IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 December 2023, accepted 2 January 2024, date of publication 5 January 2024,
date of current version 12 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350328

RESEARCH ARTICLE

Optimized Ensemble Machine Learning Models for Predicting Phytoplankton Absorption Coefficients

MD. SHAFIUL ALAM, SURYA PRAKASH TIWARI, AND SYED MASIUR RAHMAN
Applied Research Center for Environment and Marine Studies, Research Institute, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
Corresponding author: Surya Prakash Tiwari (surya.tiwari@kfupm.edu.sa)

ABSTRACT The machine learning (ML) model provides an alternative method for estimating inherent optical properties (IOPs) in clear and coastal waters. This study introduces an effective approach by employing ensemble machine learning techniques, such as random forest, gradient boosting, extra tree, adaboost, bagging, and voting model, to predict phytoplankton absorption coefficient ($a_{ph}(\lambda)$, m^{-1}) at selected key wavebands of 443, 489, 510, 555, and 670 nm in clear and coastal waters. The optimization of the hyperparameters of these models through Bayesian techniques ensured high predictive accuracy. Furthermore, this research highlights the critical importance of wavelengths 670, 489, and 510 nm through feature importance analysis. The models exhibit excellent performance in terms of the coefficient of determination (R^2) value when predicting phytoplankton at various wavelengths (e.g., 443, 489, 510, 555, and 670 nm). The R^2 value of around 0.9033 is obtained for the absorption coefficient of phytoplankton a_{ph} at the wavelength 510 nm. The lowest mean squared error (MSE) of 0.0001 was achieved at the green waveband (i.e., 555 nm). Other statistical matrices, such as mean absolute percentage error (MAPE) and mean absolute error (MAE), have shown a low error across the selected wavelengths. It is found that the predicted phytoplankton absorption coefficients are in close agreement with actual values. This study shows the success of optimized ensemble models for both global and selected regional datasets that can accurately derive $a_{ph}(\lambda)$, which will contribute to the improvement of ocean primary productivity modelling and understanding the distribution of phytoplankton blooms.

INDEX TERMS Phytoplankton absorption coefficients, remote sensing reflectance, machine learning, ensemble models, feature importance.

ABBREVIATION

ML:	Machine Learning.
MAPE:	Mean Absolute Percentage Error.
MAE:	Mean Absolute Error.
TSS:	Total Suspended Sediments.
CDOM:	Colored Dissolved Organic Matter.
GRB:	Gradient Boosting Regressor.
MSE:	Mean Squared Error.
ETR:	Extra Tree Regressor.
IOP:	Inherent Optical Properties.
AOP:	Apparent Optical Properties.

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

I. INTRODUCTION

Phytoplankton play a crucial role in marine ecosystems and the biogeochemical cycle of global environment. They contribute significantly to the oxygen content of the Earth's atmosphere through photosynthesis. They store energy in the form of carbon compounds, accounting for over half of the global primary productivity and serving as the foundation of oceanic food webs. They have a considerable influence on the ocean color biology of seawater, as observed by satellite sensors, due to their various shapes, sizes, and coloring [1], [2].

The $a_{ph}(\lambda)$ parameter changes with the variation of the chlorophyll-a concentration (Chl-a, mgm^{-3}). A related parameter to $a_{ph}(\lambda)$ is the specific absorption coefficient of

phytoplankton ($a^*_{ph}(\lambda)$), which is $a_{ph}(\lambda)$ normalized by Chl-a. As Chl-a is a primary index for understanding the tropic state of ocean and land waters, because it provides a strong relation with biomass and phytoplankton abundance [3], [4]. The level of Chl-a varies with changes in nutrient levels, environmental conditions, and seasonally with respect to temperature and precipitation [5]. It is important to investigate the multi-sensor and multi-platform retrieval performance of water Chl-a concentration by combining multi-sensor data [6], [7]. Therefore, monitoring the Chl-a concentration is crucial for evaluating and managing the aquatic ecosystems and addressing global environmental challenges.

Traditionally, the Chl-a concentration in the aquatic system was derived and monitored based on laboratory experiments and in situ measurements [8]. These procedures are only limited to the regional level and, to some extent, to the global level, with high cost and time consumption. Bio-optical algorithms have been used to estimate Chl-a from the remote sensing reflectance $R_{rs}(\lambda)$ [9], which was obtained after the atmospheric correction from the radiance at the top of atmosphere $L_t(\lambda)$ [10], [11], [12]. Chl-a values were estimated based on the spectral shape and magnitude of R_{rs}. The empirical equations were framed with the principle of R_{rs} in blue decreases (due to Chl-a absorption) with an increase in Chl-a values. Though empirical relations were much easier to implement for global and regional water like clear/moderately turbid waters (Case-1 water), but poorly retrieve Chl-a values in either total suspended sediments (TSS) or colored dissolved organic matter (CDOM) dominated waters, which are known as Case-2 waters [9]. The drawbacks of these empirical relations were rectified by semi-analytical models, which have the ability to retrieve multiple products like Chl-a, CDOM, and TSS simultaneously.

The phytoplankton specific absorption coefficients-based approach for $a_{ph}(\lambda)$ computation has shown promising progress, particularly in open ocean waters. It is worth noting that the specific absorption coefficients vary depending on the pigment content and packaging effect, both of which vary throughout phytoplankton assemblages. The viability of the stated specific absorption coefficients in several researched water locations still requires validation with additional in-situ observations. Consequently, correctly quantifying specific absorption coefficients for the water areas of interest is critical. Many techniques have been developed and are now in use for estimating $a_{ph}(\lambda)$ from remote sensing data for use in ocean color applications [13], [14]. However, there is still lots of scope to develop machine learning-based approaches for the accurate estimation of $a_{ph}(\lambda)$. Therefore, this study has developed optimized ensemble machine learning models to estimate $a_{ph}(\lambda)$ values from remote sensing reflectance. The model employs a ML framework with six distinct ensemble algorithms. The best algorithm is selected with the pipeline approach, and its parameters are optimized to improve the prediction performance further. Thus, in this work, we have also developed feature importance analysis to explicitly rank the features in predicting phytoplankton absorption coefficients. The model robustness is tested with an independent data set as part of the validation. Several statistical metrices are computed to assess the model performance. The model results exhibit a very good agreement against the in situ observations. This is mainly due to its reliance on the a_{ph} peaks at 443 and 670 nm, which are greatly influenced by phytoplankton absorption and fluorescence.

This paper is organized into four sections. Section II presents data and methods. The measured data, the background of ensemble models, the modeling approach, and performance evaluation matrices are described in this section. The performance of the proposed ensemble models in the prediction of phytoplankton is documented in Section III. The concluding remarks are highlighted in Section IV.

