Estimation of the spatiotemporal distribution of fish and fishing grounds from surveillance information using machine learning: the case of short mackerel (*Rastrelliger brachysoma*) in the Andaman Sea, Thailand

Chonlada Meeanan[a,1], Pavarot Noranarttragoon[b], Piyachoke Sinanun[c], Takahashi Yuki[d], Methee Kaewnern[e], Matsuishi Takashi Fritz[d]*

[a] *Graduate School of Global Food Resources, Hokkaido University, Kita 9, Nishi 9, Kita-ku, Sapporo 060-0809, Japan*

[b] *Marine Fisheries Research and Development Division, Department of Fisheries, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

[c] *Fish Quarantine and Fishing Vessels Inspection Division, Department of Fisheries, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

[d] *Faculty of Fisheries Sciences, Hokkaido University, 3-1-1 Minatocho, Hakodate, Hokkaido 041-8611, Japan*

[e] *Faculty of Fisheries, Kasetsart University, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

*Corresponding author.

E-mail address: catm@fish.hokudai.ac.jp (Matsuishi Takashi Fritz)

Email address

chonlada@fish.hokudai.ac.jp (C. Meeanan)

pavarotn@gmail.com (P. Noranarttragoon)

ptsinanun@yahoo.com (P. Sinanun)

yukitakahashi@fish.hokudai.ac.jp (Y. Takahashi)

ffismtk@ku.ac.th (M. Kaewnern)

catm@fish.hokudai.ac.jp (T.F. Matsuishi)

[1]Present address: Fish Quarantine and Fishing Vessels Inspection Division, Department of Fisheries, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand

Highlights

- Identifying potential purse seiner operations using machine learning and surveillance data
- Relevant predictors of fishing operations are vessel speed, operation time, and instantaneous speed
- The random forest (RF) algorithm performed best to detect the fishing operations
- Simplest visualisation of fish distribution was allocation of catch landing equally to fishing position
- CPUE distribution derived from the reallocated catch to the RF's predicted fishing positions

ABSTRACT

Rapid and accurate data gathering on fish catches and the geographical distribution of fishing activities is essential in fish stock assessments and fisheries management. Surveillance systems like the Vessel Monitoring System (VMS) have become popular globally, providing accurate data on fishing activities and fish abundance in near real-time. Our study demonstrates a procedure to find the best method in machine learning algorithms for detecting fishing operations from surveillance data. VMS data of 3,237 trips conducted by 91 purse seiners targeting short mackerel (*Rastrelliger brachysoma*) in Thai waters of the Andaman Sea during 2020 were used for the study. Twenty-five spatiotemporal, environmental, and fisheries-related parameters were investigated as predictors to detect and map the fishing activity. Eight algorithms were used to compare the model performance using cross-validation. The total catch per trip was reallocated to the predicted fishing locations of the trip from the best algorithms to depict the fishing ground. The results show that operation time and vessel speed are essential elements for determining the fishing operation of purse seiners. The most-relevant predictors are the calculated speed, operation time, and instantaneous speed. The Random Forest (RF) algorithm performed the best, with 86% accuracy, and therefore was used as the final model to detect the fishing operations. However, all reallocation approaches performed similarly well in depicting fish distribution. The allocation of catch data equally to the fishing position is recommended since it is the most straightforward technique. In contrast, reallocation catch using the RF algorithm prediction ratio for visualising the CPUE distribution achieved the best performance (with a slope of 1) compared with the training data. This study provides the first decent procedures for using surveillance information to detect and identify the prospective fishing operation of purse seiners—an approach that allows for a better understanding of short mackerel distribution. The proposed procedures should help with time–area-based management of intensively exploited migratory fish species, which requires rapid and accurate information.

Keywords: Fish distribution, Machine learning, Purse seine, Surveillance data, Vessel Monitoring System

1. Introduction

Rapid and accurate determination of the geographical distribution of fish and diverse fishing activities is essential in fish stock assessments and fisheries management, especially for an area-based approach (Effrosynidis et al., 2020; Hamel et al., 2018; McCluskey and Lewison, 2008). Area–time fishing restrictions have become an essential fisheries management tool and are receiving increasing attention. This method aims to determine hotspots for target species (Garcia et al., 2014; Su et al., 2020; Thorson et al., 2019) to protect depleted stocks during the rebuilding phase of fisheries (Cochrane and Garcia, 2009; Effrosynidis et al., 2020; Palmer and Demarest, 2018), thereby reducing the risk of uncertainty to avoid overexploitation (Cochrane and Garcia, 2009; Sumaila, 1998). The method works effectively only when the area closure is appropriately constructed, including the location, size, and shape of the reserve area for habitats of the target species (Behivoke et al., 2021; McCluskey and Lewison, 2008; Sumaila, 1998). Closures are frequently designed as static measures and do not adapt to fish migration (Smith et al., 2021). Consequently, rapid and ongoing technological advancements in real-time estimations of fish distribution need to be applied to fisheries management, particularly for threatened species (Kraak et al., 2014).

The short mackerel *Rastrelliger brachysoma* (Bleeker, 1851) is a tropical epipelagic fish that forms enormous schools throughout tropical and subtropical waters of the Indo-West Pacific, from the Red Sea east to the Andaman Sea (Collette and Nauen, 1983; Koolkalya et al., 2020). This species is one of the most-captured marine species in Southeast Asia, accounting for 2.13% (300,990 Metric tons) of the region's total marine capture fisheries production in 2009 (Kongseng et al., 2020; SEAFDEC, 2011), but the stock has declined dramatically in the last 10 years (SEAFDEC, 2022).

The coastal zone of Thailand on the Andaman Sea is an essential ecosystem for short mackerel. However, between 1995 and 2006, Thailand's mackerel production gradually decreased by 46% from the average yield (Fisheries Development Policy and Planning Division, 2021). Most fisheries management systems, including short mackerel management in Thailand, rely on a combination of regulatory measures to achieve the desired objectives (Holland, 2003). Since the seasonal closure was carried out for five years, the data for annual catches have remained unchanged. Various influences can contribute to inadequate supervision, including uncertainty of fish stock migration and overlarge management areas considered only on an annual cycle (Kraak et al., 2014).

Assessment techniques and data-collection programs are frequently inadequate for determining the expected outcomes from various management initiatives (Holland, 2003). In accordance with the current state of the short mackerel, updating the structure of the conserved area or adopting dynamic closure should be considered. Real-time closure (RTC), a fishing management measure that adjusts the closure areas based on the most recent information, should be appropriate considering the high uncertainty about the species' migration (Munehara et al., 2021a, 2021b).

In order to study the dynamics of fishing activity and fish abundance through time and space, the satellite data obtained by the Vessel Monitoring System (VMS) can be used to designate fishing grounds (Gerritsen and Lordan, 2011; Zhang et al., 2021), measure fishing activities in benthic habitats, explore fisher activity (Bastardie et al., 2010; Behivoke et al., 2021) in near real-time (De Souza et al., 2016; Zhang et al., 2021), and eventually applied to enhance fisheries management (Kraak et al., 2014; Murray et al., 2013). Thailand has had a VMS system in place since 2015 to monitor the compliance of commercial fishing vessels with the fisheries laws, a relatively new development. VMS data are increasingly accessible for scientific purposes in recent years, though using the data requires pre-processing because of legal and confidentiality requirements.

The VMS reports the vessel position without classifying the activity; therefore, it is necessary to identify the activities before using this data. The most typical method for determining fishing positions from VMS data is a filtering approach based on vessel speed, operating time, turning angles (Bez et al., 2011), or employs a Bayesian hidden Markov model (Brosset et al., 2017; Nevárez-Martnez et al., 2001; Zhang et al., 2021). Nevertheless, it is also appropriate for vessels using fishing gear with a noticeable variation in speed and a sufficient operating period, such as a trawler (Gerritsen and Lordan, 2011). According to our research, purse seines are the primary fishing gear used in the short mackerel fishery, with a short operating time of 1 to 3 hours (De Souza et al., 2016; Zhang et al., 2021). However, whether a single parameter can be used to characterise fishing activity, such as vessel speed (Behivoke et al., 2021; De Souza et al., 2016), is still debatable. Since fishing operations may correlate to other parameters, including operation time, season, fishing duration, and vessel size, they may be utilised as a parameter for improved accuracy (Watson and Haynie, 2016).

Machine learning is an essential but developing technology, quickly gaining popularity (Musa et al., 2020). It is described as the algorithms and computational techniques for learning information from data without preparing equations and explicit instructions, enhancing model performance for classification and prediction (Muhling et al., 2020). The prediction has required predictors or independent variables as a property of an item that may be measured and analysed. Machine learning has been applied to determine species distributions and fishing detection. Applied algorithms were random forest (Baba and Matsuishi, 2015; Behivoke et al., 2021; Marzuki et al., 2018; Rodriguez-Galiano et al., 2015), regression tree, support vector machine (Marzuki et al., 2018; Rodriguez-Galiano et al., 2015), logistic regression (Watson and Haynie, 2016), and neural network (Muhling et al., 2020). Moreover, machine learning would provide a more appropriate and generalisable vessel analysis framework (Behivoke et al., 2021). This study explored VMS information in compliance with additional surveillance data, such as landing data, fishing logbooks, and fishing trip reports, as a predictor to improve the prediction of purse-seine operations, though limitations emanate from the short duration of this fishing method and the VMS data-transmission gap.

A prior study determined fish distribution by combining trawler landing data with fishing operations and assigning it on a rectangular grid scale (Gerritsen and Lordan, 2011; Palmer and Wigley, 2009). Similarly, we used existing data on Thai purse-seiner behaviour, which differs between vessels in terms of the number of operations on each trip, the length of the fishing trip, and other factors. In order to achieve an appropriate estimate of fishing effort and fish abundance distribution, catch allocation should be validated by integrating the fishing positions and landing data.

This study aimed to explore and validate a procedure to find the most appropriate method for integrating VMS positions and landing data to provide real-time fishing ground information using various predictors and machine learning algorithms. This approach could be applied to similar datasets to provide the spatiotemporal distribution of CPUE as a fish abundance index for stock monitoring and management guidance.

2. Materials and methods

This study used data from the purse seine fishery for short mackerel operating in the Andaman Sea, Thailand, in 2020. We described the source of the dataset and performed preliminary processing of data from fisheries surveillance, which included Vessel Monitoring System (VMS) tracking information, date and time of the fishing trip record, fishing logbooks, and landing declaration information (Supplementary text 1; Table 1). All fishing surveillance data generate several predictors, including environmental, spatial, temporal, and fisheries predictors. The prediction algorithm's performance was then evaluated. Predicted fishing positions obtained from the best model detection were combined with catch data to reallocate and estimate short mackerel distribution. The validity of various fish allocation methods was evaluated to select the optimal distribution map (Fig. 1).

2.1. Fishing activities of short mackerel purse seiners

Purse seine nets are roughly rectangular and without a specific bag and are placed vertically in the water to surround a school of pelagic fish. Several methods are used to find or aggregate the fish schools on the same vessel. Fish aggregating devices (FADs) and light luring vessels are employed to aggregate the free-swimming schools at night, while sonar or echo sounding may be used both day and night. When fish schools are found or aggregated, the purse seine is deployed to surround the schools. Once the net is completely closed, a capstan pulls the purse line powered by the main engine while the net is hauled by hand on both sides of the boat. Thai purse seiners target short mackerel or mixed fish schools composed mainly of mackerels, sardines, scads, and neritic tunas (Southeast Asian Fisheries Development Center, 2004).

According to previous research on purse-seine fishing for similar species, the vessel speed of fishing corresponds to low-speed transiting and high-speed seine operation. For schools of tuna-like species, the purse seine is moved at high speed, around 10 knots, to avoid fish escaping (Bez et al., 2011; De Souza et al., 2016; Zhang et al., 2021). A seine operation starts when the net surrounds a fish school and ends once the net is pulled out of the water. The length of the operation might extend from an hour to several hours, depending on the potential catch. During collecting fish from the water, seiners remain more or less stationary, typically moving at a very slow speed of approximately 2.5 knots or less (De Souza et al., 2016; Zhang et al., 2021). Based on this information, the most significant aspect of purse-seine activities is time (day or night) and vessel speed.

2.2. Data source

VMS data of purse seiners (at least 30 gross tonnages) operating in the Andaman Sea were gathered with the permission of 'Thailand's Department of Fisheries (DoF). Vessel position (latitude, longitude), date, time, and the instantaneous speed of the individual vessel are provided every hour. The trip records, including the date and time each vessel departed and returned to port, as well as the duration of each trip, were provided by the 'DoF's Port In–Port Out (PIPO) centers. VMS data were integrated with the fishing trip records for the short mackerel purse seine fishery to minimise the enormous VMS dataset. Based on the landing data, short mackerel caught on each trip of the individual vessel were assigned to the VMS dataset using fishing trip numbers. All vessel tracks were identified and labelled fishing or non-fishing by the fishing logbook, which provided the date, time, and vessel position (latitude, longitude) at the beginning of the fishing operation (start of lowering the net into the water). Approximately 20% of fishing logbooks were selected and used as labelled datasets to create a machine learning and validate models (Supplementary text 1; Table 1).

2.3. Predictors preparation

2.3.1. Environmental predictors

Three predictors were chosen for this study based on characteristics that impact fish distribution (Supplementary Table S1). The data on sea surface temperature (SST) and chlorophyll-a concentration were retrieved daily with 0.03° resolution, and the blank values were filled by inverse distance weighted interpolation (Kusuma et al., 2018). Eight-day composites with overlapping sampling were employed to minimise the observation loss caused by cloud cover (Muhling et al., 2020). The bathymetry was extracted with 0.03° resolution. All extraction of environmental information was obtained using the Raster package in R (Robert and van Etten, 2012).

2.3.2. Spatial predictors

The VMS data primarily offered three spatial predictors: distance from the coast, latitude, and longitude (Supplementary Table S1) (Watson et al., 2018; Zintzen et al., 2017). The distance from the coast of each VMS position was computed using the serial capabilities provided by the code accessible in Pedregosa et al. (2011) based on the nearest proximity to the coastline. The vessel positions (latitude and longitude) were obtained from VMS every hour.

2.3.3. Temporal predictors

VMS data were used to generate five temporal predictors (De Souza et al., 2016): date, date in year, month, hour, and operation time (day/night) (Supplementary Table S1). The operation time was categorised as daytime or nighttime, beginning at 06:00 and 18:00, respectively.

2.3.4. Fisheries predictors

Fourteen fisheries predictors were derived from the instantaneous vessel speed, calculated speed (distance between two consecutive positions), vessel size/length, number of fishing days, and the ratio of target catch to total landing (Supplementary Table S1). To identify diverse forms of vessel behaviour, the standard deviation and mean of the instantaneous speed, as well as the calculated speed, were determined from three consecutive VMS positions (Chuaysi and Kiattisin, 2020). Changes in instantaneous speed ($\Delta\text{Speed}_{\text{curr}-1}$, $\Delta\text{Speed}_{\text{curr}+1}$) and calculated speed ($\Delta\text{Calculated speed}_{\text{curr}-1}$) were also provided to detect the distinct behaviour of a vessel (Chuaysi and Kiattisin, 2020; Watson et al., 2018; Watson and Haynie, 2016).

