Physicochemical Habitat Traits Preferred by Small Indigenous Fish (Chanda Nama) in Indian River Discerning through Machine Learning

Rohan Kumar Raman (rohan4741@gmail.com)
ICAR-RCER: ICAR Research Complex for Eastern Region

Archan Kanti Das
Central Inland Fisheries Research Institute

Ranjan Kumar Manna
Central Inland Fisheries Research Institute

Sanjeev Kumar Sahu
Central Inland Fisheries Research Institute

Basant Kumar Das
Central Inland Fisheries Research Institute

Research Article

Keywords: Chanda nama, physicochemical parameters, random forest, artificial neural network, support vector machine, k-nearest neighbors, feature selection, Krishna River

DOI: https://doi.org/10.21203/rs.3.rs-591781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Physicochemical habitat traits preferred by small indigenous fish (*Chanda nama*) in Indian River discerning through machine learning

Rohan Kumar Raman\(^1\)^\(^2\), Archan Kanti Das\(^1\), Ranjan Kumar Manna\(^1\), Sanjeev Kumar Sahu\(^1\) and Basanta Kumar Das\(^1\)

\(^1\)ICAR-Central Inland Fisheries Research Institute, Barrackpore, Kolkata-700120, India

\(^2\)ICAR-Research Complex for Eastern Region, Patna-800014, India

Corresponding author: Email: basantakumard@gmail.com
Abstract

Physicochemical traits of river influence the habitat of fish species in aquatic ecosystems. Fish showed a complex relationship with aquatic factors in river. Machine learning (ML) modeling is a useful tool to established relationship between complex systems. This study identified the preferred habitat indicators of *Chanda nama* (a small indigenous fish), in the Krishna River, of peninsular India, using machine learning modeling. Data were observed on *Chanda nama* fish distribution (presence/absence) and associated ten physical and chemical parameters of water at 22 sampling sites on the river during year 2001-02. Machine learning models such as random forest (RF), artificial neural network (ANN), support vector machine (SVM), k-nearest neighbors (KNN) used for the classification of *Chanda nama* distribution in the river. The ML model efficiency was evaluated using classification accuracy (CCI), Cohen’s kappa coefficient (k), sensitivity, specificity and receiver-operating-characteristics (ROC). Results showed that random forest is the best model with 82% accuracy, CCI (0.82), k (0.55), sensitivity (0.57), specificity (0.76) and ROC (0.72) for *Chanda nama* distribution (presence/absence) in the Krishna River. Random Forest model identified three preferred physicochemical habitat traits like altitude, temperature and depth for *Chanda nama* distribution in the Krishna River, India. This study will be helpful for researcher and policy maker to understand the important habitat physicochemical traits for sustainable management of small indigenous fish (*Chanda nama*) in the river system.

Key words: *Chanda nama*, physicochemical parameters, random forest, artificial neural network, support vector machine, k-nearest neighbors, feature selection, Krishna River.

1 Introduction

River and its fauna are of major concern worldwide for conservation priorities due to continuously diminishing river habitat and species. Aquatic communities are being affected by habitat degradation in river channels, in spite of persistent controlling the water pollution by many countries (Aarts et al. 2004). Fish species presence in the river is considered as biological indicators (Plafkin et al. 1989) and a key component in environmental planning (Schiemer 2000). Fishes are very much sensitive to physical and chemical changes (i.e., pH, dissolved oxygen, temperature,
etc.), flow and natural changes (Moore et al. 1995). Physical and chemical changes alter its abundance and richness in aquatic system (Oberdorff et al. 1995; Angermeier et al. 2004; Laws 2000). Fresh water modeling approach through reliable model has made the substantial improvements in ecosystems modeling to understand the relationship between ecosystems identity (Recknagel 2002). Ecosystem parameters nonlinear interactions that vary in both time and space put challenges in front of researchers to understand the complex ecological pattern (Levin 1998). Machine learning (ML) modeling in ecological informatics is an interdisciplinary framework that explains the complex ecological pattern (Lehikoinen et al. 2019 ). Machine learning are widely applied in different fields of ecology such as species distribution modeling for conservation and management planning, modeling presence absence of fish, under water fish species classification, chlorophyll concentration in fresh water (Tirelli et al 2009; Rathi et al. 2017; Maier and Keller 2019). The advantage of machine learning methods is that it does not require the restrictive assumptions of the conventional models which made it more sufficient towards prediction and explain the ecological pattern (Olden and Jackso 2002; Elithet al. 2006; Olden et al. 2008). Machine learning works in two ways (a) supervised approach where relationship between a sets of inputs and known output is developed using modeling such as artificial neural networks (Lek et al. 1996), classification and regression trees (De’ath 2000), support vector machines (Drake 2006) and wavelet analysis (Cho and Chon 2006) and (b) unsupervised approach that describes the pattern analysis using modeling including Hopfield neural networks (Hopfield 1982) and self-organizing maps (Kohonen 2001) in ecological data. These machine learning models are useful for fish classification having greater accuracy (Hu et al. 2012; Ogunlana 2015; Allken et al. 2018). Important features are identified through feature selection method in ML modeling. In this process a potential subset of data is identified relevant predicting of target values in the data sets (Hnin and Lynn 2016). Support vector machine and random forest are some commonly used for classification techniques and feature selection method in fisheries and related fields (Lin et al. 2015; Sylvester et al. 2018; Blachnik et al. 2019).

