to a report by the Ministry of Agriculture and Environment, between 2015 and 2017, rice yields in salt-contaminated fields decreased by approximately 0.9 tonnes per hectare, from 6.49 tonnes per hectare to 5.58 tonnes per hectare, a decrease of approximately 14% due to the direct impact of salinity in field water. This has a direct impact on food security in the region.

Materials

Salinity intrusion is considered a serious problem, causing significant damage to agricultural activities and food security in the country, particularly during the dry season, when discharge decreases and tides increase to their maximum, as well as rain decreases. Therefore, in this study, these factors were used to observe and predict salinity intrusion in the Red River Delta.Field data include salinity measurements, tide levels, temperature

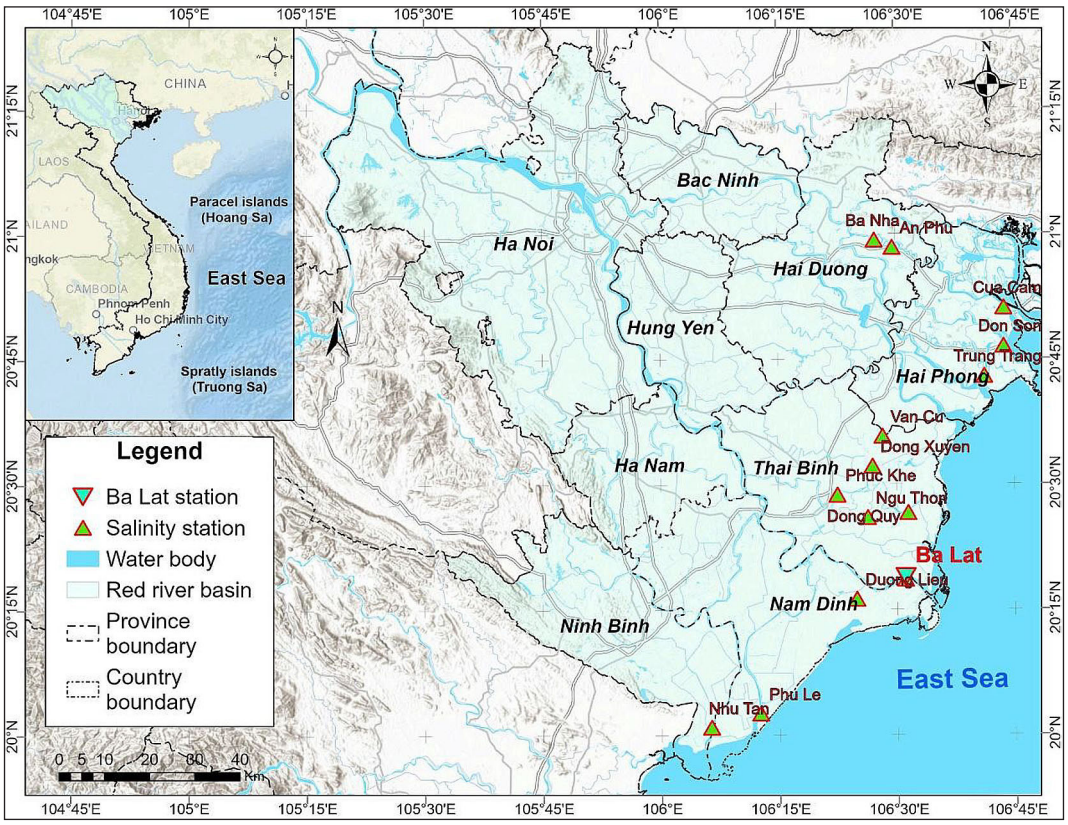

Figure 1. Location of the study area in this study

and hourly precipitation at the Ba Lat hydrological station from 2014 to 2023 (Figure 2). In this context, tides significantly affect the salinity of the Ba Lat River. When the tide rises, the seawater moves further inland, leading to an increase in salinity. Meanwhile, when the tide is low, freshwater from upstream pushes the seawater out to sea, reducing the water's salinity. During the dry season, river flow is significantly reduced, weakening the ability of rivers to push saltwater back to sea, allowing saltwater to penetrate further inland. Rainfall plays an important role in mitigating the effects of saltwater intrusion, as it provides additional freshwater to the river system, increases flow, and helps push saltwater back out to sea. Temperature is also another factor that greatly affects the level of salinity intrusion. High temperatures increase the rate of evaporation, especially during the dry season, causing a decrease in river flow. This increases saltwater intrusion. For all the models proposed in this study (XGBoost, Random Forest, LightGBM and their hybrids with decision tree or Linear Regression), this study uses the salinity value measured one day before (lag = 1) to predict one day ahead because several studies have pointed out that the salinity series depends from day to day; therefore, the best information to predict tomorrow's salinity is often today's information. In addition, model optimisation is mainly based on the one-day change, which allows us to reconcile the complexity and performance of the model. This avoids the dependency of parameter adjustment while preserving sequential variability. Therefore, this one-day delay allows all models to effectively exploit hydrological memory and temporal correlation to provide the salinity prediction on the following day.

METHODOLOGY

The objective of this study is to build machine learning models, namely XGB, RF, LightGBM, XGB-DT, RF-DT, LightGBM-DT, XGB-LR, RF-LR, LightGBM-LR, to monitor and predict salinity intrusion in the Ba Lat estuary in the Red River Delta of Vietnam. Therefore, this process was divided into four main steps: (i) the data collection and processing process; (ii) the machine learning model construction process; (iii) assessment of the accuracy of the proposed model; (iv) the analysis of salinity intrusion in the study area (Figure 3).

1. Data collection and processing – in this study, we used river discharge, water level, precipitation, and temperature from 2014 to 2023 to predict the EC value at the Ba Lat station. The data set was divided into two parts: the first part was used to build the prediction model with 80% of the data, while the second part was used to validate the proposed model with 20% of the data. The division of the data rate depends on the available data and the nature of the relationship between EC values and their influencing factors. This study tested several different ratios; however, ultimately, this ratio presented the best accuracy.
2. The machine learning model building process – was carried out in two main stages. The first is the base model and the hybrid model building. The second is the development of L2 regularisation. That is, the prediction of the base model and the hybrid model was used as input to the L2 regularisation technique The performance of the base model depends on the tuning of the parameters. In this study, the parameters were optimised using the trial-and-error method.

