

Weight Prediction for Fishes in Setiu Wetland, Terengganu, using Machine Learning Regression Model

Nurzuhrhah Hassan¹, Siti Tafzilmeriam Sheikh Abdul Kadir^{1,2*}, Mohd Lokman Husain¹
Behara Satyanarayana^{1,2}, Mohd Azmi Ambak, and Mazlan Abd.Ghaffar^{3,4}

¹Institute of Oceanography and Environment, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

²Mangrove Research Unit, Institute of Oceanography & Environment, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

³Faculty of Science and Marine Environment, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

⁴Institute of Tropical Aquaculture Tropical and Fisheries, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

Abstract. Predicting fish weight holds several essential implications in ecology, such as population assessment, trophic interactions within ecosystems, biodiversity studies of fish communities, ecosystem modelling, habitat evaluation for different fish species, climate change research, and support fisheries management practices. The objective of the studies is to analyse the prediction performance of machine learning (ML) regression models by applying different statistical analysis techniques. This study collected biometric measurements (total length and body weight) for 19 fish families from three locations in Setiu Wetland, Terengganu, captured between 2011 and 2012. The study adopts two regression types: Linear Regression (i.e., Multiple Linear, Lasso, and Ridge model) and Tree-based Regression (i.e., Decision Tree, Random Forest, and XGBoost model). Mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R^2) were used to evaluate performance. The results showed that the proposed ML regression models successfully predicted fish weight in Setiu Wetlands, and the Tree-based Regression model provides more accurate prediction results than the Linear Regression model. As a result, Random Forest is the best predictive model out of the six suggested ML regressions, with the highest accuracy at 96.1% and the lowest RMSE and MAE scores at 3.352 and 0.880, respectively. In conclusion, the use of machine learning is crucial for rapid, precise, and cost-effective fish weight measurement. By incorporating weight prediction into ecological research and management practices, we may make informed decisions supporting the conservation and sustainable use of fish populations and their habitats.

* Corresponding author: sititafzil@umt.edu.my

1 Introduction

Predicting fish weight is crucial in ecology because it helps estimate fish population size and distribution, which is vital for monitoring and managing fish. This data is essential for assessing population changes and health [1, 2], understanding ecosystem trophic interactions, energy flow, and food web structure [3], and biodiversity studies to assess species diversity and abundance [4]. Additionally, fish weight data is vital for ecosystem modelling [3] and addressing factors like fishing and climate change's impact on fish populations and habitats, as shifts in fish weight can signal changes in growth rates, productivity, and species composition [5], ultimately contributing to the sustainable management of fisheries for setting catch limits, crafting effective fishing regulations, and implementing conservation strategies, thus ensuring the long-term viability of fish stocks [6].

Typically, the process of measuring the weight of fish is conducted manually on a per-individual basis. Nevertheless, this procedure is challenging, time-intensive, and stressful for the fish. Weight measurements on fresh or frozen fish can vary significantly due to factors like fish wetness, environmental conditions, and scale capacity. Multiple scales of different sizes should be used to ensure accuracy, with an accuracy of $\pm 1\%$. Frozen fish are lighter and shorter than fresh ones [7]. Hence, the development of rapid, precise, cost-effective, and indirect measuring techniques would be significant to the field of ecology, and one of the methods is using machine learning.

Machine learning (ML) refers to the scientific investigation of algorithms and statistical models employed by computer systems to execute a designated task based on an example of a dataset without requiring explicit programming [8]. Regression is a form of supervised learning that predicts a dependent variable by estimating the correlation between independent variables [9]. Linear Regression is a widely utilised learning technique for its fundamental nature and frequent application in predictive analysis [10]. Multiple Linear Regression indicates that there is more than one independent variable. The Least Absolute Shrinkage and Selection Operator (Lasso) combines variable selection and regularisation to enhance prediction accuracy by reducing the number of variables, a process known as variable shrinkage, within a statistical model [11]. Ridge Regression mitigates the problem of multicollinearity by incorporating a regularisation term into Linear Regression models, particularly in cases where the independent variables exhibit significant correlation [11].

