Journal of Advanced Zoology

ISSN: 0253-7214

*Volume **45** Issue **-5 Year 2024** Page 242**-256***

Detection And Growth Estimation Of Indo–Pacific Eel (*Anguilla marmorata* Quoy & Gaimard, 1824) Using Machine Learning In Central Vietnam

Kieu Thi Huyen[1], Ha Nam Thang[1] and Nguyen Quang Linh[2*],

[1,2]Faculty of Fisheries, University of Agriculture and Forestry, Hue University, 102 Phung Hung, Hue city, Thua Thien Hue, 530 000, Vietnam. Email: kieuthihuyen@hueuni.edu.vn;
[4]Faculty of Animal Sciences & Veterinary Medicine, University of Agriculture and Forestry, Hue University, 102 Phung Hung, Hue City 530 000, Vietnam.*
[3]Faculty of Biology, University of Sciences, Hue University, 77 Nguyen Hue, Hue City 530000, Vietnam.

***Corresponding authors:** Nguyen Quang Linh*
**Nguyen Quang Linh, nguyenquanglinh@hueuni.edu.vn*

ABSTRACT

Context. The Indo–Pacific eel (*Anguilla marmorata*) is a widely distributed and commercially valuable species across ecological regions worldwide. Overfishing and habitat loss are leaving the Indo–Pacific eel in a risky situation and raising a high demand for conservation. Previous research has found relationships between the Indo–Pacific eel's migration patterns and environmental factors. However, there is still a need to advance the discovery of its spatial distribution by using diverse environmental and ecological datasets and modelling its growth in terms of different environmental characterizations.

Aims & **Methods.** Here, we compared machine learning (ML) CatBoost (CB) and the multivariate linear model to investigate the relationship between spatial distribution, Indo–Pacific eel development stages, and environmental factors in central Vietnam.

Key results. Our results show that CB detected the Indo–Pacific eel at high accuracy (Overall Accuracy (OA) = 0.9, F_1 = 0.88, AUC = 0.97) and estimated the total length at different confidence levels (R2 ranging from 0.51 to 0.70), demonstrating superior performance to the multivariate linear model.

Conclusions & implications. This study highlights the potential use of ML models in species distribution mapping and modelling growth patterns to support conservation efforts of Indo–Pacific eels in their natural habitats.

Keywords: Indo–Pacific eel, ecology, environment, growth stage, CatBoost, Machine learning

1. INTRODUCTION

Indo–Pacific eel (*Anguilla marmorata* Quoy & Gaimard, 1824) is the most widespread species across the Indian Ocean, the Indo–Pacific to French Polynesia in the South Pacific Ocean (Ege, 1939; Tsukamoto et al., 2020) with significant ecological, commercial, and cultural values (Itakura et al., 2020a; Kieu et al., 2020). *A. marmorata* has been in high demand in the fisheries/ aquaculture sectors, particularly in East Asia, as a replacement for temperate eels (Pike et al., 2019). As a result, *Anguilla marmorata* populations have been in over-exploitation with an observed decline trend, possibly threatening them in the wild.

A. marmorata is found in climatic regions with habits of migration and diverse living environments in different growing stages (Itakura et al., 2020a; Itakura and Wakiya, 2020; Arai and Chino, 2019; Aquino et al., 2021; Kamai et al., 2020; 2021). Accurate documentation of the eel distribution in their natural habitats across diverse environmental conditions is essential to enhance the protection and conservation of A. marmorata in the context of climate change (Itakura and Wakiya, 2020). Despite this, the literature revealed comprehensive data on eel distribution and associated environmental parameters globally, especially in Vietnam (Solomon and Ahmed, 2016; Meulenbroek et al., 2020; Neog and Konwar, 2023), where *A. marmorata* species is prevalent in various water bodies, but adequate datasets on its habitats are lacking. In addition, we discovered an unbalanced number of research works that have leveraged available ecological and environmental datasets to delineate fish habitats and growth patterns (Roberts et al., 2022; Pickens et al., 2021; Waldock et al., 2022; Sophie et al., 2022; Schickele et al., 2019; Effrosynidis et al., 2020; Chandran et al., 2023; Yin et al., 2022), while there are fewer records of similar works for eels (Itakura et al., 2020b; Matushige et al., 2022). Although various case studies have mapped the presence of different fish species in natural habitats, none have attempted to predict the presence of eels from environmental and ecological data. The estimation of fish growth (i.e., body length and weight) has been implemented in–house farms with success (Saberioon and Císař, 2018; Tengtrairat et al., 2022; Lopez–Tejeida et al., 2023). This approach, nevertheless, requires direct measurement of the fish body or indirect measurement from the camera, which is not practical in natural habitats. Environmental and ecological datasets have different levels of relationship with the spatial distribution and growth stages of eels (Clavero and Hermoso, 2015; Glova et al., 1998; Cairns et al., 2022; Itakuringa et al., 2020b), which are potential to provide an essential dataset for successful modelling of living habitat and growth in their living environment. Unfortunately, there is likely a lack of research to develop a novel approach to estimate eel's growth and spatial distribution in such conditions.

The recent situation highlights the demand for updating species distribution databases and developing advanced methods to predict species distribution in the complex context of climate change. The entire field-based mapping of eel distribution is costly and time–consuming, making successful modelling of fish presence, which only uses a small amount of environmental and ecological data, a promising approach for future automated mapping of fish distribution. Using these databases makes it reasonable to implement functional zoning to promote sustainable eel exploitation and conservation (Pike et al., 2019).

Over the past decade, there has been a significant rise in the use of artificial intelligence (AI) and machine learning (ML) in the domains of object classification and modelling (Sandhya Devi et al., 2021; Tao et al., 2023). ML, as a non–parametric and non–linear learning model, has outperformed the parametric method in detecting objects and modelling biophysical parameters (Ha et al., 2021a; b; Ha et al., 2023; Ha et al., 2020; Pham et al., 2023), offering a more efficient and rational approach to fish detection, growth modelling and spatial distribution (Syed and Weber, 2018; Gladju et al., 2022).

One such efficient ML algorithm is CatBoost (CB) in the boosting family, which works well with classification and regression tasks using numerical or categorical data (Prokhorenkova et al., 2018). Several studies have applied CB for a wide range of research topics, including the mapping of ecosystem distribution and various biophysical parameter estimations, with a high degree of success (Fan et al., 2018, Fan et al., 2018, Ha et al., 2023; Ha et al., 2021b; Pham et al., 2023; Pham et al., 2021). Compared to the popular bagging (i.e., Random Forest) and other boosting (i.e., Extreme Gradient Boost, Light Gradient Boosting Machine) methods, CB returns reliable performance with fewer hyperparameters, which is claimed as the advantage of the ordered boosting tree approach and the L_2 parameter regularization parameter (Prokhorenkova et al., 2018). The novel decision tree approaches introduced in CB help reduce overestimation, and improve prediction accuracy for any given dataset (Fan et al., 2018; Kim et al., 2023).

In this study, we validate the performance of CB machine learning in (1) evaluating the relationship between environmental and ecological factors and the distribution of Indo–Pacific eels *(A. marmorata)* in central Vietnam and (2) comparison with a multivariate linear model to estimate the growth of Indo–Pacific eel (A. marmorata) in length across different stages of development. Our results aim to diversify methods and improve the precision in tracking and modeling the growth of Indo–Pacific eels in varying environments, providing the most recent dataset on their habitats and proposing further solutions for their conservation.

2. STUDY SITE AND METHODOLOGY

2.1. STUDY SITE

Thua Thien Hue is located in the southern part of North Central Vietnam, covering a natural area of approximately 505,000 ha with a complex topography that includes diverse habitats such as rivers, estuaries,

valleys, and mountains. The river and estuary complex networks consist of the main rivers of O Lau, Huong, Truoi, Bu Lu, Lang Co lagoon, and Tam Giang – Cau Hai, which is the largest lagoon in Southeast Asia with an area of 22,000 ha (Fig. 1). The three estuaries Thuan An (Tam Giang lagoon), Tu Hien (Cau Hai lagoon), Lang Co (Lang Co lagoon) provide routes for water exchange and create unique brackishwater in the lagoon.