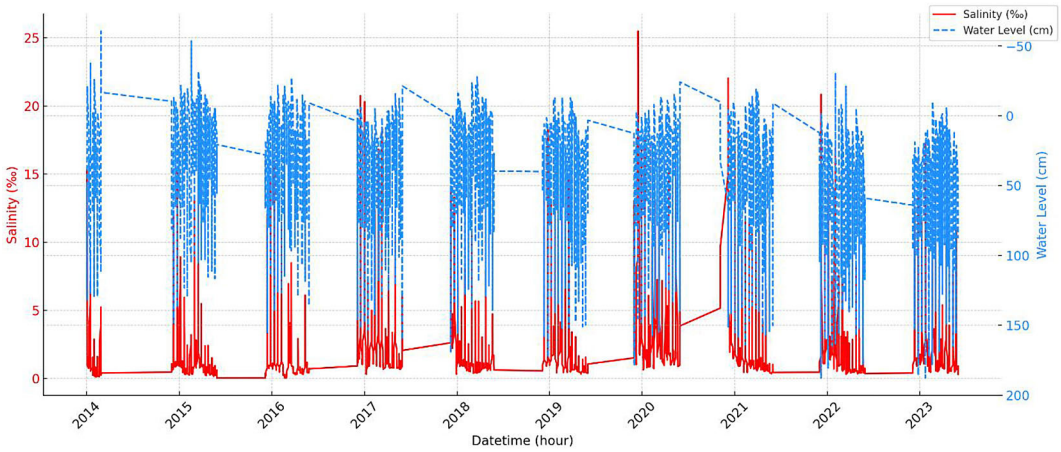

Figure 2. Dataset for the prediction of water salinity at the Ba Lat station in the Red River Delta

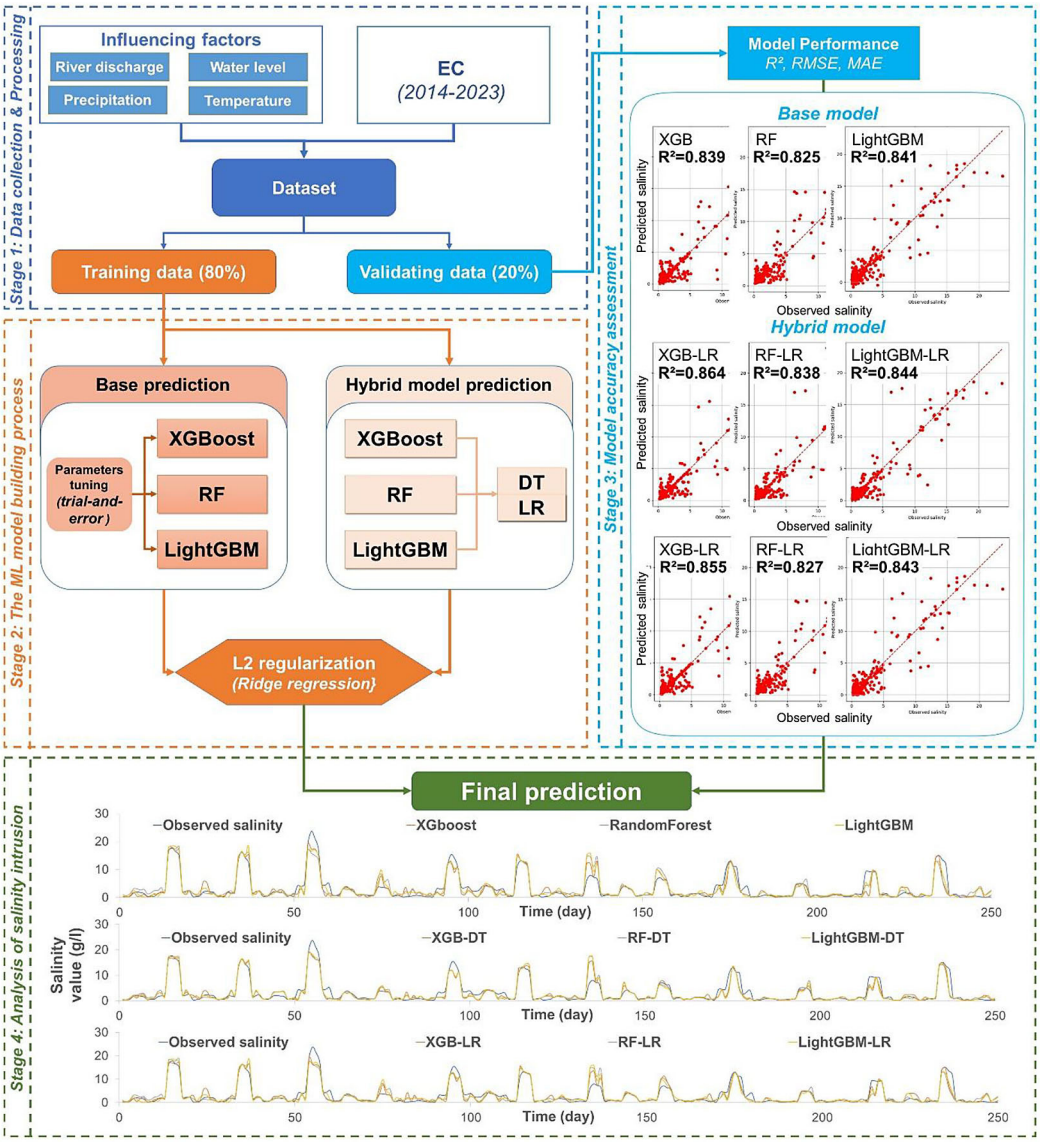

Figure 3. Methodology used for the assessment of water quality in this study

This method is chosen due to its flexibility and ease of implementation, especially in the early stages of model building and when the search space of parameters is still small. Compared to other methods such as Grid Search or Bayesian Optimisation, the trial and error method is more time-saving because it does not have to evaluate all possible parameter combinations, and this method also does not require the setup of a complex probability model and is more controllable when working with a limited number of input variables (Bianchi and Monbaliu 2024, Tran, Nguyen et al. 2025). In the end, XGBoost is configured with objective='reg:squarederror', n_estimators=200, learning_rate=0.2, max_depth=7, subsample=0.8, and colsample_bytree=0.8. Meanwhile, RandomForest is optimised with n_estimators=200, random_state=42, and max_depth=15. The LightGBM model, with the parameter objective='regression', also uses n_estimators=200, learning_rate=0.2 and max_depth=15, subsample=0.8, colsample_bytree=0.8. However, the hybrid model was developed by integrating between the base models and the algorithms (DT and LR) allows one to optimise the prediction capacity of the base model. The parameters of DT and LR are as follows: max_depth=15, min_samples_split=2, min_samples_leaf=1, splitter='best' for DT and fit_intercept: True, copy_X: True, n_jobs: None, positive: False for LR. Finally, to improve the prediction ability of the proposed models, this study used the Ridge Regression technique. Ridge regression is an L2 regularisation

technique that reduces the overfitting problem and improves the performance of the model. Ridge regression works by adding an L2 penalty to the model's loss function. This balances learning from the training data while maintaining model simplicity. The final result of Ridge Regression presents the best performance.

3. Assessment of the accuracy of the proposed model – three statistical indices, namely RMSE, MAE, and R^2, were used to evaluate the performance of the proposed models.
4. The analysis of salinity intrusion in the study area – after evaluating the proposed machine learning models, these models are used to forecast the salinity of the water 1 day in advance. The analysis focuses on determining the pattern of variation in salinity over years and seasons, based on the relationship between salinity and water level, precipitation, and temperature.

Xgboost

Xgboost is considered a widely popular machine learning model with high speed, superior accuracy, and enhanced flexibility. This model is an improved version of the Gradient Boost algorithm and was initially released in 2014 by Chen et al., 2022. XGBoost is an assembly of decision trees (weak learners) that allow predicting residuals and correcting errors from previous decision trees. This algorithm was designed to work in parallel and have the capability to process a large volume of structured datasets in a short period. XGBoost was integrated with L1 and L2 regularisation mechanisms that allow avoidance of overfitting problems in the model (Niazkar et al., 2024). In addition, this algorithm can repeatedly improve observations by combining several weak learners sequentially. It allows XGBoost to reduce the high biases that can sometimes be recurrent in machine learning models (Liu et al., 2024). During the execution process, XBoost could handle missing data efficiently by automatically learning the best imputation direction. The performance of XGBoost was proven by defeating other machine learning models in various computer science competitions (Mantena et al., 2023).