2.4. Machine learning for detection of purse seiner fishing

In order to classify fishing operations, eight algorithms were employed: Random Forest (RF), Gradient Boosting Classifier (GB), AdaBoost Classifier (AB), Decision Tree (DT), Logistic Regression (LR), Multi-layer Perceptron Classifier (ML), Gaussian Naive Bayes (GN), and K-Neighbors (KN) which are described in Supplementary text 3. The supervised classifications, a machine learning algorithm that predicts an unknown dataset using a labelled dataset called the training data (Musa et al., 2020), were carried out using the scikit-learn package (Pedregosa et al., 2011; Van et al., 2009), a Python module. In order to build and evaluate a machine learning model, the labelled dataset was split into two parts: the training set and the testing set. The training set is a subset of the labelled data required to train the machine learning model. The remaining data, called the testing set, is used to evaluate the model performance.

To comprehend the purse seiners' fishing pattern, independent of the target species, we evaluated three separate predictors in three experiments, described as follows. In the first experiment, we considered all predictors. In the second experiment, environmental predictors were eliminated. In the third experiment, both environmental and spatial experiment predictors were deactivated (Fig. 1).

The relevance of all predictors in each experiment was assessed using the feature importance technique to measure the relative importance of each predictor (Géron, 2019). The scikit-learn package and the RF were employed by fitting the attribute “feature_importances_”. This algorithm determines the relevance of a predictor by examining how much impurity is minimised on average by tree nodes that incorporate the predictor across all three in the forest (Géron, 2019; Geronimo et al., 2018; scikit-learn developers, 2017).

The averaged accuracy score across the 5-fold cross-validation was used to show the performance of all models in each experiment. The total dataset was divided into the training data with 80% of all data, and the testing data with 20%. The accuracy score indicates the ratio of correct predictions to the total number of input samples (Effrosynidis et al., 2020; Geronimo et al., 2018; scikit-learn developers, 2017). Validating a model via cross-validation decreases the risk of overfitting, resulting in a more versatile model that can perform well on unknown datasets (Effrosynidis et al., 2020). The area under the receiver operating characteristic (AUROC) curve (see Supplementary text 2) and the precision, recall, and F1 score were summarised using the confusion matrix result of 80% training data and 20% testing data (Geronimo et al., 2018; Muhling et al., 2020).

The accuracy score function in binary classification provides an accuracy score for a set of predicted labels against the true label on the testing set. The accuracy is scored as 1 (true positive or true negative) if the entire set of predicted labels matches the true set of labels; otherwise, it is scored as 0 (false positive or false negative). The accuracy score function a is the ratio of correct predictions over n, and is defined as follows (scikit-learn developers, 2017):

$$a(y, \hat{y}) = \frac{1}{n} \sum_{i=o}^{n-1} 1\,(\hat{y}_i = y_i)$$

where $\hat{y}$ is the predicted value of the i sample, and y_i represents the true labelled value.

2.5. Allocating catch to the fishing operation position

Three approaches were used to reallocate the short mackerel catch on a specific fishing trip to the predicted fishing positions from the best-performance classification model of the three experiments (described in section 2.4) using the trip ID (Fig. 1).

The first approach was to redistribute the fish among all fishing positions equally (Gerritsen and Lordan, 2011), calculated as:

$$R_{ik} = \left[\frac{L_k}{P_k}\right]$$

where R_{ik} is the reallocated catch in fishing position i of trip k; L_k is the total landing of short mackerel of trip *k*; and P_k is the sum of the fishing positions of trip k.

For the second approach, the predicted fishing positions were fitted with the logistic regression (LR) model using spatial and environmental predictors (described in Supplementary text 4). The probability of occurrence at each position from the LR model was used to reallocate proportionally with the total catch for a specific fishing trip, as follows:

$$R_{ik} = L_k \left[\frac{B_{ik}}{\sum B_k}\right]$$

where R_{ik} is the reallocated catch in operation i of trip k; L_k is the total landing of short mackerel of trip *k*; B_{ik} is the probability of occurrence obtained from the LR of operation i of trip k; and B_k is the sum of the probability of occurrence of trip k.

The third approach used a model called Random Forest Regression (RFR), a type of decision tree for approximating discrete-valued functions rather than being applied for classification (Rodriguez-Galiano et al., 2015). It is a well-known model for forecasting species distribution (Zhang et al., 2021). The RFR model was built using the same training set used to identify fishing positions. However, the output or dependent variable was changed to the catch in the fishing logbook (see Logbook section in Supplementary text 1 and Supplementary text 4). The model was created using spatial and environmental predictors. Following the creation of the RFR model, the previously predicted fishing position was used as new input data and then fitted with the newly created RFR model to predict the estimated catch. Next, the estimated catch representing the probability of occurrence in each position from the RFR was normalised proportionately with the short mackerel landings and reallocated to the forecasted fishing positions, calculated as:

$$R_{ik} = L_k \left[\frac{D_{ik}}{D_k}\right]$$

where R_{ik} is the reallocated catch in fishing operation i of trip k; L_k is the total landing of short mackerel of trip k; D_{ik} is the estimated catch obtained from RFR of fishing position i of trip k; and D_k is the sum of the estimated catch obtained from the RFR of trip k.

2.6. Estimation of fishing effort and spatiotemporal fish distribution

The fishing operations obtained from machine learning of eight models with different predictors were enumerated in grids of 0.1-degree latitude by 0.1-degree longitude. Fishing effort was calculated by counting the number of predicted fishing positions within a grid. One position was assumed as 1 h of fishing because the interval of the VMS signal transmission is 1 h. To compare the different reallocated approaches, the outcomes of assigned catch to fishing operations were likewise displayed with 0.1° × 0.1° grids (Fig. 1). Similarly compared were the values of CPUE calculated from fishing effort using RF prediction divided by reallocated catch in each grid. The value of catch and CPUE in each grid from each reallocated approach and the true catch were analysed using a pair plot and evaluated using Pearson's correlation coefficient.

3. Results

3.1. The most-important predictors

The most-relevant predictors of purse seiner fishing activity were the calculated speed and the average of calculated speed, hour, operation time (day/night), and instantaneous vessel speed (Table 2). More-important predictors show a high percentage, whereas less-critical predictors show a low percentage for forecasting fishing operations.

3.2. Predictability of purse seiner fishing detection and fishing-effort distribution

All the algorithms produced results with significantly different prediction abilities. The average accuracy of the RF model was 86.14%, with a standard error of 0.23, and was the best performing. GB and AB were the second- and third-best performing, with an accuracy of 85.06% ± 0.45% (mean ± standard error) and 83.10% ± 0.74%, respectively. The accuracy scores of the top-three models did not differ significantly. The KN and ML models had the lowest accuracy and the largest standard deviation (Fig. 2). The precision, recall, and F1-score were reported in this experiment to confirm the model's effectiveness in binary identification of fishing activities, as shown in Supplementary Table S2. The top-three AUROCs were RF (93.4%), GB (93.3%), and AB (91.6%) (Fig. 3). Three experiments revealed that binary classification performance on fishing and non-fishing activities adopted a similar pattern.

The number of fishing positions predicted by the best model (i.e., RF) in experiments 1, 2, and 3 can be seen in Fig. 4 by counting the number of fishing positions in the 0.1° × 0.1° grid; overall, the short mackerel fishery showed intense pressure when calculating the number of predicted fishing positions, including in the north of the Andaman Sea (off Ranong Province), in the middle of the Andaman Sea (off Krabi Province), and in the south of the Andaman Sea (off Trang and Satun provinces), as depicted on the grids.

However, in terms of the total number of fishing positions predicted, each model was different, as shown in Fig. 5. The top-three best-performing models (RF, GB, and AB) forecasted the number of fishing positions as being in the range of 18,003–20,420. The GN model predicted the significantly ($p < 0.05$) highest number of fishing positions in the range 37,688–37,887 of all experiments. In contrast, the ML model predicted the fewest fishing positions in all experiments, ranging from 3,560 to 17,635. The lowest significant value was determined in experiment 3 ($p < 0.01$) (Supplementary Fig. S1).

3.3. Distribution of catch reallocation and estimated CPUE

3.3.1. Catch reallocation and catch distribution

Fig. 6 depicts the catch distribution for all reallocation algorithms validated using training data in each experiment. It was revealed that all approaches could map the catch distribution in near approximation to the training data (Fig. 6). The grid with training data was compared with each redistribution approach by visualising the correspondence among all grids. Allocating catch landing to predicted fishing positions resulted in the virtual training catch with a 1.0 slope and Pearson's coefficient of 1.0 (Fig. 7).

3.3.2. Estimation of CPUE distribution

Fig. 8 depicts a comparison of the CPUE distribution derived from the reallocated catch to the RF's predicted fishing positions of each approach with the CPUE from training data. When the CPUE estimation was compared with the training data, it was observed that the fishing positions in the North Andaman Sea were similarly distributed in all techniques. However, there were minor CPUE variations over the South Andaman Sea. Additionally, the CPUE distribution results were consistent with the training data. The results were significantly associated with a slope of 0.8–1.0 and Pearson's correlation coefficient of 0.9–1.0 (Fig. 9).

4. Discussion

4.1. The influence of a predictor on forecasting purse seiner activity

The purse seiner is sped up to 10 knots for setting the net, and thereafter reduces its speed to 2.5 knots or less when hauling in the net (Bez et al., 2011; De Souza et al., 2016; Zhang et al., 2021). As a result, low-speed filtering by itself cannot capture all behaviours. Moreover, purse seiners have a tight fishing activity (1–3 h), making it challenging to detect fishing patterns using speed from VMS tracking. However, the same data set was used to detect fishing operations using the filtering approach at a speed threshold of 1.81 knots (derived from the average vessel speed of fishing operations in the training data) and nighttime operations. The accuracy score was 81%, with 0.60 precision and 0.68 recall of the fishing operation prediction.

Machine learning could capture speed patterns of fishing activities using various vessel speed predictors, such as the average and standard deviation of instantaneous speed and calculated speed obtained from three consecutive VMS positions. The filtering approach typically analyses the speed at transmission time to identify fishing activities. Because VMS has a 1-h transmission gap, actual fishing may occur during the gap. As a result, relying only on transmitted speed may not accurately represent the actual amount of fishing (O'Farrell et al., 2017). This may lead to machine learning being more productive than traditional filtering methods.

In this research, the calculated speed predictor outperformed the instantaneous speed predictor in predicting fishing activity. At the start and finish of each fishing operation, the vessel speed of short-term fishing, such as the purse seines operation, changes largely. As a result, the instantaneous speed may not accurately represent the vessel speed during the fishing operation. Previous research suggested that the timing gap between ping signals affects the accuracy of fishing-activity detection, particularly for short-term fishing activity, owing to a vessel deploying the fishing gear before or after transmission location (O'Farrell et al., 2017).

Previous research has used an average speed or the difference in speed obtained from the signal position of consecutive VMS positions to improve forecast accuracy (Muench et al., 2018; Watson et al., 2018; Watson and Haynie, 2016). Furthermore, the average speed enables the model to capture the shifting boat speed patterns of fishing that are only observable in purse seiners, which tend to speed up and then float. The shifting patterns are expected to signify the following: low average speed may indicate that it is time to collect fish on the boat, while a boat with an average speed of more than 1 knot may indicate that it is time to begin surrounding and collecting fish. This should be the time when the boat runs to search the fishing grounds if the average speed is high. Furthermore, we observed that hour and operation time might effectively predict fishing-activity classification. High-frequency fishing was seen from all training datasets at nighttime, from approximately midnight to before dawn. This result is consistent with a previous study's classification of purse seiner activities based on operation time (De Souza et al., 2016). In this study, purse seine vessels could fish both during the day and at night.

Additionally, the same vessel is available, employing various techniques to find fish, such as FADs, light, and echo sounders. FADs and light-enticing vessels are employed to aggregate the fish schools at night. When considering using light to attract fish when fishing, it implies that the natural light's brightness must be manipulated and distorted. We anticipated that the brightness value of natural light would have little impact on detecting fishing operations. In contrast, it has been demonstrated that decision-making in fishing operations throughout the operating period (day and night) is an important variable. Perhaps because of the variety of techniques that can be used on a single vessel; for example, certain vessels obviously work a lot at night, as the primary method for attracting fish is using light.

Although vessel speed and operating time are reliable indicators, our study's predictive accuracy was lower than that of De Souza et al. (2016), which also employed these variables to classify the fishing operation activities in purse seine. The fishing pattern of the purse seine described by De Souze et al. (2016) was more obvious when compared to Thai purse seiners. Firstly, with few exceptions, most purse seiners do not fish at night. Although the vessel in our study could fish both during the day and at night, it typically operates at night. As a result, identifying fishing operations becomes more challenging and requires the employment of additional variables. Therefore, it can result in lower accuracy in our study. Second, we anticipated that the sources of the labelled data used for comparison would also differ. A prior study employed the term" "expert label," which describes the fishing location from their professional experiences. The filtering criteria were determined based on experience reviews, expert interviews, and interviews with observers on board and fishermen. Therefore, the experts' labelling method forms the basis for the classification of false. In contrast, our study compared the accuracy value using label data from the logbook.

Environmental predictors, namely SST, chlorophyll-a, and bathymetry, were found to be less critical than spatial predictors (i.e., distance from shore, latitude and longitude) (Muhling et al., 2020; Watson et al., 2018) as shown in Table 2. As a result of the distribution of short mackerel, the fishing grounds were not far from shore (12.39 ±5.22 km). There were few temperature, depth, or chlorophyll-a changes near the shore. So, the environmental variables may become less important in the predictions of this fishery. In addition, this research aimed to categorise fishing positions (fishing or not fishing). The vessel's behaviour, such as its speed or duration, has a greater impact on fishing operation activities than environmental factors. In previous studies, the distribution of fish was predicted using environmental factors such as SST, chlorophyll-a, and bathymetry (Effrosynidis et al., 2020; Fabio et al., 2019; Geronimo et al., 2018; Muhling et al., 2020; Nguyen and Nguyena, 2017; Nurdin et al., 2015). These were quite different to the classified fishing activity in our study. We included them in determining if they would help identify fish-specific fishing locations, even though the previous study employed a different algorithm and had a different target.

The findings of separating the predictors into three experiments confirm this speculation. We included environmental and spatial predictors in Experiment 1, whereas we excluded environmental predictors in Experiment 2. The results showed that the predicted fishery positions from the two experiments using the best-performance model (i.e., RF) were quite close, with 19,581 and 19,582 points, respectively. Therefore, the predicted outcomes remained the same even if all three environmental factors were removed. When environmental and spatial predictors were not included in Experiment 3, the RF model predicted a larger number of fishing operations without specific species (20,420 points). Forecasts performed without considering environmental and geographic variables resulted in greater predicted fishing positions. In other words, the model could predict all fishing activities regardless of the type of fish species.