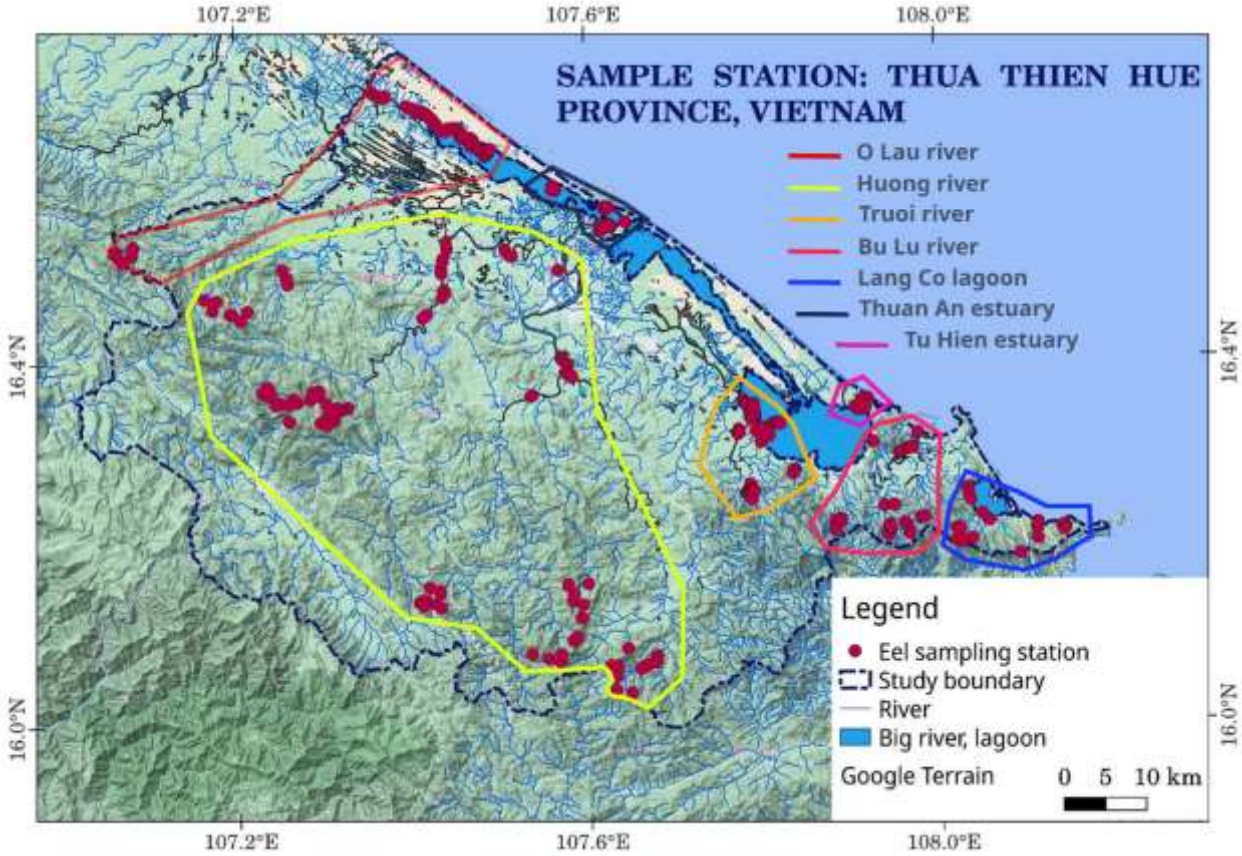


Fig. 1. Study site in Thua Thien Hue province, Vietnam, with sampling stations

In Thua Thien Hue, Vietnam, glass eels move into the lagoon through these estuaries before completing life circles in upstream rivers (Kieu et al., 2020; 2022b). Indo–Pacific eels have been found in all water bodies during the year (January – December), including phase 1 (120 – 228 mm in length – juveniles), phase 2 (187 – 410 mm – fingerling), phase 3 (387 – 1137 mm – pre–adulthood), and phase 4 (410 – 1137 mm in length – adulthood) (Kieu et al., 2022a;b).

This study identified seven (7) routes, including streams, rivers, and estuaries, where Indo–Pacific eels were present at high density and in different hydrological and ecological regions in Thua Thien Hue province, Vietnam (Table 1 and Fig. 1). Staff at the Faculty of Fisheries, Hue University, collected spatial distribution of eels and environmental and ecological parameters every month from 11/2018 to 11/2019.

Table 1. Number of study sites in Thua Thien Hue province, Vietnam

No.	Research routes	Acronym	Number of observations with eels	Number of observations without eels
1	O Lau river system	OL	34	44
2	Huong river system	SHU	105	0
3	Truoi river system	STR	49	153
4	Bu Lu river system	SBL	57	17
5	Lang Co lagoon	LC	45	23
6	Thuan An estuary	TA	30	0
7	Tu Hien estuary	TH	30	0
Total			350	240

2.2. ENVIRONMENTAL/ ECOLOGICAL PARAMETERS SAMPLING

Twelve (12) parameters were considered as the dataset for eel detection and estimation of growth length at different stages (Table 2).

Table 2. Environmental/ ecological parameters used in this study

	Parameter	Acronym	Unit	Data type	Code
1	Temperature	To	℃	Numerical data	
2	Salinity	S‰	Part per thousand (ppt)	Numerical data	
3	pH			Numerical data	
4	Dissolved oxygen	DO	mg L^{-1}	Numerical data	
5	Water depth		m	Numerical data	
6	Bottom type			Category data	Moss, flat (1), sandy (2), Stone with caves (3)
7	Tidal regime	Tide		Category data	With tide impact (1), without tide impact (2)
8	Moon phase	Moon		Category data	Moonlight (1), no moonlight (2)
9	Month			Numerical data	
10	Season			Category data	Winter (1), early spring (2)
11	Water color			Category data	Clear water with no color (1), turbid water with a change from non-color to alluvial water (2)
12	Digital elevation model	DEM	m	Numerical data	

Water samples were taken to measure temperature (℃), salinity (S‰), pH, dissolved oxygen (DO, mg L^{-1}) using the electronic thermometer, ATAGO Master S/ refractometer. MillM, Extech DO600, and Hanna HI98017, respectively. Water depth (m) was measured using a Hondex 7-depth gauge. The bottom type was collected and identified at the field site whilst tidal regime, moon phase, month, and season day were determined using meteorological and hydrological data (TTH–PSO, 2020). Eye visualization was used to identify the water color in clear or turbid water. The positions of the sampling station were recorded using a hand-held GPS Garmin 78S at the accuracy of +/- 2 m. The eel's total length (TL) was measured using a vernier scale with an accuracy of 1 mm. The developmental stages associated with migration were identified according to the description of Kieu et al. (2022a) based on morphological characteristics. Elevation data was extracted from the digital elevation model (DEM) image and sampling points in the SAGA GIS application. The DEM data was downloaded from the website https://www.eorc.jaxa.jp /ALOS/en/dataset/a w3d30/aw3d30_e.htm and was projected to the WGS–84 UTM 48N.

2.3. MACHINE LEARNING MODEL CONFIGURATION

2.3.1. CATBOOST MODEL

The CatBoost (CB) was released in 2018 as a boosting ML algorithm designed to work with numerical and category datasets (Prokhorenkova et al., 2019). Alongside the bagging ML Random Forest (RF) and the boosting ML Extreme Gradient Boost (XGB), CB has been successfully adapted to different classification and regression domains worldwide (Ha et al., 2021a,b; Ha et al., 2020). CB inherits the sequential learning of boosting algorithms, alleviating overfitting and improving prediction accuracy. CB has fewer hyperparameters, which are 3 (depth, iteration, learning rate) compared to 6 (bootstrap, maximum depth, maximum features, minimum samples leaf, minimum samples split, number of estimators) of RF and 7 (booster, gamma, learning rate, maximum depth, minimum child weight, number of estimators, subsample) of XGB, making CB cheaper to optimize and lighter in model implementation. CB uses similar binary decision trees, however, in different structures and so-called symmetric trees, proving an efficient computation and reducing prediction time. A decision tree in CB has a form as follows (Prokhorenkova et al., 2019):

$$h(x) = \sum_{j=1}^{J} b_j\{x \in R_j\} \quad (1)$$

In addition, CB introduced ordered boosting, which is a permutation-based implementation to train and evaluate the residuals in different datasets. This boosting mechanism prevents the data from leaking during the training and validation of the model, which has been a weakness of the classical boosting algorithm. In this study, CB is optimized and performed in the Python environment using the scikit–learn Python library (Pedregosa et al., 2011).