Random forest

Random forest (RF) is a machine learning algorithm proposed by Leo Breiman in 2001 that can solve both classification and regression tasks (Breiman, 2001). The idea of this algorithm is to combine multiple decision trees to make the predictions more accurate. RF constructs several independent decision trees, does separate training on each tree using a random sample dataset, and compares between outputs to select the best one (Hidayat and Astsauri, 2022, Khan et al., 2022). The RF working process was divided into three main steps. First, a replacement random sample is drawn from the data. Second, a random feature selection process was applied for each tree using the datasets from step one. During the construction of each tree, some random features are used to make decisions at each node. The last is the prediction process (Liu et al., 2013, Wang et al., 2021). Almost all parameters of RF are the same as the parameters of the decision tree algorithm, including max_depth, min_samples_split, and min_samples_leaf. However, it has two new parameters, including n_estimators and bootstrap. Which n_estimators are the number of trees in the forest, and bootstrap was used to increase diversity among trees, thereby improving the generalisability of the model and reducing overfitting (Suleymanov, Gabbasova et al. 2023). For classification, the prediction model was chosen from the individual trees. For regression, the model predicts the expected values of each tree on average.

LightGBM

LightGBM is a machine learning library that uses gradient boosting in decision trees (Ke et al., 2017). As a 'light' term, LightGBM was designed to solve problems on a large scale with high speed and low memory. LightGBM uses a technique called histogram-based to classify data into containers based on the value of the feature to reduce computational complexity and speed up training (Wang et al., 2020). The LightGBM establishes trees using the top-down method, limiting depth to control complexity. In addition, the leaf-wise mechanism was also applied to decision trees in depth by choosing to expand the leaf with the largest error reduction at each step (Dong et al., 2022). Thus, the decision trees are combined so that each new tree learns by adjusting for the differences between the current model's predictions and the actual values, helping to improve the model's overall performance. LightGBM used two techniques to process sample data, including gradient-based one-sided sampling (GOSS) and exclusive feature bundle (EFB). In

this case, GOSS allows samples with large gradients to be retained to focus on data instances with smaller gradients (Ke et al., 2017). On the other hand, EFB allows merging of less interactive features into a new feature (Ke et al., 2017, Shaker et al., 2021). From this, XGBoost could reduce the number of features to process without losing much information, reduce the amount of data to be processed, and increase efficiency. LightGBM could work well in the environment of parallel and distributed computing in big systems.

Decision tree

A decision tree is a simple supervised learning algorithm that was developed in the 1960s with various versions such as ID3, C4.5, or CART (Rokach and Maimon, 2005). The main idea of the Decision Tree algorithm is to build a tree of nodes, where each node decides to split the data based on a certain feature, in order to optimise discrimination or prediction (Kotsiantis, 2013). The structure of a decision tree includes a root node, branches, internal nodes, and leaf nodes. Based on available features, types of nodes perform evaluations to form homogeneous subsets, which are referred to as leaf nodes or terminal nodes. The leaf nodes represent all possible outcomes in the dataset. The algorithm selects the best feature and threshold to split the data in a way that increases the purity or homogeneity of the resulting subsets in relation to the target variable at each node. This splitting process is then repeated top-down and recursively until a stopping criterion is met, such as reaching a maximum depth, having a minimum number of samples in a node, or obtaining pure leaf nodes (Elnaggar and Noller, 2009, Efeoglu and Tuna, 2022). To make a prediction for a new data point, the tree is traversed from the root to a leaf node based on the results of the feature tests, and the prediction is the majority class or average value in that leaf. This algorithm has four main parameters that help to control the model, including max_depth, min_samples_split, and min_samples_leaf. In this case, max_depth (or maximum depth of the tree) helps to control complexity, and avoid overfitting. Min_samples_split (or a minimum number of samples for a node to split) helps to reduce overfitting by increasing this value. Min_samples_leaf (or a minimum number of samples in leaves) helps to smooth the model. The criterion parameter (or criteria for choosing the type of splits Gini and Entropy) affects how the tree splits. Decision trees can also maintain their accuracy by forming an ensemble using a random forest algorithm.

Linear regression

Linear regression (LR) is a continuous value prediction model that finds the linear relationship between the value of unknown data using another related, known data value. This model is the result of the study of Francis Galton in the late nineteenth century (Su et al., 2012). Basically, the core idea of LR is to plot a best-fit straight line between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. The dependent variable, y, is plotted on the vertical axis. Ultimately, predictions can prove accurate for calculating unknown dependent variables from known independent variables. This algorithm uses the ordinary least squares (OLS) method to estimate parameters (Uyanık and Güler, 2013). Unlike other machine learning algorithms, LR is simple to deploy, easy to understand, easy to implement, and very effective when the relationship between variables is linear. That is why it is one of the most popular algorithms that is selected to solve regression problems at the beginning.

RESULTS

Model performance assessment

Figure 4 presents the R^2 value of the models proposed in this study. In general, all proposed models performed well in predicting the EC one day ahead, such as the Ba Lat station in the Red River Delta. For detail, the Xgboost-DT model was more accurate than other models with an R^2 score of 0.86, followed by the Xgboost-LR model with an R^2 score of 0.85, the LightGBM-DT model with an R^2 score of 0.84, the LightGBM-LR model with an R^2 score of 0.84, the LightGBM model with an R^2 score of 0.84, the Xgboost model with an R^2 score of 0.838, the RF-DT model with an R^2 score of 0.837, the RF-LR model with an R^2 score of 0.827 and RF model with an R^2 score of 0.825, respectively.

Table 1 presents the RMSE and MAE values of the proposed models. In general, hybrid models demonstrate higher accuracy than the individual model. More accurately, for the XGB model and its hybrid, with RMSE and MAE scores of