The Krishna River is the second largest river in India and fishes in this river are under threat due to pollution and anthropogenic activities (Kharat et al. 2012). Small indigenous fish (SIF) (<25 cm standard length) is considered as cheap sources of protein, vitamins and minerals that fulfil the requirement of livelihood and nutritional security in many of the Asian countries like India, Pakistan, Bangladesh and Nepal etc. (Sarkar and Lakra 2010; Mohanty et al. 2013; Saha et al. 2018). Many SIFs were threatened and endangered due to its habitat degradation. Priority habitat identification of SIFs in aquatic systems is of major concern for its conservation and management
strategies. *Chanda nama* (Hamilton), a small indigenous fish species of ambassidae family mostly found in freshwater, brackish water, inhabiting running and standing waters (Talwar and Jhingran 1991; Nandi et al. 2013). Very rare studies were conducted for *Chanda nama* (SIF) habitat identification for presence/absence (prediction) in Indian river system using machine learning approaches. The purpose of this study is to develop a framework for prediction of habitat indicators for SIF *Chanda nama* distribution in the Krishna River, India using ML modeling. This manuscript is structured in two steps (a) at first the ML classification models such as RF, SVM, ANN, and KNN were performed and compared for prediction of *Chanda nama* and selected the best model (b) Secondly, using selected model identified priority habitat indicators for *Chanda nama* in the Krishna River.

2 Material and methods

2.1 Study area

The River Krishna (15° 45' 15.2604" N and 80° 53' 50.1720" E) is the second largest peninsular river of India has 268,786 km² drainage area covering four states, Maharashtra (25.8%), Karnataka (42.4%), and in Andhra Pradesh and Telangana (31.8%) (Figure 1). It originates from Mahabaleswar Hills (1337 m asl), Maharashtra and traverses from west to east before emptying into the Bay of Bengal. Based on the altitude and slope of the river the study area is divided into three strata (upper, middle and lower). There are total twenty two sampling stations which are distributed into these three strata in the river. Fourteen sampling stations were taken in the upper part of the river having altitude between 740 m to 515 m (slope 42 cm/km), five sampling stations in the middle part of the river having altitude between 494 to 170 m (slope 113 cm/km) and the three stations in the lower part of the river of altitude between 19 m to 5 meter (slope 11 cm/km). These sampling stations were selected to cover the variations in three strata of the selected river as well as the best representative sample and best point of gaining access to the rivers that can be suitable for easy sampling for fish and water quality.

2.2 Data structure and collection

The data were taken on *Chanda nama* fish presence/absence and associated ten physical and chemical parameters of water (temperature (°C), transparency (m), depth(m), pH, specific conductivity (µS/m), dissolved oxygen (ppm), total alkalinity (ppm), flow (cm/sec), chloride (ppm) and altitude (m) (Table 1) during post monsoon (Oct–Nov, 2001), pre-monsoon (May–June, 2002) and monsoon (September, 2002) seasons at 22 sampling sites of the Krishna river during 2001-02. Water sample analysis was carried out following APHA (1992) Standard Method.
Fisheries data were collected through in situ, landing centre and experimental fishing using different types of gear (included cast nets, gill nets, and drag nets with varying mesh sizes) at all the sampling stations. Sampling was performed two times a day, morning (06:00 to 10:00) and evening (18:00 to 21:00) at consecutive three sub-sites at an interval of 10–50 meters each of the studied sampling sites.

2.3 Methodology

The data are standardized before analysis. The complex classification models (classifiers) i.g., machine learning models such as random forest (RF), artificial neural network (ANN), support vector machine (SVM), k-nearest neighbors (KNN) based on its interpretability of parameters (Saberioon et al. 2018) were performed for classification of *Chanda nama* distribution (presence/ absence) in the river. Here complex classifier is defined as machine learning classifiers such as RF, SVM, ANN and KNN which have a number of parameters and difficult to interpret. The classifier performance was compared based on the model evaluation parameters such as classification accuracy (CCI), Cohen’s kappa coefficient (k), sensitivity, specificity and receiver-operating-characteristics (ROC) (Saberioon et al. 2018; Tirelli and Pessan 2009; Mansbridge 2018). Based on performance, the best classifier model was selected and further used for identifying the important habitat parameters for *Chanda nama* presence/absence (prediction) in the Krishna river. The classifiers are described as follows.

2.3.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a nonparametric machine learning classification algorithm where it maps its input (classification) features into a high dimensional feature space using kernel functions, where every classification feature represented by each dimension. It uses a linear hyper plane as a decision function for nonlinear problem and then it applies a back transformation in nonlinear space. It gets the optimal solution through Lagrange multiplier by partial differentiation of each feature and reduce the complexity of the training data into a significant subset known as support vectors.

