Cod Catch Forecasting through Machine Learning Algorithms at the Haul Level

Huamin Ren
School of Economics, Innovation and Technology
Kristiania University College
Oslo, Norway
email: huamin.ren@kristiania.no

Yajie Liu
Faculty of Bioscience, Fisheries and Economic
UiT The Arctic University of Norway
Tromsø, Norway
email: yajie.liu@uit.no

Keshav Prasad Paudel
Faculty of Bioscience, Fisheries and Economic
UiT The Arctic University of Norway
Tromsø, Norway
email: keshav.p.paudel@uit.no

***Abstract*—This paper leverages historical fishing data in conjunction with machine learning algorithms to uncover fishing patterns and more precisely forecast fishing catches. The introduction of Machine Learning techniques into the fishing industry holds significant promise for enhancing operational performance. Such methodologies can promote great efficiency and enhance the decision-making processes, optimizing factors such as fishing effort, location, and catch rates. Preliminary results illustrate the efficacy of three distinct machine learning algorithms: Linear Regression, RANdom SAmple Consensus (RANSAC), and Light Gradient Boosting Machine (LightGBM). Throughout our experimentation, it became evident that the modeling performance is profoundly influenced by the sampling strategy. This influence likely stems from inherent noise in the data, which degrades overall performance. Our findings offer insights into the effective employment of machine learning algorithms for data selection and modeling.**

Index Terms*—*Machine learning, big data, fishing catch, forecast.

I. Introduction

Machine Learning (ML) has emerged as a valuable tool for processing and analyzing big data [1]. Moreover, it proves to be an effective and efficient approach in tackling the key methodological issues and challenges encountered in modeling and analyzing various datasets in resource management. The integration of big data and Machine Learning can help to improve fisheries management, optimize resource allocation, enhance productivity and profitability, and overall sustainability. The machine learning can help fishers optimize their fishing efforts by analyzing historical catch data along with environmental factors such as ocean temperature, salinity, etc.

Norway has one of the world richest fishing grounds, making Norway the largest fishing country in Europe. Fishery has been an important contributor to the Norwegian Economy after the petroleum industry. Fish catches are affected by a multitude of factors including fishing effort, location, types of fishing vessels, socio-economic conditions and environmental variables. Climate is changing and the effects of climate change have been observed, including higher temperature, shrinking glaciers, altered precipitation patterns, frequent extreme weather, sea level rise and more acidic oceans. These changes are happening faster in the pole areas than the rest of the world. This climate changes have shifted the productivity of marine fisheries resources and habitats. Combing extensive data and ML algorithms to explore fishing patterns and forecast fish catches is a crucial aspect of aquatic research because of its relevance to establishing effective fishery management and resource allocation systems. In particular, it empowers fishers to make better decisions by optimizing their fishing strategies, thereby maximizing fishing productivity and profitability while reducing operational costs in a dynamic environment [2].

Research on fishery catch forecasting has considered both long-term catch forecasting on a scale of months or years and short-term forecasting on a scale of days. The fishery industry has reported challenges, particularly in short-term catch data, when faced with limitations of available data. Due to work cycles or actual work conditions, fishery practitioners responsible for catch data often do not have complete and detailed records, leading to a lack of real data and inaccuracies [3]. In a fisheries management context, a more detailed information on the catch composition including type of the fish at the actual haul may allow for better adaptations of management measures. In other words, at the scale of the individual fishing operation (with each haul or each trip considered), a better information on the type and catch distributions of target species may be learnt [6] [7], which not only helps spatial avoidance but could also increase the profit of fishing. Therefore, We investigate each catching behavior from the haul and attempt to study the fish catch in the long-term. We propose machine learning approaches into modelling the fish catch w.r.t. fishing location, vessel and gear type, the time of catch and other external factors.