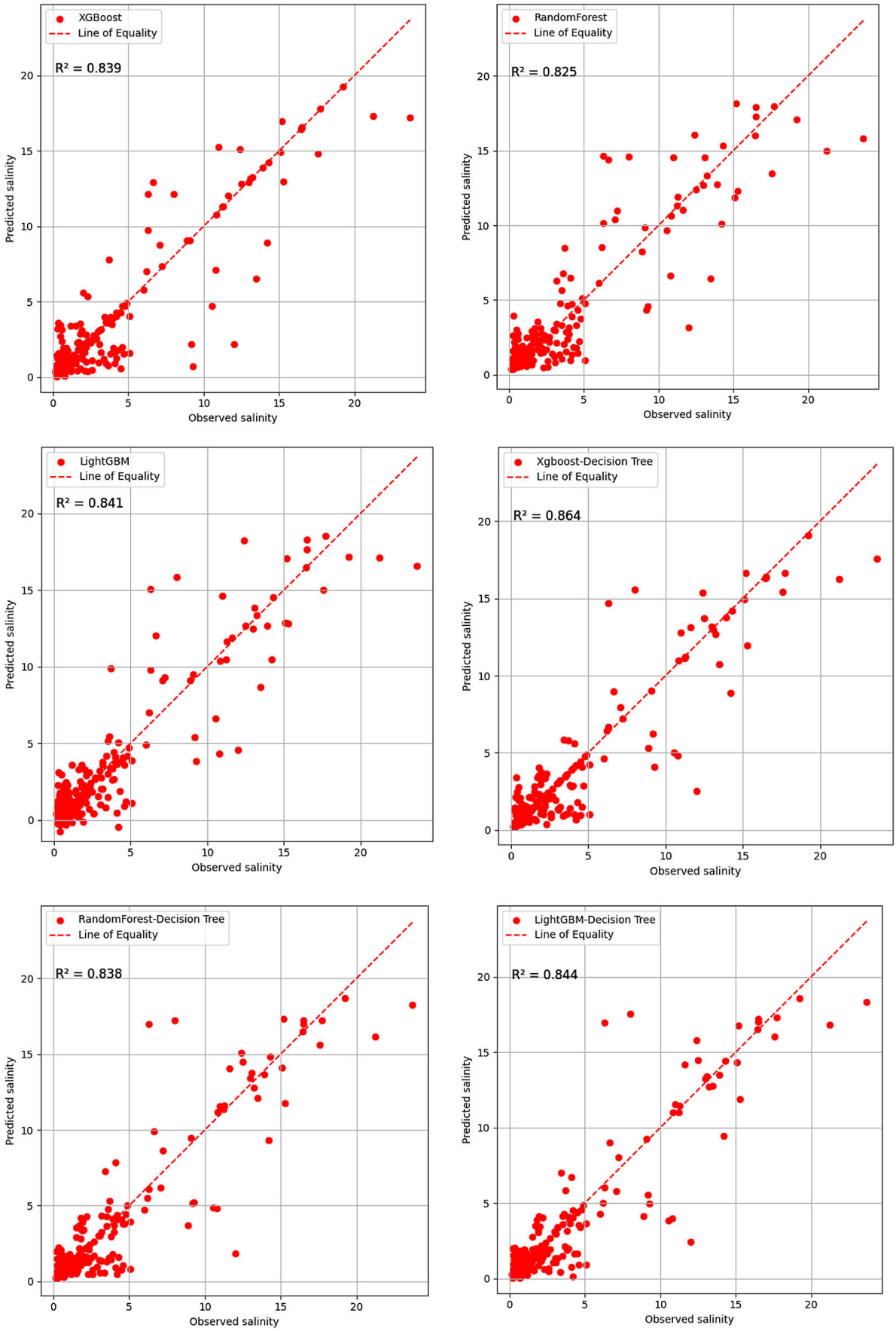

Figure 4. R^2 value for the model proposed in this study

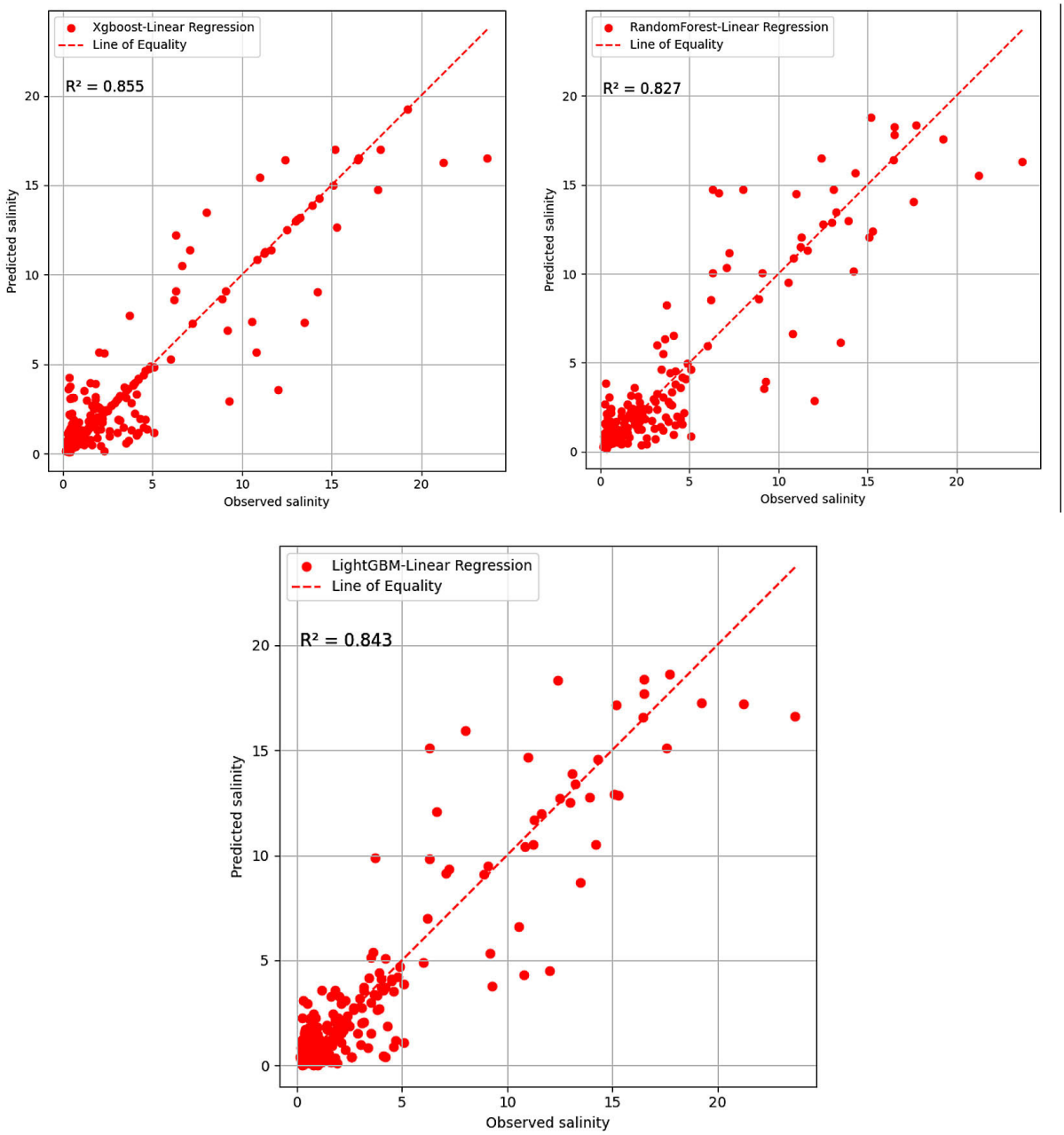

Figure 4. Cont. R^2 value for the model proposed in this study

Table 1. The value RMSE, MAE of the models proposed

Model	RMSE	MAE	R^2
B	1.8	0.93	0.83
RF	1.87	1.1	0.82
LightGBM	1.78	1.05	0.84
XGB-DT	1.65	0.87	0.86
RF-DT	1.8	0.98	0.83
LightGBM-DT	1.77	0.93	0.84
XGB-LR	1.71	0.89	0.85
RF-LR	1.86	1.05	0.82
LightGBM-LR	1.77	1.03	0.84

1.65 and 0.87, the performance of the XGB-DT model was higher than that of the XGB model (RMSE=1.8 and MAE=0.93) and XGB-LR (RMSE 1.71 and MAE 0.89). For the RF model and its hybrid, the RF-DT model performed better with an RMSE score of 1.8, MAE of 0.98, followed by the RF-LR with an RMSE score of 1.86 and an MAE of 1.05, respectively. For the LightGBM model and its hybrid, the LightGBM-DT model exhibited a better prediction performance compared to the other models with an RMSE score of 1.7 and an MAE of 0.93, followed by

LightGBM-LR (RMSE=1.77, MAE=1.03) and LightGBM (RMSE=1.78 and MAE=1.05).