2.3.2. Model configuration and implementation

Data normalization. The input data of environmental and ecological parameters were normalized to the ranges from 0 to 1 using the scikit–learn library (Pedregosa et al., 2011). The normalized data, including twelve (12) parameters of the month, season, temperature, S‰, DO, pH, water color, depth, bottom type, moon, tide, and DEM were used as the input data for the CB model during the prediction of spatial distribution and growth length modelling.

Model hyperparameter optimization. CB hyper–parameters are optimized using GridSearchCV with five–fold cross–validation in the scikit–learn library (Pedregosa et al., 2011) (Table 3). During optimization in the searching space, the GridSearchCV tests potential combinations from a defined range of hyperparameters. The best hyperparameters were returned with the highest classification accuracy for the binary mapping of eel fish distribution and the lowest root means square error (RMSE) for modelling eel growth length.

Table 3. CatBoost model hyperparameter optimization for eel fish distribution prediction and growth length modelling

Prediction of spatial distribution				
	Depth	Iteration	Learning rate	Loss function
	3	10	0.001	Log loss
Modelling of total length				
	Depth	Iteration	Learning rate	L_2 leaf regression
Phase 1	6	120	0.05	8
Phase 2	2	80	0.13	4
Phase 3	5	100	0.2	9
Phase 4	2	150	0.15	3

Model implementation. In this study, 590 and 334 observation points were used for the binary mapping of the eel fish distribution (presence/ absence) and modelling the body length at four phases of eel growth. The observation data was divided into 40 % for model training and 60 % for model testing to predict the spatial distribution, whilst a ratio of 70 % for model training and 30 % for model testing was applied to predict the growth length (Table 4). CB model was implemented and compared to the multivariate linear model using the scikit–learn library in the Python environment. The model performance was evaluated with standard metrics. The original contribution of the input factors to the CB model was measured from the function of feature importance – a built–in feature of the CB model.

Table 4. The number of observations used for spatial distribution prediction and the growth length modeling of the eel fish

Indications	No. of observation for model training	No. of observation for model testing	Total of observation
Spatial distribution prediction	236	354	590
Growth length modelling at phase 1	42	19	61
Growth length modelling at phase 2	72	32	104
Growth length modelling at phase 3	83	36	119
Growth length modelling at phase 4	35	15	50

Model evaluation. We applied various standard metrics to evaluate CB performance. For the presence/ absence binary mapping, the model skill was measured using the overall accuracy (OA), Kappa coefficient (κ), precision (P), recall (R), F_1 and the Precision-Recall curve with Area Under Curve (AUC) scores, whilst the metrics of coefficient of determination (R^2) and root mean squared error (RMSE) were used to quantify the quality of the CB model for growth length modelling. Equations (2) – (8) present the formulas of given metrics.

$$OA(y,\hat{y}) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1}(\hat{y}_i = y_i) \quad (2)$$

in which:

$\hat{y}_i$: predictedvalue; y_i: correspondingtruevalue

$n_{samples}$: the total number of validation samples

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (3)$$

in which:

p_o: the observed agreement; p_e: the expected agreement

$$P = \frac{(TP)}{(TP)+(FN)} \quad (4)$$

$$R = \frac{(TP)}{(TP)+(FN)} \quad (5)$$

$$F_1 = 2 \times \frac{P \times R}{P+R} \quad (6)$$

in which: TP: true positive; FP: false positive; FN: false negative

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

in which: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$; ϵ: the error term; n: the total number of validation samples

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n_{samples}}\sum_{i=1}^{n_{samples}-1}(y_y - \hat{y}_i)^2} \quad (8)$$

in which: $\hat{y}_i$: predictedvalue; y_i: correspondingtruevalue

$n_{samples}$: the total number of validation samples

3. RESULTS

3.1. ENVIRONMENTAL/ ECOLOGICAL PARAMETER VARIATION

We found a variation of environmental parameters in different ecological distributions of Indo–Pacific eel (Fig. 2), in which temperature ranged between 21℃ and 32 ℃, pH varied from 6.5 to 8.6, DO changed between 6.5 and 9.5 mg L^{-1}, salinity and depth had wide ranges of 0 ‰ - 15 ‰ and 0.3 – 11 m, respectively. The eel distribution was recorded at water bodies with rocky, sandy, and stone with cave bottoms (72.8 %). In addition, the data in Fig 2. indicated a higher frequency of the Indo–Pacific eel presence with disturbances of flow, water color changes, tidal regimes, moon cycles, and floods, which was assumed as related to the migratory habits in winter, and early spring.

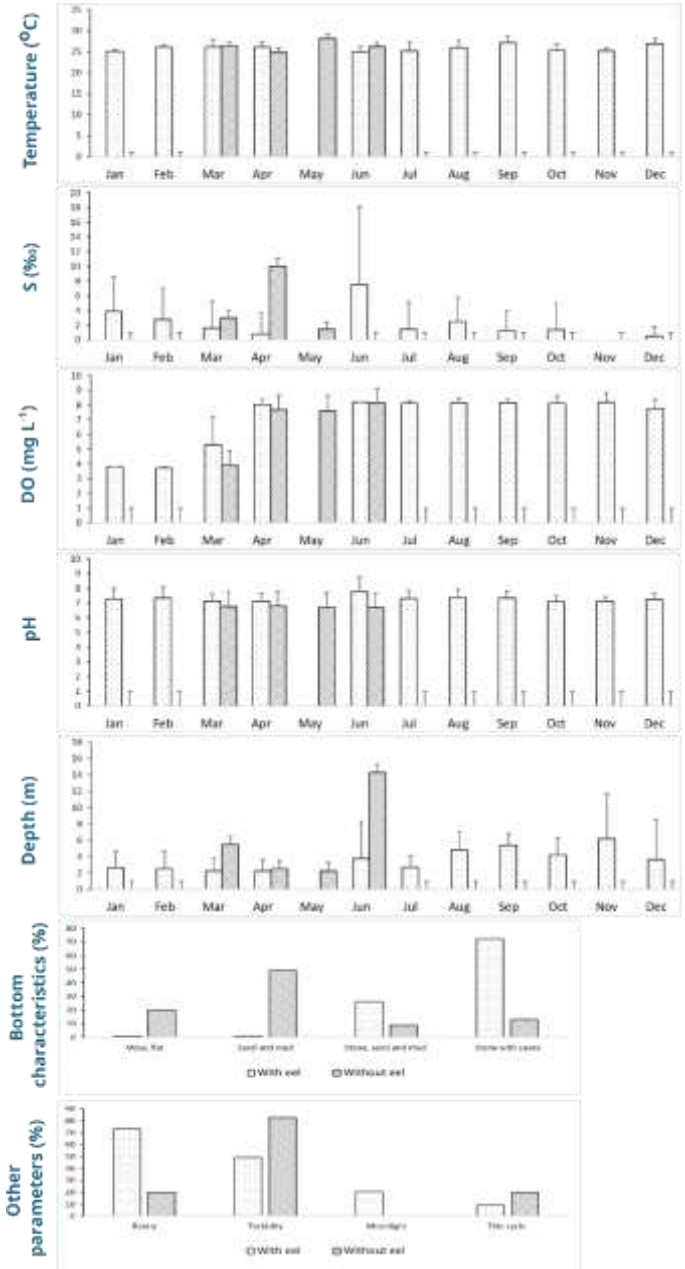


Fig. 2. Environmental/ ecological variation in study site

Correlation analysis. We examined the relationship between the total length and environmental/ ecological factors (Table 5). Accordingly, higher correlation coefficients were observed for the parameters of the month (r = 0.3), season (r = -0.29), water color (r = 0.36), and moon (r = -0.26), whilst lower values of the coefficient were obtained for S‰, DO, depth, bottom type, and tide (r ranged between -0.15 to 0.19). Temperature, pH, and DEM parameters had the lowest correlation to the total length (r varied between -0.03 to 0.09).