The main objective of the paper is to use ML to explore fishing patterns and forecast fish catches. Particularly, we investigate the application of ML methods for enhancing fish catch

ISBN: 978-1-68558-089-6

forecasting. The structure of the paper is as follows. Section II provides an overview of the most recent developments in the field, highlighting the growing importance of ML techniques in addressing fish catch forecasting challenges. In Section III, we introduce our proposed ML methods and compare them with existing approaches, and then we demonstrate their performance in Section IV. In Section V, we summarize our findings and outline the future directions of our research.

II. Ralated work

While Machine Learning (ML) and Artificial Intelligence (AI) have seen widespred application in various fields, their use in natural resource management, especially in fisheries, remains relative limited. Studies such as that of Zhang et al. [3] applied ML algorithms and ensemble learning model to predict the location of albacore tuna fishing in the South Pacific, revealing that the ensemble learning model achieves higher accuracy estimates than machine learning models. Similarly, L. F. Rahman et al. [15] developed an ML approach to predict marine capture fisheries and aquaculture production in Malaysia based on past production data and climate variables, highlighting the better performance of ensemble ML model compared to the single ML model. Compared to advancements seen in machine learning application in other fields such as computer vision and healthcare systems, the progress in employing machine learning algorithms for predicting fish catches remains relatively nascent. Nonetheless, nmerous emerging research avenues in fisheries show promise. Notable attempts, like those in [8] [9] and [10], endeavor to automatically predict fish catches using past catches and meteorological information. Anothe study, [11] emphasized that prediction errors should be evaluated in a manner that goes beyond mere consideration of absolute error, regardless of the predicted value. To illustrate this, it is important to recognize that an error of 100 kg in a predicted fish catch of 5000 kg should not be treated equivalently to the same error occurring in a prediction of 500 kg. This perspective does not align with fishers' practical understanding As a result, it is suggested that evaluation metrics should be tailored to optimize prediction errors in a way that aligns with the fishers' intuition and real-world experience.

Research exploring the impact of climate change on fish catch remains limited, but has seen recent advances. For instance, O. S. Kjesbu et al. [4] examined the effect of climate change on the migration patters of North Pacific spiny dogfish, employing a ML approach. additionally, Wikstrom [14] evaluated supervised ML algorithms to predict recreational fishing success and found that random forest algorithm proved the most effective in the experiments and a combination of variables contributes optimal predictions.

III. Method

A. Problem Formulation

Given data $D = (x_1, y_1), (x_2, y_2)...(x_i, y_i), ...(x_M, y_M,$ where M is the number of the collected data. Each x_i is the n-dimensional vector, which represents the relative attributes per haul per catch, for example, start position width, start position length, sea depth start (meters), duration - (minutes), stop position width, stop position length, sea depth stop (meters), draw distance (meters), species, round weight, etc. There are 43 attributes in our studies data after some cleaning. From the data set, we estimated the model parameter vector θ appropriately expressed as:

$$y = f(x; \theta) \tag{1}$$

Our objective is to establish the estimation of Y, represented as $\hat{Y}$, by modelling of X, so that it satisfies:

$$min\|\hat{y}_i - y_i\|_2, \text{ where } \hat{y}_i = f(x_i; \theta) \tag{2}$$

B. Proposed Pipeline

We have applied three machine learning methods to implement the modelling f in Eq. 2 and compared their performance on cod catch forecasting.

1) Linear Regression
Linear Regression learns a model by minimizing the objective function in Eq. 2:

$$f(x; \theta) = \theta_i x_i + b \tag{3}$$

Equivalently, the objective is to minimize the loss in the equation below.

$$(\theta^*, b^*) = argmin_{\theta,b} \sum_{i=1}^{M} (y_i - \theta x_i - b)^2 \tag{4}$$

2) RANSAC
RANSAC algorithm normally performs the following steps [12].

Step1 Selection of samples randomly from D and have a sample set S.
Step2 Model estimation by using S.
Step3 Counting the number of data with estimation error within parameter ϵ.
Step4 Terminate the algorithm when the number of data satisfying Step3 exceeds a threshold, and model is built using those data. Otherwise, iterate the procedure from Step1.