II. DATA AND METHODS

A. MEASURED DATA

The dataset mentioned in the paper [15] served as the primary datasets for this study. Specifically, it incorporated NOMAD (NASA Bio-Optical Marine Algorithm Data Set – SeaBASS) in situ data, CARDER in situ data, and off Point Calimere in situ data (collected in highly turbid coastal waters off Point Calimere in the Bay of Bengal, Southern India). NOMAD in situ data [16] combines pigment and optical data gathered concurrently from various geographical locations. It stands as a unique and internationally acclaimed high-quality data set, benefiting the global bio-optical community [16]. The CARDER bio-optical dataset gathered in coastal waters on the west coast of Florida from 1999 to 2006 is also used in this study [17], [18].

The above-water radiometric data (from Trios sensors), pigment data, and $a_{ph}(\lambda)$ data (determined using the methods described in [19]) were collected in highly turbid coastal waters off Point Calimere on the southeast coast of Tamil Nadu, Southern India [15]. All the data sets are merged together to create a full data set with 841 data points. Further, this data set is divided into two parts. 80 % of the total data points were used for training models, and the remaining 20 % were used for modeling validation.

Table 1 describes the statistics of the in situ measurements of remote sensing reflectance and phytoplankton absorption coefficients (at key selected wavelengths) for data set used for the model training.

The dataset covers a wide range of clear and coastal waters around the world ($a_{ph}(443)$ varying from 0.00176 to 4.443 m^{-1} and the corresponding $R_{rs}(443)$ ranging from 0.00019 to 0.0251 sr^{-1}.

B. BACKGROUND: ENSEMBLE MODELS FOR PHYTOPLANKTON ABSORPTION COEFFICIENTS PREDICTION

1) RANDOM FOREST REGRESSION

A schematic illustration of the random forest ensemble model is shown in Figure 1. The model increases the forecast probability by combining different models [20], [21]. In other

TABLE 1. Statistics of the in-situ data sets used to train models.

IOP	MIN	MAX	MEAN	N
$a_{ph}(443)$	0.0018	4.4433	0.0941	674
$a_{ph}(490)$	0.0010	3.0558	0.0622	674
$a_{ph}(510)$	0.0003	2.1888	0.0440	674
$a_{ph}(555)$	0.0001	0.8426	0.0184	674
$a_{ph}(670)$	0.0001	1.8618	0.0369	674
AOP	**MIN**	**MAX**	**MEAN**	**N**
$R_{rs}(443)$	0.0002	0.0251	0.0056	674
$R_{rs}(490)$	0.0004	0.0363	0.0070	674
$R_{rs}(510)$	0.0005	0.0335	0.0064	674
$R_{rs}(555)$	0.0006	0.0399	0.0062	674
$R_{rs}(670)$	0.0000	0.0175	0.0012	674

words, the ensemble model performs better in terms of prediction than other single models. The random forest method can be used for both classification and regression [22]. The arithmetic average of the regression results from all of the decision trees, which are collected from various regression trees, is the final model output from the regression technique. It creates decision trees from several samples, categorizing them based on their average and regressing them based on a majority vote. One of the most important features of the random forest technique is its capacity to handle data sets containing continuous variables, as in regression, and categorical variables, as in classification.

The random forest algorithm's steps are as follows:

Step 1: N randomly chosen records are chosen at random from a data collection of k records.

Step 2: A distinct decision tree is constructed for each sample.

Step 3: Each decision tree generates an output.

Step 4: The outcome for classification and regression is assessed using a majority vote or an average.

2) GRADIENT BOOSTING REGRESSION

A machine learning approach called gradient boosting regression (GBR) is employed for regression problems involving the prediction of a continuous target variable. It functions by sequentially combining a number of weak decision tree models to progressively raise prediction accuracy [23]. The approach begins by creating a straightforward decision tree model to forecast the target variable, and then iteratively fits the residual errors of the previous model to create new models, as shown in Figure 2. The mean squared error (MSE) between the predicted and actual target dataset is trained into each model in order to minimize it. The final prediction is generated by adding the individual models' predictions together. Due to its versatility, tolerance to outliers and noisy data, and ability to handle complex non-linear connections between the input features and the target variable, GBR is an effective technique that is frequently employed in many real-world applications. It can be computationally expensive for large datasets and necessitates careful hyperparameter tuning.

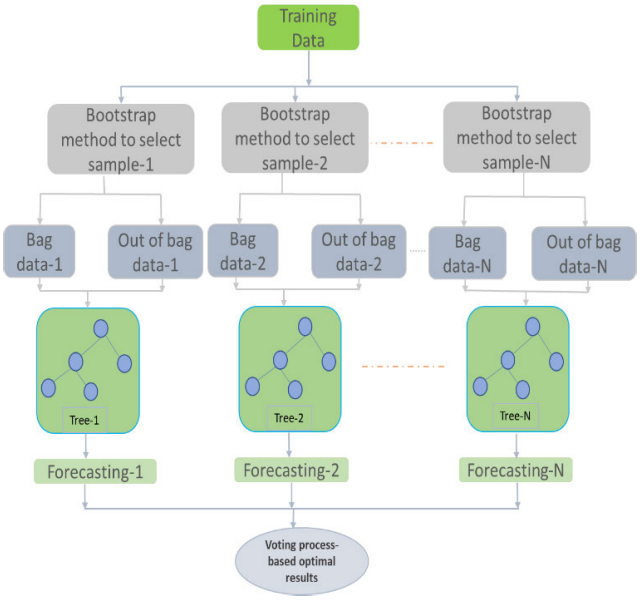

FIGURE 1. Conceptual diagram of random forest ensemble model.

3) EXTRA TREE REGRESSION

To increase the precision and resilience of regression models, the ensemble learning technique known as Extra Trees Regression (ETR) integrates numerous decision trees, as shown in Figure 3. By choosing arbitrary feature subsets and threshold values for splitting nodes, ETR adds more randomness to the tree-building process [24]. This randomization lowers overfitting and raises tree diversity, which can enhance ensemble stability and accuracy. The user has flexibility over how many trees are included in the ensemble; generally, more trees lead to higher performance but longer calculation times. Regression models' accuracy and robustness can be increased with the use of ETR, a potent and effective ensemble learning technique. ETR is a well-liked option for many machine learning applications due to its lower computing cost and effectiveness in high-dimensional data.

4) ADABOOST REGRESSION

Regression analysis is frequently performed using adaptive boosting, also known as AdaBoost. It is a technique for ensemble learning that combines a number of weak learners to produce a stronger learner to predict precisely [25]. The fundamental principle of AdaBoost regression is to fit a series of regression trees to the training data iteratively, with each regression tree acting as a weak learner that tries to predict the target variable using a collection of input features. The final prediction is the weighted average of the predictions from all the regression trees in the sequence, where the weights are based on how accurately each tree predicted the training data.