Table 5. Pearson correlation coefficient (r) between the total length and environmental/ ecological parameters

	Total length	Month	Season	To	S‰	DO	pH	Water color	Depth	Bottom type	Moon	Tide	DEM
Total length	1.00	0.30	-0.29	0.09	-0.11	0.10	-0.03	0.36	0.19	-0.16	-0.26	-0.15	0.07

3.2. EEL DISTRIBUTION DETECTION USING THE CATBOOST MODEL

For the given dataset, the CB model accurately predicts the distribution of eels (Table 6) with high accuracy (0.90) and confidence (κ = 0.80). Locations with eel can be detected with very high precision (0.96) and a bit lower recall score (0.82). Overall, the model performs well, with a high F_1 score for both locations with (F_1 = 0.88) and without the eel (F_1 = 0.93).

Tabel 6. Model performance of Indo–Pacific eel distribution prediction

OA	0.90	κ	0.80
	P	R	F_1
With eel	0.96	0.82	0.88
Without eel	0.88	0.98	0.93

As an imbalance exists between the number of "with eel" and "without eel" observations, the precision – recall curve was adapted to provide additional validation of the CB model (Fig. 3). A high number of area under the curve (AUC) of 0.974 indicated a consistent and reliable performance of the CB to detect the presence of eel in different environmental and ecological conditions.

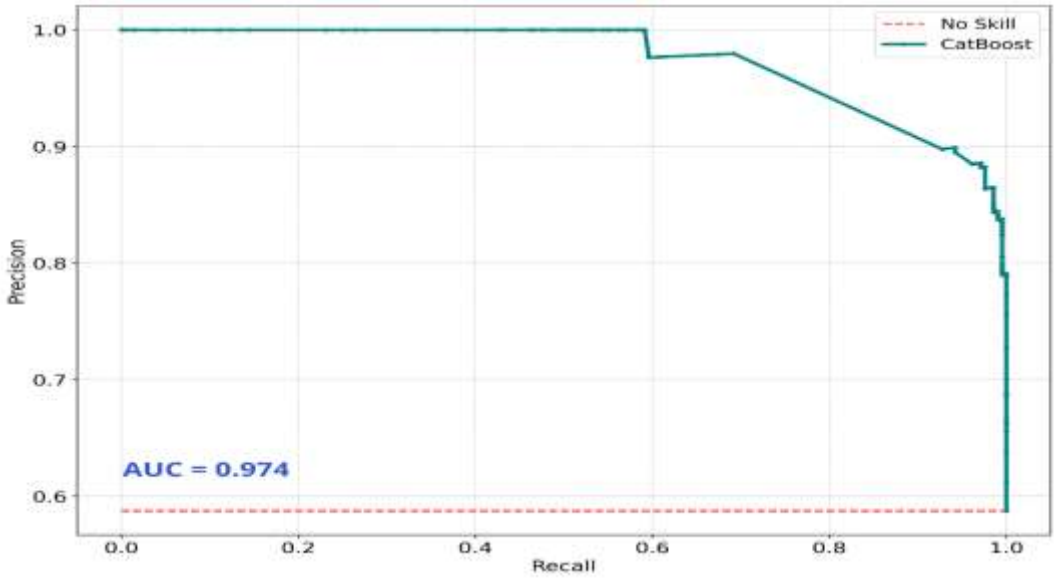


Fig. 3. Precision — recall curve of the CB model in prediction of presence/ absence of eel

In addition, we observe the differences in the contribution of input environmental and ecological parameters to the model performance (Fig. 4). Of the given dataset, the DO contributes approximately 68 % to the success of the prediction, followed by the bottom type (6.57 %), temperature (5.56 %), water color (5.23 %), elevation (4.03 %), and the season factor (3.66 %). The tide regime and moon cycle only support a very small proportion of 0.16 % and 0.53 %, respectively, whilst the factors of sampling month and salinity have no impact on the finding of places with the Indo–Pacific eel in this study.

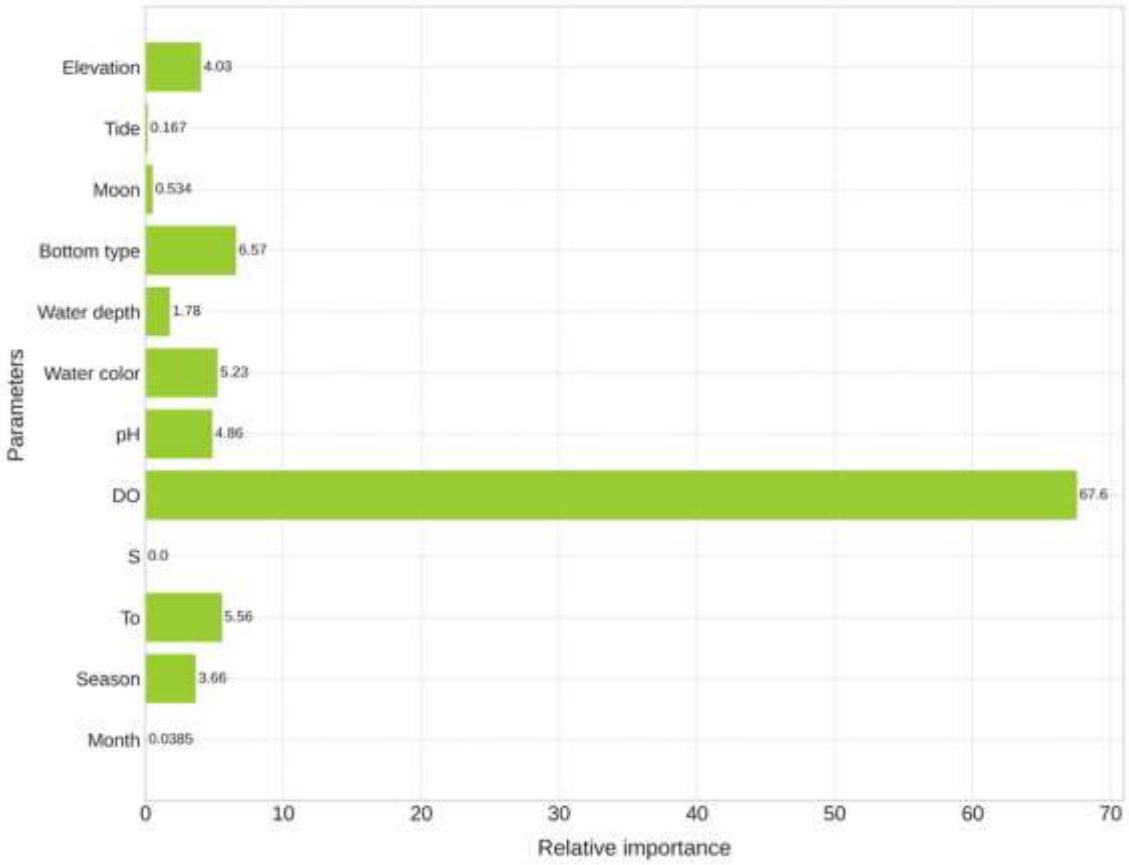


Fig. 4. The contribution of environmental and ecological parameters to the CB performance

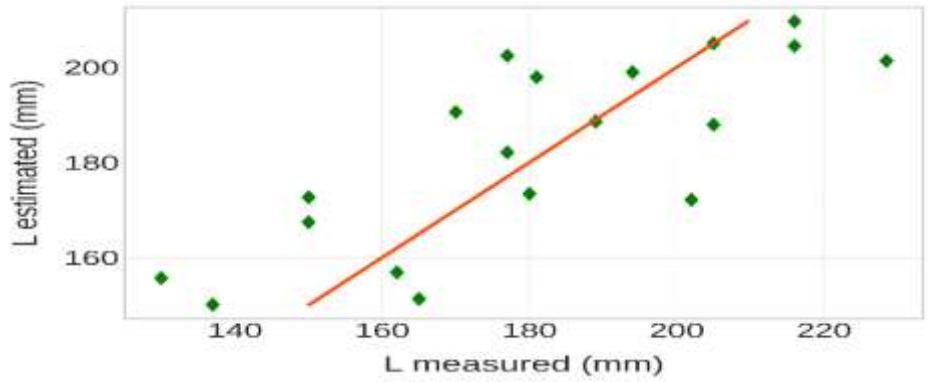
3.3. MODELING THE GROWTH LENGTH OF EEL FISH USING THE CATBOOST MODEL

Our results indicate an outperformance of the CB over the multivariate linear model for all the phases (1 – 4) (Table 7). The linear model presents the best performance in phase 1 (R^2 = 0.46), but it has an overestimation since the R^2 is only 0.29 in the training phase. In contrast, the CB model is more stable, with R^2 ranging from 0.51 to 0.70 and low RMSE values compared to the samples' mean values in the testing phases (Table 7). In phase 3 and 4, when the linear model fails to model the growth length, the CB still performs well with acceptable R^2 values of 0.51 (phase 3) and 0.70 (phase 4).