4.2. Fishing activity detection by machine learning

None of the three predictive models (RF, GB, or AB) resulted in overfitting owing to each randomisation of the predicted value achieving the same level of accuracy from the cross-validation approach. In contrast, the other models produced more variation in model performance at each of the five fittings. As it was based on bootstrap aggregation, a technique for resampling with replacement to reduce variance and enhance accuracy, RF performs well at reducing data variation. Tree-based models may also be resistant to redundant data or data with high correlation, both of which can cause overfitting in conventional learning algorithms (Kirasich et al., 2018). The classification performance of the DT model was significantly lower than that of the RF model. Although this model is based on the same principle as the RF model, it employs a single estimator; thus, the lower ability was unsurprising. Previous research suggested that the volume of training data used to train a model might affect its effectiveness, mainly through supervised machine learning. Jinseok and Jenna (2018) state that classification models for LR should ideally be based on a 1:1 ratio of positive and negative training data. The amount or fraction of training data in our investigation was inadequate for this model, as it could have reduced the accuracy. Because there were few fishing operations per all transmitted positions, we employed a ratio of about 1:3 in this experiment.

Furthermore, the amount of information training influences the KN model's accuracy. According to Musa et al. (2020), having more training data improves prediction accuracy since more data points provide us with more information. In the case of KN, we will have closer nearest neighbours, which will tell us more about the location we wish to forecast.

The proportion of fishing and non-fishing in the best-predicted RF model is consistent with the training data (Supplementary Fig. S2). In addition to considering the proportion of fishing positions and total vessel positions in the training data set, it was discovered that the proportions of RF were similar to the proportions of fishing positions of the short mackerel fishery (~20%) from the training data. However, in experiment 3, the RF model estimated the total number of fishing positions near fishing positions in the training data for non-specific species (~24%) (Supplementary Fig. S2). Therefore, considering the accuracy score, the AUROC of the prediction, and the proportion of the estimating techniques of fishing positions, the top-three models (RF, GB, and AB) were deemed capable of predicting fishing activities from all three categories of indicators. It depends on the purpose for which we require the fishing position when deciding which predictor is the best option. Only the fisheries predictor needs to be considered if we need to anticipate all fishing activities. We need to consider spatial and environmental predictors to predict fishing activities with a specific species.

We used logbook data from 16,857 positions from 654 fishing trips, accounting for 20% of all, as the practical information for training and validating machine learning. However, in some cases, inscription errors may exist in the fishing logbook. In previous studies, onboard observers or expert reviews were deployed to collect label data and verify the data to classify the 'vessel's activities (Bez et al., 2011; De Souza et al., 2016; Watson and Haynie, 2016). Currently, there are no observers onboard Thai vessels operating in Thai waters, but the vessels keep fishing logbooks to report the fishing activity. As a result, a fishing logbook is a helpful alternative for collecting label data for machine learning supervision in this study. However, this information should be cross-checked with VMS data (Palmer and Wigley, 2009) and through literature reviews and interviews with fishers. Therefore, we reviewed the information in the logbooks, one by one, before applying the logbook data.

4.3. Estimation of fish distribution by reallocation of landing data

We assert that the most straightforward and appropriate approach for mapping 0.1° × 0.1° grids is to reallocate the catch equally at each fishing station. This conclusion is consistent with the work of Gerritsen and Lordan (2011), who distributed landing data uniformly with a similar rectangle grid resolution. The average catch assigned to each haul position was approximately the same as the training weight, resulting in a strong correlation with a slope of 1. Even though trawlers appear to occupy a larger area per haul than purse seiners, the redistribution findings were close to the training data. The purse seiner as well can benefit from this equal-allocation approach.

Nevertheless, it would be better if we first used the predicted fisheries positions to predict the fish distribution by the RFR model with the environmental and spatial predictors again. The estimated fish numbers were then normalised proportionally with the short mackerel landings and redistributed to predicted fishing positions. This approach should be more successful than using data to estimate CPUE, confirming the distribution outcomes in Fig. 9, which indicate the greatest slope and coefficient correlation. Thus, we will get more accurate information for fish distribution using a 0.1° × 0.1° grid size.

4.4. Applications for fisheries management

The application of geographically explicit data, such as the combination of VMS and landing data, to estimate fishing operations and fish distribution can considerably improve spatially dynamic fisheries management. Due to the limited data, technology, and specialised expertise at the time, most fisheries management was previously done on an annual or multi-year basis and referred to large management areas, resulting in an inadequate spatial and temporal resolution (Kraak et al., 2015). Additionally, fish abundance changes on considerably smaller spatiotemporal scales. Therefore managing fishing mortality at a finer scale may help to regulate it more successfully (Kraak et al., 2014). The development of VMS and logbook data, along with properly recorded fishing (Bez et al., 2011; Gerritsen and Lordan, 2011; Jennings and Lee, 2012), has made it possible to manage fisheries at a considerably more realistic level of geographic analysis leading to the spatially dynamic fisheries management approach than the conventional management.

Real-time spatial management for controlling bycatch and discards have been made recently accessible in the US and Europe (Little et al., 2015). Implementing a real-time closure in the North Sea to protect cod (Gadus morhua), which are caught as bycatch, was mentioned as an example of applying real-time spatially explicit data. The hotspots of cod abundance were determined in a small area (7×7 nautical miles square) using the integration of VMS and landing data from the preceding 14 days, which were then designated as real-time closures (21 days intervals) for minimising fish mortality and discarding (Holmes et al., 2011). Therefore, the designation of fishing grounds, like in the present study, could be more useful in capturing both the biological dynamics of stocks and the associated activities of the vessels. This could help spatially inspired management systems like real-time closure. Our approach would provide the ability to track fishing activity and fish distribution at a fine spatiotemporal scale, allowing us to target the species with abundant agreed-upon fishing possibilities while avoiding concentrations of the "endangered species" and/or other sensitive areas.

Currently, the management of short mackerel is the seasonal closures. Imposing protected areas could be an alternative solution for protecting spawning habitat, providing recruitment sources throughout the area, and maintaining the natural population when other fisheries management methods are impracticable (Sumaila, 1998). The quantity of short mackerel caught, however, is still demonstrated to have remained constant. Despite the management measures implemented during the last five years, this may indicate that they need to be reviewed. This can be related to the uncertainty regarding the migration of fish stocks. According to the current state of this species, the existing area closure design must be considered since it may not reflect the proper design. As a suggestion, updating the conserved area structure or adopting dynamic closure should be considered.

5. Conclusions

By using key predictors and machine learning, the fishing position from VMS can be integrated with landing data to provide spatiotemporal distribution of CPUE as a proxy of fish abundance. Alternatively, the approach could be applied to existing data to allow for near-real-time monitoring of fleet dynamics and fish distribution. This information will be precious in adopting effective time–area management designs, such as dynamic closures and rebuilding of depleted fish stocks or migration species, which require immediate scientific data.

Acknowledgements

We thank the Department of Fisheries (Thailand) and all administrative staff for providing the surveillance data, as all data needed to be pre-matched before our analyses owing to regulations around confidentiality. Cynthia Kulongowski with Edanz (https://jp.edanz.com/ac) edited the language of a draft of this manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at DOI: …

References

Baba, S., Matsuishi, T., 2015. Pacific saury fishing forecasting by using random forest. Nippon Suisan Gakkaishi 81, 2–9. https://doi.org/10.2331/suisan.81.2

Bastardie, F., Nielsen, J.R., Ulrich, C., Egekvist, J., Degel, H., 2010. Detailed mapping of fishing effort and landings by coupling fishing logbooks with satellite-recorded vessel geo-location. Fish. Res. https://doi.org/10.1016/j.fishres.2010.06.016

Behivoke, F., Etienne, M.P., Guitton, J., Randriatsara, R.M., Ranaivoson, E., Léopold, M., 2021. Estimating fishing effort in small-scale fisheries using GPS tracking data and random forests. Ecol. Indic. 123. https://doi.org/10.1016/j.ecolind.2020.107321

Bez, N., Walker, E., Gaertner, D., Rivoirard, J., Gaspar, P., 2011. Fishing activity of tuna purse seiners estimated from vessel monitoring system (VMS) data. Can. J. Fish. Aquat. Sci. 68, 1998–2010. https://doi.org/10.1139/F2011-114

Brosset, P., Fromentin, J.M., Van Beveren, E., Lloret, J., Marques, V., Basilone, G., Bonanno, A., Carpi, P., Donato, F., Čikeš Keč, V., De Felice, A., Ferreri, R., Gašparević, D., Giráldez, A., Gücü, A., Iglesias, M., Leonori, I., Palomera, I., Somarakis, S., Tičina, V., Torres, P., Ventero, A., Zorica, B., Ménard, F., Saraux, C., 2017. Spatio-temporal patterns and environmental controls of small pelagic fish body condition from contrasted Mediterranean areas. Prog. Oceanogr. 151, 149–162. https://doi.org/10.1016/j.pocean.2016.12.002

Chuaysi, B., Kiattisin, S., 2020. Fishing vessels behavior identification for combating IUU fishing: enable traceability at sea. Wirel. Pers. Commun. https://doi.org/10.1007/s11277-020-07200-w

Cochrane, K.L., Garcia, S.M., 2009. A fishery manager's guidebook, 2nd ed. Food and Agriculture Organization of the United Nations and Wiley-Blackwell.

Collette, B, B., Nauen, E, C., 1983. FAO Species Catalogue, Vol. 2: Scombrids of the world. An annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date. Rome.

De Souza, E.N., Boerder, K., Matwin, S., Worm, B., 2016. Improving fishing pattern detection from satellite AIS using data mining and machine learning. PLoS One 11, 1–20. https://doi.org/10.1371/journal.pone.0158248

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. Appl. Sci. 10, 1–23. https://doi.org/10.3390/app10248900

Fabio, P., Alessandra, R., Elisabetta, R., Claudio, S., Marta, S., Amilcar, S., Stan, M., 2019. Predicting Fishing Effort and Catch Using Semantic Trajectories and Machine Learning, in: Multiple-Aspect Analysis of Semantic Trajectories. Springer, Cham, pp. 88–99. https://doi.org/doi.org/10.1007/978-3-030-38081-6_7

Fisheries Development Policy and Planning Division, 2021. Fisheries statistics of Thailand 2019. Bangkok.

Garcia, M.S., Rice, J., Charles, A., 2014. Governance of Marine Fisheries and Biodiversity Conservation Interaction and Coevolution. John Wiley & Sons, Ltd, West Sussex.

Géron, A., 2019. Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow. O'Reilly Media, Inc.

Geronimo, R.C., Franklin, E.C., Brainard, R.E., Elvidge, C.D., Santos, M.D., Venegas, R., Mora, C., 2018. Mapping fishing activities and suitable fishing grounds using nighttime satellite images and maximum entropy modelling. Remote Sens. 10. https://doi.org/10.3390/rs10101604

Gerritsen, H., Lordan, C., 2011. Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution. ICES J. Mar. Sci. 68, 245–252. https://doi.org/10.1093/icesjms/fsq137

Hamel, M.A., Pressey, R.L., Evans, L.S., Andréfouët, S., 2018. The importance of fishing grounds as perceived by local communities can be undervalued by measures of socioeconomic cost used in conservation planning. Conserv. Lett. 11, 1–9. https://doi.org/doi: 10.1111/conl.12352

Holland, D.S., 2003. Integrating spatial management measures into traditional fishery management systems: the case of the Georges Bank multispecies groundfish fishery. ICES J. Mar. Sci. 60, 915–929. https://doi.org/10.1016/S1054-3139(03)00097-3

Holmes, S.J., Bailey, N., Campbell, N., Catarino, R., Barratt, K., Gibb, A., Fernandes, P.G., 2011. Using fishery-dependent data to inform the development and operation of a co-management initiative to reduce cod mortality and cut discards. ICES J. Mar. Sci. 68, 1679–1688. https://doi.org/https://doi.org/10.1093/icesjms/fsr101

Jennings, S., Lee, J., 2012. Defining fishing grounds with vessel monitoring system data. ICES J. Mar. Sci. 69, 51–63. https://doi.org/10.1093/icesjms/fsr173

Jinseok, K., Jenna, K., 2018. The impact of imbalanced training data on machine learning for author name disambiguation. Scientometrics. https://doi.org/DOI:10.1007/s11192-018-2865-9

Kirasich, K., Smith, T., Sadler, B., 2018. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Sci. Rev. 1, 1–24.

Kongseng, S., Phoonsawat, R., Swatdipong, A., 2020. Individual assignment and mixed-stock analysis of short mackerel (*Rastrelliger brachysoma*) in the inner and eastern Gulf of Thailand: contrast migratory behavior among the fishery stocks. Fish. Res. 221, 105372. https://doi.org/10.1016/j.fishres.2019.105372

Koolkalya, S., Trueman, C., Sawusdee, A., Jutagate, T., 2020. Determination of short mackerel *Rastrelliger brachysoma* (Bleeker, 1851) stocks in the Gulf of Thailand using otolith microchemistry. Asian Fish. Sci. 33, 249–257. https://doi.org/10.33997/j.afs.2020.33.3.006

Kraak, S.B.M., Reid, D.G., Bal, G., Barkai, A., Codling, E.A., Kelly, C.J., Rogan, E., 2015. RTI ("Real-Time Incentives") outperforms traditional management in a simulated mixed fishery and cases incorporating protection of vulnerable species and areas. Fish. Res. 172, 209–224. https://doi.org/10.1016/j.fishres.2015.07.014

Kraak, S.B.M., Reid, D.G., Codling, E.A., 2014. Exploring the RTI (real-time incentive) tariff-based approach to single-species fisheries management. Fish. Res. 155, 90–102. https://doi.org/10.1016/j.fishres.2014.02.014

Kusuma, Denny, W., Murdimanto, A., Sukresno, B., Jatisworo, D., Hanintyo, dan, R., 2018. Comparison of interpolation methods for sea surface temperature data. J. Fish. Mar. Sci. 2, 102–115. https://doi.org/http://dx.doi.org/10.21776/ub.jfmr.2018.002.02.7

Little, A.S., Needle, C.L., Hilborn, R., Holland, D.S., Marshall, C.T., 2015. Real-time spatial management approaches to reduce bycatch and discards: Experiences from Europe and the United States. Fish Fish. 16, 576–602. https://doi.org/10.1111/faf.12080

Marzuki, M.I., Gaspar, P., Garello, R., Kerbaol, V., Fablet, R., 2018. Fishing Gear Identification from Vessel-Monitoring-System-Based Fishing Vessel Trajectories. IEEE J. Ocean. Eng. https://doi.org/10.1109/JOE.2017.2723278

McCluskey, S.M., Lewison, R.L., 2008. Quantifying fishing effort: A synthesis of current methods and their applications. Fish Fish. 9, 188–200. https://doi.org/10.1111/j.1467-2979.2008.00283.x

Muench, A., Depiper, G.S., Demarest, C., 2018. On the precision of predicting fishing location using data from the vessel monitoring system (VMS). Can. J. Fish. Aquat. Sci. https://doi.org/10.1139/cjfas-2016-0446

Muhling, B.A., Brodie, S., Smith, J.A., Tommasi, D., Gaitan, C.F., Hazen, E.L., Jacox, M.G., Auth, T.D., Brodeur, R.D., 2020. Predictability of Species Distributions Deteriorates Under Novel Environmental Conditions in the California Current System. Front. Mar. Sci. 7, 1–22. https://doi.org/10.3389/fmars.2020.00589

Munehara, M., Kaewnern, M., Noranarttragoon, P., Matsuishi, T.F., 2021a. Effectiveness of Adaptive Real-Time Closures for Managing Migratory Species in Data-Limited Multispecies Fisheries. Asian Fish. Sci. 34, 365–372. https://doi.org/10.33997/j.afs.2021.34.4.009

Munehara, M., Kaewnern, M., Noranarttragoon, P., Matsuishi, T.F., 2021b. Simulations of fixed closure and real-time closure to manage migratory fish species for data-limited fisheries. Ocean Coast. Manag. 211, 105736. https://doi.org/10.1016/j.ocecoaman.2021.105736

Murray, L.G., Hinz, H., Hold, N., Kaiser, M.J., 2013. The effectiveness of using CPUE data derived from Vessel Monitoring Systems and fisheries logbooks to estimate scallop biomass. ICES J. Mar. Sci. 70, 1330–1340. https://doi.org/doi:10.1093/icesjms/fst099 The

Musa, E., Elnazeer, A., Amshi, H.A., 2020. Evaluation of the Performance of K-Nearest Neighbor Algorithm in Determining Student Learning Styles.