It is quite adaptable and works with a variety of input features and target variables, but it can be sensitive to outliers and data noise. All things considered, AdaBoost regression is a strong and adaptable machine learning technique that can

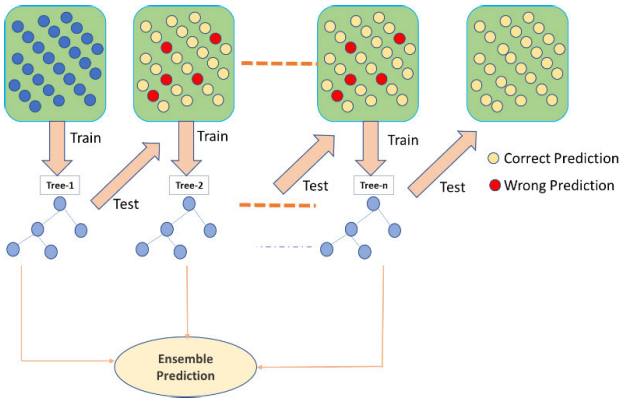

FIGURE 2. Implementation of gradient boosting regression.

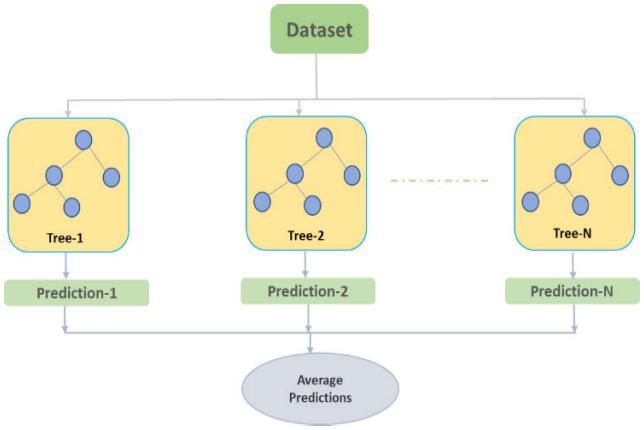

FIGURE 3. Implementation of extra tree regression.

be applied to a variety of regression applications. The main step of AdaBoost regression is given below.

Initialize the weights: Each sample in the training set is given an equal weight at the beginning of the algorithm. These weights are used to give misclassified samples extra weight throughout each iteration.

Step-1: Using the current weights, the first regression tree is fitted to the training data. The misclassified samples are given larger weights when the tree is used to make predictions based on the training data.

Step-2: The same method used to fit the first tree is used to fit additional regression trees to the weighted data. The user chooses how many trees to utilize; more trees typically result in higher performance but longer training timeframes.

Step-3: The final prediction is a weighted average of the predictions from all the regression trees in the sequence, where the weights depend on how well each tree performed on the training set of data.

Step-4: The validation set is used to assess the model's performance. This enables you to assess the model's ability to generalize to new data and make any necessary changes to enhance performance.

5) BAGGING REGRESSION

The accuracy and stability of regression models are enhanced by the robust machine learning algorithm known as Bagging

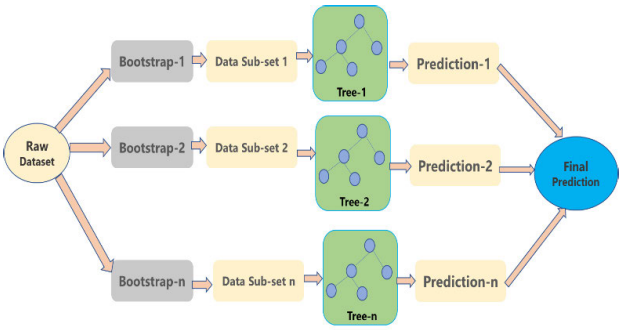

FIGURE 4. Implementation of bagging regression.

Regression, which applies the bootstrap aggregation (bagging) method as shown in Figure 4. Several models are built using various subsets of the training data in this ensemble learning technique, and the predictions from each model are then combined to get the final prediction [26], [27]. The Bagging Regression procedure entails a number of processes, such as dividing the data into a training set and a validation set, producing numerous subsets of the training data, and aggregating the predictions from all the models to provide a final prediction. The Bagging Regression technique can also make use of feature bagging to increase the model's stability and accuracy. Bagging Regression is an all-purpose machine learning approach that is effective and adaptable for a variety of regression tasks.

6) VOTING REGRESSION

A machine learning approach called voting regression combines the results of various regression models to produce a single forecast. It is an ensemble learning technique for enhancing the precision and consistency of regression models [28], [29]. It entails building numerous regression models with various hyperparameters, methods, and training data subsets. Then, using a voting method, the predictions from all the models are combined to determine the final prediction. The Voting Regression approach can also include ensemble methods like bagging and boosting to increase the model's stability and accuracy. Voting Regression is an all-around effective and adaptable machine learning technique that may be applied to a variety of regression tasks.

C. MODELING APPROACH

The experimental setup for the proposed ensemble models to predict the absorption coefficient of phytoplankton is envisioned in Figure 5. The data has 843 samples, five input features such as remote sensing reflectance $R_{rs}(443)$, $R_{rs}(489)$, $R_{rs}(510)$, $R_{rs}(555)$, and $R_{rs}(670)$ and five output features such as $a_{ph}(443)$, $a_{ph}(489)$, $a_{ph}(510)$, $a_{ph}(555)$, $a_{ph}(670)$.

The data are cleaned, filled in where necessary, normalized, and split tested and trained sets. Based on the testing data's smallest standard deviation, the optimal model is chosen. Hyperparameters like the number of trees, maximum depth, minimum samples split, and minimum samples leaf