Let us assume a training set of N data points, \( \{x_k, y_k\}_{k=1}^n \) with input data, in a n-dimensional data vector \( x_k \in \mathbb{R}^N \) and output, which is the one dimensional vector space \( y_k \in r \); SVM creates the classifier as shown in equation (1)

\[
y(x) = \text{sign}\left[ \sum_{k=1}^N \alpha_k y_k \psi(x, x_k) + b \right]
\]

\( \ldots (1) \)
where $\alpha_k$ is positive real constants and $b$ is a real constant. In this present study, most frequently used SVM with radial basis function kernel was used which can be calculated as using equation (2),

$$
\Psi(x,x_k) = \exp\left(-\frac{(x-x_k)^2}{2\sigma^2}\right), ~ k = 1, ..., N \tag{2}
$$

where $\sigma$ is the width of the radial function determined by a grid search method using repeated cross validation approach. For details please see (Vapni 1998; Hsu et al. 2003). Here, R package “caret” used for SVM classification model and feature selection (Kuhn 2008).

### 2.3.2 Random Forest (RF)

Random Forest (RF) is a supervised and tree based ensemble machine learning approach that works on mixture of decision trees; $\{T_1(X), ..., T_B(X)\}$, where $X=\{x_1, ..., x_p\}$ is a $p$-dimensional vector of variables. Each split of the tree follows the concept of bootstrap aggregation (i.e., subsampling input samples with replacement) and random subspace method (i.e., subsampling the variables without replacement). The ensemble produces $B$ outputs $\{\hat{Y}_1 = T_1(X), ..., \hat{Y}_B = T_B(X)\}$, where $\hat{Y}_b, b = 1, ..., B$ is the predicted weight by the $b^{th}$ tree. The class $\hat{Y}$ is predicted by the majority of trees (Breiman 2001). RF has advantage that it does not over-fit and has robustness to noise and irrelevant features to produce good predictors. Here, R package “random Forest” used for prediction modeling.

### 2.3.3 Artificial Neural Network (ANN)

Artificial neural network (ANN), a machine learning nonlinear and nonparametric model inspired from the biological neural networks, consists of a set of connected cells (neurons) based on multi-Layer Perceptron (MLP) having at least three layers i.e., input layer, hidden layer and output layer arranged in leftmost, middle and the right most positions (Leray and Gallinari 1999). The Input layer feeds the values of the features to the hidden layer impulses from the input cells or neurons. The hidden layer does not have any direct input. The neurons of neural network are known as sigmoid neurons having multiple inputs $x_1, x_2, ..., x_n$, but the output is on a scale of 0 and 1. The sigmoid neuron has weight for each input and over all bias. The neurons in the input layer receive input from the input cell perform some kind of transformation by assigning weights to the input and transmit outcome to the other neurons in the output layer. Commonly, each of the class has one output unit.

The discriminate function having $k^{th}$ output unit for the neural network can be written as:
$$g_k(\mathbf{X}) = \sigma\left[\sum_{j=1}^{J} \alpha_{j,k} \sigma\left(\sum_{i=1}^{n} x_i w_{i,j} + b_j^h\right) + b_k^o\right]. \quad (3)$$

Here, $\mathbf{X}$ is a feature vector, $w_{i,j}$ is the weight assign to the $j^{th}$ hidden node by the $i^{th}$ input unit, $\alpha_{j,k}$ is the weight assign to the $k^{th}$ output node by the $j$th hidden unit, $b_j^h$ bias term of the $j$th hidden unit, $b_k^o$ is the bias term of the $k$th output unit. These are the adjustable parameters that were estimated during the training process by minimizing the loss function. Let us assume that the training samples $N_T$ are available to train a neural network with the $K$ output units, then the error of the neural network can be computed as $E(\boldsymbol{\omega}) = \sum_{s=1}^{N_T} e_s(\boldsymbol{\omega})$, and $e_s = \frac{1}{2} \sum_{k=1}^{K} (d_{s,k} - g_{s,k})^2$, here $\boldsymbol{\omega}$ represents all the adjustable parameters of the neural network (weights and biases) which are initialized with small random values, $d_{s,k}$ represents the expected output of the unit $k$ for the samples $s$ and $g_{s,k}$ is the actual output value for the same sample. During learning process the commonly used learning back propagation (Rumelhart 1985) is practiced for adjusting the parameter $\boldsymbol{\omega}$ for targeting the lowest training error $E(\boldsymbol{\omega})$. For details in the theory and application please see (Venables and Ripley 2002; Tarca et al. 2007). The decision over the number of hidden layer and numbers of neurons in each layer is determined by trial and error methods to avoid the under and/or over fitting problems (Valipour 2013).