Nevárez-Martínez, M.O., Lluch-Belda, D., Cisneros-Mata, M.A., Pablo Santos-Molina, J., De los Angeles Martínez-Zavala, M., Lluch-Cota, S.E., 2001. Distribution and abundance of the Pacific sardine (*Sardinops sagax*) in the Gulf of California and their relation with the environment. Prog. Oceanogr. 49, 565–580. https://doi.org/10.1016/S0079-6611(01)00041-6

Nguyen, K.Q., Nguyena, V.Y., 2017. Changing of sea surface temperature affects catch of spanish mackerel *Scomberomorus commerson* in the set-net fishery. Fish. Aquac. J. 08. https://doi.org/10.4172/2150-3508.1000231

Nurdin, S., Mustapha, M.A., Lihan, T., Ghaffar, M.A., 2015. Determination of potential fishing grounds of *Rastrelliger kanagurta* using satellite remote sensing and GIS technique. Sains Malaysiana. https://doi.org/10.17576/jsm-2015-4402-09

O'Farrell, S., Sanchirico, J.N., Chollett, I., Cockrell, M., Murawski, S.A., Watson, J.T., Haynie, A., Strelcheck, A., Perruso, L., 2017. Improving detection of short-duration fishing behaviour in vessel tracks by feature engineering of training data. ICES J. Mar. Sci. https://doi.org/10.1093/icesjms/fsw244

Palmer, M.C., Demarest, C., 2018. Getting to good enough: Performance of a suite of methods for spatially allocating fishing effort to management areas. Fish. Res. 204, 26–32.

https://doi.org/https://doi.org/10.1016/j.fishres.2018.02.003

Palmer, M.C., Wigley, S.E., 2009. Using positional data from vessel monitoring systems to validate the logbook-reported area fished and the stock allocation of commercial fisheries landings. North Am. J. Fish. Manag. 29, 928–942. https://doi.org/10.1577/m08-135.1

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. J. Mach. Learn. Res. 12, 2825–2830.

Robert, J, H., Etten, van, J., 2012. Raster: Geographic analysis and modeling with raster data. R package version 2.0-12. https://doi.org/http://cran.r-project.org/package=raster

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.

scikit-learn developers, 2017. scikit-learn user guide, Release 0.18.2 scikit-learn developers.

SEAFDEC, 2022. Fishery Statistical Bulletin 2019. Bangkok.

SEAFDEC, 2011. Fishery Statistical Bulletin 2009. Bangkok.

Smith, J.A., Tommasi, D., Welch, H., Hazen, E.L., Sweeney, J., Brodie, S., Muhling, B., Stohs, S.M., Jacox, M.G., 2021. Comparing Dynamic and Static Time-Area Closures for Bycatch Mitigation: A Management Strategy Evaluation of a Swordfish Fishery. Front. Mar. Sci. 8, 1–19. https://doi.org/10.3389/fmars.2021.630607

Southeast Asian Fisheries Development Center, 2004. Fishing gear and methods in Southeast Asia: I. Thailand (revisional edition), 2nd ed.

Su, N.J., Chang, C.H., Hu, Y.T., Chiang, W.C., Tseng, C. Te, 2020. Modeling the spatial distribution of swordfish (Xiphias Gladius) using fishery and remote sensing data: Approach and resolution. Remote Sens. 12. https://doi.org/10.3390/rs12060947

Sumaila, R. U., 1998. Protected marine resurves as hedges against uncertainty: an economist's perspective, in: Pitcher, T.J., Hart, P., Pauly, D. (Eds.), Reinventing Fisheries Management. Springer, Dordrecht, Boston, pp. 303–309. https://doi.org/https://doi.org/10.1007/978-94-011-4433-9

Thorson, J.T., Adams, G., Holsman, K., 2019. Spatio-temporal models of intermediate complexity for ecosystem assessments: A new tool for spatial fisheries management. Fish Fish. 20, 1083–1099. https://doi.org/10.1111/faf.12398

Van, R., Guido, Drake, Fred, L., 2009. Python 3 Reference Manual.

Watson, J.T., Haynie, A.C., 2016. Using vessel monitoring system data to identify and characterise trips made by fishing vessels in the United States North Pacific. PLoS One 11, 1–20. https://doi.org/10.1371/journal.pone.0165173

Watson, J.T., Haynie, A.C., Sullivan, P.J., Perruso, L., O'Farrell, S., Sanchirico, J.N., Mueter, F.J., 2018. Vessel monitoring systems (VMS) reveal an increase in fishing efficiency following regulatory changes in a demersal longline fishery. Fish. Res. 207, 85–94. https://doi.org/10.1016/j.fishres.2018.06.006

Zhang, H., Yang, S.L., Fan, W., Shi, H.M., Yuan, S.L., 2021. Spatial analysis of the fishing behaviour of tuna purse seiners in the western and central Pacific based on vessel trajectory date. J. Mar. Sci. Eng. 9. https://doi.org/10.3390/jmse9030322

Zintzen, V., Anderson, M.J., Roberts, C.D., Harvey, E.S., Stewart, A.L., 2017. Effects of latitude and depth on the beta diversity of New Zealand fish communities. Sci. Rep. https://doi.org/10.1038/s41598-017-08427-7

Table 1. Data sources for compiling a relevant dataset on the short mackerel purse seine fishery in the Andaman Sea, Thailand.

Data source	Description	Number of vessels	Number of trips	Number of positions	Coverage data
VMS	• Date time • Vessel positions every hour • Instantaneous speed every hour	91	3,237	86,495	100% of trip
Fishing trip	• Date and time of port-out • Date and time of port-in • Fishing trip duration	91	3,237	–	100% of trips
Logbook	• Date and time of the beginning of each fishing operation • Duration of each fishing operation • Position of each fishing operation	63	654	16,857	20.2% of trips
Landing	• Catch weight by species of each fishing trip	91	3,237	–	100% of trips

Table 2. The important predictors determined in each experiment. The number represents the relevance score in percentage of each predictor for predicting fishing operations by purse seiners in the short mackerel fishery in the Andaman Sea, Thailand. Total importance in each experiment is equal to 100; more-relevant predictors have a larger number, and less-important predictors have a lower number.

Predictor name	Experiment 1	Experiment 2	Experiment 3
Environmental predictors			
Sea surface temperature	2.2%		
Chlorophyll-a	2.3%		
Bathymetry	2.5%		
Spatial predictors			
Distance from shore	3.8%	4.7%	
Latitude	3.1%	3.5%	
Longitude	3.0%	3.6%	
Temporal predictors			
Date	1.8%	2.1%	2.6%
Date in year	2.0%	2.4%	3.0%
Month	0.9%	1.2%	1.4%
Hour	8.3%	8.3%	8.8%
Day/night	6.6%	6.6%	6.9%
Fisheries predictors			
Vessel speed	5.0%	5.3%	6.0%
$\Delta Speed_{curr-1}$	4.7%	4.8%	5.4%
$\Delta Speed_{curr+1}$	3.7%	4.0%	4.7%
Standard deviation speed	2.8%	3.2%	3.8%
Average speed	3.7%	4.2%	4.6%
Calculated speed	13.2%	12.8%	14.8%
ΔCalculated $speed_{curr-1}$	3.3%	3.6%	4.3%
Standard deviation of calculated speed	3.5%	4.0%	4.5%
Average of calculated speed	11.2%	12.0%	12.9%
Vessel size	2.2%	2.5%	3.0%
Vessel length	2.3%	2.6%	3.1%
Fishing day	2.8%	3.1%	3.7%
Ratio of target catch to total landings	2.4%	2.7%	3.1%
Total target catch landing	2.5%	2.8%	3.4%

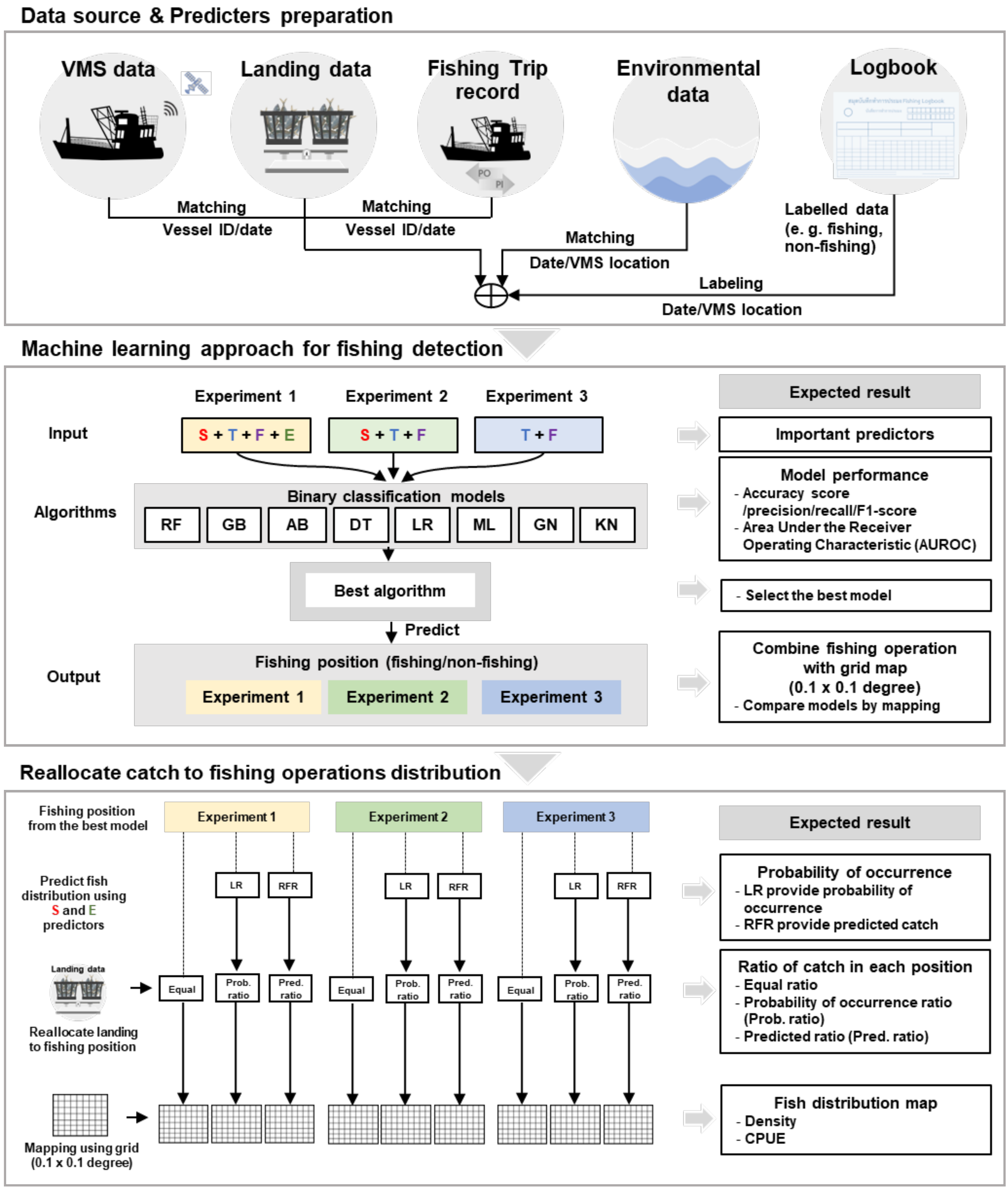


Fig. 1. Processes framework for using logistic regression (LR) or random forest regression (RFR) to predict fishing operations and fish distribution in Thailand's short mackerel purse seine fishery. The letters "S," "T," "F," and "E" denote spatial, temporal, fisheries, and environmental predictors, respectively. Eight algorithms were employed: RF = Random Forest; GB = Gradient Boosting Classifier; AB= AdaBoost; DT = Decision Tree; LR = Logistic Regression; ML = Multi-layer Perceptron Classifier; GN = Gaussian Naive Bayes; KN = K-Neighbors.

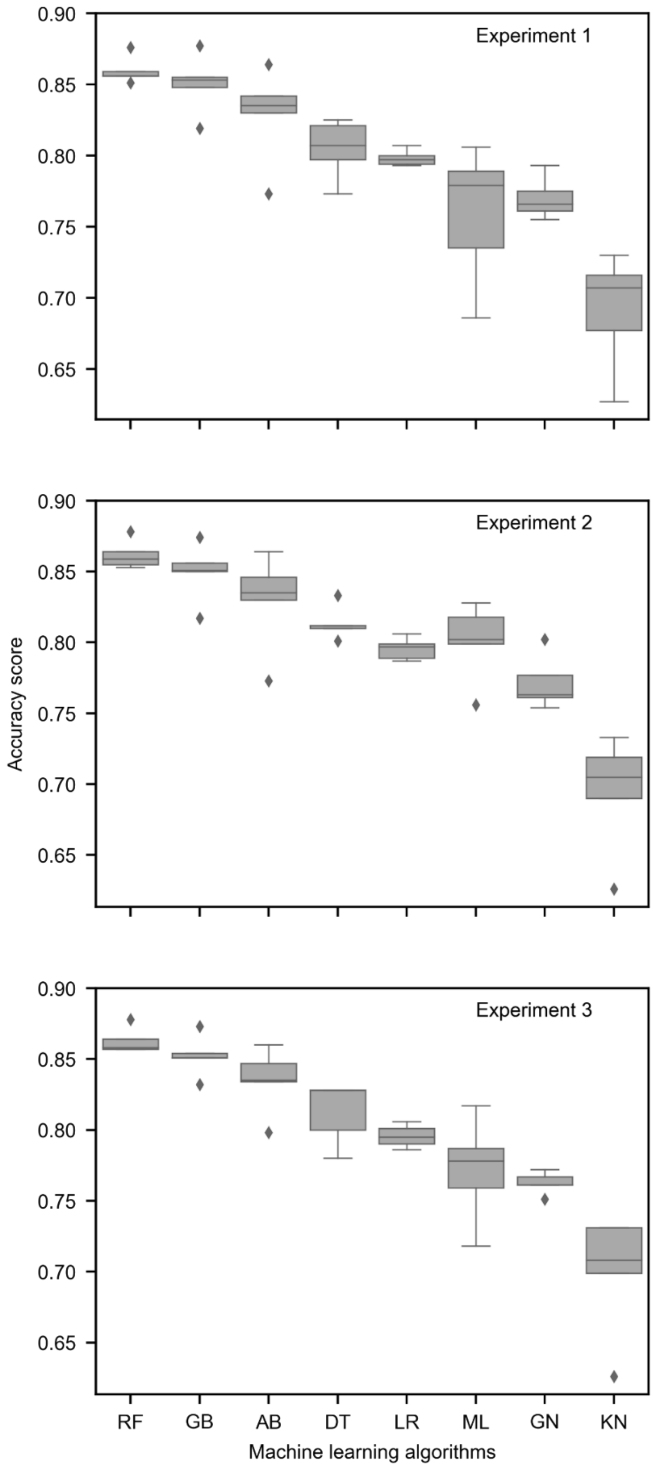


Fig. 2. The accuracy score of each algorithm in three different experiments. See the caption to Fig. 1 for abbreviations.

