Journal of Scientific & Industrial Research
Vol. 83, February 2024, pp. 174-182
DOI: 10.56042/jsir.v83i2.3684

Application of Machine Learning Techniques on Multivariate Ocean Parameters

Sivasankari M[1]*, R Anandan[1] & G Rajesh[2]

[1]Department of CSE, Vels Institute of Science, Technology and Advanced Studies, Chennai 600 117, India

[2]Department of Information Technology, Anna University, Chennai 600 044, India

Received 06 July 2023; revised 11 September 2023; accepted 19 January 2024

Locating potential fishing zones is a requirement for aquaculture. The existence of Potential Fishing Zones is dependent on several ocean parameters. The goal of this paper is to analyze the various techniques to identify the Potential and Non-Potential Fishing Zones based on multivariate parameters like Sea Surface Temperature, Chlorophyll and Salinity. Regression-based model, that is derived from Random Forest methodology has been developed in order to process the dependent parameters, and the outcome is compared with other methodologies namely Support Vector Method (SVM), k-Nearest Neighbor (k-NN), and Decision Trees. The data used for this analysis is the California Cooperative Oceanic Fisheries Investigations (CalCOFI) dataset, which represents the hydrographic data since 1949, of the Californian Current System. The overall efficiency of each method is captured using Accuracy, Prediction Precision, and Area under the ROC Curve (AUC), F1 Score and Recall values. The test accuracy of the proposed system based on Random Forest has been recorded as 96.21 as compared to other methodology. The SVM, k-NN and Decision Tree methods have recorded 79.21, 93.14 and 96.11, respectively. The evidence based on the prediction outcome has affirmed the relationship between chlorophyll and SST, as well as with the Salinity data.

Keywords: Chlorophyll, Fishing zone, Regression analysis, Sea surface temperature

Introduction

A large population in the world relies on Fisheries as it plays a vital role in sustenance for many individuals and communities. As per, Food and Agriculture Organization (FAO) of the United Nations, the fish production globally had reached a peak of 171 million tons in 2016, out of which most of it was from aquaculture.[1,2] The capture fishery has been static and the average global fish food consumption has increased by 3.2% which has exceeded the population growth (1.6%). The official update from FAO indicates that around 59.6 million people are involved in capture and aquaculture fisheries, out of which 40.3 million in capture fisheries. Hence there have been several researches[3,4] in identifying the Potential Fishing Zones, as this will eventually help us to tap the resources in an efficient manner. Recent developments in Remote Sensing area and also availability of cutting-edge sensors along with ocean parameters collection techniques have aided us in development of efficient methods to perform ocean parameter analysis there by mapping the potential fishing zones (PFZ). The PFZs are affected by the parameters like Sea Surface Temperature, Chlorophyll, Salinity, Oxygen nutrients, Currents, pH and even algal types. Concentration levels of Chlorophyll can be measured via Remote Sensing. The institute data has an increased reflection at the sensors, when there is increased chlorophyll content.

In the initial days of research, the potential fishing zones were identified by relying on Sea Surface Temperature. The National Oceanic and Atmospheric Administration- Advanced Very High-Resolution Radiometer (NOAA - AVHRR) had applied the data for identification of Thermal Fronts, which in turn was projected as the Potential Fishing Zone. Several PFZ validation programs were then introduced, which eventually received a boost when the techniques on retrieval and analysis of Ocean Color Monitoring (OCM) were developed.[5] This gave way to potential identification of fishing zone, based on the turbid conditions, sediment concentration, Chlorophyll content and the presence of aquatic weeds and algae. The PFZ data was generated based on the Sea Surface Temperature (SST) and Chlorophyl content. In this paper, Salinity (SSS) has also been taken into consideration which improves the accuracy of the data. Based on the multi variants, the regression model has been developed and the potential fishing zones has been identified.

*Author for Correspondence
E-mail: sivasankari21@gmail.com

Related Work

As more developments have occurred in Remote Sensing and Sensors technology, several researches have been done in the retrieval of the ocean parameters, analyzing it and making future prediction. Park *et al*.[6] studied on the retrieval of Sea surface temperature using empirical method. The fast Radiative Model (RTM) has been developed based on the regression between the incremental data and scaling procedures. This gave good results for the nighttime data. The fish forecast research done by Anis *et al*.[7] proposed a Data Assimilation (DA) based Daily Forecast for Fish Catch which exhibited significant perfection over ML based methods, but it gave better results for smaller amount of data. As per Goldstein[8], there are some guiding principles to select the ML methods. As per his suggestion, Bayesian Networks can be considered for Probabilistic answers and Genetic Algorithm for multiple free parameters and a fixed equation. Unfortunately, we cannot provide definitive answers for the methods those needs to be empirically analyzed to get an optimal solution. The amount of data that needs to be selected to develop a strong predictor is also a point for research. As per Beuzen *et al*.[9] this would be dependent on various factors like network complexity, degree of freedom/independent variable and finally the signal to noise ratio of the dataset. The various works related to SVM, k-NN, Decision Tree and Random Forest is given in the methodology section.