### 2.3.4 k-Nearest Neighbors (k-NN)

The k-Nearest Neighbor (k-NN) is also a machine learning nonparametric model which predicts the class of an object according to the class of its $k$ nearest neighbors. It computes the distance ($N_0$) from an observation $y_i$ to all the other observation $y_j$ using the distance function. Here Euclidean distance function was used. Further it estimates the conditional probability for class $j$ as the fraction of points in $N_0$ whose response value equal $j$:

$$\Pr(Y = j|X=x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j) \quad (4)$$

Finally, class is determined based on neighbors follows Bays rule (James et al. 2013).

### 3 Model validation

Validation of model is an important part to test the accuracy of the existing model. For this purpose total data sets are divided into a training set (80% of the samples) and validation set (20% of the samples). The training set is used for model buildings for classification and validation set is used to assess the prediction accuracy of the model. The
samples are selected randomly in training and validation sets both. Further, following five model evaluation criteria based on confusion matrices are used to access the performances of the models (Fielding and Bell 1997) (a) the classification accuracy i.e., percentage of correctly classification instances (CCI) (b) sensitivity of the model (ability to accurately predict species presence at sampling sites) (c) model specificity (the ability to predict species absences at different sites (Tirelli et al. 2009) (d) Cohen’s kappa coefficient (k), a reliable performance measure of presence/absence models because it is negligible affected due to prevalence. However, for freshwater ecology k value ranges between 0 to 1 and higher values give a better model classification. The range of k from 0.20 to 0.40 considered as fair, 0.40-0.60 as moderate, 0.60-0.80 substantial and 0.80 to 1.00 as excellent and (e) ROC curve (the area under the receiver-operating-characteristics) generally 0.7 indicates satisfactory discrimination, 0.8 good and 0.9 very good discrimination (Hosmer and Lemeshow 2000). Here 10-fold cross validation is taken into consideration for the average predictive performance of the model. The best selected model is used to identify the important habitat parameters for classification for the presence/absence of fish in the river.

Here all the five classification models are developed on training data sets and validated on validation data sets and compared based on five model evaluation criteria.

4 Results

Habitat parameters i.e., water temperature (Temp), Transparency (Trans), depth, pH, Conductivity (Cond.), dissolved oxygen (DO), water flow (Flow), total alkalinity (TA), chloride (Cl) and altitude (Alt) of the Krishna River mean and standard deviation were presented in table 1. The variation was found to be higher in depth, conductivity, flow, total alkalinity, chloride and altitude. The correlation between these parameters was presented in table 2. All the parameters showed low correlation with each other except chloride and conductivity. The classification model performance measures such as accuracy of classification (CCI), kappa coefficient, sensitivity, specificity and ROC were presented in the table 3. The model performance measures with average (avg.) and standard deviation (std.) were presented in table 3. Results showed that accuracy of ML models lies between 0.59 (59%) to 0.82 (82%). Random forest showed maximum accuracy of (0.82) 82% followed by ANN (0.68) 68%, SVM (0.59) 59% and k-NN (0.62) 62%. The distribution of model’s accuracy was presented in figure 2. The kappa value for RF (0.55) showed maximum followed by k-NN (0.36), ANN (0.33) and SVM (0.28). The kappa distribution of models was presented in figure 3. However, sensitivity was observed to be higher for ANN (0.61) followed by RF (0.57), SVM (0.43) and k-NN (0.38). The sensitivity distribution of models was presented in figure...
4. The specificity measure of RF and k-NN showed similar (0.76) followed by SVM (0.75) and ANN (0.66). The distribution of specificity of different models was presented in figure 5. ROC value for RF (0.72) was observed to be maximum followed by SVM (0.68), k-NN (0.68) and ANN (0.66). The distribution of ROC of different models was presented in figure 6. Out of five model performance measures criteria four measures were in favors of Random Forest ML model, the best among the all used models for classification of Chanda nama presence/absence in the Krishna River. Only ANN performs better than other models based on only sensitivity criteria. Hence, RF was selected as the best classification model for Chanda nama presence/absence in the Krishna River. Further, habitat features for Chanda nama distribution in Krishna River were identified using the RF model (Figure 7). RF model identified three important habitat parameters in priority order i.e., altitude (Alt), temperature (temp), depth for presence/absence of the Chanda nama in the Krishna River. Habitat parameters (mean and std.) of Chanda nama presence (Y) and absence (N) at distinct sites on the river was presented in the table 4. Identified habitat parameters i.e., altitude, temperature, and depth were observed to be lower at Chanda nama presence sites as compared to absence sites. Besides this, DO and alkalinity was observed to be higher and conductivity lower at Chanda nama presence sites.