Tree-based Regression models utilise one or more explanatory variables to explain variation in a numeric response variable and systematically divide the training data into progressively smaller groups, each characterised by the mean value of the response variable, group size, and the defining values of the explanatory variables [12]. The Decision Tree algorithm constructs a tree-based model that utilising nodes to represent tests on input data, branches for possible outcomes, and leaf nodes for final predictions, with the ability to select the best features for data splitting at each node and options for pruning to prevent overfitting [13]. Random Forest combines multiple Decision Tree predictions that create a "forest", and each is trained on a random subset of the data and features and then aggregates their predictions [14]. Extreme Gradient Boosting (XGBoost) combines the predictions of multiple Decision Tree models and incorporates techniques such as gradient boosting, regularisation, handling missing data, and parallel processing for efficient training [15].

Some existing studies applying the ML regression model to predict fish weight included the use of Linear Regression, Random Forest Regression, and Support Vector Regression to estimate the weight of Tilapia fish with a coefficient of determination (R^2) value of 0.70 [16], weight prediction of Rainbow Trout (*Oncorhynchus mykiss*) using Linear, Power, and Second Order Polynomial Regression with R^2 values 0.98, 0.99, and 0.98 respectively [17], Symbolic Regression was used in the context of Perch, Bream, Roach, and Pike weight prediction with accuracy within the range of 0.98 to 0.99 [18], Linear Regression models for

the prediction of Nile Tilapia achieved R^2 values ranging between 0.95 and 0.96 for body weight and between 0.92 and 0.95 for carcass weight [19], and *Sparus aurata* weight estimation using Linear Regression and Nonlinear Regression, and coefficients of correlation (R) were assessed, and the values ranged from 0.95 to 0.98 and 0.96 to 0.97, respectively [20].

In the Setiu Wetland, the fish growth patterns of 22 families have been studied using the fish length-weight relationship (LWR) method. In the Power Regression model, denoted as $W = aL^b$, the length of the fish (L, measured in centimetres) is utilised as a predictor for the weight of the fish (W, measured in grams). The values of the intercept coefficient (a) and the exponential coefficient (b) exhibit variation based on the specific fish species and their respective environmental conditions [21].

In this study, we use fish sampling data to predict the fish weight of multi-family assemblages in the Setiu Wetland. The objective is to analyse the prediction performance of six widely used regression MLs using different statistical analysis techniques. As a result, this study creates an automated method that is straightforward and inexpensive to measure the fish's weight. By using the default model specifications offered by the statistics package, our study only explores the fundamental interpretations of each model rather than attempting to increase the performances of each ML by changing or adding model parameters.

2 Methods

2.1 Data Collection

Biometric measurements (total length and body weight) were collected for 19 fish families, viz., Siganidae, Leiognathidae, Ambassidae, Lutjanidae, Tetraodontidae, Apogonidae, Ariidae, Terapontidae, Belonidae, Tetraogidae, Lethrinidae, Hemiramphidae, Latidae, Paralichthyidae, Haemulidae, Serranidae, Gerreidae, Gobiidae, and Eleotridae (n=8662), from three different locations in Setiu Wetland, Terengganu, as mentioned in Table 1.

Table 1. Location coordinates for three sample sites

Site No.	Location Name	Latitude	Longitude
Site 1 (North)	Pulau Telaga Tujuh	5° 41' 43.13" N	102° 41' 52.01" E
Site 2 (Middle)	Pulau Che Him	5° 41' 0.07" N	102° 42' 35.54" E
Site 3 (South)	Pulau Tok Haji	5° 39' 42.38" N	102° 44' 5.31" E

From June 2011 to July 2012, sampling was conducted during wet and dry seasons. The sampling techniques employed encompassed multifilament gill nets, trawls, and fish traps. All fish were weighed using a portable electronic balance and electronic crane scale, while a digimatic calliper and ruler board were used to measure total length. The sampling and measuring protocols were described in Siti Tafzilmeriam et al. [21, 22].

2.2 Data Preparation

The data are pre-processed for analysis. Python programming by Jupyter Notebook of the Anaconda Navigator Software was used as a tool. Initially, the data-cleaning process is carried out on the dataset by wrangling, renaming the header, and deleting any missing values using the Pandas Python library. The outlier values were removed by Interquartile Range

(IQR) Analysis for total length and body weight variables. Removing outliers helps the model capture the underlying patterns and relationships in the data more effectively, leading to better generalisation and predictive performance. We can generate more reliable estimates and reduce the potential bias the extreme values introduce [23, 24].