Fuzzy logic methods have been used[10] in Sea Surface Chlorophyll image against the Sea Surface Temperature Image. The Fuzzy C Means algorithm which is an unsupervised classification method has improved the decision-making process. The image fusion concept has been implemented by the logical AND Operator. Random Forest Methods have been used along with Multi Type Predictor Variables (MTVRF) which helps in establishing the nonlinear relationship between the attributes.[11] Random Forest regressions have acceptance to multi collinearity which helps in processing high dimension data sets, but this study was involved in a single land cover types and was not covered for different latitude and longitude, which will expand the generalization capability.

Gilerson *et al*. used a detailed artificial data set of spectral reflectance and intrinsic optical characteristics pertaining to multiple sampling points, and also a very coherent in situ data set of several lakes in Nebraska, USA, to test MERIS' two-band and three-band red-NIR methodologies.[12]

Marine science and machine intelligence have been correlated, which has helped in data driven decision making using de nova data. Newer sensor technologies also help in quicker resolution of huge data.[18]

Li *et al*.[19] proposed the usage of remote sensing technique to understand the Chlorophyll concentration fluctuation and the various factors accompanying it. This used the MODIS data, using Hovmeoller data analysis method. The study done by Wu & Li [20] has also shown how the random forest aids in downscaling while using the multi scale parameters. However, there have been few discrepancies in outcome, which has a negative effect on downscaling results.

Daqamseh *et al*.[21] has also utilized MODIS Data and tried to compare the parameters to understand the fish aggregation patterns. This was more of a seasonal pattern study, the data for round the year has not been identified individually.

From the literature, it is observed that the processing of multivariate parameters for the Ocean parameters has not been effective. The amount of data that needs to be selected to develop a strong predictor is also a point for research to be solved. There is also lack of trustworthy data pertaining to fishery sources. In fisheries and aquaculture, we also notice a lack of a multidisciplinary approach in fish culture in terms of inadequate attention to environmental, economic, social, and gender issues in fisheries and aquaculture, as well as insufficient HRD and highly specialized manpower in various disciplines.[17]

Scope of Work from the Research Gap

Contribution of this work is to improve the performance of prediction algorithm for high volume of data with multi variant parameters. A state space model has been created for the Data assimilation, which is further processed using Gradient Boost Decision Trees. Further to it, the clustering would enable to predict the target variables based on the distant metric.

Data Sets and Methods

Data Sets

The data has been taken from the CalCOFI dataset (1949–2019) which includes the information from the 500 sampling stations as shown in Table 1. The