3) LightGBM
LightGBM is based on Gradient Boosting Decision Tree (GBDT) [13], which is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability. However, GBDT is facing challenges, especially in the tradeoff between accuracy and efficiency, due to the reason that conventional implementations of GBDT need to scan all the data instances to estimate the information gain of all the possible split points. Therefore, their computational complexities will be proportional to both the number of features and the number of instances. To address such limitation, LightGBM was proposed by applying two new techniques called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), see more details in [5].

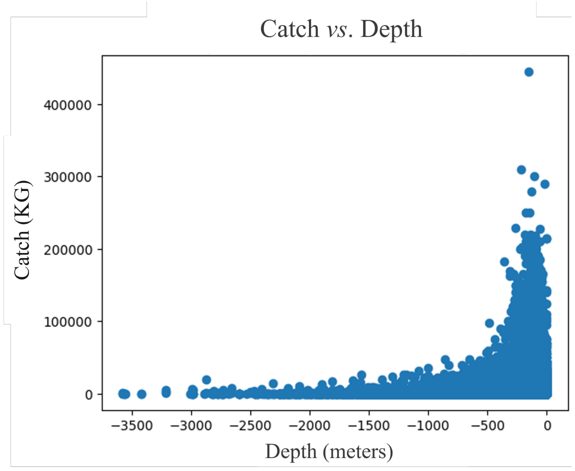

Fig. 1: Catch by depth.

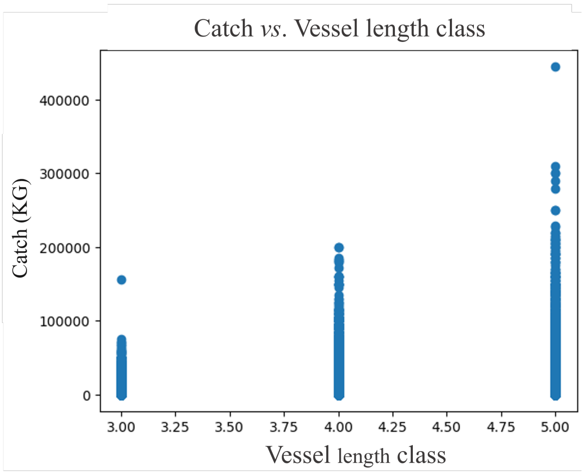

Fig. 2: Catch by vessel length class.

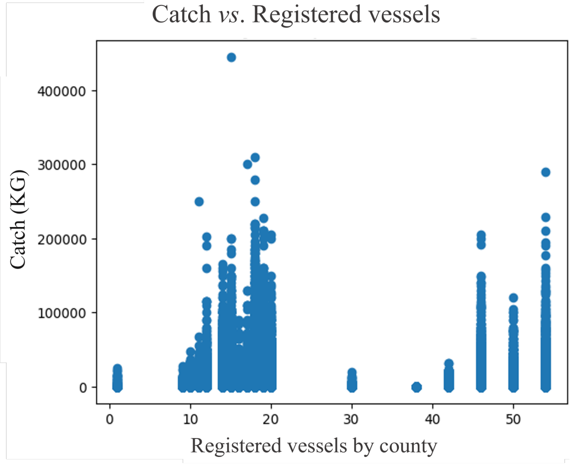

Fig. 3: Catch with registered vessels by county.

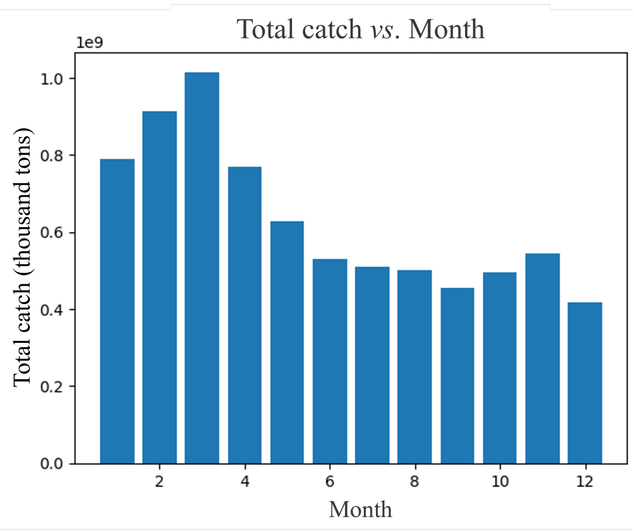

Fig. 4: Catch by the harvesting month.