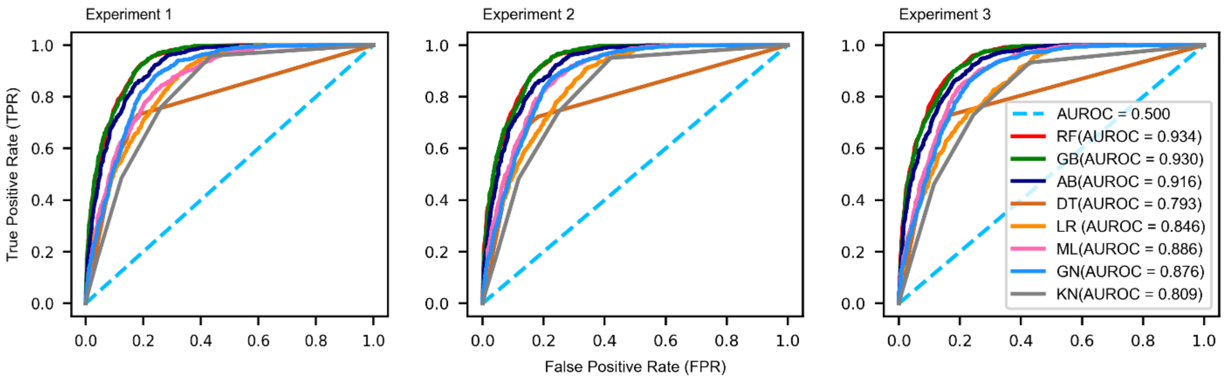


Fig. 3. The area under the receiver operating characteristic (AUROC) curve of each algorithm in three different experiments. See the caption to Fig. 1 for abbreviations.

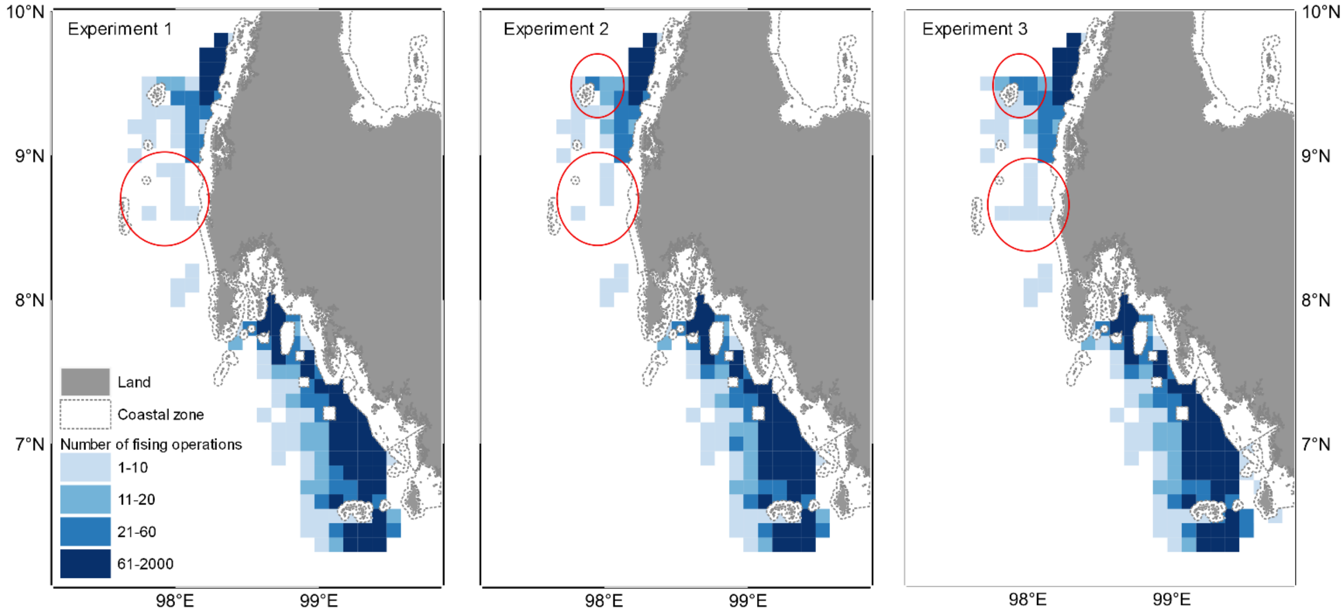


Fig. 4. Depiction of fishing operations in Thailand's short mackerel purse seine fishery, as predicted in experiments 1, 2, and 3 from the best algorithm, Random Forest.

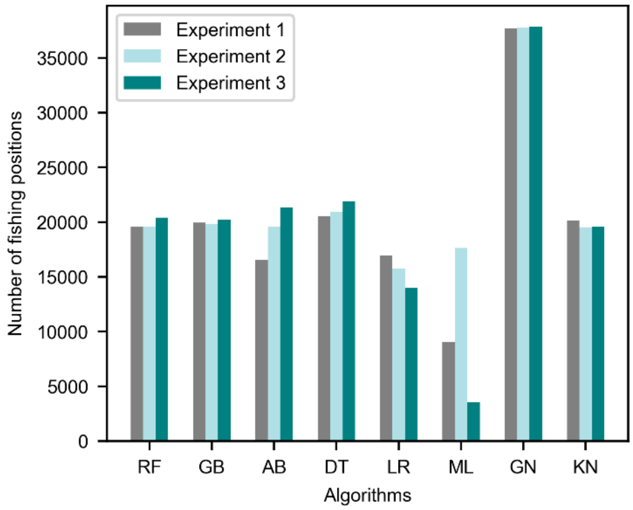


Fig. 5. The number of fishing positions predicted by each experiment for Thailand's short mackerel purse seine fishery. Algorithms: RF = Random Forest; GB = Gradient Boosting; AB= AdaBoost; DT = Decision Tree; LR = Logistic Regression; ML = Multi-layer Perceptron Classifier; GN = Gaussian Naive Bayes; KN = K-Neighbors.

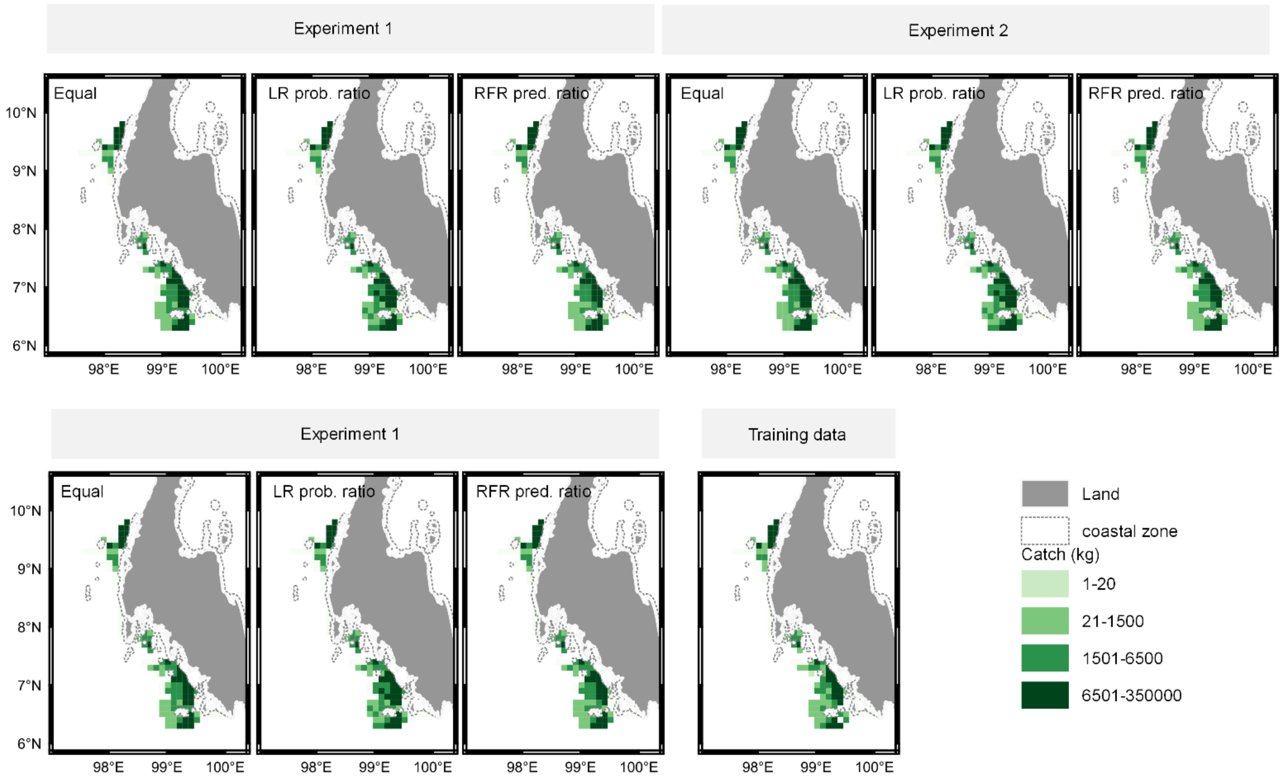


Fig. 6. Catch distribution map calculated using the reallocated catch technique and the number of fishing hours obtained from forecasted fishing positions after aggregating the catch data into 0.1° × 0.1° grids. LR prob. = the second allocation method using logistic regression probability; RFR pred. = the third allocation method using random forest regression prediction.

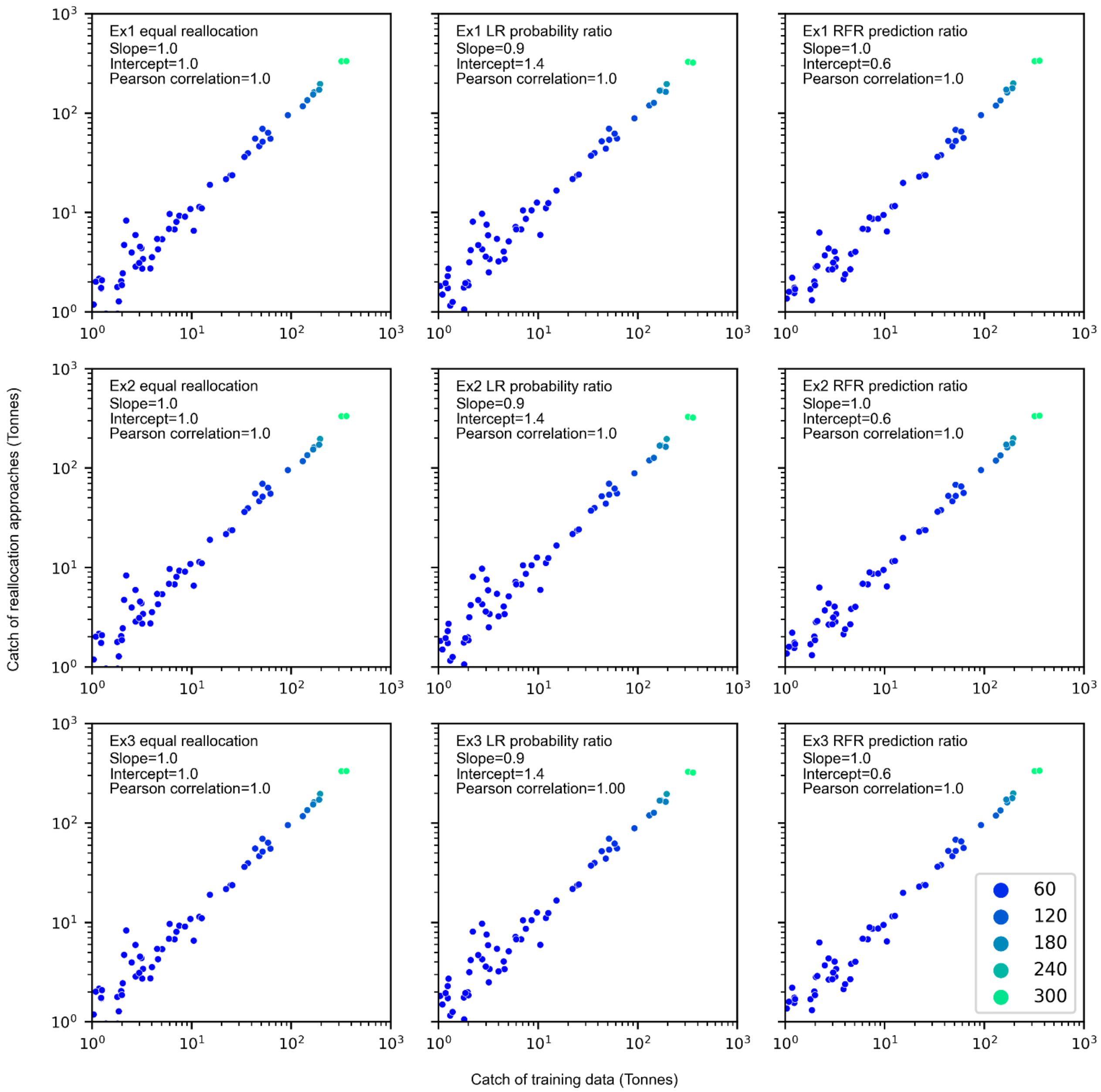


Fig. 7. Comparison of three approaches of reallocation catch to predicted fishing positions in each experiment. The vertical axis represents the number of catches in each fishing position as collected from the logbooks used for training data. The horizontal axis corresponds to the number of catch landings reallocated to predicted fishing positions based on the Random Forest model. Each point represents the number of catches after aggregating the catch data into 0.1° × 0.1° grids. Ex1 = experiment 1; Ex2 = experiment 2; Ex3 = experiment 3; LR = logistic regression; RFR = random forest regression.