Table 1 — CalCOFI cruise 1011 Hydrographic report

DEPTH		TEMP		POT TEMP	SALINITY		SIGMA	SVA	DYN HT	OXYGEN	OXY		SiO_3	PO_4	NO_3	NO_2	NH_4	CHL-A	PHAEO	PRES	SAMP
m		°C		°C			THETA			mL/L	%		uM/L	uM/L	uM/L	uM/L	uM/L	ug/L	ug/L	db	
0	ISL	15.57		15.57	33.233		24.486	343.7	0	5.83	102.6		1.6	0.32	0.8	0.07	0.08	0.58	0.2	0	—
2		15.57		15.57	33.233		24.486	343.8	0.007	5.83	102.6		1.6	0.32	0.8	0.07	0.08	0.58	0.2	2	220
10		15.57		15.57	33.232		24.485	344.1	0.034	5.84	102.7		1.6	0.33	0.8	0.07	0.1	0.68	0.12	10	219
20		15.57		15.57	33.233		24.486	344.3	0.069	5.82	102.4		1.5	0.32	0.8	0.07	0.1	0.61	0.17	20	218
30	ISL	15.52	D	15.52	33.233	D	24.498	343.5	0.103	5.82	102.3		1.4	0.34	0.8	0.07	0.1	0.63	0.17	30	—
31		15.53		15.53	33.235		24.497	343.6	0.107	5.82	102.3		1.4	0.34	0.8	0.07	0.1	0.63	0.17	31	217
40		13.33		13.32	33.086		24.847	310.5	0.136	5.77	96.9		2.9	0.58	3.4	0.43	0.21	0.53	0.34	40	216
50	ISL	12.75	D	12.74	33,133	D	24.998	296.3	0.166	5.52	91.6		4.2	0.75	6.4	0.48	0.1	0.36	0.35	50	—
51		12.81		12.8	33.138		24.990	297.1	0.169	5.49	91.2		4.4	0.76	6.7	0.49	0.08	0.34	0.35	51	215
60		11.89		11.88	33.093		25.131	283.8	0.196	5.3	86.3		6.5	0.88	9.2	0.03	0.05	0.24	0.22	60	214
71		10.84		10.83	33.105		25.330	265.0	0.226	5.12	81.5		9.9	1.05	12.1	0.02	0.05	0.14	0.12	71	213
75	ISL	10.45	D	10.44	33.128	D	25.416	256.8	0.236	4.87	76.9		12	1.18	14.3	0.02	0.04	0.11	0.1	75	—
85		10.2		10.19	33.413		25.681	231.9	0.261	4.21	66.2		17.2	1.5	19.6	0.01	0	0.04	0.07	85	212
100		9.24		9.23	33.498		25.906	210.6	0.294	3.92	60.4		21.2	1.6	21.7	0	0	0.01	0.05	100	211
120		9.25		9.24	33.681		26.048	197.6	0.335	3.24	50		26.1	1.87	25.8	0.01	0	0.01	0.04	121	210
125	ISL	9.23	D	9.22	33.733	D	26.092	193.5	0.344	3.09	47.7		27	1.91	26.5	0.01	0	0.01	0.04	126	—
140		9.08		9.06	33.800		26.168	186.5	0.373	2.75	42.3		29.2	2	27.8	0	0	0.01	0.03	141	209
150	ISL	8.86	D	8.84	33.861	D	26.251	178.8	0.391	2.73	41.8		30.8	2	28.1	0	0.02	0.01	0.03	151	—
171		8.31		8.29	33.919		26.381	166.7	0.427	2.69	40.7		33.7	2.01	28.6	0	0.05	0	0.03	172	208
200		8.03		8.01	33.983		26.474	158.3	0.475	2.74	41.2		36	2	28.6	0	0	0	0.02	201	207
231		7.77		7.75	34.014		26.537	152.8	0.523	2.3	34.4		41	2.17	30.7	0	0	—	—	232	206
250	ISL	7.64	D	7.62	34.054	D	26.587	148.3	0.551	1.9	28.3		44.5	2.31	32.4	0	0	—	—	251	
271		7.5		7.47	34.072		26.622	145.3	0.582	1.5	22.3		48.3	2.46	34.2	0	0	—	—	273	205
300	ISL	7.05	D	7.02	34.069	D	26.683	139.8	0.623	1.42	20.9		52.3	2.53	35.3	0	0	—	—	302	
320		6.87		6.84	34.068		26.707	137.7	0.651	1.37	20.1	D	55	2.56	35.8	0	0	—	—	322	204
381		6.12		6.09	34.089		26.822	127.1	0.732	1.06	15.3		66.1	2.74	38.6	0	0	—	—	383	203
400	ISL	5.87	D	5.84	34.088	D	26.853	124.2	0.756	1.02	14.6		69	2.77	39.1	0	0	—	—	403	
440		5.55		5.51	34.097		26.899	120.0	0.805	0.92	13.1		74.8	2.84	40.1	0	0	—	—	443	202
500	ISL	5.24	D	5.2	34.167	D	26.992	111.7	0.874	0.57	8		84.1	2.98	41.7	0	0	—	—	503	
516		5.15		5.11	34.183		27.015	109.6	0.892	0.48	6.8		86.6	3.02	42.1	0	0	—	—	520	201

California Cooperative Oceanic Fisheries Investigations (CalCOFI) works in partnership with California Department of Fish and Wildlife, NOAA Fisheries Services and Scripps Institution of Oceanography. It has the longest and extensive Oceanographic data collection. The latitude spanning from 31° 10.1' N to 37° 50.7' N and the longitude spans from 117° 17.0' W to 124° 54.2' W. It covers the pacific region. Conductivity Temperature Depth (CTD) instrument with an Emblem, also called Rosette is deployed in each cruise assigned to a station. The depth of installation ranges from 20 to 55 meters at the close interval of nearly 10 meters.

Salinity samples are collected using rosette bottles and examined at sea using a Guildline model 8410 Portasal salinometer. The salinity value is then calculated using Practical Salinity Scale and rounded to 3 decimal places. The temperature reported has been rounded to the nearest hundredth of a degree Celsius. Chlorophyll data (μg/L) was measured at sea from the samples taken from the top 200 meters. Sample Data is shown below in Table 1 for reference (Reference from the CalCOFI Report)[13] and Fig. 1 represents the Location of the Sampling Stations for the CalCOFI Data collected.

There has been significant growth in the area of Ocean Parameter Collection and analysis as an outcome of Geographic Information System (GIS), Sea Data Collection approach, Sensors and Image Processing techniques. Several researches are in place to model the relationship between the different parameters. The Salinity data measurement looks at models like MODIS Aqua SSS Algorithm[13] which was mainly focused on MODIS Image Radiance and measured Salinity. The Salinity was retrieved using a Multiple Linear Regression Model which factors the relation between Ocean Salinity Brightness and In-situ data measurements, i.e., the predicted value using the Brightness Bands were validated using in-situ measurements.