Compared to the proposed models, the XGB-DT model was more accurate with RMSE and MAE scores of 1.65 and 1.05. The XGB-LR model was second, with RMSE (1.71) and MAE (0.89) slightly higher than the XGB-DT model. The LightGBM-LR model was third with RMSE and MAE scores of 1.77 and 1.03, respectively. The LightGBM-DT model was ranked fourth with an RMSE score of 1.77 and an MAE of 0.93. The LightGBM model was ranked fifth with an RMSE score of 1.78 and an MAE of 1.05. The RF-DT model was ranked sixth with an RMSE score of 1.8 and an MAE of 0.98. The Xgboost model was ranked seventh with an RMSE score of 1.8 and an MAE of 0.93. The RF-LR model was ranked eighth with an RMSE score of 1.86 and an MAE of 1.05. The Random Forest model was ranked lower with an RMSE score of 1.87 and an MAE of 1.1.

Salinity intrusion analysis

Figure 5 presents the EC value at the Ba Lat station by the XGB, RF, and LightGBM models

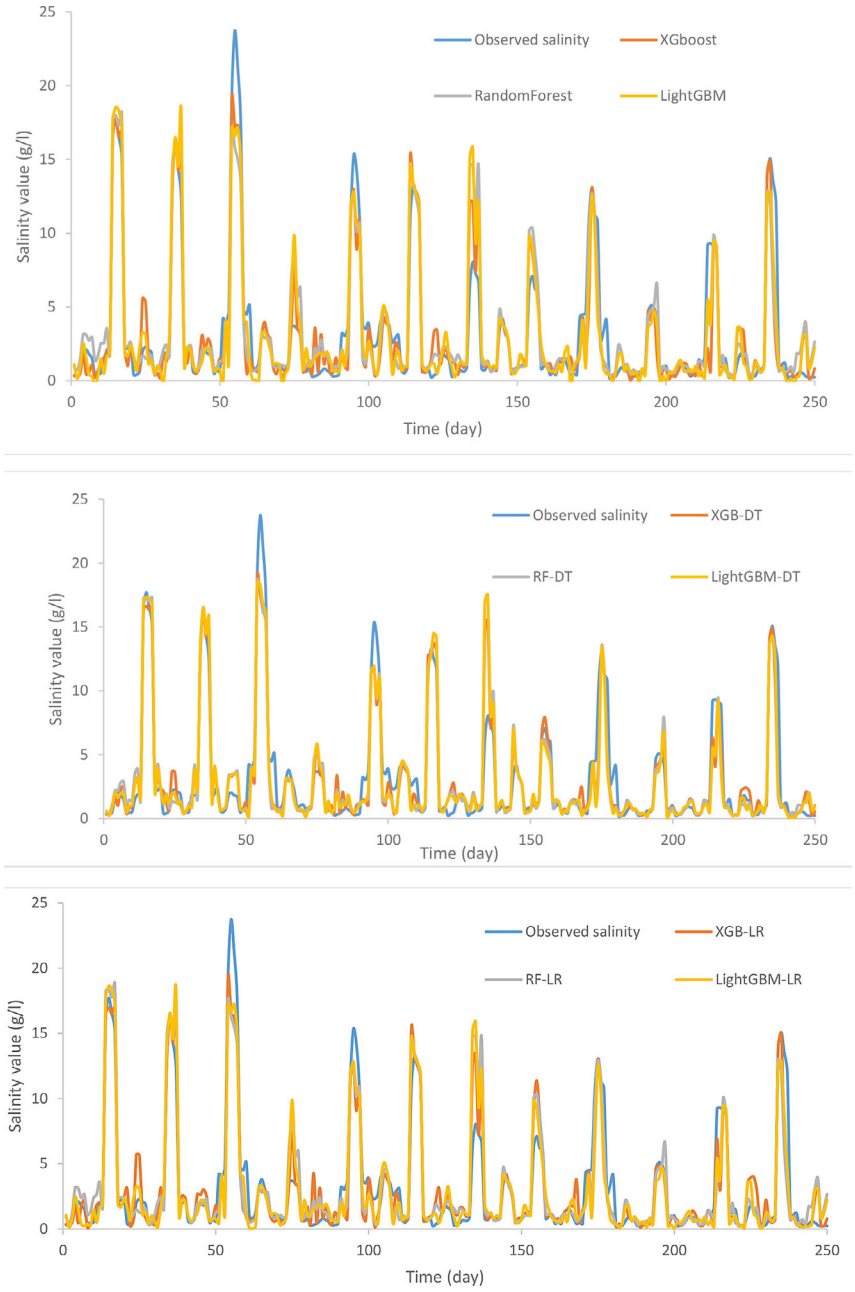

Figure 5. Salinity prediction using the models proposed

and their hybrids with the Decision tree and LR. Specifically, during high tide days, the salinity (EC) at Ba Lat station reaches a very high peak, up to approximately 23–24 g/L. All models correctly captured the general trend of salinity fluctuations, but there were clear differences in amplitude and timing, especially the XGB model, which accurately captured the locations of salinity peaks but often predicted them 1–2 days earlier and overestimated the peak amplitude by 0.5–1 g/L. On the contrary, the RF model tended to respond approximately 1–2 days later than observed and reduced the amplitude of salinity peaks compared to the actual values, usually 1–2 g/L less. LightGBM yielded the most accurate results, with the prediction curve matching the timing and amplitude of salinity peaks almost perfectly, including the largest peak around 23–24 g/L.

For the group of hybrid models combined with the decision tree (DT), XGB-DT maintained the correct timing of the salinity peaks but still underestimated the amplitude by approximately 0.5–1 g/L compared to observations; RF-DT again correctly captured large peaks but exaggerated small peaks (< 5 g/L). Meanwhile, LightGBM-DT was the best choice, as it matched both the timing and the amplitude of the high and low peaks with an error of less than 0.3 g/L. Moving to the hybrid group with linear regression (LR), XGB-LR, RF-LR, and LightGBM-LR accurately reproduce the general trend, but all exhibit a lag of about 1 to 3 days compared to the observed value, which is particularly evident in RF-LR and XGB-LR. In this case, LightGBM-LR, although still closely following the general fluctuations, underestimates the salinity peaks by more than 1 g/L, thus having difficulty catching up with the actual amplitude.

DISCUSSION

Salinity intrusion is considered a major environmental problem, negatively influencing agricultural development and affecting food security in the country. This problem is increasingly serious in the context of climate change and rising sea levels (Vineis et al., 2011, Katende and Sagala, 2019, Musie and Gonfa, 2023). Therefore, monitoring and predicting salinity intrusion is an essential task, helping policymakers and farmers propose appropriate measures to reduce the effects of salinity intrusion. The objective of this study is to develop a machine learning model to predict the intrusion of salinity of the Ba Lat station in the Red River Delta of Vietnam.