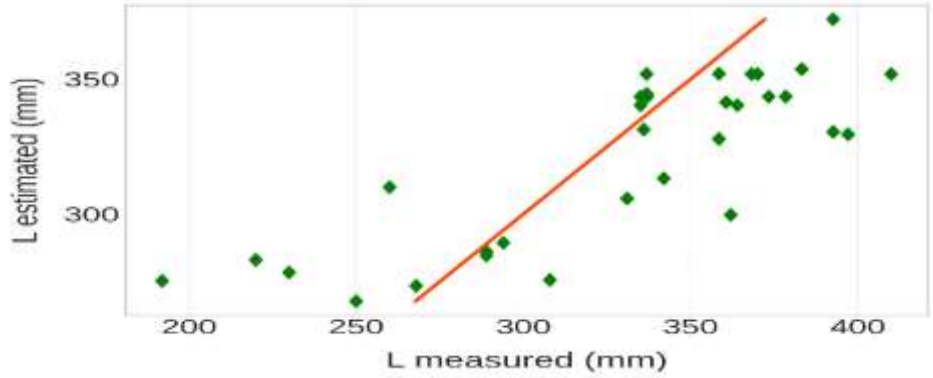
Table 7. CatBoost and multivariate linear model performance for phase 1 — phase 4

Phase	CatBoost			Multi-variate linear		Mean value of test samples (mm)
	R^2 train	R^2 test	RMSE (mm)	R^2 train	R^2 test	
Phase 1	0.74	0.60	16.86	0.29	0.46	180.76
Phase 2	0.54	0.57	35.65	0.31	0.40	329.96
Phase 3	0.79	0.51	99.30	0.22	-0.88	569.87
Phase 4	0.85	0.70	113.49	0.37	0.10	711.67

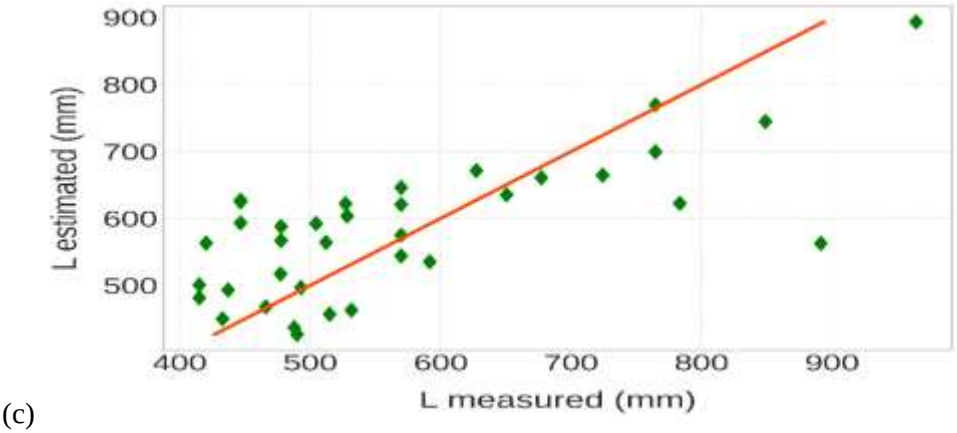
We also observe the optimal ranges of predicted growth lengths for each of the growth phases (Fig. 5 (a), (b), (c), (d)). For growth phase 1 and phase 2, the CB predicts well for the lengths between 150 – 200 mm and 260 – 360 mm, whilst wider ranges of growth lengths are predictable at higher growth phases (400 – 900 mm for phase 3 and 600 – 1100 mm for phase 4).


L estimated (mm)
200
180
160
140
160
180
200
220
L measured (mm)

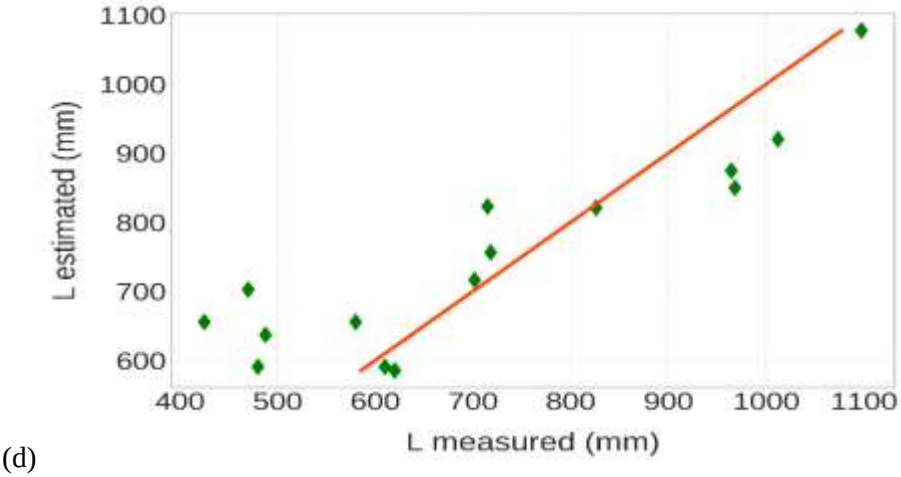
(a)


L estimated (mm)
350
300
200
250
300
350
400
L measured (mm)

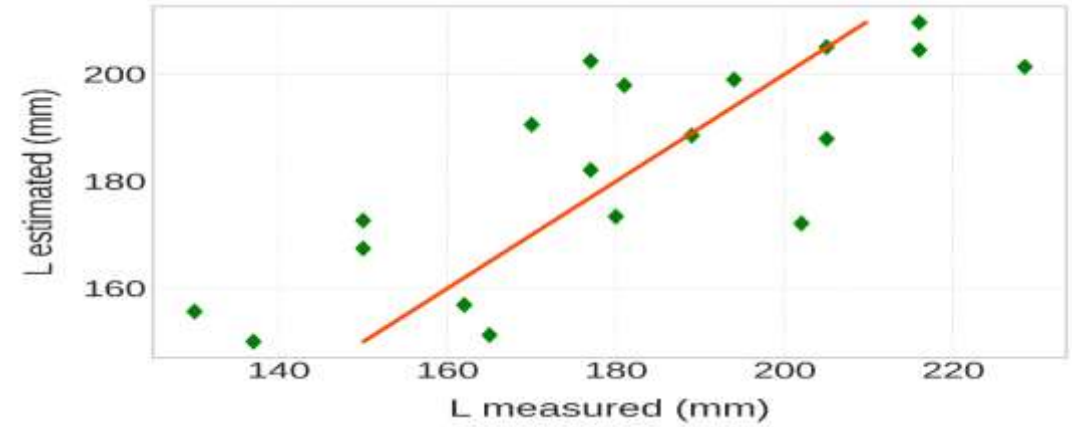
(b)


L estimated (mm)
900
800
700
600
500
400
500
600
700
800
900
L measured (mm)

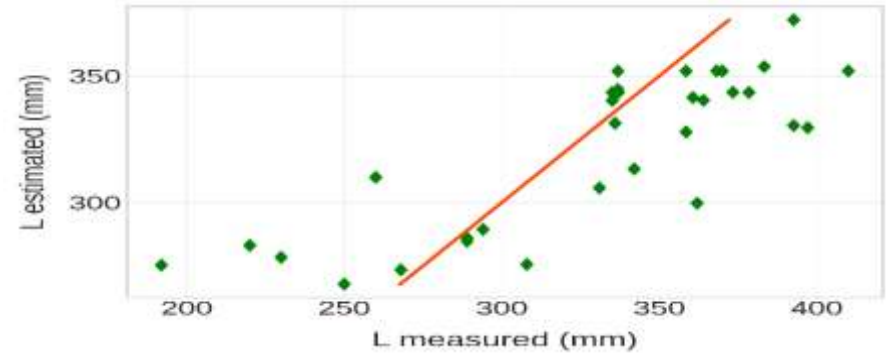
(c)