5 Discussion

The purpose of this study was to identify the important habitat parameters to predict the presence/absence of Chanda nama in the Krishna River using machine learning model. Random forest (RF) machine learning algorithm models was found to be the best classifiers (table 3) for prediction of Chanda nama presence/absence in the Krishna River based on ML model evaluation criteria (Jinadasa, 2006). Over the last decades, the applications of machine learning become more popular to understand the complexity of ecological ecosystem (Valletta et al. 2017; Willcock et al. 2018). Machine learning algorithms were suitable for aquatic ecosystem modeling in relation to fisheries where habitat suitability modeling, ANN and decision tree modeling showed high potential (Goethal et al. 2007; Dakou et al. 2007), k-NN and SVM accuracy more than 95% for fish taxonomic classification (Noda et al. 216). Guisan and Zimmerman (2000) in a review paper highlighted the use of machine learning techniques for species distribution. Random Forest showed the best accurate classification model and used for important features selection (Jinadasa et al. 2006, Lonzarich and Quinn 1995; Sylvester et al. 2018). Some previous study also showed that RF is a better classifier than the SVM, ANN and CART (Lonzarich and Quinn 1995; Jinadasa et al. 2006; Sylvester et al. 2018)
since it works on pattern recognition method using “ensemble learning” where it generates several classifiers and by combining all the results produce the final prediction. The beauty of the RF is that it is simpler to tune, faster (depends on data) and precisely works for the categorical input than the SVM and ANN. Hence, it is an ensemble method that works better in some situations (Breiman 2001). However, generally the machine learning algorithm performance varies based on the nature and complexities of the data sets. Some studies showed SVM is better than ANN and RF (Tirelli et al. 2009), ANN performs better than SVM and RF (Rosenfeld 2003), k-NN better accuracy than RF (Noda et al. 2016). The machine learning modeling approach such as classification trees and artificial neural networks showed more robust for modeling species presence/absence distribution (Olden and Jackso 2002; Guisan and Zimmermann 2000).

Here, RF identified the three important habitat parameters i.e., altitude, temperature and, depth for Chanda nama presence/absence prediction in the Krishna River. The first and most important habitat feature is the altitudes of the river. Chanda nama is a small indigenous fish and in this study it was observed maximum in numbers at the higher altitudes in the upper portion of the study sites on the river between 740 m and 515 m with the slope at 42 cm/km and very less in the sites in the middle altitudes 494 to 170 m (slope 113 cm/km) and negligible in the down-stream of the river having altitudes between 19 m and 5 m (slope 11 cm/km). The presence of Chanda nama in the Krishna river was observed in less sloppy zone. But in downstream of the river even slope is very less but Chanda nama was absent due to estuarine condition of the river estuarine condition. The identified habitat features showed that the preferable average altitude 484.13 m and 472.75 m for the presence (Y) and absence (N) of Chanda nama in the Krishna River (table 4). Hence, Chanda nama prefers its presence in the upper portion less sloppy area on the Krishna River. The average temperature and depth was observed to be 28.56±2.27 °C and 5.03±9.07 m for the presence (Y) in comparison to the absence (N) where average temperature and depth was 29.57 ±1.97 °C and 6.50±8.88 m in the river (table 4). In this river Chanda nama habitat preference was low temperature and less depth.

Some previous studies also reported that fish distribution is governed by stream gradient, zones, altitude and temperature (Lonzarich and Quinn 1995; Jinadasa et al. 2006). A review paper (Rosenfeld 2003) showed that the habitat features such as velocity, depth, temperature and conductivity were associated with different fish species. Besides the preferable habitat features of individuals the persistence of populations also depends on landscape-scale features in relation to immigration and emigration rates, broader regional abiotic constraints and habitat fitness consequences (Pulliam 1988; Dunning et al. 1992; Poff 1997).
Thus, this ML modeling approach for identification of important physical and chemical parameters of water will provide a good understanding to the researchers and policy makers for determining the important habitat parameters of *Chanda nama* for decision making and sustainable management in the Krishna river. This study is the first modest attempt to apply the application of machine learning modeling like random forest for habitat identification of the prediction (presence/absence) of small indigenous fish like *Chanda nama* distribution in an Indian river.

6. Conclusion

Machine learning modeling is a useful tool for understanding the complex aquatic habitat parameters relationship with fish prediction in river system. Random Forest, machine learning model is observed as the best model for the identification of priority habitat parameters for *Chanda nama* distribution with 82% accuracy on the Krishna River, India. Random Forest model is better than the SVM, ANN, and k-NN models for *Chanda nama* prediction in the Krishna River. Altitude, temperature and depth were identified as the three major physical and chemical habitat parameters for *Chanda nama* distribution on the Krishna River. Higher altitude with less slope, low temperature and low depth, were the preferable habitat indicators for the *Chnada nama* presence on the Kishna River. The Machine learning model was observed to be the more robust for establishing relationship between complex aquatic ecosystem and fish species distribution in the Indian River system. Hence, the machine learning modeling an alternative approach study opens the scope for habitat parameter identification for fish and other aquatic species distribution in different aquatic systems.

Acknowledgement

Authors are thankful to Director, ICAR-Central Inland Fisheries Research Institute, Barrackpore, Kolkata, India for extending the facilities and encouragement during the research work.

Author contributions RKR developed ML model framework study, manuscript writing, data analysis, editing; AKD and RKM data collection and editing; SKS GIS mapping of study area and editing; BKD framework designed, editing and overall guidelines; all authors read and approved the final manuscript.
Funding: This research did not receive any specific grant from funding agencies. This study is conducted at ICAR Central Inland Fisheries Research Institute, Barrackpore, India.