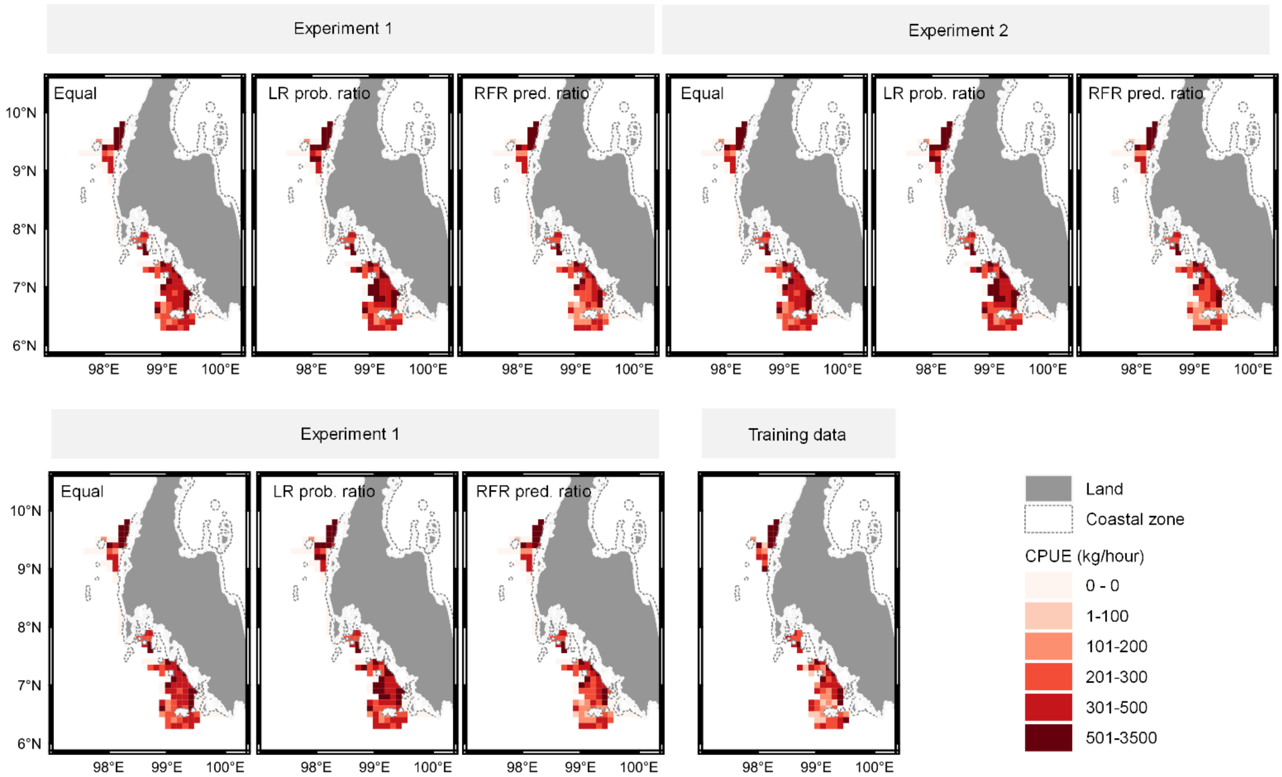


Fig. 8. The distribution map of catch per unit effort (CPUE) calculated using the reallocated catch technique and the number of fishing hours obtained from forecasted fishing positions after aggregating the catch data into 0.1° × 0.1° grids. LR prob. = the second allocation method using logistic regression probability; RFR pred. = the third allocation method using random forest regression prediction.

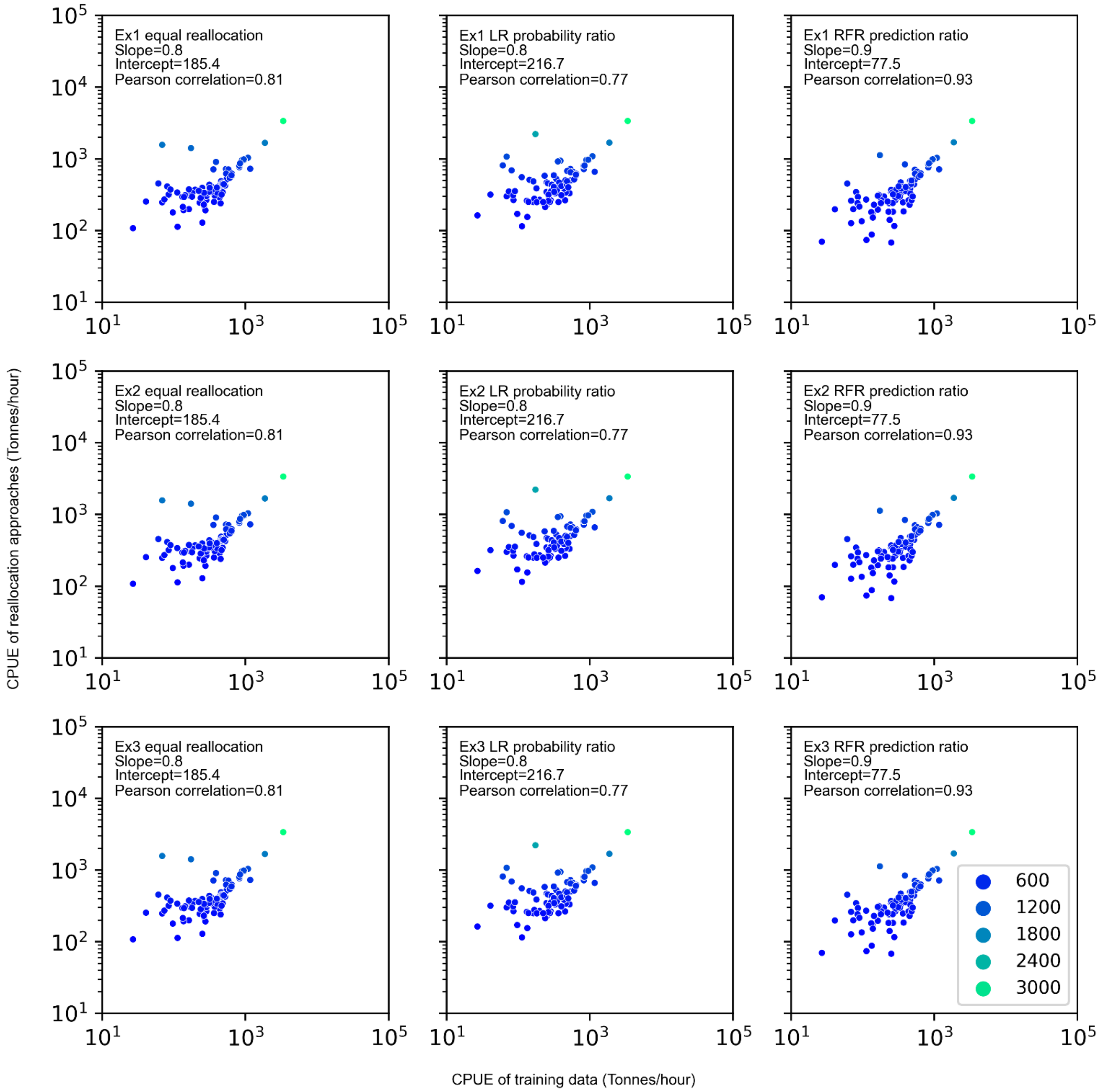


Fig. 9. Three approaches for redistributing catch to predicted fishing spots for estimating catch per unit effort (CPUE) were compared in each experiment. The vertical axis represents the estimation of CPUE from the vessel logbooks used for the training data. The horizontal axis corresponds to the estimation of CPUE from reallocation approaches of catch landings to the predicted fishing position based on the Random Forest model. Each point represents the CPUE after aggregating the catch data into 0.1° × 0.1° grids. Ex1 = experiment 1; Ex2 = experiment 2; Ex3 = experiment 3; LR = logistic regression; RFR = random forest regression.

Supplementary Material

Estimation of the spatiotemporal distribution of fish and fishing grounds from surveillance information using machine learning: the case of short mackerel (*Rastrelliger brachysoma)* in the Andaman Sea, Thailand

Chonlada Meeanan[a], Pavarot Noranarttragoon[b], Piyachoke Sinanun[c], Takahashi Yuki[d], Methee Kaewnern[e], Matsuishi Takashi Fritz[d]*

[a] *Graduate School of Global Food Resources, Hokkaido University, Kita 9, Nishi 9, Kita-ku, Sapporo 060-0809, Japan*

[b] *Marine Fisheries Research and Development Division, Department of Fisheries, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

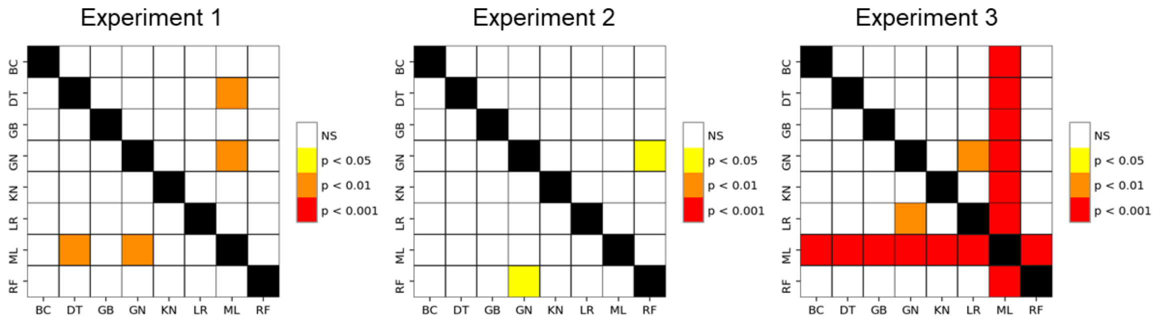
[c] *Fish Quarantine and Fishing Vessels Inspection Division, Department of Fisheries, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

[d] *Faculty of Fisheries Sciences, Hokkaido University, 3 Chome-1-1 Minatocho, Hakodate, Hokkaido 041-8611, Japan*

[e] *Faculty of Fisheries, Kasetsart University, 50 Paholyothin Road Kaset Klang, Chatuchak, Bangkok 10900, Thailand*

*Corresponding author.

E-mail address: catm@fish.hokudai.ac.jp (Matsuishi Takashi Fritz)



Supplementary Fig. S1. Pairwise comparisons of predicted fishing positions of each model using the Conover posthoc test. NS denotes no significant difference. Algorithms: RF = Random Forest; GB = Gradient Boosting; AB = AdaBoost Classifier; DT = Decision Tree; LR = Logistic Regression; ML = Multilayer Perceptron Classifier; GN = Gaussian Naive Bayes; KN = K-Neighbors.

Supplementary Table S1. The 25 predictors used for predicting the potential of the fishing operations and fish distribution using machine learning.

Predictor name	Description	Expectations with the fishing operation
Environmental predictors		
1. SST	Daily sea surface temperature (SST, °C), with a 4-km resolution, obtained from the MODIS-Aqua/Terra (NASA's Goddard Space Flight Center, https://oceancolor.gsfc.nasa.gov)	Influence on species distribution
2. Chlorophyll-a	Daily chlorophyll-a, with a 4-km resolution, obtained from the MODIS-Aqua/Terra (NASA's Goddard Space Flight Center, https://oceancolor.gsfc.nasa.gov)	Influence on species distribution
3. Bathymetry	Sea depth (m) of a spatial resolution of 15 arc-seconds (about 0.004°) obtained from the British Oceanographic Data Centre, National Oceanography Centre (https://www.gebco.net)	Fishing opportunities will be affected by depth characteristics and depth-specific fish habitats (Watson et al., 2018)
Spatial predictors		
1. Distance from shore	Distance from the nearest country line (km)	A proxy of the fishing location to identify the potential of fishing grounds

Predictor name	Description	Expectations with the fishing operation
2. Latitude	Latitude from VMS tracking	Identifies the potential fishing grounds
3. Longitude	Longitude from VMS tracking	Identifies the potential fishing grounds
Temporal predictors		
1. Date	Number of days in a particular month, ranging from 1 to 31	Dates that may affect an operation decision
2. Date in year	Date in the year, ranging from 1 to 366	A proxy of date integrated with months It may have an impact on the abundance in each season, as well as the different regulations (prohibited areas) (De Souza et al., 2016; Watson et al., 2018)
3. Month	Month in the year, ranging from 1 to 12	It may have an impact on the abundance in each season, as well as the different regulations (prohibited areas); some vessels target different fishing locations during the different seasons (De Souza et al., 2016; Watson et al., 2018)

Predictor name	Description	Expectations with the fishing operation
4. Hour	24 hours in a day	Hour is a precise timescale; the fishing methods of purse seine operations are unique to the time of day
5. Operation time (day/night)	Defining the time at 6 a.m. and 6 p.m. to denote sunrise and sunset to divide between day and night	Day and night may influence fishing operations; the fishing operations usually occur at nighttime
Fisheries predictors		
1. Vessel speed	Instantaneous vessel speed from VMS data (knot)	The speed for conducting fishing operations
2. $\Delta Speed_{curr-1}$	The different vessel speeds calculated by the current and the previous VMS record	The new speed that denotes the difference in speed between arriving at the fishing grounds and beginning operations
3. $\Delta Speed_{curr+1}$	The different vessel speeds calculated by the current and the subsequent VMS record	The new speed that denotes the difference in speed between arriving at the fishing grounds and beginning operations

Predictor name	Description	Expectations with the fishing operation
4. Standard deviation speed	The standard deviation of the vessel's current speeds in forward and backward directions	High variance affects high-fishing opportunities
5. Average speed	The average of the vessel's current speeds in forward and backward directions	Reduce the VMS transmission gap by calculating the average speed for 3 hours, which is the length of time between the start to the end of fishing; low and moderate average speeds may have corresponded to fishing operation
6. Calculated speed	Distance between two consecutive positions (km)	A proxy of average speed between two consecutive positions
7. ΔCalculated speed$_{curr-1}$	The different distances calculated by the current and the previous VMS record	The new distance that denotes the difference in distance between the previous; low distance differentiation may indicate the potential of consecutive fishing operations

Predictor name	Description	Expectations with the fishing operation
8. Standard deviation of calculated speed	The standard deviation of the current distances in forward and backward directions	The new distance that represents the difference in distance between the previous; the consecutive areas of the possible fishing activities may be indicated by a low distance difference
9. Average of calculated speed	The average of current distances in forwarding and backward directions	A proxy of average speed Reduce the VMS transmission gap by calculating the average distance for 3 hours, which is the length of time between the start to the end of fishing; low and moderate average distances may correspond to fishing operations
10. Vessel size	Registration vessel size (gross tonnage)	Fishing patterns may vary depending on the specific vessel size
11. Vessel length	Registration vessel length (m)	Fishing patterns may vary depending on the specific vessel length

Predictor name	Description	Expectations with the fishing operation
12. Fishing day	Fishing trip duration (day)	
13. Ratio of target catch to total landings	The proportion of short mackerel landings to the total landings	Related to location and identifies the density of target species in the fishing grounds
14. Total target-catch landing	Landings of short mackerel on a fishing trip	Related to location and identifies the density of target species in the fishing grounds

Supplementary text 1. Surveillance system in Thailand's fisheries

Vessel Monitoring System (VMS)

VMS data of purse seiners (30 gross tonnages and above) were obtained under the permission of Thailand's Department of Fisheries. According to the VMS data, which contains confidential information regarding the Royal Ordinance on Fisheries B.E. 2558 (2015), a proxy of vessel identification (ID) for each vessel was prepared for research purposes to classify individual vessels without revealing personal information. The vessel position (latitude, longitude), dates, times, and the instantaneous speed of the individual vessel were provided at 1-hour intervals. Before being linked to fishing trips, all data were checked for duplicate records. The beginning and end of fishing trips are not identified in VMS transmission records. Thus, the fishing trip numbers were assigned to the individual VMS data using the proxy vessel ID and the date and time. Based on this information, we could use the date, time, and speed as predicters to anticipate the potential of the fishing operations.