The Red River Delta is considered the second most productive agricultural region in Vietnam, providing approximately 18% of Vietnam's rice production, 26% of its vegetable production, and 20% of its seafood production. With its vast area, the Red River Delta not only ensures food security for the entire country but also provides a livelihood for millions of people, who rely mainly on agricultural and livestock activities (Yuen et al., 2021, Phung and Dao, 2024). However, this area is currently considered one of the most affected by saltwater intrusion, especially in the context of climate change and sea level rise. The decline in the dry season flow of the Red River due to the impact of the construction of the upstream dam, combined with the rise in sea level, has made saline intrusion increasingly serious in coastal estuaries such as Ba Lat (Nguyen et al., 2017, Hien et al., 2023). According to previous studies, salinity in the Ba Lat estuary has increased beyond the threshold of 4 during peak tide hours. This threshold exceeds the tolerance of rice plants (Phan and Kamoshita, 2020, Do et al., 2024). Previous studies have also reported that in some coastal localities in Nam Dinh and Thai Binh provinces, approximately 15–20% of rice yield in winter and spring crops was affected by saline intrusion (Nguyen et al., 2017). In the context of increasing saline intrusion in this region, the development of a warning system is essential, which can help planners and managers make decisions to minimise the impact of saline intrusion on agricultural production and water resource management. However, currently, studies on saltwater intrusion prediction in the Red River Delta are still limited, focusing mainly on the use of traditional models, while few studies use machine learning models. This approach is considered promising and has the potential to solve non-linear problems, especially in the current context of climate change and sea level rise.

In this study, all proposed models have high accuracy ($R^2 > 0.8$) in predicting saline intrusion in the Ba Lat Estuary, Red River Delta. It can be seen that hybrid models have higher accuracy than simple models because each model has its own strengths and weaknesses. For example, the Xgboost model is able to solve nonlinear problems and problems related to overfitting (Kiriakidou et al., 2024), while the DT model is composed

of weak learners to become stronger learners, but the DT model also often encounters the problem of overfitting when the data has large fluctuations and noise (Huang, 2024, Parhi and Patro, 2024, Saputra et al., 2024). Thus, when these models are combined with each other, they can complement each other and limit the weaknesses of each model. This allows the hybrid model to have higher accuracy and avoid overfitting problems. In addition, hybrid models have the ability to reduce bias and variance, which allows hybrid models to perform better than simple models. Hybrid models also have higher generalisation ability than simple models. Specifically, linear regression models often exhibit high bias, especially in complex data sets and nonlinear relationships (Azevedo et al., 2024, Fan et al., 2024). Therefore, combining the Xgboost, Random Forest, or LightGBM models with LR models can help reduce bias and variance, and hybrid models have greater stability than simple models, especially in cases where the data contain noise and are not continuous, such as in saltwater intrusion prediction (Priyadharshini and Karpagam, 2024, Wu et al., 2024). In the case of random forest models, although they have good classification and prediction capabilities, they often encounter overfitting problems, especially in cases of very complex data such as salinity intrusion data (Khan et al., 2024, Salman et al., 2024). Therefore, by combining the random forest model with LR or DT models, the hybrid models can reduce the overfitting problem and the random forest models can also take advantage of the generalisability of LR and DT models. Finally, one of the notable advantages of hybrid models is their ability to globalise and generalise the prediction of problems based on linear and non-linear relationships between independent and dependent variables. This allows the hybrid model to produce more accurate forecasts and to more accurately reflect the complexity of the data.

The use of machine learning models to predict saltwater intrusion is becoming increasingly popular in countries around the world, particularly as these phenomena become increasingly severe due to the impact of climate change and human activities (Nguyen et al., 2025, Yu et al., 2025). Machine learning models are capable of handling complex and nonlinear problems, which is particularly important in the case of salinity intrusion prediction. Saltwater intrusion is influenced by many different factors, such as tides, currents, precipitation, and human activities such as dam construction and groundwater exploitation (Barzegar and Moghaddam, 2016, Mahmoud et al., 2025). Therefore, machine learning models can integrate climate change scenarios into the prediction of future saltwater intrusion (Zennaro et al., 2021). However, even though the proposed model can successfully predict salinity intrusion in the study area, this result may be related to the random selection of samples in the study area and may not be accurate in other areas. For example, the Red River Delta is less affected by upstream dam systems than other regions, such as the Mekong Delta. In these areas, salinity intrusion data are often closely correlated with inflow and tides. For areas heavily impacted by upstream dam systems, the effectiveness of machine learning models in accurately predicting salinity intrusion may be questionable, and more specific and detailed information is required to train the models. This is because the performance of a machine learning model is largely dependent on the quality and quantity of data on which it is trained. Therefore, data collection is of great importance in building machine learning models, especially information related to human impacts.

In recent years, local authorities and the people of the Red River Delta have implemented many strategies to minimize the impact of saline intrusion on agriculture. These strategies include strengthening the dyke system and saline barriers, adjusting the seasonal farming schedule to avoid periods of high salinity, and searching for plant varieties that are more resistant to saline intrusion. Additionally, local authorities have also actively deployed real-time monitoring and warning technology for saline intrusion, helping people adjust their pepper planting schedules appropriately (Nguyen et al., 2017, Nguyen et al., 2019). The Ministry of Agriculture and Environment has also issued many programmes and techniques to guide farmers in adapting to saline intrusion and integrating these contents into national programmes such as the "National Target Programme on Climate Change Response" and the "New Rural Development Programme". Therefore, integrating advanced tools such as machine learning into water resource forecasting and planning plays an important role in enhancing resilience to salinity intrusion, especially in the context of climate change.

Although this study successfully predicted salinity intrusion, it also had data limitations. Currently, saline intrusion in the Red River Delta is affected by the upstream dam system. However,

due to the limited funding and data sharing policies of the relevant regions, this study is limited in integrating activity-related data into the saline intrusion prediction model. Furthermore, saline intrusion is significantly affected by climate change and the rise in sea level. Therefore, future research will attempt to integrate the impacts of dam systems, climate change and sea level rise scenarios into the assessment and prediction of saline intrusion. The results of this study play an important role in supporting policy-makers or farmers in establishing effective measures to reduce the effects of salinity intrusion.

CONCLUSIONS

Salinity intrusion is considered significant environmental degradation, influencing agricultural development and food security in the country, particularly in the context of climate change and rising sea levels. The Red River Delta is considered one of the regions most affected by the salinity intrusion problem, a problem that is increasingly serious under the effects of climate change and rising sea levels. Therefore, assessing salinity intrusion is one of the essential tasks to support decision makers or farmers in optimising water resource management to reduce negative effects on agricultural development. This study presents the following results:

1) This study demonstrates the capacity of machine learning and hybrid machine learning in the assessment of salinity intrusion. This is one of the tools that helps local authorities and farmers quickly assess and predict salinity intrusion.
2) All the models proposed in this study performed well in predicting salinity intrusion in the Red River Delta with the value of R^2 plus 0.8. Among them, the Xgboost-DT model was more accurate than the other models, with an R^2 score of 0.86. This model can be applicable to other estuarine or coastal regions with similar hydrological and climatic conditions.