L estimated (mm)
1100
1000
900
800
700
600
400
500
600
700
800
900
1000
1100
L measured (mm)

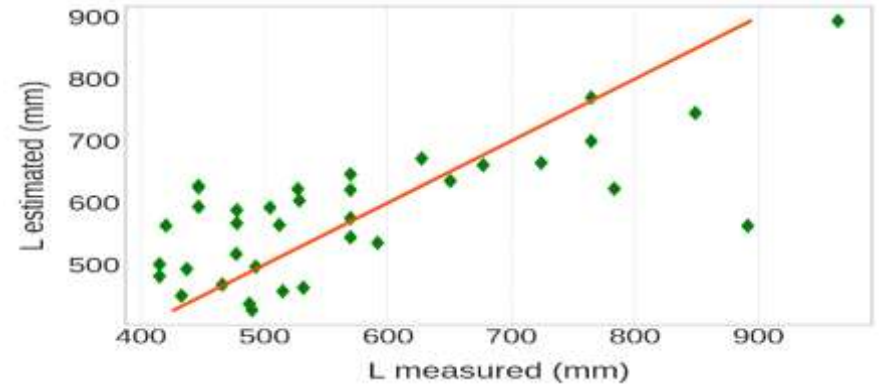
(d)



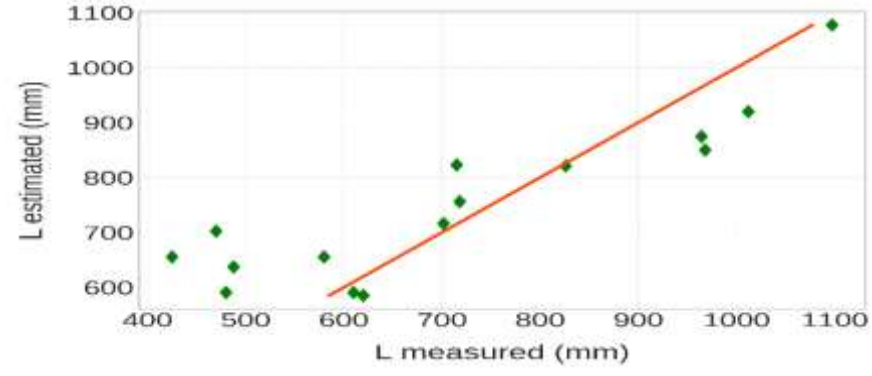
(a)



(b)



(c)



(d)

Fig. 5. Model performance (a – d) and feature importance (e – h) of input environmental and ecological parameters for the modelling of growth length phase 1 – phase 4

At the early growth stage, the moon cycle (9.6 %) and the sampling month (19 %) are the most critical factors, whilst additional factors of elevation (13.3 %), bottom type (13.2 %), DO (16.1 %), and the sampling month (14.4 %) contribute the most impact in phase 2 (Fig. 5 (e), (f), (g), (h)). The subsequent growth phases seem to be influenced by different factors of both environmental and ecological groups. In phase 3, the sampling month (22.2 %) still has a substantial impact on the modelling of growth length, together with other factors of water depth (12.4 %), pH (12.2 %), and sampling season (11.6 %). A more significant number of influential factors contribute more to the growth phase modelling in phase 4 with additional contributions of water color (22.1 %), water depth (17.6 %), bottom type (11.1 %), sampling month (12.8 %), and temperature (10.7 %).

4. DISCUSSION

The spatial distribution of Indo–Pacific eel varies greatly depending on environmental and ecological conditions and development stages, as evidenced by previous studies (Hagihara et al., 2012; Kieu et al., 2022b). Both endogenous (circadian clock) and exogenous (tidal direction) rhythms have been observed to be related to different environmental and ecological conditions, such as ocean currents (Han et al., 2012; Aoyama et al., 2018); tidal regime (Arai et al., 2020a); specific habitats (freshwater tidal limits, water body structure, depth, velocity, sediment, aquatic vegetation) (Itakura and Wakiya, 2020; Kume et al., 2019).

In this study, we first report using the ML model (CB) and environmental and ecological dataset to detect the presence and model the total length of Indo–Pacific eel in a wide range of topography, ranging from estuaries to mountainous areas. We proposed a machine learning–based method to spatially predict the distribution of Indo–Pacific eel in a broad ecological range (i.e., estuary, lagoon, freshwater rivers, streams) and then to model the fish body length at different growth phases (1 – 4). Detection Indo–Pacific eel distribution in different ecosystems shows the influence of environmental factors such as DO (68 %), bottom type (6.57 %), temperature (5.56 %), watercolor (5.23 %), elevation (4.03 %), season factors (3.66 %), however tidal regime (0.16 %) and moon phase (0.53 %) had less impact on the occurrence of Indo–Pacific eel (Table 6 and Fig. 5). The CB was able to identify the places with Indo–Pacific eel with up to 96 % precision and 88 % F_1 score, however, the recall metric was lower at 88 %, implying that the CB might overlook a few observation points and results in misclassification of the fish presence/ absence. Comparing our approach to similar works, we found that the CB model achieves higher accuracy than the results for Chanda mana (Raman et al., 2023).

Our analysis also indicates the challenges for eel growth modelling, especially in growth phases 2 and 3. The CB was superior to the multivariate linear model in estimating the total length (R^2 ranging 0.57 – 0.70) of Indo–Pacific eel using a variety of environmental and ecological variables at all the phases. Specifically, the CB is capable of estimating the growth length with confidence in phases 1 and 4 (R^2 = 0.60 and 0.70, respectively, Table 6); however, the performance might need to be improved with the estimation in phases 2 and 3 (R^2 = 0.57 and 0.51, respectively, Table 6). We discovered other works using environmental parameters as explainable variables for modelling the relationship between the length and weight of the fish *Mugil cephalus* (linear model with sea surface temperature, DO, salinity, and nutrient parameters) (Chandran et al., 2023) and the scallop *Placopecten magellanicus* (spatiotemporal model with temperature and depth parameters) (Yin et al., 2022). To our knowledge, no studies are implementing the state–of–the–art ML (CB model) to estimate the fish growth (length) from a wide range of environmental parameters, which indicates a novel contribution of our works to the field of fish ecology analysis. In addition, feature importance analysis revealed the complex living environment in different growth phases with a combined impact of environmental and ecological parameters on the growth length. The parameters of the moon, depth, season, and month had the most impact on phase 1 growth, the DEM, bottom type, water color, DO, and month had more outstanding contribution at phase 2, whilst fewer parameters of depth, pH, season, and month were attributed to phase 3 and significant influences were recorded for the parameters of bottom type, depth, water color, temperature, month in phase 4. The unclear biological development of fish individuals between phases 2 and 3 might create an overlapping of growth in phases together with the diverse living environment, explaining the underperformance of the model in these phases.

This unavoidable limitation may be due to the diversity in the topography of the study site and the deficiency of high spatial resolution data on environmental and ecological variables. The number of Indo–Pacific eel samples is different from sampling areas for the phases, which might lead to an insufficient number of data points for the learning of the model and, therefore, potentially impact the accuracy of growth length modelling at the study site. Ongoing research will focus on collecting higher spatial resolution data on environmental and ecological factors, increasing the number of sampling points in different areas, and implementing advanced feature selection methods such as Genetic Algorithms and Particle Swarm Optimization. The successful

application of the CB model for eel detection and growth estimation from a variety of environmental and ecological conditions holds promise for the development of automated mapping and modelling methods, which reduces time and cost in field surveys. These mapping databases serve as the basis for implementing functional zoning towards sustainable fishing and the conservation of eel species.

5. CONCLUSION

The Alguilla species has important economic, cultural, and ecological values. However, the community has been overfished globally, leading to significant degradation of its habitat and the number of eel individuals. Here, we propose an advanced method that utilizes the ML model (CB) to accurately detect the presence and quantify the growth total length of A. marmorata using environmental and ecological datasets across geographical ranges.

The CB model, with optimized hyperparameters, archives high accuracy in spatial distribution detection of *A. marmorata* (OA = 0.9, F_1 = 0.88, κ = 0.88) and a high value of the AUC (0.974). The parameters of DO, bottom type, DEM, water color, pH, To, and the season contributed the most information to detecting Indo–Pacific eel in the study site.