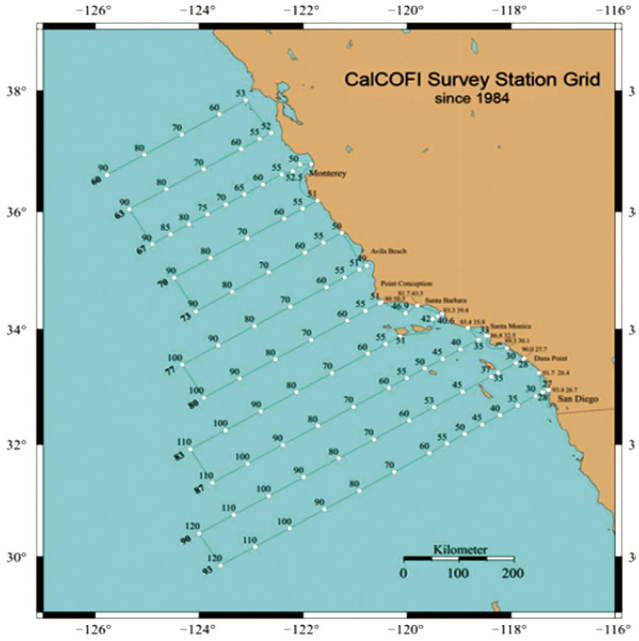

Fig. 1 — The location of the sampling stations for the CalCOFI data

$$\begin{bmatrix} SSS_1 \\ SSS_2 \\ \vdots \\ SSS_n \end{bmatrix} = \begin{bmatrix} 1 & B_{81} & B_{82} & \dots & \dots & B_{8P} \\ 1 & B_{91} & B_{92} & \dots & \dots & B_{9P} \\ \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 1 & B_{n1} & B_{n2} & \dots & \dots & B_{9P} \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_P \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad \dots (1)$$

where, SSS is the observed Sea Salinity, B_nP is the value of the P^{th} Band, α_p Co-efficient of P^{th} Predictor, e_n Error term i.e. the aggregate value of the difference between Observed and Predicted value.

Several algorithms are in place to extract the SST Data. The research on Empirical Regression algorithm has been widely implemented. Another method used is a Hybrid SST Algorithm[6] which has better accuracy in night data. This takes into considerations the incremental values and scaling procedures. It is a combination of RTM Inversion method & regression methods. However, the day time measurements and mean square value still remained the same and not much improvement in efficiency.

Data Assimilation[14] can be used in cases where preceding knowledge about the monitoring target is available and integrated into modeling. Machine learning has improved performance when the amount of data is high. In case of Data Assimilation, a State Space Model (SSM) has to be created, which will incorporate the prior knowledge. This is then processed using Gradient Boost Decision Tree process, that creates a group of weak decision trees through boosting method which are the learners. The output from each tree is then aggregated, based on their weights and computed sum is derived using the below formula

$$S = \sum_{n=1}^{N} T\,(x, \theta_n) \quad \dots (2)$$

where, S – Sum of the weighted trees, T (x,θ_n) – Out of the n^{th} tree

This paper deals with Multiple Regression model, handling three variables SST, SSS (Salinity) and Chlorophyll.[5,7] Here the clustering can be used to predict most of the target variables/attributes and is based on distant metric. This model has three target variables, hence the last node i.e. the leaf node would be a vector with length 3.

Methodology

The regression methods involve the statistical technique to estimate the relation between the variables which has cause-effect relationship. When we have one dependent variable and multiple independent variables, then we use the technique of multivariate regression method.

The basic formula underlying the multivariate model is

$$y = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n + \varepsilon \quad \dots (3)$$

where, y is the dependent variable, x_i is the independent variable out of n variables, α is the parameter, ε is the error.

Where there is an assumption of linearity, absence of extreme values and lack of dependency ties between the independent variables results in the normal distribution of the variables. The frequency analysis is performed to check the availability of missing data. Then proceed to perform the univariate analysis for each variable and then proceed with the multivariate analysis. Scatter diagram was set up for linearity assumptions. Four models have been taken into consideration to predict the fishing zones based on the independent variable.

Support Vector Machine

This SVM algorithm helps in obtaining the ideal boundary among the possible output value. It has been proven to be effective in data mining.[3] The elements that make it effective is the Dual Theory, Maximal boundary and Kernel (that defines the dot product) trick. SVM makes use of multiple hyper planes and the optimal one is selected. The selection is based on the optimal split up of the data between the classes and there should increase split among the classes. The binary classification method is converted to multi

class classification using one against one method which was developed by Knerr. For a K way problem, we have to train $k(k-1)/2$ binary classifiers. Each of them will get samples from training set of both the classes and prediction is done. The selection of the best approach is done in terms of voting after applying it to some testing samples. The underlying assumption is that some classes cannot be divided.

Consider two subsets m, n and the training data is taken from these two classes. The below optimization question is resolved

$$\min_{w^{mn} b^{mn} \xi^{mn}} \frac{1}{2} (w^{mn})\hat{}Tw^{mn} + C \sum_t \xi^{mn} \quad \ldots(4)$$

Samples of m are positive and that of n is negative. The optimal weight coefficient vector (w) being the linear combination of training sample vector, α is optimal solution.