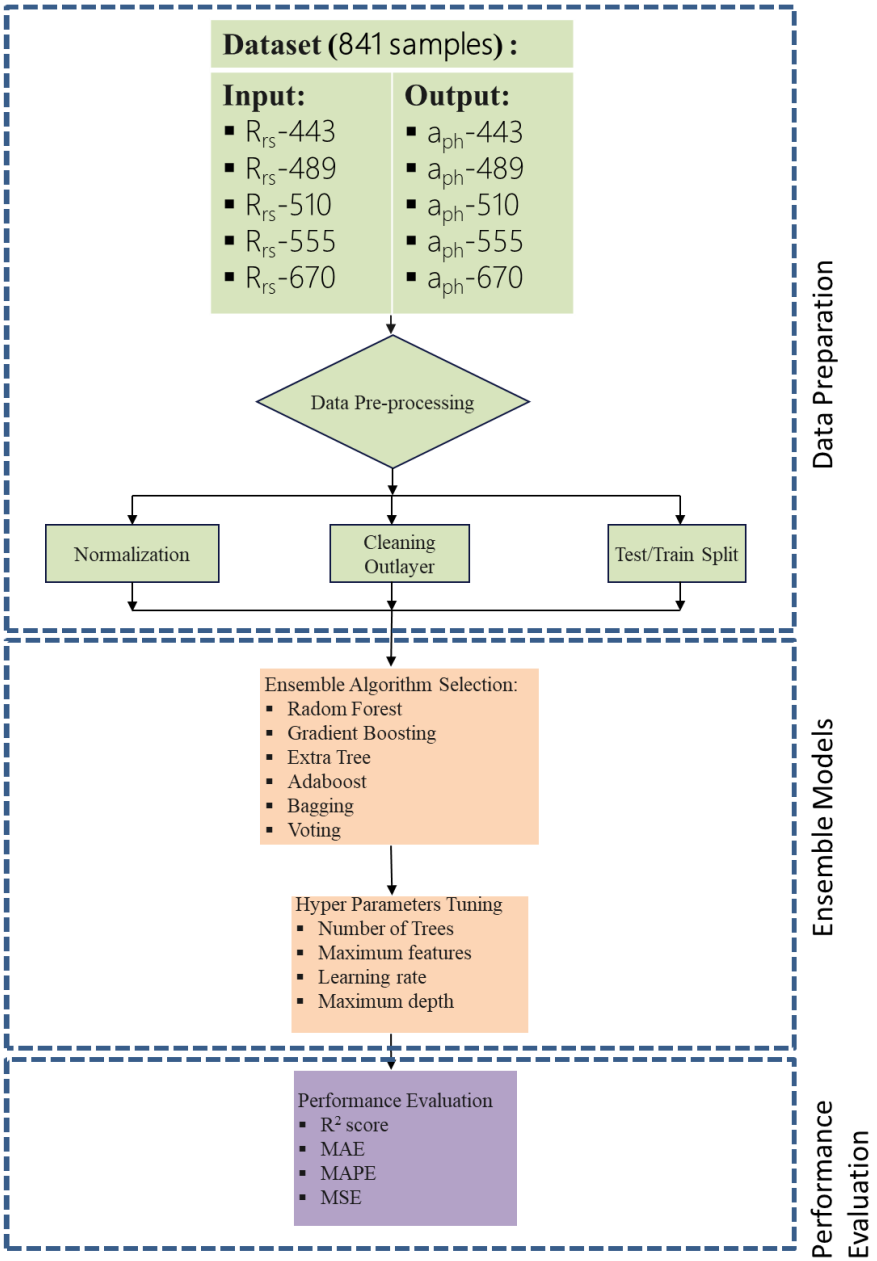

FIGURE 5. Experimental setup of ensemble model for predicting the phytoplankton absorption coefficients.

are tuned after the best model has been chosen. In this work, the method of parameter tuning known as "Bayesian Optimization" is utilized to evaluate the model. It is a method that chooses the ideal set of hyperparameters for a particular model using probabilistic modeling. In recent years, ensemble ML hyperparameter optimization has been successfully accomplished via Bayesian optimization. The approach operates by simulating the ensemble model's performance on a validation set, the objective function. The model is updated based on the discrepancy between predicted and actual data, and the process is repeated until the best set of hyperparameters is identified. Bayesian optimization is an excellent technique for fine-tuning ensemble models because it can handle noisy or non-smooth objective functions and effectively search a wide hyperparameter field.

D. DISTINCTIVENESS OF THE PROPOSED APPROACH

In this study, different ensemble models were meticulously developed, employing a systematic approach to select the most suitable one. These models were refined by adjusting their parameters to ensure optimal configurations. In the domain of hyperparameter tuning for ensemble models, Bayesian optimization was employed, which is an advanced technique that follows an iterative process of selecting

TABLE 2. Optimized hyperparameters of extra tree regressor.

$a_{ph}(\lambda)$	Hyperparameters			
	Number of estimators (n_estimators)	Maximum depth (max_depth)	Minimum sample split (min_samples_split)	Minimum sample leaf (Min_samples_leaf)
$a_{ph}(443)$	84	15	2	1
$a_{ph}(489)$	81	17	4	3
$a_{ph}(510)$	66	14	2	1
$a_{ph}(555)$	191	15	3	1
$a_{ph}(670)$	167	17	5	2

hyperparameter combinations, assessing their performance, and updating a probabilistic surrogate model. This approach facilitates a systematic exploration of the hyperparameter space, with a particular emphasis on configurations that have promising potential. This strategy enhances the predictive accuracy and overall performance of these models in predicting absorption coefficients. To enhance our understanding of the models, advanced techniques, including Shapley values and permutation-based assessments, were applied. Shapley values, originating from game theory, provided insights into the impact of individual features on predictions, while permutation analysis tested model stability by shuffling features and identifying the most influential ones. This detailed analysis not only increased the credibility of the results but also deepened comprehension of the underlying data. By integrating these advanced methods, the study revolutionized phytoplankton absorption predictions, setting a new benchmark for accuracy and interpretability in this field.

E. MODEL ACCURACY ASSESSMENT

Several performance matrices, including R^2 score, mean-squared error (MSE), mean absolute percentage error (MAPE), and mean absolute error (MAE), were used to assess the accuracy of models. MAPE is the ratio of the mean absolute value of the actual data to the mean absolute value of the prediction errors. The model performs better when the MAPE, MAE, and MSE scores are lower. On the other hand, a higher R^2 score indicates better model performance. R^2 is a statistical index that shows the proportion of a dependent variable's variation in a regression model that can be accounted for by one or more independent variables.

$$R^2 = 1 - \frac{SSR}{SST}$$

where, SST represents the total sum of squares and SSR represents sum of squared regression.

The MAPE is defined by the following equation.

$$MAPE = \frac{\sum_{i=1}^{N} |dA_i - dF_i|}{dA_i} \times 100\%$$

where dF_i is the prediction data, dA_i is the actual data, and N is the number of samples.

TABLE 3. Optimized hyperparameters of random forest.

$a_{ph}(\lambda)$	Hyperparameters			
	n_estimators	max_depth	min_samples_split	Min_samples_leaf
$a_{ph}(443)$	157	19	2	1
$a_{ph}(489)$	192	19	19	5
$a_{ph}(510)$	145	20	2	1
$a_{ph}(555)$	200	18	2	2
$a_{ph}(670)$	211	20	2	1

TABLE 4. Optimized hyperparameters of gradient boosting.

$a_{ph}(\lambda)$	Hyperparameters				
	n_estimators	max_depth	min_samples_split	Min_samples_leaf	Learning_rate
$a_{ph}(443)$	125	17	8	1	0.262
$a_{ph}(489)$	218	20	20	1	0.01
$a_{ph}(510)$	123	18	19	1	0.011
$a_{ph}(555)$	211	7	16	10	0.01
$a_{ph}(670)$	138	17	7	1	0.191

The MSE indicates the model error, and it is 0 if the predicted and the observed values are same. The value of MSE rises in proportion to the model error, and it is generally calculated by the equation given below.

$$MSE = \frac{\sum_{i=1}^{N} (dA_i - dF_i)^2}{N}$$

where dA_i is the i^{th} observed value and dF_i is the corresponding predicted value.