Fishing trip records

The owner of a fishing vessel to depart on a commercial fishing trip should report every port-in and port-out (PIPO) operation from and into a fishing port. Dates, times, and purposes of entry into and departure from fishing ports of specific vessels were recorded and identified by trip number as an electronic entry into the Fishing Info system of PIPO Controlling Centers (ePIPO) under supervision by Thailand's Department of Fisheries. The fishing trip information is integrated with VMS data based on the date and time of entry and departure. In subsequent steps, the fishing trip number is applied to connect the landing data.

Logbooks

Logbook declarations are provided by skippers, following the requirement of fishery legislation to record activity in a fishing logbook for each fishing trip and submit it to the Department of Fisheries every time the vessel returns for berthing at a port. The logbook should be immediately handed to the competent officer stationed at the PIPO center where the vessel is berthed for inspection.

Fishing activities are written down on a paper logbook while at sea, and each recorded row represents a fishing operation. Each row contains information on the date, time, and locations (latitude, longitude) of the start of the fishing operation (start of lowering the net into the water). The fishing period (the start of pulling the net into the water until the entire catch is hauled out and collected) is defined in hours. The estimated catch in each operation is reported in kilograms per the main species caught.

In order to prepare the potential of fishing operations as data labels for training and validating the binary classification in machine learning, approximately 20.2% of the logbooks from all fishing trips of short mackerel purse seiners were used. Based on the date and fishing periods reported in the logbook of the individual vessel, all fishing and non-fishing potential is linked in the VMS data as 1 and 0, respectively. Logbooks were randomly examined at the port for consistency in the reporting of fishing positions using VMS data, as legally required. However, before using the data as a training data set necessitates additional verification. Each element of fishing information in the logbook was scrutinized individually. Written reports that are unclear were omitted. Impossible location records or impossible speeds were considered by comparing the distance recorded in the logbook to the distance reported in the VMS system and then reviewed based on information on purse seine fishery characteristics, as acquired from the literature and interviews with fishers.

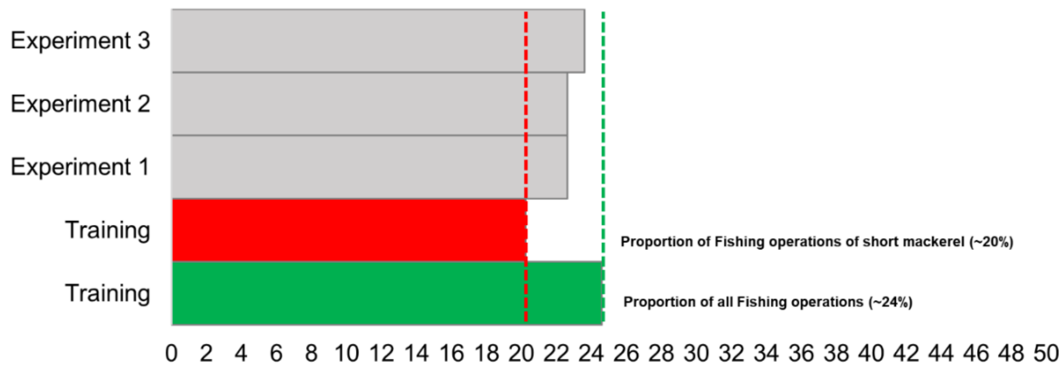
We also used data from the logbooks of fish caught in each haul as output data for forecasting fish distribution. Skippers evaluate the catch weight in the logbooks. Therefore, before using the weight of the catch, we standardized it to avoid over- or underestimating when comparing it with the actual landings. This approach included determining the catch proportion of each fishing operation in the same fishing trip using the number of catches obtained from a logbook's estimations. We multiplied the proportional value at each operation with the sum of the landing of mackerel during the current trip, using the formula:

$$ls_{ik} = ld_k * \left[\frac{lb_{ik}}{\sum lb_k}\right]$$

where ls_{ik} is the calculated catch in operation i of trip k in kg; ld_k is total landing of short mackerel of trip k in kg; lb_{ik} is the total estimated catch, in which short mackerel is the major species, in operation i of trip k in kg, as entered in the logbook; lb_k is the sum of estimated catch, in which short mackerel is the major species, of trip k in kg, as entered in the logbook.

Landing data

All commercial fishing vessels entering the fishing port for landing must report the actual catch weight by species after weighing at the landing site. The landing data of each trip of the individual vessel was integrated with VMS data using fishing trip numbers. After using machine learning and training data from the logbook to anticipate the potential of fishing effort from the VMS data, the landing data were utilized to reallocate the catch to each fishing operation area based on the trip number of individual vessels.



Supplementary Fig. S2. The proportion of fishing positions (of short mackerel purse seiners) and all fishing positions in relation to the total vessel positions used in the training dataset in each experiment.

Supplementary Table 2. The performance of eight machine learning algorithms of fishing operations in Thailand’s short mackerel purse seine fishery: RF = Random Forest; GB = Gradient Boosting; AB = AdaBoost Classifier; DT = Decision Tree; LR = Logistic Regression; ML = Multilayer Perceptron Classifier; GN = Gaussian Naive Bayes; KN = K-Neighbors.

Model	Experiment 1					Experiment 2					Experiment 3				
	Mean accuracy	SD	Precision	Recall	F1-score	Mean accuracy	SD	Precision	Recall	F1-score	Mean accuracy	SD	Precision	Recall	F1-score
RF	0.86	0.01	0.74	0.71	0.72	0.86	0.01	0.74	0.71	0.72	0.86	0.01	0.74	0.7	0.72
GB	0.85	0.02	0.74	0.71	0.72	0.85	0.02	0.73	0.71	0.72	0.85	0.01	0.73	0.71	0.72
AB	0.83	0.03	0.70	0.69	0.70	0.83	0.03	0.70	0.69	0.70	0.83	0.02	0.70	0.67	0.68
DT	0.80	0.02	0.62	0.61	0.62	0.81	0.01	0.63	0.61	0.62	0.81	0.02	0.63	0.62	0.62
LR	0.80	0.01	0.64	0.41	0.5	0.80	0.01	0.65	0.44	0.44	0.80	0.01	0.66	0.44	0.53
ML	0.76	0.05	0.65	0.59	0.59	0.80	0.03	0.7	0.46	0.56	0.77	0.04	0.35	0.99	0.52
GN	0.77	0.01	0.52	0.9	0.66	0.77	0.02	0.52	0.9	0.66	0.76	0.01	0.52	0.9	0.66
KM	0.69	0.04	0.56	0.88	0.52	0.69	0.04	0.57	0.48	0.52	0.70	0.04	0.57	0.46	0.51

Supplementary text 2. The area under the receiver operating characteristic (AUROC) curve: sensitivity and specificity

The area under the receiver operating characteristic (AUROC) curve: sensitivity and specificity were summarized using confusion matrix results of 80% training set and 20% testing set (Geronimo et al., 2018). A confusion matrix summarises a classification algorithm's performance (Effrosynidis et al., 2020; Muhling et al., 2020; scikit-learn developers, 2017).

The roc_auc_score function in scikit-learn computed the AUROC curve. By computing the area under the curve, the curve information is summarized in one number. The curve is a binary classification challenge evaluation metric. It is a probability curve that compares the true positive rate (TPR) against the false positive rate (FPR) at various thresholds (Pedregosa et al., 2011; scikit-learn developers, 2017):

Actual class (observation)		**Predicted class (expectation)**	
		Negative	Positive
	Negative	TN (true negative) Correct absence of result	FP (False positive) unexpected result
	Positive	FN (False negative) missing result	TP (true positive) Correct result

The **sensitivity** indicates how many individuals in the positive class were accurately categorized. It means determining what percentage of the true positive were identified by the model:

$$Sensivity\ (True\ Positive\ Rate)\ = \frac{TP}{TP + FN}$$

The **False Negative Rate (FNR)** indicates how much of the positive class was incorrectly categorized by the classifier. We expected that a higher TPR and a lower FNR would accurately categorize the positive class:

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{TP + FN}$$

Specificity shows how much of the negative class was correctly categorized. In the same manner that sensitivity determines the proportion of true positives that the model properly identified, specificity determines the proportion of true negatives that the model correctly detected:

$$Specificty\ (True\ Negative\ Rate) = \frac{TN}{TN + FP}$$

The **False Positive Rate (FPR)** provides what percentage of the negative class was categorized incorrectly by the classifier. We wanted to classify the negative class correctly. Hence a higher TNR and a lower FPR were preferable:

$$False\ Positive\ Rate\ (1 - Specificity) = \frac{FP}{TN + FP}$$

Supplementary text 3. Summary of machine learning models

Decision tree

Decision tree (DT) models are non-parametric supervised learning algorithms, often used in classification and regression applications to predict the value of a target variable by learning basic decision rules learned from data attributes. In general, they learn a sequence of if/else questions that leads to a decision (Géron, 2019; Müller and Guido, 2015). A decision tree classifier is a model with a tree structure in which the class prediction is contained in the tree's leaves. When constructing a decision tree, a question on the features of the training data is addressed at each node, setting the standard by which data is to be recursively split. Each criterion provides to identify a path for unclassified data to follow to infer labels from leaf nodes (Charbuty and Abdulazeez, 2021). If you wanted to design a decision tree to identify the behavior of fishing vessels (fishing or not fishing) using vessel speed and operation time, for example, you might create the one shown in Fig. 1. DT is especially effective because of the binary splitting; in a well-built tree, each query will reduce the number of alternatives by around half, quickly reducing the choices even among a huge number of classes (VanderPlas, 2016). A decision tree split corresponds to the predictor with the strongest ability to separate variables. In other words, the best split creates nodes where a single class predominates the best (Sundhari, 2011). Furthermore, DT are the core elements of

RF, one of the most potent machine learning algorithms accessible today (Géron, 2019). The two most important disadvantages of using DTs are overfitting and the generation of inadequate solutions. DT may not generate the best final model because it only uses single tree (Jafarzadeh et al., 2021).

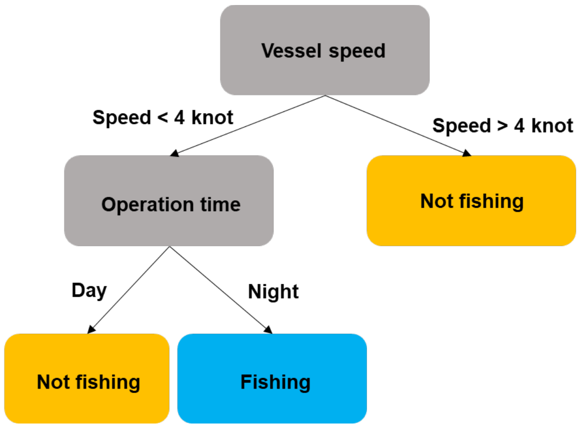


Figure 1 An example of a binary DT classification

Random forest

Random forest (RF) is a reliable ensemble approach technique that combines a number of DT classifiers to overcome the limitations of using single classifier to get the optimal solution and increase predictive performance (Jafarzadeh et al., 2021). RF used the concept of bootstrap aggregation (or bagging method) (Yoon et al., 2020), which is a method for reducing variance by resampling with replacement (Joo et al., 2013; Kirasich et al., 2018). The foundational components of RF are basically DT (Géron, 2019). RF works by training a large number of DT on different subsets of the features, including a bootstrap to assess the suitability of the sample distribution, and it performs resampling as necessary and then average (or majority vote) their predictions (Fig. 2) (Géron, 2019; Yoon et al., 2020). The number of trees can be decided to build a RF model using estimators parameter in scikit learn package (Müller and Guido, 2015). The key advantages of using the RF approach are handling overfitting to increase accuracy and processing huge, higher dimensional data sets. There are some disadvantages, such as the possibility that a minor change in the data might have a significant influence and the difficulty in understanding how to classify due to the use of many estimators and challenging visualization (Jafarzadeh et al., 2021).

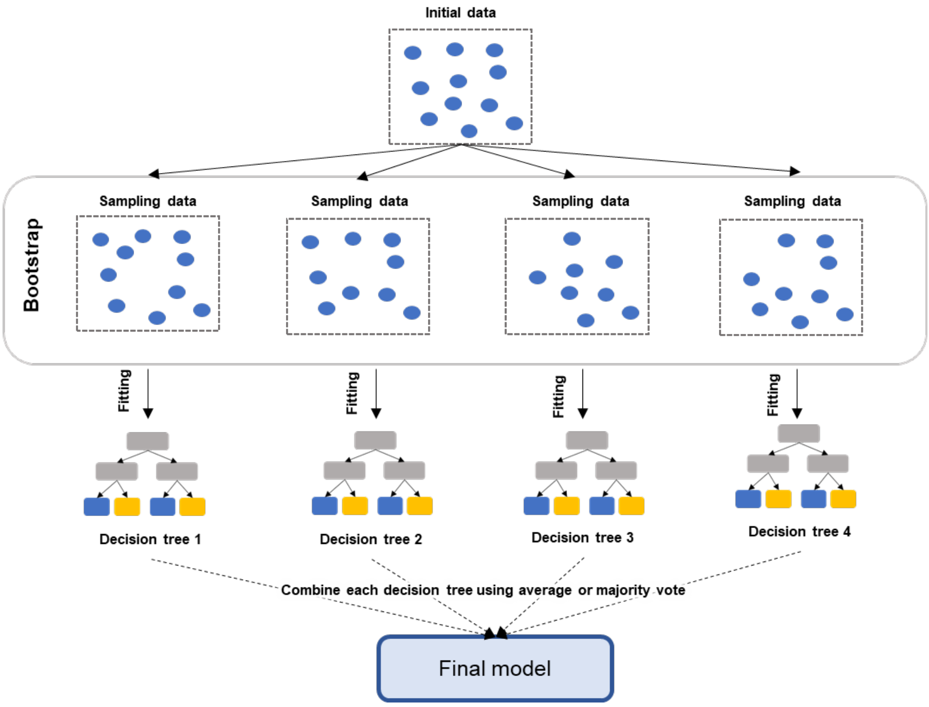


Figure 2 Conceptual framework of RF

Gradient Boosting Classifier

Gradient Boosting Classifier (GB) uses boosting method which is one of the popular learning ensemble modeling techniques used to reduce the errors and build strong classifiers from various weak classifiers. This algorithm's fundamental principle is to create models consecutively while attempting to minimize the errors of the previous model (Jafarzadeh et al., 2021). The expression "gradient boosting" decided to come up because each case's intended outcomes are determined by the gradient's error relative to the predictions. Every model improves prediction accuracy by moving in the right direction. With this technique, the new predictor is fitted to the residual errors created by the previous predictor. Three components can minimize the error and improve the model's level of prediction precision: weak learner, additive model, and loss function. Weak learners are required to reach the most advantageous split point, and trees must be developed eagerly. Loss functions should be minimized from weak learners when obtaining output averages in order to eliminate prediction-related mistakes. In order to reduce the error and enhance the performance of the model, the additive model is described as adding trees to the model one at a time without modifying the existing trees. In this process, a final prediction is obtained by combining the prediction from the recently added tree with the prediction from the previous series of trees (Jafarzadeh et al., 2021; Mayr et al., 2014). GB, however, might not perform well enough for extremely noisy data since it can result to overfitting (Jafarzadeh et al., 2021).