The total growth length of Indo–Pacific eel is estimated to have varying levels of success for different growth phases (R^2 ranging from 0.51 to 0.70). The CB estimates the total length in phases 1 and 4 more accurately than in phases 2 and 3. The contribution of environmental and ecological parameters differed from the growth phases, which may be attributed to differences in living habitats. The superior performance of the CB over the linear model (R^2 ranging from -0.88 to 0.46) suggests the potential application of machine learning models for various domains of fish ecology analysis and functional zoning–based conservation efforts in the future.

Acknowledgements: We thank the support from students at the Faculty of Fisheries, University of Agriculture and Forestry, Hue University; Local officials and fishermen participated during the field survey.

Conflicts of Interest: The authors declare no conflict of interest

Author Contributions: Kieu Thi Huyen, Ha Nam Thang, and Nguyen Quang–Linh developed the research ideas, methodology, and model implementation together. Kieu Thi Huyen collected the environmental and distribution data in the field. Nam Thang Ha implemented the machine learning model. All the authors have read and approved the final manuscript by Nguyen Quang Linh.

Declaration of funding: This research received no external funding.

Data availability statement: The data that supports this study will be shared with the corresponding author upon reasonable request.

REFERENCES

1. Aoyama, J., Wouthuyzen, S., Miller, M.J., Sugeha, H.Y., Kuroki, M., Watanabe, S., Syahailatua, S., Tantu, F.Y., Triyanto, H.S., Otake, T., Tsukamoto, K. 2018. Reproductive Ecology and Biodiversity of Freshwater Eels around Sulawesi Island Indonesia. Zoological Studies 57(30): 5–11.
2. Aquino, G.A.G, Cabaitan, P.C., Secor, D.H. 2021. Locomotor activity and growth response of glass eel Anguilla marmorata exposed to different salinity levels. Fish Sci 87:253–262. https://doi.org/10.1007/s12562-021-01493-x
3. Arai, T., Hussein. T. 2021. Contrasting patterns of genetic population structure in tropical freshwater eels of genus Anguilla in the Indo–Pacific. Heliyon 7(5):e07 097. https://doi.org/10.1016/j.heliyon.2021.e07097
4. Arai, T., Chino, N. 2019. Variations in the migratory history of the tropical catadromous eels Anguilla bicolor and A. bicolor pacifica in south-east Asian waters. Journal of Fish Biology. 94 (5). DOI: 10.1111/jfb.13 952
5. Arai, T., Sugeha, H.Y., Limbong, D., Tsukamoto, K. 2020a. Rhythmic activity of inshore migration of tropical freshwater glass eels of the genus Anguilla. Environ Biol Fish 103:1295–1308. https://doi.org/10.1007/s10641-020-01023-1

6. Arai, T., Taha, H., Mohd-Riduan, M.N., Mokti, S.S.A. 2020b. Molecular and morphological evidence for the identity of the giant mottled eel, Anguilla marmorata in Southeast Asia, Trop Ecol. 61(3):429–436. https://doi.org/10.1007/s42965-020-00096-4
7. Cairns, D.K., Benchetrit, J., Bernatchez, L., Bornarel, V., Casselman, J.M., Castonguay, M., Zhu, X. 2022. Thirteen novel ideas and underutilised resources to support progress towards a range-wide American eel stock assessment. Fisheries Management and Ecology 29(5):516– 541.
8. Chandran, R., Singh, R. K., Singh, A., Ganesan, K., Thangappan, A. K. T., Lal, K. K., & Mohindra, V. 2023. Evaluating the influence of environmental variables on the length-weight relationship and prediction modelling in flathead grey mullet, Mugil cephalus Linnaeus, 1758. PeerJ 11:e14884. https://doi.org/10.7717/peerj.14884
9. Clavero, M, Hermoso, V. 2015. Historical data to plan the recovery of the European eel. Journal of Applied Ecology 52:960–968.
10. Effrosynidis, D., Tsikliras, A., Arampatzis, A., Sylaios, G. 2020. Species Distribution Modelling via Feature Engineering and Machine Learning for Pelagic Fishes in the Mediterranean Sea. Applied Sciences. 10(24): 8900. https://doi.org/10.3390/app10248900
11. Ege, V. 1939. A revision of the genus Anguilla Shaw, a systematic, phylogenetic and geographical study. Dana Rep. 16:1-256
12. Elliott, S.A.M., Acou, A., Beaulaton, L., Guitton, J., Réveillac, E., Rivot, E. 2023. Modelling the distribution of rare and data-poor diadromous fish at sea for protected area management, Progress in Oceanography 210:102 924, https://doi.org/10.1016/j.pocean.2022.102924.
13. Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., Xiang, Y. 2018. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy Convers. Manage. 164:102–111. https://doi.org/10.1016/j.enconman.2018.02.087
14. Glova, G.J., Jellyman, D.J., Bonnett, M.L. 1998. Factors associated with the distribution and habitat of eels (Anguilla spp.) in three New Zealand lowland streams. New Zealand Journal of Marine and Freshwater Research. 32(:2):255-269. https://doi.org/10.1080/00288330.1998.951682
15. Ha, N.T., Manley-Harris, M., Pham, T.D., Hawes, I. 2020. A Comparative Assessment of Ensemble-Based Machine Learning and Maximum Likelihood Methods for Mapping Seagrass Using Sentinel-2 Imagery in Tauranga Harbor, New Zealand. Remote Sensing. 12(3):355. https://doi.org/10.3390/rs12030355
16. Ha, N.T., Manley-Harris, M., Pham, T.D., Hawes, I. 2021a. The use of radar and optical satellite imagery combined with advanced machine learning and metaheuristic optimization techniques to detect and quantify above ground biomass of intertidal seagrass in a New Zealand estuary, International Journal of Remote Sensing, 42(12):4712-4738. https://doi.org/10.1080/01431161.2021.1899335
17. Ha, N.T., Manley-Harris, M., Pham, T.D., Hawes, I. 2021b. Detecting Multi-Decadal Changes in Seagrass Cover in Tauranga Harbour, New Zealand, Using Landsat Imagery and Boosting Ensemble Classification Techniques. ISPRS International Journal of Geo-Information. 10(6): 371. https://doi.org/10.3390/ijgi10060371
18. Hagihara, S., Aoyama, J., Limbong, D., Tsukamoto, K. 2012. Morphological and physiological changes of female tropical eels, Anguilla celebesensis and Anguilla marmorata, in relation to downstream migration. Journal of fish biology. 81:408-26. https://doi.org/10.1111/j.1095-8649.2012.03332.x
19. Han, Y.S., Yambot, A.V., Zhang, H., Hung, C.L. 2012. Sympatric spawning but allopatric distribution of Anguilla japonica and Anguilla marmorata: temperatureand oceanic current-dependent sieving. PLoS One 7(6):374 — 384
20. Itakura, H., Wakiya, R., Gollock, M., Kaifu, K. 2020a. Anguillid eels as a surrogate species for conservation of freshwater biodiversity in Japan. Scientific Reports. 10:8790. https://doi.org/10.1038/s41598-020-65883-4
21. Itakura, H., Wakiya, R., Sakata, M.K., Hsu, H.Y., Chen, S.C., Yang, C.C., Huang, Y.C., Han, Y.S., Yamamoto, S., Minamoto, T. 2020b. Estimations of Riverine Distribution, Abundance, and Biomass of Anguillid Eels in Japan and Taiwan Using Environmental DNA Analysis. Zool Stud. 59:e17. doi: 10.6620/ZS.2020.59-17
22. Itakura, H., Wakiya, R. 2020. Habitat preference, movements and growth of giant mottled eels, Anguilla marmorata, in a small subtropical Amami — Oshima Island river. PeerJ 8:e10 187. https://doi.org/10.7717/peerj.10187