Correlation analysis between two variables was used to determine the direction and strength of the linear association. Correlation analysis attempts to measure and understand the effect of two continuous variables on linear or nonlinear interactions. Association coefficients assume values that vary from negative (-1) to uncorrelated (0) to positive (+1) correlations. The sign of the correlation coefficient (i.e., positive or negative) defines the orientation of the relationship.

2.3 ML Model Development

In this study, we implemented regression algorithms as a method for prediction. The study adopts two regression types: Linear Regression (i.e., Multiple Linear, Lasso, and Ridge) and Tree-based Regression (i.e., Decision Tree, Random Forest, and XGBoost model).

Two encoding methods in Scikit-Learn were used to convert categorical data into numbers. OneHotEncoder class is required for Linear Regression, and the OrdinalEncoder class is functional for tree-based models. Hyperparameter tuning was performed to the Lasso and Ridge regression model to determine the optimal values of parameters using the GridSearchCV function in Scikit-Learn.

The data were split into the training and testing sets with a split ratio of 70:30, respectively, using the Scikit-learn Python library via the `train_test_split()` function. The training set is used for training the model, and the testing set is used for testing the model's accuracy. The split ratio value was chosen according to the study by Picard and Berk [25], Dobbin and Simon [26], Pham et al. [27], and Nguyen et al. [28].

2.4 Models Validation

A k-fold cross-validation (CV) approach was used to validate all developed models. This approach divides the whole dataset into k non-overlapping folds at random, then utilises k-1 folds for model construction and the remaining 1-fold for validation. This step is performed k times such that each fold is used as the validation set once at the end. The k-fold CV is a robust approach for estimating model accuracy that involves averaging the k outcomes to create a single estimate for a model. It is common practice to replicate the k-fold CV to maintain stability in model performance. The 15-fold CV was adopted since most model performance estimates were almost unbiased when k was between 10 and 20 [29].

2.5 Model Performance Evaluation

We evaluated and compared the overall performance of all regression models by employing three model performance measures: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The MAE informs us of the average size of an error from the forecast, whereas the RMSE displays the average distances between predicted and observed data [30]. R^2 measures how effectively predictors account for variation in response variables. R^2 values vary from 0 to 1, with 1 indicating perfect model fit. The best model has the lowest MAE and RMSE values and the highest R^2 .

3 Result and Discussions

3.1 Dataset Overview

The data consists of 8662 rows and three variable components, as shown in Fig. 1, with no missing value.

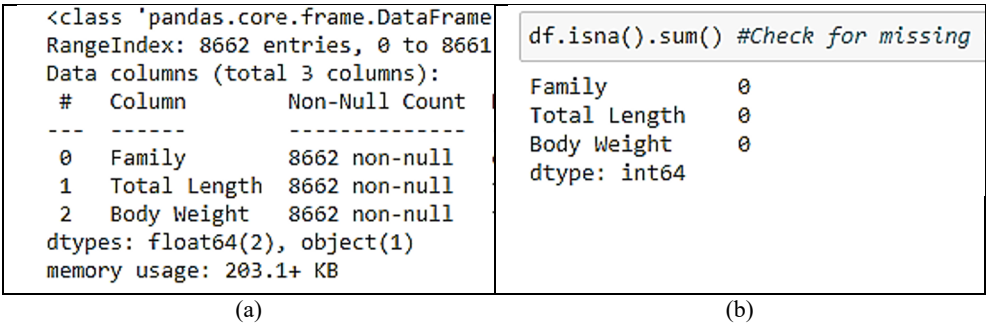


Fig. 1. Dataset overview (a) number of rows and features and (b) missing values data

Correlation analysis helps in selecting relevant variables for a predictive model. The correlation between variables was examined, and results in Table 2 show that all the variables did not exhibit multicollinearity (R-value = |1|), which results in unsatisfactory predictions. So, all three variables were fitted with the Linear and Tree-Based Regression models.