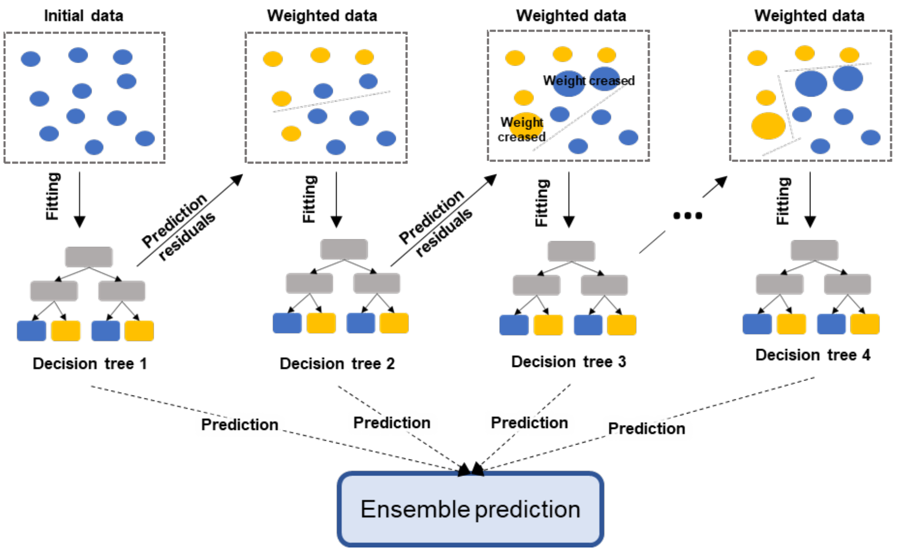


Figure 3 Conceptual framework of BC

AdaBoost Classifier (AB)

The AdaBoost Classifier (Adaptive Boosting) is members of ensemble learning family. AB and BC are established using the booting approach based on the same principle (see Fig. 3). The most significant distinction between the AB and GB approaches is how they resolve the drawbacks of poor classifiers. As described in the previous algorithm, weaknesses in GB are determined using gradients, but in AB, weaknesses are determined utilizing high-weight data points that are hard to fit. AB begins by creating a decision tree, after which all the data points are given equal weights. The weights for all the improperly classified points are then increased, while those that are simple to classify or are correctly classified have their weights decreased. These weighted data points are used to create a new decision tree (Jafarzadeh et al., 2021; Mayr et al., 2014).

Logistic Regression (LR)

Logistic Regression (LR) (also called Logit Regression) is widely used to calculate the probability that a certain instance belongs to a class. The statistical technique like logistic regression is used to create machine learning models when the dependant variable is binary. It provides information on the relationship between one dependent variable and one or more independent variables (Kirasich et al., 2018). The logistic function, commonly referred to as the sigmoid function, is a notion utilized in LR. The response variable is transformed to provide a continuous probability distribution across the output classes that is limited between 0 and 1; this transformation is known as the “logistic” or “sigmoid” function, where “z” refers for log odds divided by the logit. The logistic—noted $\sigma(\cdot)$—is a sigmoid function (i.e., S-shaped) that outputs a number between 0 and 1. It is defined as shown in equation below

$$\sigma\,(Z) = \frac{1}{1+exp^{-z}}$$

For example, in fig. 4, the model predicts whether belongs to the positive class (fishing), denoted by the letter “1” or the negative class (not fishing), denoted by the letter “0”, depending on whether the estimated probability is larger than 50% (Géron, 2019; scikit-learn developers, 2017).

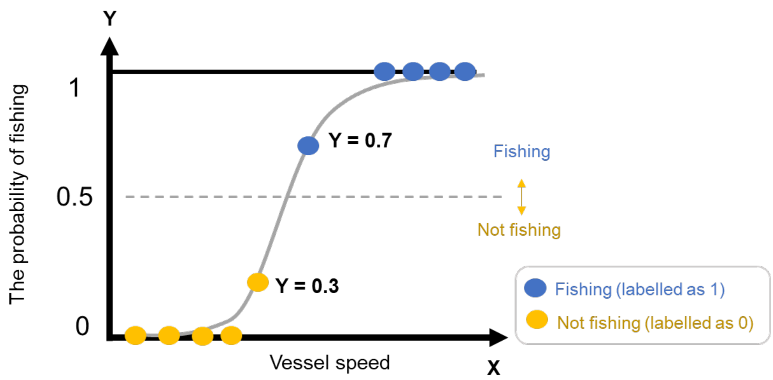


Figure 4 An example of LR for binary classification

Gaussian Naive Bayes (GN)

One of the effective classification methods in data mining and machine learning is Naive Bayes. Despite being effective, naive Bayes learners struggle with the faulty assumption of conditional independence between the qualities (Jahromi and Taheri, 2018). The Naive Bayes classifier is a simplistic probabilistic classifier that applies the Bayes theorem (Kamel et al., 2019). GN recognizes each attribute variable as independent variable. This classifier can be efficiently learned by supervised learning and used in complicated instances. The Bayes principle is implemented to classify data using the formula $P(C = c|X_1 = x_1, \dots, X_n = x_n)$ to determine the probability that a specific instance $X_n = x_n$ falls into the class C. A common assumption when working with continuous data is that the continuous values corresponding to each class are distributed according to the Gaussian distribution. The training data are divided into classes, and for each class, the mean and standard deviation are generated. Therefore, the following equation could be employed to estimate the probability of a continuous data set (Kamel et al., 2019).

$$P(X = x|C = c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where X= variable, c= class, μ =mean, σ =stander deviation. The main benefit of Naive Bayes is that it requires relatively little in the amount of training data, which is important for characterizing and required for classifying.

K-Neighbors (KN)

The K-Nearest (KN) algorithm is one of the most popular supervised machine learning methods used in a variety of applications worldwide (Musa et al., 2020). The concept behind nearest neighbour approaches is to select a set number of training samples that are geographically closest to the new location and then estimate the label based on them. The known output for this training point is then the prediction (Müller and Guido, 2015). The training set of pattern vectors for each class is provided as a set of instance prototypes. The k closest neighbors of an unknown vector are identified from all of the prototype vectors, and a majority rule is used to determine the class label.

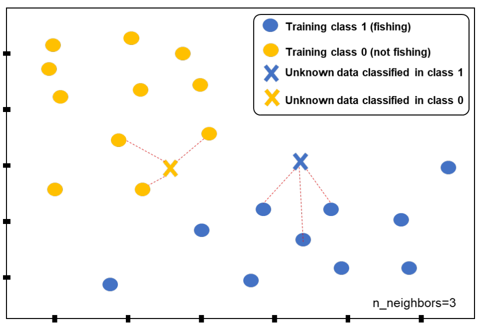


Figure 5 An example predictions made by the three-nearest-neighbors model

For example, we added two unknown class data points, which are represented by a cross. We identified the nearest location for each of them in the training set using the three closest neighbors (n neighbors=3). Then, voting is used to assign a label when taking into account more than one neighbors. This implies that we count how many neighbors are in class 0 and how many neighbors are in class 1 for each unknown data item. The more frequent class, or the class that has the majority of the k-nearest neighbors, is subsequently assigned. Finally, the prediction of the one-nearest-neighbor algorithm is the label of unknown data point (shown by the color of the cross) (Fig. 5).

Multilayer Perceptron (MLP) Classifier

Multilayer Perceptron (ML), a specific type of Artificial Neural Network machine learning model, are supervised learning algorithms (Muhling et al., 2020). A function $f(\cdot): R^m \rightarrow R^o$ is used to learn by ML by training on a dataset where m is the number of input dimensions and o is the number of output dimensions. It can learn an approximator for a non-linear function for either classification or regression given a set of variables $X = x_1, x_2, \ldots, x_n$ and a target y. There may be one or more non-linear layers, known as hidden layers, between the input layer and the output layer. A single hidden layer ML with scalar output is shown in Fig. 6. The advantages of machine learning (ML) include the capacity to learn non-linear models and the ability to learn models in real-time, but it is sensitive to feature scaling and requires tuning a variety of hyperparameters, including the number of hidden neurons, layers, and iterations (scikit-learn developers, 2017).

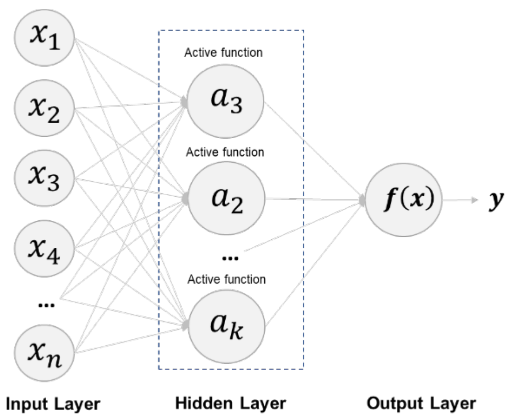


Figure 6 An example of one hidden layer ML

Supplementary text 4. Logistic Regression and Random Forest Regression fitting and performance

Logistic Regression (LR)

The LR for using in allocating catch to the fishing operation position had the form:

Fishing operation = Latitude + Longitude + Bathymetry + Chlorophyll-a + Distance from shore

Based on logbook data, fishing operations are labelled into two categories: fishing potential and non-fishing potential:

$$\text{Fishing operation} \begin{cases} 0 < 500\ kg\ of\ catch\ in\ each\ VMS\ position \\ 1\ \geq 500\ kg\ of\ catch\ in\ each\ VMS\ position \end{cases}$$

The entire dataset was split into training and testing data, with 80% of the total data being used for training and 20% being used for testing, before the model was fitted. The 5-fold cross-validation accuracy score, on average, is provided as an assessment of the model's performance. The average accuracy of the LR model was 81.21%, with a standard error of 0.022. Then, the LR model's probability of occurring at each position was applied to reallocate proportionally to the overall catch for a particular fishing trip.

Random Forest Regression (RFR)

The RFR for using in allocating catch to the fishing operation position had the form:

Catch = Latitude + Longitude + Bathymetry + Chlorophyll-a + Distance from shore

The same training set used to identify fishing positions in Section 2.4 was used for creating the RFR model. The output was changed to reflect the catch in the fishing logbook (Supplementary text 1). Root Mean Square Error (RMSE) and Negative Mean Absolute Error (MAE) were used to calculate the model performance by dividing the total dataset into training and testing data in an 80:20 ratio. The results showed that the RMSE and MAE values were 462.068 and -208.774, respectively. The output of the RFR, which is described as an expected catch expressing the probability of occurrence in each position, was then proportionally adjusted according to the landings of short mackerel and redistributed to the predicted fishing positions.

References

Charbuty, B., Abdulazeez, A., 2021. Classification Based on Decision Tree Algorithm for Machine Learning. J. Appl. Sci. Technol. Trends 2, 20–28. https://doi.org/10.38094/jastt20165

De Souza, E.N., Boerder, K., Matwin, S., Worm, B., 2016. Improving fishing pattern detection from satellite AIS using data mining and machine learning. PLoS One 11, 1–20. https://doi.org/10.1371/journal.pone.0158248

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. Appl. Sci. 10, 8900. https://doi.org/10.3390/app10248900

Géron, A., 2019. Hands-on Machine Learning whith Scikit-Learing, Keras and Tensorfow. O’Reilly Media, Inc.

Geronimo, R.C., Franklin, E.C., Brainard, R.E., Elvidge, C.D., Santos, M.D., Venegas, R., Mora, C., 2018. Mapping fishing activities and suitable fishing grounds using nighttime satellite images and maximum entropy modelling. Remote Sens. 10. https://doi.org/10.3390/rs10101604

Jafarzadeh, H., Eric Gill, F.M.M.M., Homayouni, S., 2021. Bagging and Boosting Ensemble Classifiers for Classification of Comparative Evaluation. Remote sensing, MDPI.

Jahromi, A.H., Taheri, M., 2018. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. 19th CSI Int. Symp. Artif. Intell. Signal Process. AISP 2017 2018-Janua, 209–212. https://doi.org/10.1109/AISP.2017.8324083

Joo, R., Bertrand, S., Tam, J., Fablet, R., 2013. Hidden Markov Models: The Best Models for Forager Movements? PLoS One. https://doi.org/10.1371/journal.pone.0071246

Kamel, H., Abdulah, D., Al-Tuwaijari, J.M., 2019. Cancer Classification Using Gaussian Naive Bayes Algorithm. Proc. 5th Int. Eng. Conf. IEC 2019 165–170. https://doi.org/10.1109/IEC47844.2019.8950650

Kirasich, K., Smith, T., Sadler, B., 2018. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Sci. Rev. 1, 1–24.

Mayr, A., Binder, H., Gefeller, O., Schmid, M., 2014. The evolution of boosting algorithms: From machine learning to statistical modelling. Methods Inf. Med. 53, 419–427. https://doi.org/10.3414/ME13-01-0122

Muhling, B.A., Brodie, S., Smith, J.A., Tommasi, D., Gaitan, C.F., Hazen, E.L., Jacox, M.G., Auth, T.D., Brodeur, R.D., 2020. Predictability of species distributions deteriorates under novel environmental conditions in the California Current System. Front. Mar. Sci. 7, 1–22. https://doi.org/10.3389/fmars.2020.00589

Müller, A.C., Guido, S., 2015. Introduction to Machine Learning with Python and Scikit-Learn. O'Reilly Media, Inc.

Musa, E., Elnazeer, A., Amshi, H.A., 2020. Evaluation of the Performance of K-Nearest Neighbor Algorithm in Determining Student Learning Styles.

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. J. Mach. Learn. Res. 12, 2825–2830.

scikit-learn developers, 2017. scikit-learn user guide, Release 0.18.2 scikit-learn developers.

Sundhari, S.S., 2011. A knowledge discovery using decision tree by Gini coefficient. ICBEIA 2011 - 2011 Int. Conf. Business, Eng. Ind. Appl. 232–235. https://doi.org/10.1109/ICBEIA.2011.5994250

VanderPlas, J., 2016. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, Inc.

Watson, J.T., Haynie, A.C., Sullivan, P.J., Perruso, L., O'Farrell, S., Sanchirico, J.N., Mueter, F.J., 2018. Vessel monitoring systems (VMS) reveal an increase in fishing efficiency following regulatory changes in a demersal longline fishery. Fish. Res. 207, 85–94. https://doi.org/10.1016/j.fishres.2018.06.006

Yoon, Y.J., Cho, S., Kim, S., Kim, N., Lee, S.J., Ahn, J., Lee, E., Joh, S., Lee, Y.W., 2020. An artificial intelligence method for the prediction of near- And off-shore fish catch using satellite and numerical model data. Korean J. Remote Sens. 36, 41–53. https://doi.org/10.7780/kjrs.2020.36.1.4