In this proposed work, the Sea Surface Temperature, the Salinity and Chlorophyll data are taken as sub vectors, hence we will be having 3 binary classifiers, predicting the Potential Fishing Zone.

K Nearest Neighbors

The next approach proposed in this paper is a rank-based k-NN method as shown in Fig. 2, for the multi label classification. For a test instance x, we identify the k nearest neighbors. These neighbors are then processed via a ranking model, and re-ranked based on their proximity to the true label set. These re-ranked neighbors are then assigned weights, using a weighted voting method which is an optimization method. Most commonly used method is the Hamming Method[4]

Let us assume, X as the domain of the instance in a multi label classification problem.

L is the set of labels denoted by $\{\lambda_1, \lambda_2 \ldots, \lambda_m\}$

The multi label training set is S = $\{x_1\ Y(x_1), x_2\ Y(x_2), \ldots., x_n Y(x_n)\}$

Y(x) is a subset of the label L

For a test instance x, we find the k-neighbors, N $\{N_k(x,j) \mid j = 1,2\ldots k\}$

This is then re-ranked to produce $\acute{N}_k(x,j)$ for j=1…k. This is then processed through weighted voting strategy to produce Y(x), the final prediction.

Algorithm for Prediction using k-NN:

Input:

Test Instance: x
No. of Neighbors: k
Ranking Model: M
S is the Training Set
Weight Scores – w

Output:

1. $\{N_k(x,j) \mid j = 1,2,..k\}$= k-NN Search (x,k,S)
2. $\{N_k(x,j) \mid j = 1,2,..k\}$= Re-rank the output from step 1
3. The result label set Y(x) = WeightedVote of the output of Step 2

Decision Tree

Decision Tree is looked at as a recursive partitioning activity of the instance space, which consists of root and test node. The accuracy of the model directly depends on the tree complexity, which in turn depends on the pruning method and the stopping criteria. The goal is also to minimize the generalization error. Here as well, we divide the instance space into hyperplanes. The algorithm used to implement this is C4.5 induction method[8] that aids in finding the threshold value of the continuous attribute. The splitting attribute is selected using information gain ratio. The training is performed until the training cases of the current node is in one single class, which initially starts from the root node holding the entire training set. The decision tree output is shown in Fig. 3.

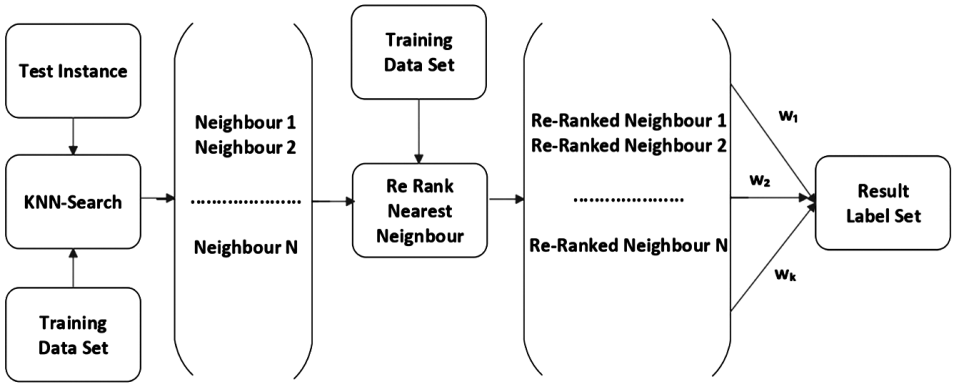

Fig. 2 — Testing strategy for k-NN

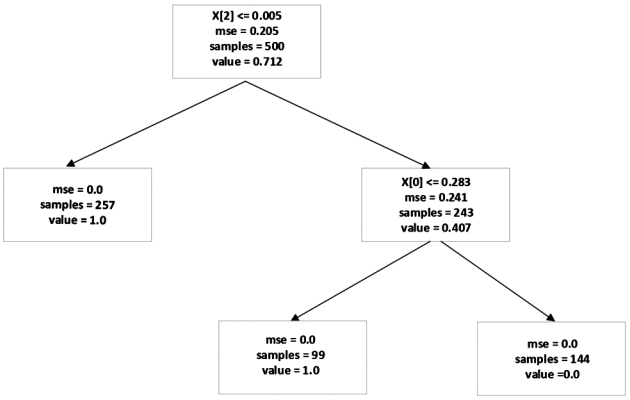

Fig. 3 — The decision tree output

Candidate Cut Point (CP) is identified using the below formula

$$CP_{ij} = \left\{\frac{x_{ij}+x_{(i+1)j}}{2}, i = 1, \ldots., n-1\right\} \quad \ldots (5)$$

A_j represents the continuous attribute $\{x_{ij}\ldots, x_{nj}\}$

Find Threshold Algorithm For C4.5:

1. A_{mn} is the Continuous Attribute Matrix
2. For each attribute in A_j, m = 1 to m

a. Sort the attribute values and find the cut off points (CP_{ij} to CP_{kj}

b. For each of the Cut Off point, CP_{ij}, i = 1…k

Calculate the information gain

c. Select the Optimal Cut off point

d. Calculate the Splitting Performance and Gain Ratio

3. Select the Optimal Attribute and its Cut Point

Modified Random Forest

A Supervised learning method can be used for both regression and classification. The trees are built and then trained using bagging method. It always searches for the best amongst the random sub set of features. It constructs the decision trees based on the features which are randomly selected and averages the result. The resulting output is then combined with an un-weighted voting[2]. The proposed algorithm is given below.