III. RESULTS AND DISCUSSIONS

A. RESULTS

Using the default parameters, the pipeline approach compares six different ensemble models and chooses the best one. A machine learning pipeline is a complete design that manages the input and output of a variety of models. With the extra tree regressor, the height score is obtained for predicting absorption coefficients of phytoplankton at selected wavebands such as 443, 490, 510, 555, and 670 nm. After determining the best ensemble models (extra tress regressor), the hyperparameters are further tuned with Bayesian optimization. The Bayesian optimization is used to improve expensive-to-evaluate black-box functions. The approach creates a surrogate model, a probabilistic representation of the unknown function. The cost function values are predicted using the surrogate model for unexplored regions of the search space. Based on the predictions of the surrogate model and the recent status of the search, the acquisition function is utilized to determine the next point to examine. The number of estimators, maximum depths, minimum sample splits, and minimum samples leaf are optimized for the extra tree regressor. These data-driven models do not consider the physics of the underlying fact. The optimized parameters for predicting different absorption coefficients of phytoplankton are provided in Table 2.

TABLE 5. Optimized hyperparameters of adaboost.

$a_{ph}(\lambda)$		
	n_estimators	Learning_rate
$a_{ph}(443)$	194	0.835
$a_{ph}(489)$	125	0.315
$a_{ph}(510)$	156	0.085
$a_{ph}(555)$	119	0.177
$a_{ph}(670)$	140	0.089

TABLE 6. Optimized hyperparameters of bagging.

$a_{ph}(\lambda)$	Hyperparameters		
	n_estimators	max_samples	max_features
$a_{ph}(443)$	172	0.988	0.833
$a_{ph}(489)$	173	0.985	0.833
$a_{ph}(510)$	88	0.97	0.838
$a_{ph}(555)$	171	0.938	0.803
$a_{ph}(670)$	244	0.618	0.913

TABLE 7. Optimized hyperparameters of voting.

$a_{ph}(\lambda)$	Hyperparameters		
	n_estimators	max_samples	max_features
$a_{ph}(443)$	95	0.964	0.849
$a_{ph}(489)$	172	0.988	0.833
$a_{ph}(510)$	244	0.900	0.862
$a_{ph}(555)$	243	0.868	0.942
$a_{ph}(670)$	244	0.965	0.988

TABLE 8. Performance matrices for phytoplankton absorption coefficient prediction.

$a_{ph}(\lambda)$	Performance Matrices			
	R^2 Score	*MSE*	*MAPE*	*MAE*
$a_{ph}(443)$	0.8920	0.0021	0.2784	0.0204
$a_{ph}(489)$	0.8574	0.0013	0.3123	0.0158
$a_{ph}(510)$	0.9033	0.0005	0.4909	0.0100
$a_{ph}(555)$	0.8615	0.0001	0.4195	0.0052
$a_{ph}(670)$	0.8151	0.0010	0.3950	0.0125

Table 2 shows that the maximum number of n_estimators is required to predict the absorption coefficients of phytoplankton at a wavelength of 555 nm. The optimized maximum depths indicate that the phytoplankton prediction at a wavelength of 489 nm and 610 nm has the same depth of 17. The minimum sample split ranges between 2 to 5 for the selected wavebands.

The hyperparameters of the other ensemble models are provided in Tables 3, 4, 5, 6, and 7. The models are trained with optimal hyperparameters once the main parameters for the extra tree regressor are tuned. The trained ensemble machine-learning models are then pickled, a standard Python method for serializing objects, and saved to a file for later use and prediction of phytoplankton absorption coefficients the five (443, 490, 510, 555, and 670 nm) selected wavebands with unobserved data. Table 8 displays the developed model's statistical performance using

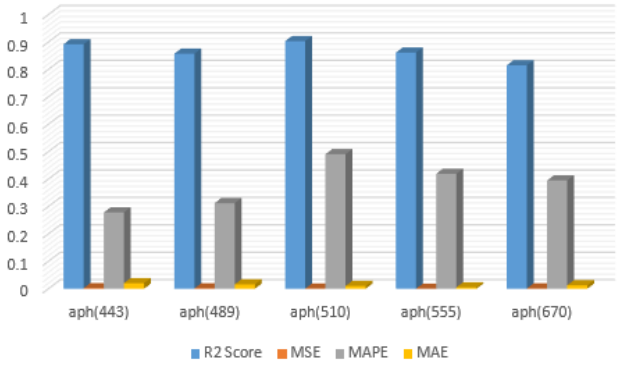

FIGURE 6. Performance index for several a_{ph} predictions.

the validation dataset. It demonstrates that the developed models perform superbly in terms of R^2 value when predicting phytoplankton at various wavelengths. The highest R^2 is obtained for the absorption coefficient of phytoplankton $a_{ph}(510)$ prediction, which is around 0.9033. The lowest MSE of about 0.0001 for 169 observations also suggests strong statistical performance. Table 8 also includes the other performance matrices, such as MAPE and MAE. The minimum MAPE and MAE are 0.2784 and 0.0052 for $a_{ph}(443)$ and $a_{ph}(555)$ predictions, respectively. The performance index values for phytoplankton absorption coefficient predictions with different wavelengths are provided in Figure 6. Several visualization techniques are used to better investigate the extra tree ensemble modes' performance in absorption coefficient prediction. Figure 7 shows the scatter plot for comparisons between actual $a_{ph}(\lambda)$ from in-situ and predicted $a_{ph}(\lambda)$ from the model for the five (443, 490, 510, 555, and 670 nm wavebands) selected wavebands. For almost all in situ values, it is discovered that the predicted phytoplankton absorption coefficients are quite close to the actual values. Linear regression parameters (i.e., slope and intercept) for the predicted and actual phytoplankton absorption coefficients are shown in Figure 7.

Though a wide variety of models with varying degrees of complexity ranging from empirical to complex semi analytical approaches for determination of the $a_{ph}(\lambda)$ coefficient were developed in the past, however, no model has potential to estimate the $a_{ph}(\lambda)$ accurately in coastal waters.

B. FEATURE IMPORTANCE ANALYSIS

Even though machine learning-based regression studies may produce highly accurate predictions, they are frequently criticized for their lack of explicit interpretability. Models are useful for making predictions, but they do not provide much information on how various input factors alter phytoplankton absorption coefficients. To address this issue, a feature importance analysis is carried out. This study sheds more light on the feature-importance techniques of the models.