Moreover, this study contributes to the advancement of technology to predict water quality in general, specifically, its salinity. The developed models in this study can be deployed to predict water salinity in different regions of the world. The results of this study can help decision-makers or farmers build the necessary activities to reduce the effects of water salinity on agricultural development.

REFERENCES

1. Azevedo, B. F., A. M. A. Rocha and A. I. Pereira (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning 113*(7): 4055–4097.
2. Barzegar, R. and A. Asghari Moghaddam (2016). Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. *Modeling Earth Systems and Environment 2*: 1–13.
3. Bianchi, P. and J. C. M. Monbaliu (2024). Revisiting the paradigm of reaction optimization in flow with a priori computational reaction intelligence. *Angewandte Chemie International Edition 63*(5): e202311526.
4. Breiman, L. (2001). Random forests. *Machine learning 45*: 5–32.
5. Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano and T. Zhou (2022). xgboost: Extreme gradient boosting. R package version 1.6. 0.1.
6. Chong, Y. J., A. Khan, P. Scheelbeek, A. Butler, D. Bowers and P. Vineis (2014). Climate change and salinity in drinking water as a global problem: using remote-sensing methods to monitor surface water salinity. *International journal of remote sensing 35*(4): 1585–1599.
7. Do, A. N. T., T. A. T. Do and H. D. Tran (2024). Distribution of fish larvae and juveniles on salinity in an estuary predicted from remote sensing and fuzzy logic approach. *Aquatic Ecology 58*(3): 983–998.
8. Dong, L., J. Qi, B. Yin, H. Zhi, D. Li, S. Yang, W. Wang, H. Cai and B. Xie (2022). Reconstruction of subsurface salinity structure in the south China Sea using satellite observations: A LightGBM-based deep forest method. *Remote Sensing 14*(14): 3494.
9. Efeoglu, E. and G. Tuna (2022). Determination of salt concentration in water using decision trees and electromagnetic waves. *Journal of Water and Health 20*(5): 803–815.
10. Elhag, M. (2016). Evaluation of different soil salinity mapping using remote sensing techniques in arid ecosystems, Saudi Arabia. *Journal of Sensors 2016*(1): 7596175.
11. Elnaggar, A. A. and J. S. Noller (2009). Application of remote-sensing data and decision-tree analysis to mapping salt-affected soils over large areas. *Remote Sensing 2*(1): 151–165.
12. Fan, G.-F., Y.-Y. Han, J.-W. Li, L.-L. Peng, Y.-H. Yeh and W.-C. Hong (2024). A hybrid model for deep learning short-term power load forecasting based on feature extraction statistics techniques. *Expert Systems with Applications 238*: 122012.
13. Geng, X. and M. C. Boufadel (2015). Numerical

modeling of water flow and salt transport in bare saline soil subjected to evaporation. *Journal of Hydrology 524*: 427–438.

14. Gül, G. O., N. Harmancıoğlu and A. Gül (2010). A combined hydrologic and hydraulic modeling approach for testing efficiency of structural flood control measures. *Natural hazards 54*: 245–260.
15. Hidayat, F. and T. M. S. Astsauri (2022). Applied random forest for parameter sensitivity of low salinity water Injection (LSWI) implementation on carbonate reservoir. *Alexandria Engineering Journal 61*(3): 2408–2417.
16. Hien, N. T., N. H. Yen, M. Balistrocchi, M. Peli, V. M. Cat and R. Ranzi (2023). Salinity dynamics under different water management plans coupled with sea level rise scenarios in the Red River Delta, Vietnam. *Journal of Hydro-environment Research 51*: 67–81.
17. Huang, X. (2024). Predictive models: regression, decision trees, and clustering. *Applied and Computational Engineering 79*: 124–133.
18. Ireland, G., M. Volpi and G. P. Petropoulos (2015). Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: a case study from a Mediterranean flood. *Remote sensing 7*(3): 3372–3399.
19. Katende, A. and F. Sagala (2019). A critical review of low salinity water flooding: Mechanism, laboratory and field application. *Journal of Molecular Liquids 278*: 627–649.
20. Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems 30*.
21. Khan, M. A., M. I. Shah, M. F. Javed, M. I. Khan, S. Rasheed, M. El-Shorbagy, E. R. El-Zahar and M. Malik (2022). Application of random forest for modelling of surface water salinity. *Ain Shams Engineering Journal 13*(4): 101635.
22. Khan, T. A., R. Sadiq, Z. Shahid, M. M. Alam and M. B. M. Su'ud (2024). Sentiment analysis using support vector machine and random forest. *Journal of Informatics and Web Engineering 3*(1): 67–75.
23. Khosravi, R., M. Simjoo and M. Chahardowli (2024). Low salinity water flooding: estimating relative permeability and capillary pressure using coupling of particle swarm optimization and machine learning technique. *Scientific Reports 14*(1): 13213.
24. Khullar, S. and N. Singh (2022). Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. *Environmental Science and Pollution Research 29*(9): 12875–12889.
25. Kiriakidou, N., I. E. Livieris and C. Diou (2024). *C-XGBoost: A tree boosting model for causal effect estimation*. IFIP international conference on artificial intelligence applications and innovations, Springer.
26. Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review 39*: 261–283.
27. Kraiem, Z., K. Zouari and N. Chkir (2024). Accurate prediction of salinity in Chott Djerid shallow aquifers, southern Tunisia: Machine learning model development. *Water Science 38*(1): 33–47.
28. Liu, M., X. Liu, J. Jiang and X. Xia (2013). Artificial neural network and random forest approaches for modeling of sea surface salinity. *International Journal of Remote Sensing Applications 3*(4): 229–235.
29. Liu, X., Y. Hu, X. Li, R. Du, Y. Xiang and F. Zhang (2024). An Interpretable Model for Salinity Inversion Assessment of the South Bank of the Yellow River Based on Optuna Hyperparameter Optimization and XGBoost. *Agronomy 15*(1): 18.
30. Mahmoud, M. F., M. Arabi and S. Pallickara (2025). Harnessing ensemble Machine learning models for improved salinity prediction in large river basin scales. *Journal of Hydrology 652*: 132691.
31. Mantena, S., V. Mahammood and K. N. Rao (2023). Prediction of soil salinity in the Upputeru river estuary catchment, India, using machine learning techniques. *Environmental Monitoring and Assessment 195*(8): 1006.
32. Mishra, A. P., H. Khali, S. Singh, C. B. Pande, R. Singh and S. K. Chaurasia (2023). An assessment of in-situ water quality parameters and its variation with Landsat 8 level 1 surface reflectance datasets. *International Journal of Environmental Analytical Chemistry 103*(18): 6344–6366.
33. Musie, W. and G. Gonfa (2023). Fresh water resource, scarcity, water salinity challenges and possible remedies: A review. *Heliyon 9*(8).
34. Nasir, N., A. Kansal, O. Alshaltone, F. Barneih, M. Sameer, A. Shanableh and A. Al-Shamma'a (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering 48*: 102920.
35. Nguyen, H. D., D. K. Dang and Q.-T. Bui (2025). Estuary salinity prediction using machine learning: case study in the Hau estuary in Mekong River, Vietnam. *Water Supply:* ws2025007.
36. Nguyen, M. T., F. G. Renaud and Z. Sebesvari (2019). Drivers of change and adaptation pathways of agricultural systems facing increased salinity intrusion in coastal areas of the Mekong and Red River deltas in Vietnam. *Environmental Science & Policy 92*: 331–348.
37. Nguyen, Y. T. B., A. Kamoshita, V. T. H. Dinh, H. Matsuda and H. Kurokura (2017). Salinity intrusion and rice production in Red River Delta under changing climate conditions. *Paddy and Water Environment 15*: 37–48.
38. Niazkar, M., A. Menapace, B. Brentan, R. Piraei, D.