The training data is D = $\{(x_1y_1), (x_2y_2),.(x_Ny_N)\}$

$h_j(x)$ is the prediction of the response variable at j^{th} tree

Random Forest Algorithm:

For j = 1 to J

1. Select the Bootstrap sample (D_j)of size N from the training data
2. Run the Binary Recursive process in D_j and fit the tree

a. Then start with the single node observation

b. Repeat for each unsplit node, until it reaches the stooping criteria

1 Select r random predictors from p available predictors

2 Find the best fit among the r predictors

3 Split it further into two nodes using split method in step ii

Prediction at a new point x is as given below

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \quad \ldots (6)$$

The total number of randomly selected predictor variable (m) is based on the sample size (N) i.e., m = N/3. However, random forest is not sensitive to m, hence there is no necessity to worry about fine tuning.

Regression Analysis has been used to quantify the possible statistical impacts of the parameters in consideration namely SST, SSS and Chlorophyll on the determination of PFZ in the CalCOFI Dataset site. It has been tried to estimate the conditional outcome of the dependent variable created based on independent variables namely (X_1, X_2…. Xn). Thereby, the dependent variable's average value has been derived, assuming that both have a linear relationship. The following formula is used to calculate the Potential Fishing Zone, PFZ

$$PFZ = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \cdots.. + \varepsilon_i \; 1 \le i \le n_i \quad \ldots (7)$$

n is the sample size

$\alpha_0, \alpha_1, \alpha_{3,\ldots\ldots}\alpha_n$ are the regression coefficients

$$\begin{bmatrix} PFZ_1 \\ PFZ_2 \\ \vdots \\ PFZ_n \end{bmatrix} = \alpha_0 + \alpha_1 \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix} + \alpha_2 \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix} + \alpha_3 \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \ldots (8)$$

The matrix form of the prediction is given above where, T – Temperature in ℃, C – Chlorophyll in μg/L, S – Salinity, ε – Error in estimate generated by the model

Results and Discussion

The results are concluded based on the performance of the prediction models. The parameters taken into consideration are the ROC (which is the graphical representation of the threshold values) and AUC values. The comparison of the different methods can also be done using the parameters namely Accuracy, Recall, F1 Score and Precision. These are also calculated using the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as shown in Fig. 4.

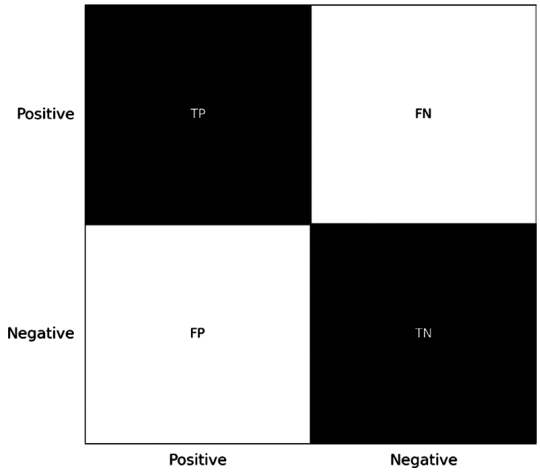

Fig. 4 — Confusion matrix with classification metrics: [Accuracy = (TP + TN)/(TP + TN + FP + FN), Precision = TP/(TP + FP), Recall = TP/(TP + FN), F-Score = (Precision × Recall)/ (Precision + Recall)]

Precision is referred to as the ability of the classifier to identify the negative data, whereas Recall would be for the positive samples. Accuracy is the ability to predict the correct outcome. The higher the value of these three, the better the predictivity and the accuracy of the model will be higher. F-Score depicts the harmonic mean value of the Precision and Recall values i.e., the weighted average. The Confusion matrix of all the four models researched in this paper is given in Fig. 5.

AUC is the Area under the ROC Curve and is a classification analysis metric which will help in determining the prediction efficiency of the models used. It identifies the average performance over all decision thresholds. The higher the AUC value and better is the model in prediction. For example, when AUC value is 0.8, it means that there is 80% chance in distinguishing the positive and negative classes.