IV. Experimental Results

A. Dataset Description

We used cod fishery as a case study to test the modeling and prediction. The historical fishing data were extracted from the Vessel Monitoring System (VMS) from the Norwegian Fisheries Directorate, ranging from 2000 - 2022. The dataset compromises haul time, draw distance, fishing location (width, length, depth), catch weight, and vessel characteristics (e.g., length, tonnage, engine power, etc.). We visualized, explored, and analyzed spatial relationships in ArcGIS Pro [16] to identify and remove records with unreliable position (i.e., vessel positions and/or fishing activities on land). The einvironmental variables included two oceanographic variables: Sea Surface Temperature (SST) and sea surface CHlorophyll a concentration (CHLa), three bathymetric and/or topographic variables: depth, slope and terrain ruggedness (rugosity), and two distance related variables: distance to coast and distance to nearest port were used in this study. SST and CHLa were derived from Moderate Resolution Imaging Spectroradiometer (MODIS) satellite measurements. The level 3 (4 km resolution) monthly average SST data for both Aqua and Terra satellites from the NASA Ocean Color [17] were downloaded. Observations from the four temperature images for each month (both Terra and Aqua, day – 10:30, 13:30, respectively – and night – 22:30, 01:30 respectively) were combined to calculate the monthly mean SST. Similarly, MODIS/Aqua monthly level 3 data of chlorophyll concentration were obtained from the NASA [17], Goddard Earth Science [18], and Distributed Active Archive Center [19]. The General Bathymetric Chart of the Oceans (GEBCO) gridded bathymetric data were sued. Slope was calculated from the depth data (GBECO 2023 grid). Rugosity, a measure of terrain complexity or the seabed roughness, was derived using the Benthic Terrain Modeler (BTM version 3.0). It is woth noting that seabed roughness is strongly correlated to biodiversity in marine environments.