23. Kieu, T.H., Nguyen, Q.L. 2020. Phylogenetic analysis of Anguilla marmorata population in Thua Thien Hue, Vietnam based on the cytochrome C oxidase I (COI) gene fragments. AMB Expr. 10:122. https://doi.org/10.1186/s13568-020-01059-7
24. Kieu, T.H., Vo, D.N., Tran, N.N,, Truong, V.D., Vo, V.P., Tran, Q.D., Nguyen, Q.L. 2020. Using DNA barcodes based on mitochondrial COI and 16S rRNA genes to identify Anguilla eels in Thua Thien Hue province, Vietnam. Genet. Mol. Res. 19 (4):gmr18 722, http://dx.doi.org/10.4238/gmr18722
25. Kieu, T.H., Tran, N.N., Ha, T.H., Vo, V.Q., Nguyen, Q.L. 2022a. Morphological characteristics and population structure of Marbled Eel (Anguilla marmorata) in Thua Thien Hue, Vietnam. Journal of Applied Animal Research 50:1:54-60, DOI: 10.1080/09 712 119.2021.2 018 326
26. Kieu, T.H., Truong, V.D., Ha, N.T., Nguyen, Q.L. 2022b. Distribution of marbled eel (Anguilla marmorata Quoy & Gaimard, 1824) in Thua Thien Hue, Vietnam, HUAF Journal of Agricultural science and technology. 3 (6): 3142-3152. https://tapchi.huaf.edu.vn/index.php/id20194/article/view/922. (In Vietnamese with English abtracst)
27. Kim, H., Park, S., Park, H.J., Son, H.G., Kim, S. 2023. Solar Radiation Forecasting Based on the Hybrid CNN-CatBoost Model. IEEE Access 11:13 492-13 500, doi: 10.1109/ACCESS.2023.3 243 252
28. Kumai, Y., Tsukamoto, K., & Kuroki, M. (2020). Growth and habitat use of two anguillid eels, Anguilla marmorata and A. japonica, on Yakushima Island, Japan. Ichthyological Research, 67, 375-384.
29. Kumai, Y., Kuroki, M., & Morita, K. (2021). Influence of environmental parameters on habitat use by sympatric freshwater eels Anguilla marmorata and Anguilla japonica on Yakushima Island, Japan. Canadian Journal of Zoology, 99(12), 1020-102
30. Kuroki, M., Aoyama, J., Miller, M.J., Watanabe, S., Shinoda, A., Jellyman, D.J., Feunteun, E., Tsukamoto, K. 2008. Distribution and early life history characteristics of Anguillid leptocephali in the western south Pacific. Marine and Freshwater Research 59:1035-1047.
31. Lee, S.K., Jung, S.W., Son, S.J., Hwang, H.S., Kim, C.H., Oh, J.W., Hyun, B.R., Kim, D.H., Min, H.K., Cho, S.H., Kang, J.H., Byun, S.H., Han, J.H. 2020. A Study of a Conservation and Management Plan for Natural Monument No. 27 Jeju Anguilla marmorata via Landscape Analysis and Food Source Analysis. Journal of Korean Institute of Traditional Landscape Architecture 18: 28-41.
32. Lopez-Tejeida, S., Soto-Zarazua, G.M., Toledano-Ayala, M., Contreras-Medina, L.M., Rivas-Araiza, E.A., Flores-Aguilar, P.S. 2023. An Improved Method to Obtain Fish Weight Using Machine Learning and NIR Camera with Haar Cascade Classifier. Appl. Sci. 13:69. https://doi.org/10.3390/ app13 010 069
33. Matsushige, K., Yasutake, Y., Mochioka, N. 2022. Contrasting riverine distribution and habitat use of the Japanese eel, Anguilla japonica, and the giant mottled eel, Anguilla marmorata, in a sympatric river. Journal of Fish Biology 101(6):1617-1622.
34. Meulenbroek, P., Hammerschmied, U., Schmutz, S., Weiss, S., Schabuss, M., Zornig, H., Shumka, S., Schiemer, F. 2020. Conservation Requirements of European Eel (Anquilla anquilla) in a Balkan Catchment. Sustainability 12(20): 8535. https://doi.org/10.3390/su12208535
35. Minegishi, Y., Aoyama, J., Tsukamoto, K. 2008. Multiple population structure of the giant mottled eel Anguilla marmorata. Molec Ecol. 17:3109–3122.
36. Neog, P.R., Konwar, B.K. 2023. The distribution, economic aspects, nutritional, and therapeutic potential of swamp eel Monopterus cuchia: A review. Fisheries Research 261:106 635, https://doi.org/10.1016/j.fishres.2023.106635
37. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:2825–2830.
38. Pham, H.T., Nguyen, H.Q., Le, K.P., Tran, T.P., Ha, N.T. 2023. Automated Mapping of Wetland Ecosystems: A Study Using Google Earth Engine and Machine Learning for Lotus Mapping in Central Vietnam. Water 15(5):854. https://doi.org/10.3390/w15050854
39. Pickens, B.A., Carroll, R., Schirripa, M.J., Forrestal, F., Friedland, K.D., Taylor, J.C. 2021. A systematic review of spatial habitat associations and modeling of marine fish distribution: A guide to predictors, methods, and knowledge gaps. PLoS ONE 16(5): e0 251 818. https://doi.org/10.1371/journal.pone.0251818
40. Pike, C., Crook, V., Gollock, M., Jacoby, D. 2019. Anguilla marmorata (Errata Version Published in 2020). The IUCN Red List of Threatened Species 2019: E.T166 189A167 699 312. [Downloaded on 08 May 2020]. https://doi.org/https://dx.doi.org/10.2305/IUCN.UK.2019-3.RLTS.T166189A167699312.en.]

41. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems 31. https://doi.org/10.48550/arXiv.1706.09516
42. Raman, R.K., Das, A.K., Manna, R.K., Sahu, S.K., Das, B.K. 2023. Ability of machine learning models to identify preferred habitat traits of a small indigenous fish (Chanda nama) in a large river of peninsular India. Environ Sci Pollut Res Int. 30(6):16499-16509. doi: 10.1007/s11356-022-23396-9
43. Roberts, S.M., Halpin, P.N., Clark, J.S. 2022. Jointly modeling marine species to inform the effects of environmental change on an ecological community in the Northwest Atlantic. Sci Rep 12:132. https://doi.org/10.1038/s41598-021-04110-0
44. Saberioon, M., Císař, P. 2018. Automated within tank fish mass estimation using infrared reflection system. Computers and Electronics in Agriculture 150:484-492, https://doi.org/10.1016/j.compag.2018.05.025
45. Schickele, A., Leroy, B., Beaugrand, G., Goberville, E., Hattab, T., Francour, P., Raybaud, V. 2020. Modelling European small pelagic fish distribution: Methodological insights, Ecological Modelling: 416:108 902, https://doi.org/10.1016/j.ecolmodel.2019.108902
46. Tsukamoto, K., Kuroki, M., Watanabe, S. 2020. Common names for all species and subspecies of the genus Anguilla. Environmental Biology of Fishes 103(8):985-991
47. Tengtrairat, N., Woo, W.L., Parathai, P., Rinchumphu, D., Chaichana, C. 2022. Non-Intrusive Fish Weight Estimation in Turbid Water Using Deep Learning and Regression Models. Sensors 22:5161. https://doi.org/10.3390/s22145161
48. TTH-PSO 2020. Thua Thien Hue Provincial Statistical Yearbook 2019, Statistical Publishing House.
49. Tsukamoto, K. 2009. Oceanic migration and spawning of Anguillid eels. J Fish Biol., 74:1833–1852.
50. Yin, Y., Sameoto, J.A., Keith, D.M., Flemming, J.M. 2022. Improving estimation of length-weight relationships using spatiotemporal models. Canadian Journal of Fisheries and Aquatic Sciences 79(11):1896–1910. https://doi.org/10.1139/cjfas-2021-0317
51. Waldock, C., Stuart-Smith, R.D., Albouy, C., Cheung, W., Edgar, G.J., Mouillot, D., Tjiputra, J., Pellissier, L. 2022. A quantitative review of abundance-based species distribution models. Ecography 1:e05 694.
52. Wang, F.Y., Fu, W.C., Wang, I.L., Yan, H.Y., Wang, T.Y. 2014. The giant mottled eel, Anguilla marmorata, uses Blue-shifted rod photoreceptors during upstream migration. PLoS ONE 9(8):1–11. https://doi.org/10.1371/journal.pone.0103953
53. Watanabe, S., Aoyama, J., Tsukamoto, K. 2008. The use of morphological and molecular genetic variations to evaluate subspecies issues in the genus Anguilla. Coast. Mar. Sci. 32:19–29.