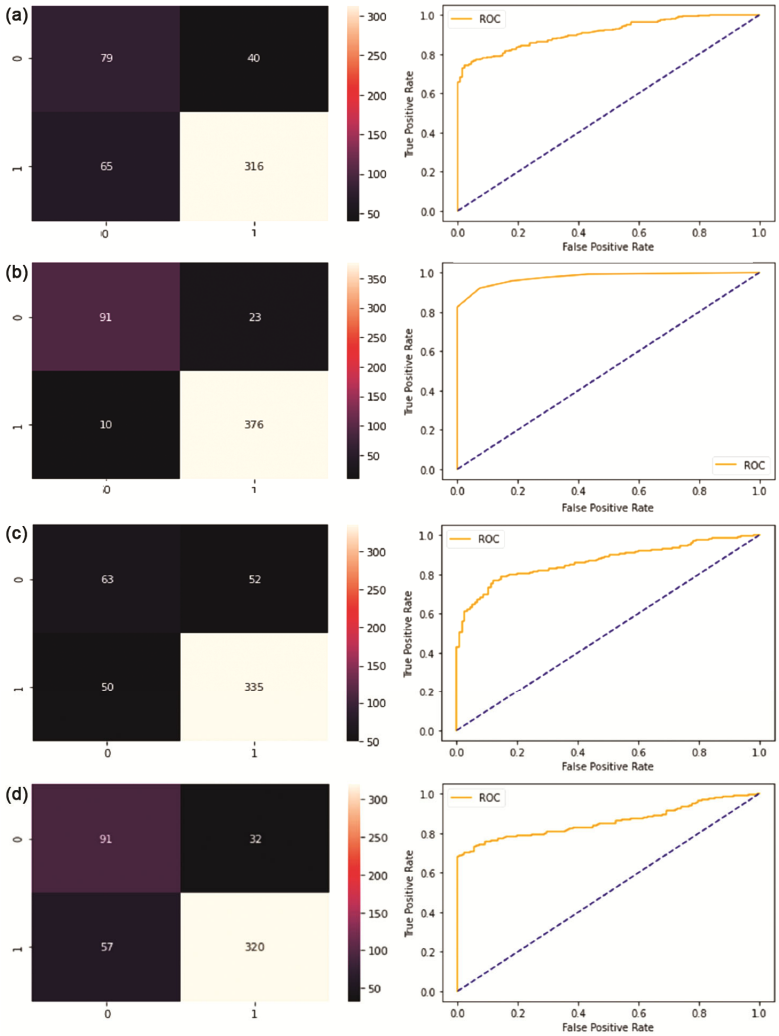

Fig. 5 — Confusion matrix and its ROC Curve Details for all the four models: (a) SVM, (b) Neighbor classifier, (c) Decision Tree, and (d) Random Forest

Table 2 — Performance analysis based on training data

Algorithm	AUC	Train Accuracy (%)	Recall	Precision	F1-Score
SVM	0.91	79.17	0.99	0.99	0.99
kNN	0.98	93.1	0.9	0.97	0.93369
Decision Tree	0.99	96.11	0.99	0.99	0.99
Random Forest	1	96.22	0.98	0.99	0 984975

Table 3 — Performance analysis based on test data

Algorithm	AUC	Test accuracy (%)	Recall	Precision	F1-Score
SVM	0.91	79.21	0.99	0.99	0.99
kNN	0.98	93.14	0.9	0.97	0.93369
Decision Tree	0.99	96.09	0.99	0.99	0.99
Random Forest	1	96.21	0.98	0.99	0.984975

Table 4 — Classification metric outcome for the models analyzed

Algorithm	TP	FP	TN	FN
SVM	79	65	316	40
kNN	91	10	376	23
Decision Tree	148	2	349	1
Random Forest	256	4	737	3

The sharpness of the curve's 'elbow' indicates a better separation between the two classes for a binary classifier. The ROC Curve of our prediction models is given in Fig. 5, the AUC values and test accuracy are mentioned in Table 2 and 3. The true positive, true negative, false positive and false negative is given in Table 4.

SVMs are popular for their exceptional performance with limited data. However, we observe a poor performance from SVM in our case. For larger amounts of data, SVM might not perform better against random forests.

The train and test metrics (refer Table 2 and 3) are not very different. This indicates better generalization by the models and hence, overall, better performance.

Conclusions

The analysis has been done for the identification of the potential fishing zone based on the multivariate Ocean parameters. Regression analysis has been used here to predict the output. The metric output for the four models, i.e. SVM, k-NN, Decision Tree and Modified Random Forest, AUC value for Random Forest is 1.00 which is best among the four. The precision, recall and F1 Score are 0.98, 0.99 and 0.9849 respectively, which is also the best among the four models. Hence the prediction output is best for the Modified Random Forest Method. The higher the SST value and Chlorophyll in the final model, the possibility of good PFZ is more, based on the positive coefficient values generated by the model. This is also based on the fish catch history of the location. On the contrary, lower salinity value is associated with Non-Potential Fishing Zone (NPFZ) and is based on the negative coefficient parameter. The model is based on the reflectance range and the intrinsic optical properties of the study site. Hence forecasted values of the calibrated model, i.e. the predicted PFZ mapping will vary as location varies. The result of the study confirmed a relationship between chlorophyll and SST, as well as with the Salinity data. The prediction ability of the proposed model is at 96.21%. The results have been processed based on the data collected so far whose size is limited and availability is restricted. Further refinement will be targeted for future scope of research.