Jimenez, P. Dhawan and M. Righetti (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software 174*: 105971.

39. Parhi, S. K. and S. K. Patro (2024). Compressive strength prediction of PET fiber-reinforced concrete using Dolphin echolocation optimized decision tree-based machine learning algorithms. *Asian Journal of Civil Engineering 25*(1): 977–996.
40. Phan, L. T. and A. Kamoshita (2020). Salinity intrusion reduces grain yield in coastal paddy fields: case study in two estuaries in the Red River Delta, Vietnam. *Paddy and Water Environment 18*: 399–416.
41. Phung, Q. A. and N. Dao (2024). Farmers' perceptions of sustainable agriculture in the Red River Delta, Vietnam. *Heliyon 10*(7).
42. Priyadharshini, P. and V. Karpagam (2024). Biomedical advantages of magnetohydrodynamics williamson nanofluid: optimization of multiple linear regression and multilayer perceptron. *Journal of Nanofluids 13*(6): 1350–1363.
43. Raheli, B., N. Talebbeydokhti, S. Saadat and V. Nourani (2024). Uncertainty assessment of surface water salinity using standalone, ensemble, and deep machine learning methods: A case study of Lake Urmia. *Iranian Journal of Science and Technology, Transactions of Civil Engineering 48*(2): 1029–1047.
44. Rajeev, A., R. Shah, P. Shah, M. Shah and R. Nanavaty (2025). The potential of big data and machine learning for ground water quality assessment and prediction. *Archives of Computational Methods in Engineering 32*(2): 927–941.
45. Rokach, L. and O. Maimon (2005). Decision trees. *Data mining and knowledge discovery handbook*: 165–192.
46. Sakai, T., K. Omori, A. N. Oo and Y. N. Zaw (2021). Monitoring saline intrusion in the Ayeyarwady Delta, Myanmar, using data from the Sentinel-2 satellite mission. *Paddy and Water Environment 19*: 283–294.
47. Salman, H. A., A. Kalakech and A. Steiti (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning 2024*: 69–79.
48. Saputra, N., R. Antoni, A. Widodo, E. M. Solissa and I. Arief (2024). *Improving foreign language proficiency in society by decision tree classification.* AIP Conference Proceedings, AIP Publishing.
49. Shaker, B., M.-S. Yu, J. S. Song, S. Ahn, J. Y. Ryu, K.-S. Oh and D. Na (2021). LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM. *Bioinformatics 37*(8): 1135–1139.
50. Simons, M., G. Podger and R. Cooke (1996). IQQM—a hydrologic modelling tool for water resource and salinity management. *Environmental Software 11*(1–3): 185–192.
51. Su, X., X. Yan and C. L. Tsai (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics 4*(3): 275–294.
52. Suleymanov, A., I. Gabbasova, M. Komissarov, R. Suleymanov, T. Garipov, I. Tuktarova and L. Belan (2023). Random forest modeling of soil properties in saline semi-arid areas. *Agriculture 13*(5): 976.
53. Tran, D. D., M. M. Dang, B. Du Duong, W. Sea and T. T. Vo (2021). Livelihood vulnerability and adaptability of coastal communities to extreme drought and salinity intrusion in the Vietnamese Mekong Delta. *International Journal of Disaster Risk Reduction 57*: 102183.
54. Tran, V. N., H. D. Nguyen, H. Van Khuong, H. B. Dao, Q. H. M. Le, C. Q. Nguyen and G. T. Nguyen (2025). Reconstructing Long-Term Daily Streamflow Data at the Discontinuous Monitoring Station in the Ungauged Transboundary Basin Using Machine Learning. *Water Resources Management:* 1–22.
55. Uyanık, G. K. and N. Güler (2013). A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences 106*: 234–240.
56. Vineis, P., Q. Chan and A. Khan (2011). Climate change impacts on water salinity and health. *Journal of epidemiology and global health 1*(1): 5–10.
57. Wang, F., S. Yang, Y. Wei, Q. Shi and J. Ding (2021). Characterizing soil salinity at multiple depth using electromagnetic induction and remote sensing data with random forests: A case study in Tarim River Basin of southern Xinjiang, China. *Science of the Total Environment 754*: 142030.
58. Wang, S., H. Peng and S. Liang (2022). Prediction of estuarine water quality using interpretable machine learning approach. *Journal of Hydrology 605*: 127320.
59. Wang, X., G. Liu, J. Yang, G. Huang and R. Yao (2017). Evaluating the effects of irrigation water salinity on water movement, crop yield and water use efficiency by means of a coupled hydrologic/crop growth model. *Agricultural water management 185*: 13–26.
60. Wang, Y., J. Chen, X. Chen, X. Zeng, Y. Kong, S. Sun, Y. Guo and Y. Liu (2020). Short-term load forecasting for industrial customers based on TCN-LightGBM. *IEEE Transactions on Power Systems 36*(3): 1984–1997.
61. Wu, S., P. Cheng and F. Yang (2024). Study on the impact of digital transformation on green competitive advantage: The role of green innovation and government regulation. *Plos one 19*(8): e0306603.
62. Yan, X., T. Zhang, W. Du, Q. Meng, X. Xu and X. Zhao (2024). A comprehensive review of machine learning for water quality prediction over the past five years. *Journal of Marine Science and Engineering 12*(1): 159.
63. Yu, D., Z. Wang, C. Yue and J. Wang (2025). Spatial

modeling of brine level and salinity in the Qarhan Salt Lake using GIS and automated machine learning algorithms. *Journal of Hydrology: Regional Studies 58*: 102195.

64. Yu, J. W., S. Kim, J. H. Ryu, W. B. Lee and T. J. Yoon (2024). Spatiotemporal characterization of water diffusion anomalies in saline solutions using machine learning force field. *Science Advances 10*(50): eadp9662.
65. Yuen, K. W., T. T. Hanh, V. D. Quynh, A. D. Switzer, P. Teng and J. S. H. Lee (2021). Interacting effects of land-use change and natural hazards on rice agriculture in the Mekong and Red River deltas in Vietnam. *Natural Hazards and Earth System Sciences 21*(5): 1473–1493.
66. Zennaro, F., E. Furlan, C. Simeoni, S. Torresan, S. Aslan, A. Critto and A. Marcomini (2021). Exploring machine learning potential for climate change risk assessment. *Earth-Science Reviews 220*: 103752.
67. Zhang, M., N. Liu, R. Harper, Q. Li, K. Liu, X. Wei, D. Ning, Y. Hou and S. Liu (2017). A global review on hydrological responses to forest change across multiple spatial scales: Importance of scale, climate, forest type and hydrological regime. *Journal of Hydrology 546*: 44–59.
68. Zhu, M., J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu and L. Ye (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health 1*(2): 107–116.