Data availability: The datasets used for the current study are available and received from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests. Consent for publication Not applicable.

Ethical Statement

The submitted manuscript is not submitted in any other journal.

This study has been approved by the ICAR- Central Inland Fisheries Research Institute institute ethical committee, Barrackpore-700120, India.

Consent to Participate: Not applicable

Consent to Publish: The approval for submitting manuscript received from ICAR- Central Inland Fisheries Research Institute, Barrackpore-700120, India.

Open Access: Not applicable

References:

Allken V, Handegard N O, Rosen S, Schreyeck T, Mahiout T, Malde K, Handling editor: Richard O’Driscoll (2018) Fish species identification using a convolutional neural network trained on synthetic data. ICES Journal of Marine Science 76(1): 342-349

Angermeier P L, Davideanu G (2004) Using fish communities to assess streams in Romania: initial development of an index of biotic integrity. Hydrobiologia, 511(1-3):65-78

Aarts B GW, Van den Brink FWB, Nienhuis PH (2004) Habitat loss as the main cause of the slow recovery of fish faunas of regulated large rivers in Europe: the transversal floodplain gradient. River Research and Applications, 20: 3–23

Blachnik M, Soltysiak M, Dąbrowska D (2019) Predicting Presence of Amphibian Species Using Features Obtained from GIS and Satellite Images. ISPRS International Journal of Geo-Information 8(3): 123

Breiman L (2001) Random Forests. Mach. Learn. 45, 5–32

Cho E, Chon T S (2006) Application of wavelet analysis to ecological data. Ecological Informatics 1(3):229–233
Dakou E, D'heygere T, Dedecker A P, Goethals P L M, Lazaridou-Dimitriadou M, DePauw N (2007) Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). Aquatic Ecology 41:399–411

De’ath G, Fabricius K E (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11):3178–3192

Drake J M, Randin C, Guisan A (2006). Modelling ecological niches with support vector machines. Journal of Applied Ecology. 43(3): 424–432

Dunning J B, Danielson B J, Pulliam H R (1992) Ecological processes that affect populations in complex landscapes. Oikos 65:169–175

Elithet J, Graham C, Anderson P, Dudik R, Ferrier M, Guisan S, Hijmans A J, Huettmann R, Leathwick F R, Lehmann J A, Li J (2006) Novel methods improve prediction of species’ distributions from occurrence data. Ecography: 29(2), 129-151

Fielding A H, Bell J F (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24(1):38–49

Goethals P LM, Dedecker A P, Gabriels W, Lek S, DePauw N (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology 41: 491–508

Guisan A, Zimmerman N E (2000) Predictive habitat distributions models in ecology. Ecological Modelling 135:147–186.

Hnin T T, Lynn K T (2016) Fish Classification Based on Robust Features Selection Using Machine Learning Techniques. In Genetic and Evolutionary Computing. Springer, Cham. 237-245

Hopfield J J (1982) Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences USA79(8): 2554 –2558

Hosmer D, Lemeshow S (2000) Applied Logistic Regression Second Edition. A. Wiley-Interscience Publication

Hsu C, Chang C, Lin C (2003) A Practical Guide to Support Vector Classification; Department of Computer Science, National Taiwan University: Taipei, Taiwan

Hu J, Li D, Duan Q, Han Y, Chen G, Si X (2012) Fish species classification by color, texture and multi-class support vector machine using computer vision. Comput. Electron. Agric 88:133–140

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. New York. Springer. Vol. 112, 18
Jinadasa J, Herath H M B K, Samarasinghe J M H S K (2006) Species Composition and Altitudinal Distribution of Fish in Upper Walawe River Basin in Sri Lanka. Sri Lanka Journal of Aquatic Sciences 8:1

Kharat S S, Paingankar M, Dahanukar N (2012) Freshwater fish fauna of Krishna River at Wai, Northern Western Ghats. India Journal of Threatened Taxa 4(6):2644–2652

Kohonen T (2001) Self-Organizing Maps. Berlin (Germany) and New York: Springer-Verlag

Kuhn M (2008) Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28.

Laws E A (2000) Aquatic pollution: an introductory text. United States of America: John Wiley & Sons

Lek S, Delacoste M, Baran P, Dimopoulos I, Laugå J, Aulagnier S (1996) Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90(1), 39–52

Leray P, Gallinari P (1999). Feature Selection With Neural Networks. Behaviormetrika 26: 145–166 https://doi.org/10.2333/bhmk.26.145

Levin S A (1998) Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1(5):431–436

Lehikoinen A, Olsson J, Bergström L, Bergström U, Bryhn A, Fredriksson R, Uusitalo L, (2019) Evaluating complex relationships between ecological indicators and environmental factors in the Baltic Sea: A machine learning approach. Ecological indicators 101:117-125

Lin K C, Chen SY, Hung J C (2015) Feature selection and parameter optimization of support vector machines based on modified artificial fish swarm algorithms. Mathematical Problems in Engineering