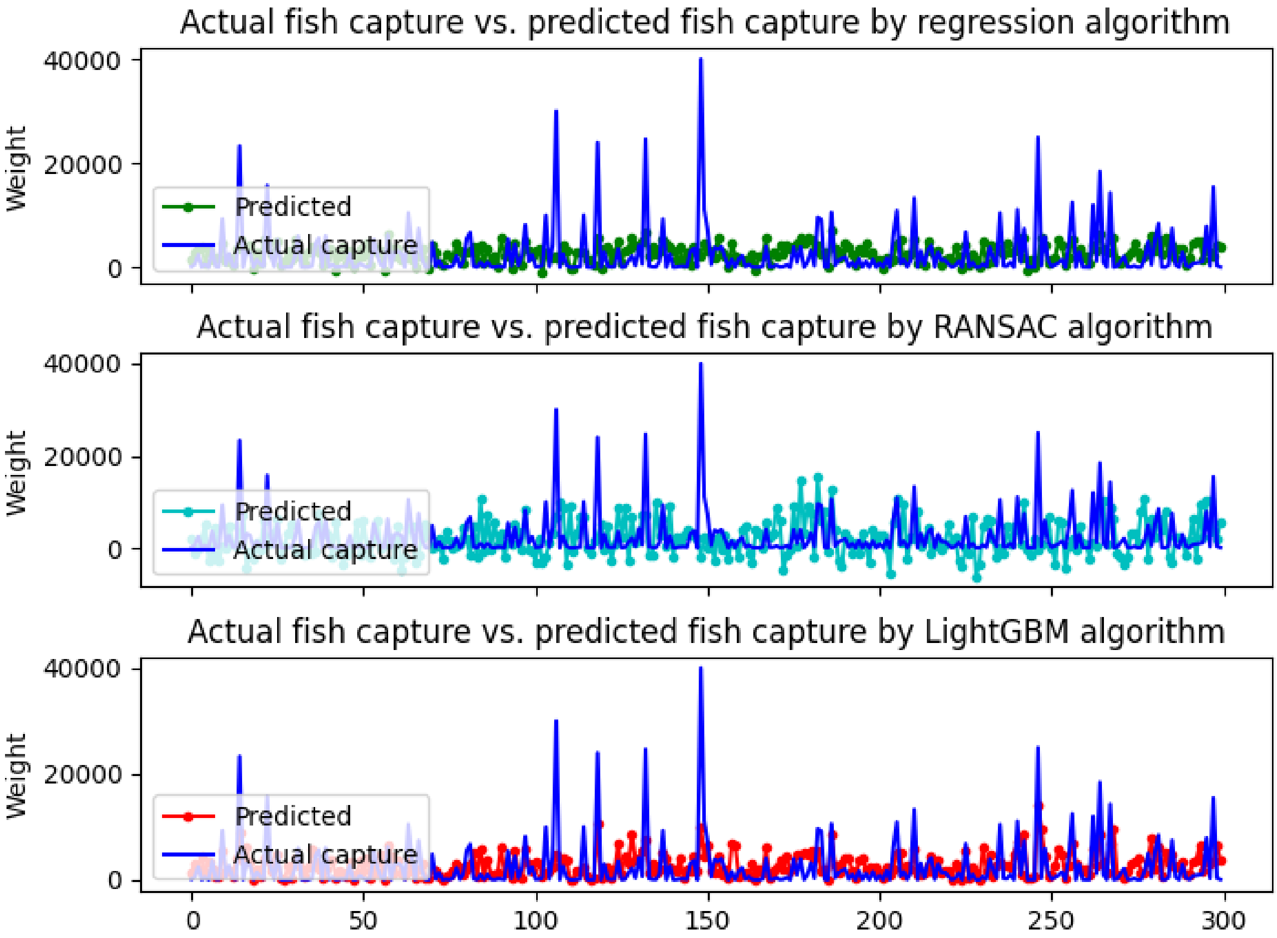

Fig. 5: Predictions of Fish catches with Regression/RANSAC/LightGBM algorithm.

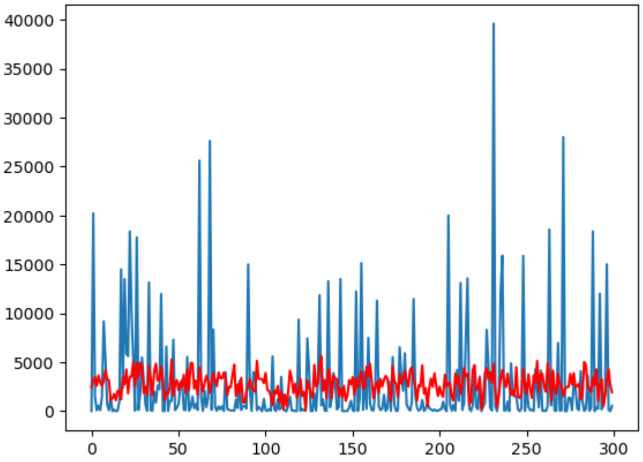

Fig. 6: Fish catch performance with Linear Regression.

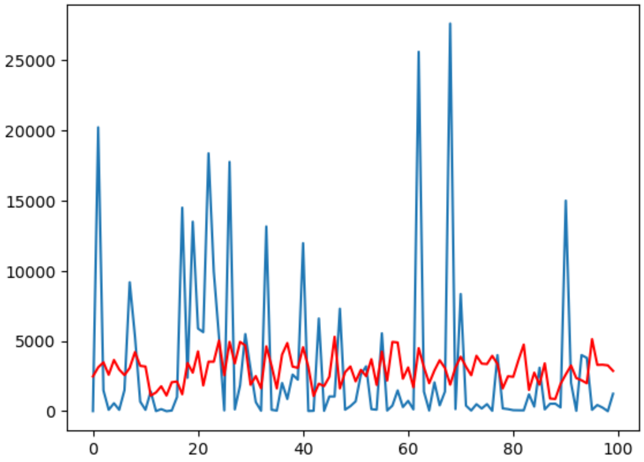

Fig. 7: Fish catch performance with RANSAC algorithm.