Reference

1 Gilerson A, Gitelson A, Zhou J, Gurlin D, Moses W, Ioannou I & Ahmed S, Algorithms for remote estimation of chlorophyll-α in coastal and inland waters using red and near infrared bands, *Optics Express*, **18** (2010) 24109–24125, doi: 10.1364/OE.18.024109.

2 Cutler A, Cutler D & Stevens J, Random Forests, **411** (2006) 422–432, doi: 10.1016/S0076-6879(06)11023-X.

3 Zhang Y, Support vector machine classification algorithm and its application, information computing and applications, *ICICA 2012*, *Commun Comput Inform Sci*, **308** (2012) 179–186, doi: 10.1007/978-3-642-34041-3_27.

4 Chiang T-H, Lo H-Y & Lin S-D, A Ranking-based KNN approach for multi-label classification, *Proc Asian Conf Machine Learning, PMLR* **25** (2012) 81–96.

5 Hu C, Lee Z & Franz B, Chlorophyll a algorithm for oligotrophic oceans: A novel approach based on three-band reflectance difference, *J Geophys Res*, **117** (2012), doi: 10.1029/2011JC007395.

6 Park K, Lee E & Woo H, Comparison of hybrid sea surface temperature (SST) with empirical regression SST in the seas around Korea, *IEEE Int Geosci Remote Sens Sympos* (IGARSS), Beijing, (2016) 4016–4018, doi: 10.1109/IGARSS.2016.7730044.

7 Cherfi A, Nouira K & Ferchichi A, Very fast c4.5 decision tree algorithm, *Appl Artif Intell*, **32(2)** (2018) 119–137.

8 Goldstein E B, Coco G & Plant N G, A review of machine learning applications to coastal sediment transport and morphodynamics, *Earth-Sci Rev*, **194** (2019) 97–108, doi: 10.31223/osf.io/cgzvs.

9 Beuzen T, Splinter K D, Marshall L A, Turner I L, Harley M D & Palmsten M L, Bayesian networks in coastal engineering: Distinguishing descriptive and predictive applications, *Coastal Eng*, **135** (2018) 16–30.

10 El Abidi Z, Minaoui K, Tamim A & Laanaya H, A simple fusion approach of chlorophyll images and sea surface temperature images for improving the detection of Moroccan coastal upwelling, *IEEE Int Geosci Remote Sens Sympos*, (2018) 7208–7211, doi: 10.1109/IGARSS.2018.8518521.

11 Cui S, Zhou Y, Wang Y & Zhai L, Fish detection using deep learning, *Appl Comput Intell Soft Comput*, (2020) 1–13, doi: 10.1155/2020/3738108.

12 Majumder S, Maity S, Balakrishnan Nair T M, Bright R P, Kumar M N, Shwetha N & Kumar N, Potential fishing zone characterization in the Indian ocean by machine learning approach, *Adv Intell Syst Comput*, **2** (2021) 43–54.

13 Horiuchi Y, Kokaki T, Kobayashi T & Ogawa T, Data assimilation versus machine learning: Comparative study of fish catch forecasting, *OCEANS 2019*, (2019) 1–5, doi: 10.1109/OCEANSE.2019.8867066.

14 www.calcofi.com (accessed on 30 October 2022).

15 Gladju J, Kamalam B S & Kanagaraj A K, Applications of data mining and machine learning framework in aquaculture and fisheries*:* A Review, *Smart Agric Technol*, **2** (2022) 100061–14, doi: 10.1016/ j.atech.2022.100061.

16 ACTION, S.I. World fisheries and aquaculture, *Food Agric Organization*, (2020), 1–244.

17 Beyan C & Browman H I, Setting the stage for the machine intelligence era in marine science, *ICES J Mar Sci*, **77(4)** (2020) 1267–1273, doi: 10.1093/icesjms/fsaa084.

18 Li W, El-Askary H, ManiKandan K P, Qurban M A, Garay M J & Kalashnikova O V, Synergistic use of remote sensing and modeling to assess an anomalously high chlorophyll-a event during summer 2015 in the south-central red sea, *Remote Sens*, **9(8)** (2017) 778, doi: 10.3390/rs9080778.

19 Wu H & Li W, Downscaling land surface temperatures using a random forest regression model with multitype predictor variables, *IEEE Access*, **7** (2019) 21904–21916, doi: 10.1109/ACCESS.2019.2896241.

20 Daqamseh S T, Al-Fugara A, Pradhan B, Al-Oraiqat A & Habib M, MODIS derived sea surface salinity, temperature, and chlorophyll-a data for potential fish zone mapping: West red sea coastal areas, Saudi Arabia, *Sensors*, **19(9)** (2019) 2069, doi: 10.3390/s19092069.