Lonzarich D G, Quinn T P (1995) Experimental evidence for the effect of depth and structure on the distribution, growth, and survival of fishes. Canadian Journal of Zoology 73:2223–2230

Maier P M, Keller S (2019) Estimating Chlorophyll a Concentrations of Several Inland Waters with Hyperspectral Data and Machine Learning Models. arXiv preprint arXiv:1904.02052

Mansbridge N, Mitsch J, Bollard N, Ellis K, Miguel-Pacheco G, Dottorini T, Kaler J (2018) Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep. Sensors 18(10) 3532

Mohanty B P, Pati M K, Bhattacharjee S, Hajra A, Sharma A P (2013) Small indigenous fishes and their importance in human health. Advances in Fish Research. 5: 257-278

Moore J A, Lew A A, Cunningham J A, Kachuba M (1995) Fish communities as indicators of environmental quality in the West River watershed. Yale Forestry and Environmental Studies Bulletin, 100:178-196
Nandi S, Majumder S, Saikia S K (2013) Small freshwater fish species (SFFs) culture: issues from nutrient security, carp-SFF integration and feeding ecology. Reviews in fish biology and fisheries 23(3):283-291

Noda J, Travieso C, Sánchez-Rodríguez D (2016) Automatic taxonomic classification of fish based on their acoustic signals. Applied Sciences 6(12), 443

Oberdorff, T, Guegan, J F, Hugueny B (1995) Global scale patterns of fish species richness in rivers. Ecaogrophy 18(4): 345–352

Ogunlana S O, Olabode O, Oluwadare S A A, Iwasokun G B (2015) Fish classification using support vector machine. African Journal of Computing & ICT 8(2):75-82

Olden J D, Jackson D A (2002) A comparison of statistical approaches for modeling fish species distributions. Freshwater Biology 47:1976–1995

Olden J D, Lawler J J, Poff N L (2008) Machine learning methods without tears: a primer for ecologists. The Quarterly review of biology 83(2): 171-193

Poff N L (1997) Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology Journal of the North American Benthological Society 16:391–409

Pulliam H R (1988). Sources, sinks, and population regulation. American Naturalist 132:652–661.

Rathi D, Jain S, Indu S (2017) Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning. In 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR) pp. 1-6 IEEE

Recknagel F (2001) Application of machine learning to ecological modelling. Ecological Modelling 146 (1–3):303–310

Rosenfeld J (2003) Assessing the Habitat Requirements of Stream Fishes: An Overview and Evaluation of Different Approaches, Transactions of the American Fisheries Society 132:5 953-968, DOI: 10.1577/T01-126

Rumelhart D E, Hinton GE, Williams R J (1985) Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science

Saberiion M, Cisař P, Labbé L, Souček P, Pelissier P, Kerneis T (2018) Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. Sensors 18(4): 1027

Saha D, Pal S, Mukherjee S, Nandy G, Chakraborty A, Rahaman S H, Aditya G (2018). Abundance and biomass of assorted small indigenous fish species: Observations from rural fish markets of West Bengal, India. Aquaculture and fisheries 3(3): 129-134
Sarkar U K, Lakra W S (2010) Small indigenous freshwater fish species of India: significance, conservation and utilization. Aquaculture Asia 15(3): 34-35

Schiemer F (2000) Fish as indicators for the assessment of the ecological integrity of large rivers. Hydrobiologia 422/423: 271–278

Sylvester V, Bentzen P, Bradbur I R, Clément M, Pearce J, Horne J, Beiko R G (2018) Applications of random forest feature selection for fine-scale genetic population assignment. Evolutionary applications 11(2): 153-165

Talwar P K, Jhingran A G (1991) Inland Fisheries of India and Adjacent Countries, Vol 2. Oxford and IBH Publishing, New Delhi

Tarca A L, Vincent J C, Xue-wen C, Roberto R, SorinDrăghici (2007) Machine learning and its applications to biology. PLoS computational biology 3 no. 6 : e116

Tirelli T, Pessan D (2009) Use of decision tree and artificial neural network approaches to model presence/absence of Telestesmuticellus in piedmont (North-Western Italy). River Res. Appl. 25(8):1001-1012

Tirelli T, Pozzi L, Pessani D (2009) Use of different approaches to model presence/absence of Salmo marmoratus in Piedmont (Northwestern Italy). Ecological Informatics 4(4):234-242

Valipour M, Banihabib M E, Behbahani S M R (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural. J. Hydrol. 476, 433–441

Valletta J J, Torney C, Kings M, Thornton A, Madden J (2017) Applications of machine learning in animal behaviour studies. Anim. Behav. 124:203–220

Vapnik V (1998) Statistical Learning Theory John Wiley & Sons: Hoboken, NJ, USA ISBN 0-471-03003-1

Venables B, Ripley B (2002) Modern applied statistics with S. 4th edition. New York: Springer. 495 p.