Table 2. Correlation between variables

	Family	Total Length	Body Weight
Family	1.000000	-0.474689	-0.235571
Total Length	-0.474689	1.000000	0.754893
Body Weight	-0.235571	0.754893	1.000000

3.2 Outlier Removal

The variables presented in Fig. 2 have undergone outlier predictions, and afterwards, the outliers have been identified and removed using IQR Analysis, as depicted in Fig. 3.

3.3 Model Performance Result

The best model performance has high R^2 values and low RMSE and MAE values. The performance evaluations are summarised in Table 3 for the Linear Regression Model and Table 4 for the Tree-based Regression model. The Tree-based Regression model has a better R^2 score than the Linear Regression model. When evaluating the RMSE, it was observed that the Tree-based Regression model had a lower score value than the Linear Regression model. Concerning the MAE metric, the Tree-based Regression model scored lower than the Linear Regression model. The empirical findings indicate that the Tree-based Regression model performed better than the Linear Regression model.

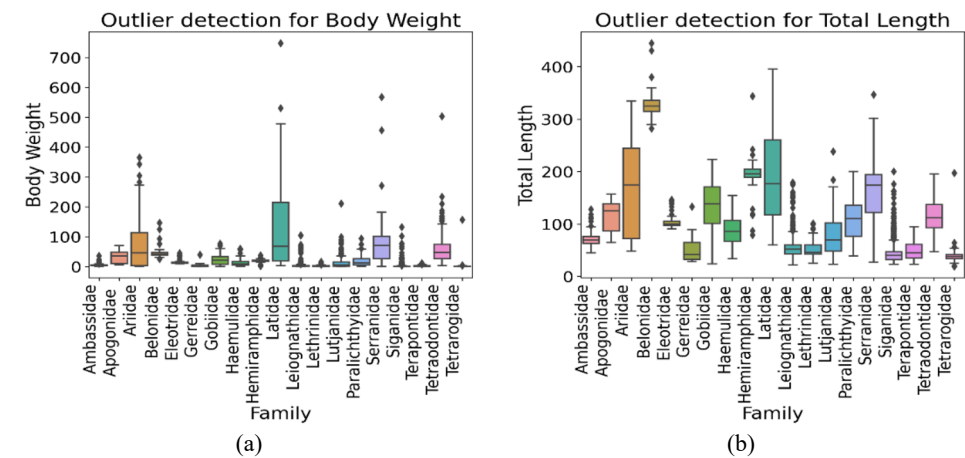


Fig. 2. Outlier detection for (a) Body weight and (b) Total length

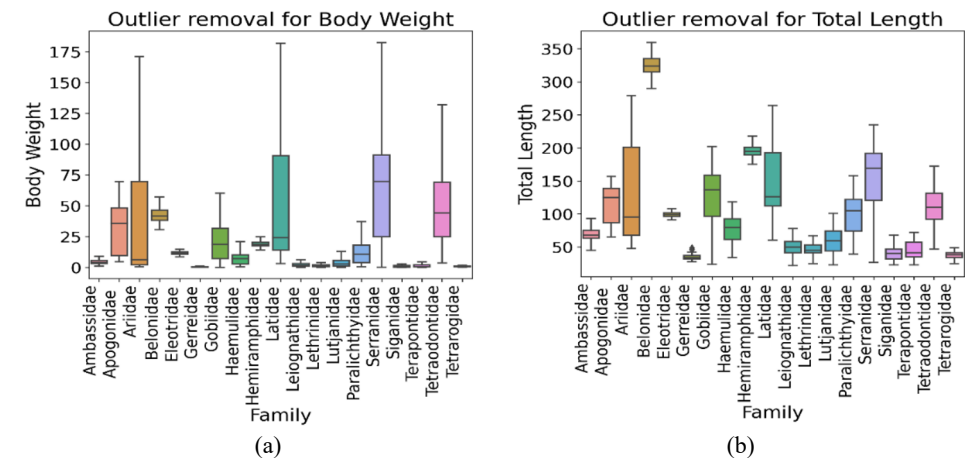


Fig. 3. Outlier removal for (a) Body weight and (b) Total length

This performance is because the Tree-based Regression is adept at capturing nonlinear patterns and interactions between variables by partitioning the feature space into regions and fitting distinct linear models within each region, providing flexibility in modelling. Tree-based Regression models can capture complex relationships among variables effectively. Interactions can be automatically detected and modelled without explicit specification [31].