The significance of each feature can be determined by the degree to which it contributed to a reduction in the loss

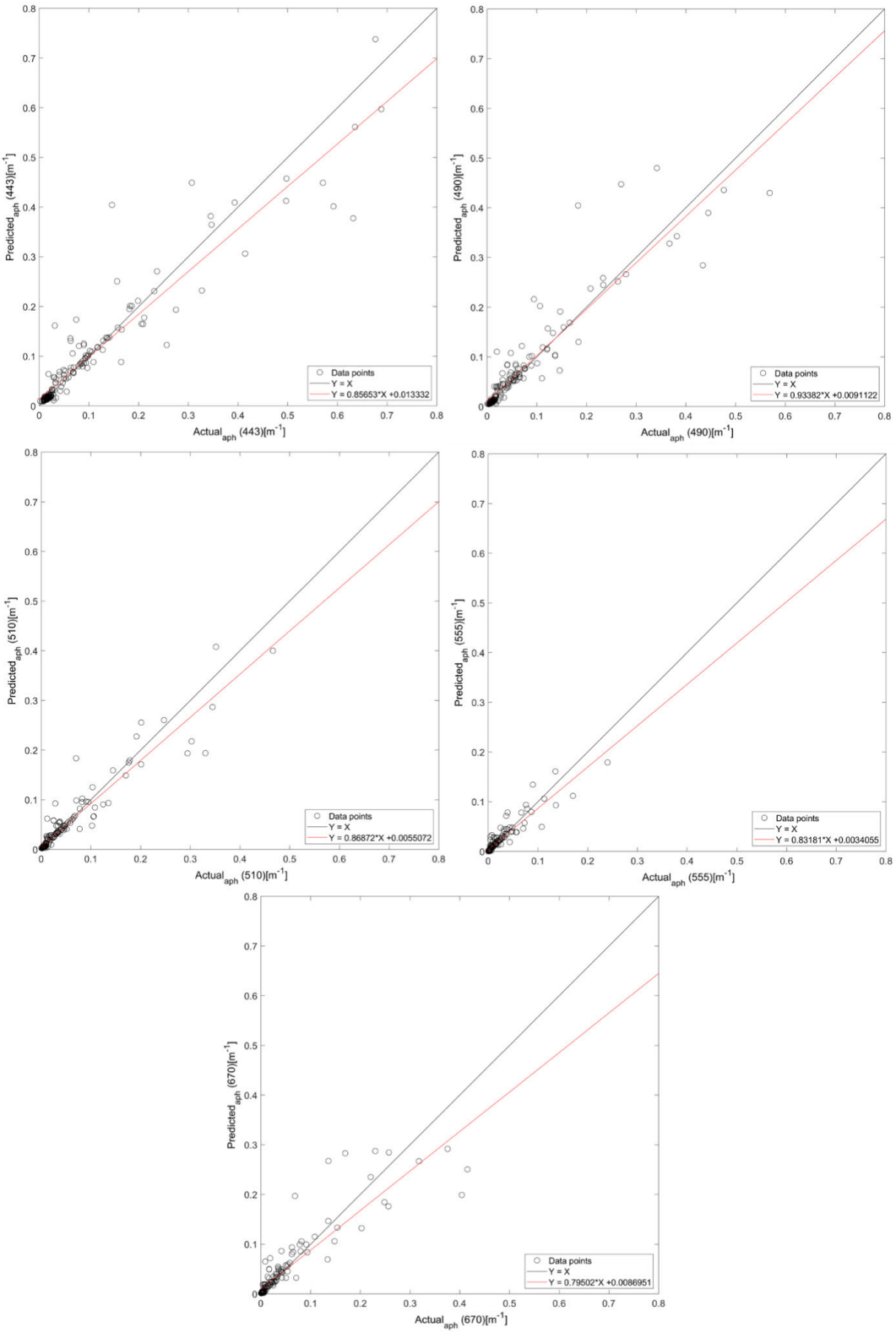

FIGURE 7. Scatter plots for comparisons between actual $a_{ph}(\lambda)$ from in-situ samples and predicted $a_{ph}(\lambda)$ from the model for the five (443, 490, 510, 555, and 670 nm wavebands) selected wavebands. The solid black line is the 1:1 line and solid red line is the linear regression fitted line.

function throughout the training phase. Shapley values and permutation importance are two techniques used to assess the importance of a feature. When a feature's values are randomly permuted to determine its permutation significance, the effect on the performance of the model is examined. Changing the value of a feature that is essential will have a catastrophic

TABLE 9. Feature importance analysis for phytoplankton absorption coefficient prediction.

Feature Rank	Output [importance]				
	$a_{ph}(443)$	$a_{ph}(490)$	$a_{ph}(510)$	$a_{ph}(555)$	$a_{ph}(670)$
1	$R_{rs}(670)$ [33.38]	$R_{rs}(670)$ [32.72]	$R_{rs}(670)$ [34.59]	$R_{rs}(670)$ [32.26]	$R_{rs}(670)$ [35.46]
2	$R_{rs}(489)$ [22.67]	$R_{rs}(489)$ [23.03]	$R_{rs}(489)$ [19.82]	$R_{rs}(489)$ [22.58]	$R_{rs}(489)$ [21.04]
3	$R_{rs}(510)$ [16.76]	$R_{rs}(510)$ [16.01]	$R_{rs}(510)$ [16.89]	$R_{rs}(510)$ [17.71]	$R_{rs}(510)$ [16.48]
4	$R_{rs}(443)$ [16.58]	$R_{rs}(555)$ [15.88]	$R_{rs}(443)$ [15.54]	$R_{rs}(555)$ [14.27]	$R_{rs}(443)$ [15.96]
5	$R_{rs}(555)$ [10.58]	$R_{rs}(443)$ [12.34]	$R_{rs}(555)$ [13.13]	$R_{rs}(443)$ [13.16]	$R_{rs}(555)$ [11.03]

impact on the model's accuracy. Shapley values, on the other hand, estimate the marginal contribution of a feature to the prediction by considering all plausible combinations of features. Table 9 ranks five features in predicting phytoplankton absorption coefficient. It is observed that $R_{rs}(670)$, $R_{rs}(489)$, and $R_{rs}(510)$ are top-ranked features for predicting phytoplankton absorption coefficients phytoplankton at various wavelengths (e.g., 443, 489, 510, 555, and 670 nm). This demonstrates that while the investigated models differ on the relative importance of other variables, they do share a commonality in their features for three of the five features. This implies that despite having slightly different performances, these models might have an advantage in some situations. This could mean that one type of model is better than another in the real world, but only in certain conditions. A decision support system can also leverage the data gathered from the deployment of several models.

IV. CONCLUSION

We have developed optimized ensemble machine learning models for predicting phytoplankton absorption coefficient with enhanced accuracy at five selected wavebands, which are common wavebands for most of the ocean color satellite sensors. The utilization of advanced techniques, including Bayesian optimization and feature importance analysis, significantly enhanced the accuracy and interpretability of predictions. Five input features are ranked to clearly indicate the most influential inputs in predicting $a_{ph}(\lambda)$. It has been found that the $R_{rs}(670)$, $R_{rs}(489)$, and $R_{rs}(510)$ impact $a_{ph}(\lambda)$ prediction at a variety of wavelengths. Models have estimated a_{ph} values with high accuracy in terms of statistical matrices. We also noticed close agreement between the estimated and actual values for all five wavebands. The best R^2 score (0.9033) and MSE (0.0005) are obtained for $a_{ph}(510)$ out of five selected wavebands. The developed models and comprehensive analysis not only enhanced result applicability but also enhanced understanding of underlying data, transforming phytoplankton absorption predictions and setting a new benchmark for accuracy and interpretability in the field. In future, a range of different ensemble models will be investigated for modeling phytoplankton absorption coefficients in the clear and coastal waters considering large scale global datasets.