B. Data visualization

We conducted an in-depth analysis of the data to determine the correlation between catch weight (kg) and various influencing factors. Initially, we investigated the relationship between locations and fish capture rates. The correlation between catch and depth is illustrated in Fig. 1, and the relationship between catch and vessel length class can be observed in Fig. 2. Further insights related to catch and registered vessels by county are presented in Fig. 3. Then, the monthly total catch is depicted in Fig. 4 (calculated by total weight in thousand tons). The fish catch shows a strong correlation with factors, such as the county where the vessel is registered, the traveling distance, and the fishing month.

C. Performance of machine learning algorithms without parameter tuning

To visualize the forecasting of fish catch based on different modeling approaches, refer to the accompanying figures. Fig. 6 illustrates the performance of a linear regression model, while Fig. 7 depicts the performance using the RANSAC algorithm. A comparative assessment of both figures clearly indicates that RANSAC generally demonstrates superior performance over linear regression. The x-axis represents sampled data points. Additionally, it should be noted that the initial 100 data points in Fig. 6 correspond exactly to the first 100 data points in Fig. 7. These outcomes may be due to considerable noise in the training data, which adversely impacts the forecasting accuracy. Employing RANSAC allows for a more selective use of data for training, potentially mitigating this issue.

The comparative analysis in Fig. 5 shows that LightGBM outperforms the other two models. Due to misreporting data per trip, or per vessel often contains noise, which can substantially degrade model performance. LightGBM's data sampling strategy results in improved forecasting accuracy and performance. In gradient boosting, data points with larger gradients (errors) are crucial for calculating information gain. The Gradient-based One-Side Sampling (GOSS) technique in LightGBM retains these critical data points and conducts random sampling on the remaining data.

V. Conclusion and future work

We have applied three machine learning methods on historical fishing data in this paper to tackle a fish catch forecasting problem. Specifically, we conducted preliminary analyses to showcase the effectiveness of linear regression, RANSAC, and LightGBM, comparing their performances in fish catch predictions. Our current method still exhibits limitations in terms of model performance, particularly when dealing with data that contains a significant amount of noise. Throughout our experiments, it became evident that the influence of the sampling strategy should not be underestimated. Therefore, a more robust fish catch forecasting model that integrates advanced data sampling techniques will be one of our future research directions.

As our research evolves, several promising directions have captured our attention. A notable focus is the transformation of haul-level data into time series formats, targeting more vessel-focused or trajectory-driven models. Furthermore, we will delve into the influence of psychological factors and introduce a novel metric for assessing the accuracy of fish catch forecasting. This is especially crucial since existing error-based metrics may not fully integrate external variables and the perspectives of the fishermen.

References

[1] L. Wang and C. A. Alexander. ”Machine learning in big data”, International Journal of Mathematical, Engineering and Management Sciences, 1(2), pp.52, 2016.

[2] K. Sakaguchi and N. Yamashita, “Method to forecast the catches of Japanese common squid Todarodes pacificus in the Sea of Okhotsk off Hokkaido,” Bull. Japanese Soc. Fisheries Oceanogr., vol. 79, no. 2, pp. 43–45, 2015.

[3] Y. Zhang, M. Yamamoto, G. Suzuki and H. Shioya, ”Collaborative Forecasting and Analysis of Fish Catch in Hokkaido From Multiple Scales by Using Neural Network and ARIMA Model,” IEEE Access, vol. 10, pp. 7823-7833, 2022. doi: 10.1109/ACCESS.2022.3141767.