Willcock S, Martínez-López J, Hooftman D A, Bagstad K J, Balbi S, Marzo A, Prato C, Sciandrello S, Signorello G, Voigt B, Villa F (2018) Machine learning for ecosystem services. Ecosystem services 33:165-174
Table 1: Physical and chemical parameters (mean ± std.) of Krishna river, India, during year 2001-02.

| Physical and chemical parameters | Temp ±Std | Trans ±0.60 | depth ±8.96 | pH ±1.9 | Cond ±34.06 | DO ±1.37 | Flow ±23.91 | TA ±52.89 | Cl ±68.16 | Alt ±224.80 |
|---------------------------------|-----------|-------------|-------------|---------|-------------|---------|-------------|---------|-----------|-------------|
| Avg ±Std                        | 28.99±2.20| 1.12±0.60  | 5.66±8.96  | 8.12±0.19 | 561.9±342.06 | 7.38±1.37 | 24.68±23.91 | 147.86±52.89 | 72.24±68.16 | 415.32±224.80 |

Table 2: Correlation matrix between physical and chemical parameters of the Krishna river during year 2001-02.

| Physicochemical parameters | Temp | Trans | depth | pH | Cond | DO | Flow | TA | Cl | Alt |
|----------------------------|------|-------|-------|----|------|----|------|----|----|-----|
| Temp                       | 1.00 |       |       |    |      |    |      |    |    |     |
| Trans                      | 0.37 | 1.00  |       |    |      |    |      |    |    |     |
| depth                      | -0.14| 0.28  | 1.00  |    |      |    |      |    |    |     |
| pH                         | 0.07 | -0.06 | 0.00  | 1.00|      |    |      |    |    |     |
| Cond                       | 0.23 | 0.16  | -0.08 | -0.24| 1.00|    |      |    |    |     |
| DO                         | -0.39| -0.38 | -0.02 | -0.01| -0.23| 1.00|      |    |    |     |
| Flow                       | -0.06| 0.04  | -0.19 | -0.07| -0.03| 0.02| 1.00|    |    |     |
| TA                         | -0.08| -0.32 | -0.27 | 0.20 | 0.09 | 0.32| 0.26| 1.00|    |     |
| Cl                         | 0.20 | 0.16  | -0.07 | -0.26| 0.80 | -0.23| -0.04| 0.03| 1.00|     |
| Alt                        | -0.20| -0.04 | 0.05  | -0.01| 0.05 | 0.13| -0.24| 0.00| 0.06| 1.00|
Table 3: Model performance on validation sets for identification of presence and absence of *Chanda nama* fish in Krishna river, India

| Models                          | Accuracy of classification (avg±std) | Kappa (avg±std) | Sensitivity (avg±std) | Specificity (avg±std) | ROC (avg±std) |
|--------------------------------|-------------------------------------|-----------------|-----------------------|------------------------|---------------|
| Random Forest (RF)             | 0.82±0.03                           | 0.55±0.06       | 0.57±0.17             | 0.76±0.03             | 0.72±0.06     |
| Neural Network (ANN)           | 0.68±0.06                           | 0.33±0.09       | 0.61±0.09             | 0.66±0.02             | 0.66±0.02     |
| Support vector machine (SVM)   | 0.59±0.03                           | 0.28±0.09       | 0.43±0.10             | 0.75±0.04             | 0.68±0.10     |
| k-Nearest Neighbors (k-NN)     | 0.62±0.10                           | 0.36±0.04       | 0.38±0.06             | 0.76±0.03             | 0.68±0.03     |

Table 4: Physico chemical parameters (mean ± std.) at sites for *Chanda nama* presence (Y) and absence (N) in Krishna river, India, during 2001-02.

| Fish (Chanda nama) | Temp | Trans | depth | pH | Cond | DO | Flow | TA | Cl | Alt |
|-------------------|------|-------|-------|----|------|----|------|----|----|-----|
| N (Chanda nama)    | 29.57±1.97 | 6.50±8.88 | 8.15±0.19 | 647.36±355.31 | 7.19±1.14 | 21.12±16.78 | 145.74±40.69 | 88.09±81.03 | 472.75±180.04 |
| Y (Chanda nama)    | 28.56±2.27 | 1.09±0.53 | 5.03±9.07 | 8.09±0.18 | 498.58±222.07 | 7.54±1.52 | 26.50±28.09 | 149.42±60.82 | 60.57±155.11 | 484.12±161.6 |
Figure 1: Study area of the Krishna River, India
Figure 2: Model accuracy of k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) for classification of presence and absence of *Chanda nama* fish in Krishna River, India.
Figure 3: Kappa values of k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) for classification of presence and absence of *Chanda nama* fish in Krishna River, India.
Figure 4: Sensitivity of k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) for classification of presence and absence of *Chanda nama* fish in Krishna River, India.
Figure 5: Specificity of k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) for classification of presence and absence of *Chanda nama* fish in Krishna River, India.
Figure 6: ROC of k-nearest neighbors (KNN), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) for classification of presence and absence of Chanda nama fish in Krishna River, India.
Figure 7: Importance of habitat features (altitude, temperature and depth) for *Chanda nama* fish classification of presence and absence in Krishna river, India using Random Forest modeling.