ACKNOWLEDGMENT

The authors would like to thank the Deanship of Research Oversight and Coordination (DROC), King Fahd University of Petroleum and Minerals, Saudi Arabia, for the support to conduct this study. They would also like to thank all the institutions and researchers for making the data publicly available via the NASA SeaBASS website and thank Prof. P. Shanmugam, Dr. C. Hu, J. P. Cannizzaro, and Prof. K. L. Carder for providing the bio-optical dataset.

REFERENCES

[1] A. Morel and L. Prieur, "Analysis of variations in ocean color," *Limnol. Oceanogr.*, vol. 22, no. 4, pp. 709–722, Jul. 1977, doi: 10.4319/LO.1977.22.4.0709.

[2] A. Morel and S. Maritorena, "Bio-optical properties of oceanic waters: A reappraisal," *J. Geophys. Res. Oceans*, vol. 106, no. C4, pp. 7163–7180, Apr. 2001, doi: 10.1029/2000JC000319.

[3] W. J. Moses, A. A. Gitelson, S. Berdnikov, and V. Povazhnyy, "Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—Successes and challenges," *Environ. Res. Lett.*, vol. 4, no. 4, Oct. 2009, Art. no. 045005, doi: 10.1088/1748-9326/4/4/045005.

[4] K. Toming, T. Kutser, R. Uiboupin, A. Arikas, K. Vahter, and B. Paavel, "Mapping water quality parameters with sentinel-3 ocean and land colour instrument imagery in the Baltic Sea," *Remote Sens.*, vol. 9, no. 10, p. 1070, Oct. 2017, doi: 10.3390/RS9101070.

[5] P. Garnesson, A. Mangin, O. Fanton d'Andon, J. Demaria, and M. Bretagnon, "The CMEMS GlobColour chlorophyll a product based on satellite observation: Multi-sensor merging and flagging strategies," *Ocean Sci.*, vol. 15, no. 3, pp. 819–830, Jun. 2019, doi: 10.5194/OS-15-819-2019.

[6] X. Zhao, Y. Li, Y. Chen, X. Qiao, and W. Qian, "Water chlorophyll a estimation using UAV-based multispectral data and machine learning," *Drones*, vol. 7, no. 1, p. 2, Dec. 2022, doi: 10.3390/DRONES7010002.

[7] B. Fu, S. Li, Z. Lao, B. Yuan, Y. Liang, W. He, W. Sun, and H. He, "Multi-sensor and multi-platform retrieval of water chlorophyll a concentration in Karst wetlands using transfer learning frameworks with ASD, UAV, and planet CubeSate reflectance data," *Sci. Total Environ.*, vol. 901, Nov. 2023, Art. no. 165963, doi: 10.1016/J.SCITOTENV.2023.165963.

[8] S. Clay, A. Peña, B. DeTracey, and E. Devred, "Evaluation of satellite-based algorithms to retrieve chlorophyll—A concentration in the Canadian Atlantic and Pacific oceans," *Remote Sens.*, vol. 11, no. 22, p. 2609, Nov. 2019, doi: 10.3390/RS11222609.

[9] J. E. O'Reilly and P. J. Werdell, "Chlorophyll algorithms for ocean color sensors–OC4, OC5 & OC6," *Remote Sens. Environ.*, vol. 229, pp. 32–47, Aug. 2019, doi: 10.1016/J.RSE.2019.04.021.

[10] B. A. Franz, S. W. Bailey, N. Kuring, and P. J. Werdell, "Ocean color measurements with the operational land imager on Landsat-8: Implementation and evaluation in SeaDAS," *J. Appl. Remote Sens.*, vol. 9, no. 1, Mar. 2015, Art. no. 096070, doi: 10.1117/1.JRS.9.096070.

[11] N. Pahlevan et al., "ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters," *Remote Sens. Environ.*, vol. 258, Jun. 2021, Art. no. 112366, doi: 10.1016/J.RSE.2021.112366.

[12] Q. Vanhellemont and K. Ruddick, "Advantages of high quality SWIR bands for ocean colour processing: Examples from Landsat-8," *Remote Sens. Environ.*, vol. 161, pp. 89–106, May 2015, doi: 10.1016/J.RSE.2015.02.007.

[13] I. D. Joshi, D. Stramski, R. A. Reynolds, and D. H. Robinson, "Performance assessment and validation of ocean color sensor-specific algorithms for estimating the concentration of particulate organic carbon in oceanic surface waters from satellite observations," *Remote Sens. Environ.*, vol. 286, Mar. 2023, Art. no. 113417, doi: 10.1016/J.RSE.2022.113417.

[14] J. Ryu, S. Son, C. O. Jo, H. Kim, Y. Kim, S. H. Lee, and H. Joo, "Revised chlorophyll–A algorithms for satellite ocean color sensors in the East/Japan sea," *Regional Stud. Mar. Sci.*, vol. 60, Jun. 2023, Art. no. 102876, doi: 10.1016/J.RSMA.2023.102876.

[15] S. P. Tiwari and P. Shanmugam, "An evaluation of models for the satellite-estimation of phytoplankton absorption coefficients in coastal/oceanic waters," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 364–371, Jan. 2014, doi: 10.1109/JSTARS.2013.2252151.

[16] P. J. Werdell and S. W. Bailey, "An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation," *Remote Sens. Environ.*, vol. 98, no. 1, pp. 122–140, Sep. 2005, doi: 10.1016/J.RSE.2005.07.001.

[17] P. Shanmugam, "New models for retrieving and partitioning the colored dissolved organic matter in the global ocean: Implications for remote sensing," *Remote Sens. Environ.*, vol. 115, no. 6, pp. 1501–1521, Jun. 2011, doi: 10.1016/J.RSE.2011.02.009.

[18] S. P. Tiwari and P. Shanmugam, "An optical model for the remote sensing of coloured dissolved organic matter in coastal/ocean waters," *Estuarine, Coastal Shelf Sci.*, vol. 93, no. 4, pp. 396–402, Jul. 2011, doi: 10.1016/J.ECSS.2011.05.010.

[19] Y.-H. Ahn and P. Shanmugam, "Derivation and analysis of the fluorescence algorithms to estimate phytoplankton pigment concentrations in optically complex coastal waters," *J. Opt. A, Pure Appl. Opt.*, vol. 9, no. 4, pp. 352–362, Mar. 2007, doi: 10.1088/1464-4258/9/4/008.