[4] O. S. Kjesbu et al., ”Highly mixed impacts of near-future climate change on stock productivity proxies in the North East Atlantic. Fish and Fisheries”, pp. 1–15. 2021. https://doi.org/10.1111/faf.12635

[5] G. Ke et al., ”LightGBM: A highly efficient gradient boosting decision tree”, Proc. Adv. Neural Inf. Process. Syst., vol. 30, pp. 3146–3154, 2017.

[6] M. Robert et al., ”Spatial distribution of discards in mixed fisheries: species trade-offs, potential spatial avoidance and national contrasts”, Reviews in Fish Biology and Fisheries, Vol. 29, pp. 917–934, 2019.

[7] S. Kristian, B. Plet-Hansen and C. U. François, ”The value of commercial fish size distribution recorded at haul by haul compared to trip by trip”, ICES Journal of Marine Science, vol. 77, Issue 7-8, pp. 2729–2740, December 2020. https://doi.org/10.1093/icesjms/fsaa141

[8] T. Komatsu, I. Aoki, I. Mitani, and T. Ishii, “Prediction of the catch of Japanese sardine larvae in Sagami Bay using a neural network,” Fisheries Science, vol. 60, no. 4, pp. 385-391, Dec. 1994.

[9] J. R. Leathwick, J. Elith, M. P. Francis, T. Hastie, and P. Taylor, “Variation in demersal fish species richness in the oceans surrounding New Zealand: an Analysis using boosted regression trees,” Marine Ecology Progress Series, vol. 321, pp. 267-281, Sept. 2006.

[10] Y. Kokaki, N. Tawara, T. Kobayashi, K. Hashimoto, and T. Ogawa, “Sequential fish catch forecasting using Bayesian state space models”, Proc. ICPR2018, pp. 776-781, Aug. 2018.

[11] Y. Kokaki, T. Kobayashi and T. Ogawa, ”Psychological measure on fish catches and its application to optimization criterion for machine-learning-based predictors”, OCEANS 2019 - Marseille, Marseille, France, pp. 1-5, 2019. doi: 10.1109/OCEANSE.2019.8867405.

[12] T. Watanabe, ”Initial Performance Improvement for Fuzzy RANSAC Algorithm Based on Weighted Estimation Model”, International Conference on Image Processing and Robotics (ICIP), Negombo, Sri Lanka, pp. 1-6, 2020. doi: 10.1109/ICIP48927.2020.9367332.

[13] J. H Friedman. ”Greedy function approximation: a gradient boosting machine”, Annals of statistics, pp 1189–1232, 2001.

[14] J. Wikström. ”Evaluating supervised machine learning algorithms to predict recreational fishing success : A multiple species, multiple algorithms approach”, 2015. https://api.semanticscholar.org/CorpusID:109388862.

[15] L. F. Rahman et al., ”Developing an Ensembled Machine Learning Prediction Model for Marine Fish and Aquaculture Production”, Sustainability , 13 (16), 9124, 2021. https://doi.org/10.3390/su13169124.

[16] Environmental Systems Research Institute (ESRI). ESRI Spatial Analyst. Retrieved from https://www.esri.com/en-us/arcgis/products/arcgis-spatial-analyst/ArcGIS. [Last accessed Oct. 2023]

[17] NASA Ocean Biology Processing Group. Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Level-3 Ocean Color Data. Retrieved from https://oceancolor.gsfc.nasa.gov/l3/, [Last accessed Oct. 2023]

[18] Goddard Earth Science (GES)- Distributed Active Archive Center (DAAC). Retrieved from https://daac.gsfc.nasa.gov/, [Last accessed Oct. 2023]

[19] Distributed Active Archive Center (DAAC). Retrieved from https://www.earthdata.nasa.gov/eosdis/daacs, [Last accessed Oct. 2023]

[20] FAO Major Fishing Areas. Retrieved from https://www.fao.org/fishery/en/area/27/en, [Last accessed Oct. 2023]