Out of the six machine learning regression models that were considered, it was found that the Random Forest model had the highest predictive capability, achieving an accuracy rate of 96.1%. Additionally, this model demonstrated the lowest values for the RMSE and MAE, which were determined to be 3.352 and 0.880, respectively.

Table 3. Performance score for Linear Regression Model

Model Name	RMSE	MAE	R ²
Multiple Linear	6.510	3.932	0.841
Lasso	6.529	3.852	0.836
Ridge	6.521	3.890	0.839

Table 4. Performance score for Tree-based Regression model

Model Name	RMSE	MAE	R ²
XGBoost	3.642	0.903	0.954
Random Forest	3.352	0.880	0.961
Decision Tree	4.241	0.986	0.942

3.4 Cross-Validation Result

Cross-validation is a statistical technique that offers skill estimates to compare and select a prediction model. The results presented in Table 5 indicate that the ML regression models proposed in this study were effective in predicting fish weight, as evidenced by their high accuracy values ranging from 0.866 to 0.964. Additionally, the models exhibited low RMSE and MAE values, ranging from 3.417 to 6.529 and from 0.864 to 3.962, respectively.

Table 5. The performance value of cross-validation in each proposed ML model

Model Name	RMSE	MAE	R ²
Multiple Linear	6.406	3.962	0.866
Lasso	6.529	3.872	0.866
Ridge	6.389	3.925	0.866
XGBoost	3.785	0.890	0.960
Random Forest	3.417	0.864	0.964
Decision Tree	3.785	0.915	0.957

4 Conclusion, Limitation, and Future Works

This research aims to demonstrate the potential use of ML regression algorithms for predicting fish weight in the Setiu Wetland, Terengganu area. The motivation behind this methodology is to analyse the prediction performance of ML regression models by applying different statistical analysis techniques. This paper selected two regression types, Linear Regression (i.e., Multiple Linear, Lasso, and Ridge model) and Tree-based Regression (i.e., XGBoost, Random Forest, and Decision Tree model) for prediction purposes. Regarding the model performance and k-fold CV result (k=15), the ML proposed successfully predicted fish weight with low MSE and MAE values and high R² value. However, the Tree-based Regression model provides more accurate prediction results than the Linear Regression model. Among the six proposed ML regressions, Random Forest is the best predictive model with the highest accuracy value of 96.1%, while the lowest RMSE and MAE score values are 3.352 and 0.880, respectively.

The model possesses limitations that require consideration in intending to incorporate other independent variables or information, including but not limited to fork length, standard length, species, age, and sex. In addition, the modelling method does not account for environmental factors such as water temperature, salinity, pH level, dissolved oxygen, or

condition factor, indicating the fish's general well-being and health. These factors are known to influence fish growth and weight significantly.

In the future, developing more sophisticated ML algorithms and techniques can contribute to more accurate and robust prediction models. Exploring advanced deep learning and ensemble learning methods or combining ML algorithms with other techniques, such as image segmentation and processing, may yield improved prediction accuracy, generalisation, and model interpretability. By incorporating weight prediction into ecological research and management practices, we may make informed decisions supporting the conservation and sustainable use of fish populations and their habitats.

Acknowledgement

The authors would like to thank the Institute of Oceanography and Environment (INOS) management and Universiti Malaysia Terengganu (UMT) funds, for allowing us to utilise their facility and staff for assisting during fieldwork and specimen collection.