[20] Y. He, C. Chen, B. Li, and Z. Zhang, "Prediction of near-surface air temperature in glacier regions using ERA5 data and the random forest regression method," *Remote Sens. Appl., Soc. Environ.*, vol. 28, Nov. 2022, Art. no. 100824, doi: 10.1016/J.RSASE.2022.100824.

[21] S. Giri, Y. Kang, K. MacDonald, M. Tippett, Z. Qiu, R. G. Lathrop, and C. C. Obropta, "Revealing the sources of arsenic in private well water using random forest classification and regression," *Sci. Total Environ.*, vol. 857, Jan. 2023, Art. no. 159360, doi: 10.1016/J.SCITOTENV.2022.159360.

[22] M. M. Rahman, M. Shafiullah, M. S. Alam, M. S. Rahman, M. A. Alsanad, M. M. Islam, M. K. Islam, and S. M. Rahman, "Decision tree-based ensemble model for predicting national greenhouse gas emissions in Saudi Arabia," *Appl. Sci.*, vol. 13, no. 6, p. 3832, Mar. 2023, doi: 10.3390/APP13063832.

[23] T. Zhang, W. Lin, A. M. Vogelmann, M. Zhang, S. Xie, Y. Qin, and J. Golaz, "Improving convection trigger functions in deep convective parameterization schemes using machine learning," *J. Adv. Model. Earth Syst.*, vol. 13, no. 5, May 2021, Art. no. e2020MS002365, doi: 10.1029/2020MS002365.

[24] X. N. Bui, H. Nguyen, and P. Soukhanouvong, "Extra trees ensemble: A machine learning model for predicting blast-induced ground vibration based on the bagging and sibling of random forest algorithm," in *Proc. Int. Conf. Geotechnical Challenges Mining, Tunneling Underground Infrastructures*, vol. 228, Dec. 2022, pp. 643–652, doi: 10.1007/978-981-16-9770-8_43.

[25] M. S. Alam, F. S. Al-Ismail, M. S. Hossain, and S. M. Rahman, "Ensemble machine-learning models for accurate prediction of solar irradiation in Bangladesh," *Processes*, vol. 11, no. 3, p. 908, Mar. 2023, doi: 10.3390/PR11030908.

[26] M. Shafiullah, K. A. AlShumayri, and M. S. Alam, "Machine learning tools for active distribution grid fault diagnosis," *Adv. Eng. Softw.*, vol. 173, Nov. 2022, Art. no. 103279, doi: 10.1016/J.ADVENGSOFT.2022.103279.

[27] J. Pérez-Rodríguez, F. Fernández-Navarro, and T. Ashley, "Estimating ensemble weights for bagging regressors based on the mean–variance portfolio framework," *Expert Syst. Appl.*, vol. 229, Nov. 2023, Art. no. 120462, doi: 10.1016/J.ESWA.2023.120462.

[28] T. Zhang, J. Zheng, and Y. Zou, "Weighted voting ensemble method for predicting workpiece imaging dimensional deviation based on monocular vision systems," *Opt. Laser Technol.*, vol. 159, Apr. 2023, Art. no. 109012, doi: 10.1016/J.OPTLASTEC.2022.109012.

[29] G. Chen, X. Jiang, Q. Lv, X. Tan, Z. Yang, and C. Y.-C. Chen, "VAERHNN: Voting-averaged ensemble regression and hybrid neural network to investigate potent leads against colorectal cancer," *Knowl.-Based Syst.*, vol. 257, Dec. 2022, Art. no. 109925, doi: 10.1016/J.KNOSYS.2022.109925.

MD. SHAFIUL ALAM received the B.Sc. degree in electrical and electronic engineering (EEE) from the Dhaka University of Engineering and Technology, Gazipur, Bangladesh, the M.Sc. degree in EEE from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, and the Ph.D. degree in electrical engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. In August 2008, he started his career as a Faculty Member with the Department of EEE, International Islamic University Chittagong (IIUC), Bangladesh, where his highest rank was an Associate Professor. From March 2020 to March 2022, he was a Postdoctoral Fellow with the K. A. CARE Energy Research and Innovation Center (ERIC), KFUPM. He is currently a Research Engineer III with the Applied Research Center for Environment and Marine Studies (ARCEMS), Research Institute, KFUPM. He worked on several funded projects during the Ph.D. research. His research interests include energy and environment, greenhouse gas emission management, data analysis, renewable energy sources integration into the utility grids, ac/dc microgrids, high-voltage dc transmission, voltage source converter control, fault current limiter, optimization algorithms, fuzzy logic, neural networks, and machine learning. He is a member of the Institution of Engineers Bangladesh. He was a recipient of the best paper awards at many IEEE international conferences.

SURYA PRAKASH TIWARI received the B.Sc. degree in mathematics and physics and the M.Sc. degree in physics from Gorakhpur University, Uttar Pradesh, India, in 2005 and 2007, respectively, the M.Tech. degree in remote sensing and GIS from SRM University, Chennai, India, in 2009, and the Ph.D. degree in ocean engineering from the Indian Institute of Technology Madras, Chennai, India, in 2013. From 2013 to 2014, he was a Postdoctoral Research Associate with The City University of New York, USA. From 2014 to 2017, he was a Postdoctoral Fellow with the Biological and Environmental Science and Engineering Division, Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. He is currently a Research Scientist-II with the Applied Research Center for Environment and Marine Studies, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He has published more than 43 research articles which include peer-reviewed journal articles, full conference proceedings, and other conference papers. His research interests include bio-geo-optics, marine biology, oceanography, ocean color remote sensing, radiative transfer theory, underwater visibility, phytoplankton blooms, primary productivity, instrument design, and other subdisciplines. He was a recipient of the IOCCG Scholarship for the IOCCG Summer Lecture Series, in 2014. He is a peer-reviewer for several international journals, such as *Remote Sensing of Environment*, *Ocean Science Journal*, IEEE TRANSACTION ON GEOSCIENCES, *Optics Express*, and *Remote Sensing*.

SYED MASIUR RAHMAN received the B.S. degree in civil engineering from the Bangladesh University of Engineering and Technology, Bangladesh, in 2000, and the master's degree in city and regional planning and the Ph.D. degree in civil engineering from the King Fahd University of Petroleum and Minerals, Saudi Arabia, in 2004 and 2010, respectively.

Since 2022, he has been a Research Engineer I with the Applied Research Center for Environment and Marine Studies, King Fahd University of Petroleum and Minerals. He is the author of three books, more than 60 articles, and one invention. His research interests include sustainable transportation, machine learning-based modeling, environmental impact assessment, and climate change studies. He is an Editorial Board Member of the *Journal of Transportation and Logistics*. He was awarded the Certificate of Distinction-King Fahd University of Petroleum and Minerals, from 2005 to 2006, for his contribution to the project "Environmental Impact Assessment–Contract Area A in Northern Part of Rub Al-Khali, Stage 1-2D and 3D Seismic Operation." He was also awarded the same title for the academic year, from 2011 to 2012.