References

1. R. Froese, *Journal of Applied Ichthyology* **22**, 241 (2006)
2. R. E. Jones, R. J. Petrell, and D. Pauly, *Aquac Eng* **20**, 261 (1999)
3. L. Xing, Y. Chen, B. Li, K. R. Tanaka, R. Boenish, Y. Ren, and B. Xu, *Front Mar Sci* **8**, 1 (2021)
4. A. Maureaud, D. Hodapp, P. Daniël Van Denderen, H. Hillebrand, H. Gislason, T. S. Dencker, E. Beukhof, and M. Lindegren, *Biodiversity-ecosystem functioning relationships in fish communities: biomass is related to evenness and the environment, not to species richness*, in *Proceedings of the Royal Society B: Biological Sciences* (2019)
5. V. W. Y. Lam, E. H. Allison, J. D. Bell, J. Blythe, W. W. L. Cheung, T. L. Frölicher, M. A. Gasalla, and U. R. Sumaila, *Nat Rev Earth Environ* **1**, 440 (2020)
6. B. G. Blackwell, M. L. Brown, and D. W. Willis, *Reviews in Fisheries Science* **8**, 1 (2000)
7. D. H. Ogle, *Fish Res* **99**, 244 (2009)
8. B. Mahesh, *International Journal of Science and Research* **9**, 381 (2018)
9. S. Zhao, S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li, and R. Zhao, *Aquaculture* **540**, 1 (2021)
10. J. Li, *Theoretical and Natural Science* **5**, 658 (2023)
11. H. Saleh, J. A. Layous, and S. A. Republic, *Machine Learning-Regression*, 2022
12. G. De'ath and K. E. Fabricius, *Ecology* **81**, 2000 (2002)
13. L. Rokach and O. Maimon, *Data Mining and Knowledge Discovery Handbook* (Springer, Boston, 2006)
14. A. Cutler, D. R. Cutler, and J. R. Stevens, *Ensemble Machine Learning* (Springer, New York, 2012)
15. T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, San Francisco, California, USA (2016)
16. N. Tengtrairat, W. L. Woo, P. Parathai, D. Rinchumphu, and C. Chaichana, *Sensors* **22**, 1 (2022)
17. B. Gümüş and M. O. Balaban, *Journal of Aquatic Food Product Technology* **19**, 227 (2010)

18. Y. Yang, B. Xue, L. Jesson, and M. Zhang, *Genetic Programming for Symbolic Regression: A Study on Fish Weight Prediction*, in 2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings (Institute of Electrical and Electronics Engineers Inc.) (2021)
19. A. F. A. Fernandes, E. M. Turra, É. R. de Alvarenga, T. L. Passafaro, F. B. Lopes, G. F. O. Alves, V. Singh, and G. J. M. Rosa, *Comput Electron Agric* **170**, 1 (2020)
20. L. Sangün, O. I. Güney, P. özalp, and N. Başusta, *Iran J Fish Sci* **19**, 541 (2020)
21. S. A. K. Siti Tafzilmeriam, S. H. Nur Farah, M. L. Husain, M. A. Ambak, H. Motomura, and M. A. Ghaffar, *Length-weight relationship for 30 dominant and commonest fish species*, in 3rd Tropical Ocean and Marine Science International Symposium (TOMSY2022), Institute Oceanography and Environment (INOS) (2022)
22. S. T. Sheikh Abdul Kadir, M. Mohamad-Norizam, N. B. Baharim, T. Arai, H. Motomura, M. L. Husain, M. A. Ghaffar, and M. A. Ambak, *Greater Kenyir Landscapes: Social Development and Environmental Sustainability: From Ridge to Reef* (Springer International Publishing, 2019)
23. D. Cousineau and S. Chartier, *Int J Psychol Res (Medellin)* **3**, 58 (2010)
24. T. Yehia, A. Wahba, S. Mostafa, and O. Mahmoud, *Fundamental Research and Application of Physical Science Vol. 4* (B P International, 2023)
25. R. R. Picard and K. N. Berk, *Source: The American Statistician* **44**, 140 (1990)
26. K. K. Dobbin and R. M. Simon, *BMC Med Genomics* **4**, 1 (2011)
27. B. T. Pham, I. Prakash, A. Jaafari, and D. T. Bui, *Journal of the Indian Society of Remote Sensing* **46**, 1457 (2018)
28. Q. H. Nguyen, H. B. Ly, L. S. Ho, N. Al-Ansari, H. Van Le, V. Q. Tran, I. Prakash, and B. T. Pham, *Math Probl Eng* **2021**, 1 (2021)
29. R. Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, in Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995)
30. P. B. Ngor, S. Uy, R. Sor, B. Chan, J. Holway, S. E. Null, N. So, G. Grenouillet, S. Chandra, Z. S. Hogan, and S. Lek, *Front Ecol Evol* **11**, 1 (2023)
31. A. Abdulhafedh, *OALib* **09**, 1 (2022)