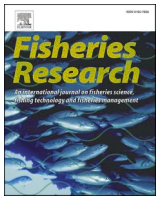
Fisheries Research 250 (2022) 106263

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



TUN-AI: Tuna biomass estimation with Machine Learning models trained on oceanography and echosounder FAD data

Daniel Precioso [a], Manuel Navarro-García [b,d], Kathryn Gavira-O'Neill [c], Alberto Torres-Barrán [d], David Gordo [d], Víctor Gallego [d,e], David Gómez-Ullate [a,*]

[a] Universidad de Cádiz, Av. Universidad de Cádiz, 10, Puerto Real, 11519 Cádiz, Spain
[b] Universidad Carlos III de Madrid, Calle Madrid, 126, Getafe, 28093 Madrid, Spain
[c] Satlink, Carretera de Fuencarral, Arbea Campus Empresarial Edificio 5. Planta Baja, Alcobendas, 28108 Madrid, Spain
[d] Komorebi AI Technologies, Avenida General Perón N 26, Planta 4, 28020 Madrid, Spain
[e] Instituto de Ciencias Matemáticas (CSIC-UAM-UCM-UC3M), Calle Nicolás Cabrera, 13-15, 28049 Madrid, Spain

ARTICLE INFO

Handled by A.E. Punt

Keywords:
Tunas
Direct abundance index
Machine Learning
Echo-sounder buoys
Fish aggregating devices
Purse seiner

ABSTRACT

The use of dFADs by tuna purse-seine fisheries is widespread across oceans, and the echo-sounder buoys attached to these dFADs provide fishermen with estimates of tuna biomass aggregated to them. This information has potential for gaining insight into tuna behaviour and abundance, but has traditionally been difficult to process and use. The current study combines FAD logbook data, oceanographic data and echo-sounder buoy data to evaluate different Machine Learning models and establish a pipeline, named TUN-AI, for processing echo-sounder buoy data and estimating tuna biomass (in metric tons, t) at various levels of complexity: binary classification, ternary classification and regression. Models were trained and tested on over 5000 sets and over 6000 deployments. Of all the models evaluated, the best performing one uses a 3-day window of echo-sounder data, oceanographic data and position/time derived features. This model is able to estimate if tuna biomass was higher than 10 t or lower than 10 t with an F1-score of 0.925. When directly estimating tuna biomass, the best model (Gradient Boosting) has an error (MAE) of 21.6 t and a relative error (SMAPE) of 29.5%, when evaluated over sets. All models tested improved when enriched with oceanographic and position-derived features, highlighting the importance of these features when using echo-sounder buoy data. Potential applications of this methodology, and future improvements, are discussed.

1. Introduction

Throughout tropical and sub-tropical oceans, a variety of fish species are known to aggregate around objects drifting on the surface, a behavior which fishermen have learned to exploit for centuries (Castro et al., 2002; Maufroy et al., 2015). In tropical tuna purse-seine fisheries, targeting mainly skipjack tuna (*Katsuwonus pelamis*), yellowfin tuna (*Thunnus albacares*) and bigeye tuna (*T. obesus*), these drifting objects, known as drifting Fish Aggregating Devices (dFADs), have become an essential tool for locating tuna-schools and increasing fishing efficiency. Today, more than 55% of tropical tuna caught by industrial purse-seine vessels in the Indian, Atlantic and Pacific oceans is caught using dFADs, accounting for 36% of the world's total tropical tuna catch (Wain et al., 2021; Anon. ISSF, 2021).

Initially, dFADs were of natural origin, such as floating logs or objects, that fishermen would come across while searching for free-swimming schools of tuna. In the mid-1980s, tools began to be developed to allow for tracking of these dFADs, and fishermen themselves designed purpose-built dFADs that could be attached to tracking beacons: first based on radar reflectors or radio, and later satellite connected GPS buoys, allowing the dFADs to be located remotely (Davies et al., 2014; Lopez et al., 2014). The use of these tracking buoys has been considered "the most significant technological development that has occurred (...) for increasing the efficiency of dFAD tuna fishing" (Lopez and Scott, 2014). Nowadays, most dFADs are equipped with satellite-linked instrumented buoys which include both GPS and an echo-sounder, providing fishermen with accurate geolocation information as well as an estimate of associated tuna biomass. These buoys allow

* Corresponding author.
E-mail addresses: daniel.precioso@uca.es (D. Precioso), mannavar@est-econ.uc3m.es (M. Navarro-García), kgo@satlink.es (K. Gavira-O'Neill), alberto.torres@komorebi.ai (A. Torres-Barrán), david.gordo@komorebi.ai (D. Gordo), victor.gallego@komorebi.ai (V. Gallego), david.gomezullate@uca.es (D. Gómez-Ullate).

https://doi.org/10.1016/j.fishres.2022.106263
Received 13 September 2021; Received in revised form 31 January 2022; Accepted 1 February 2022
Available online 17 February 2022

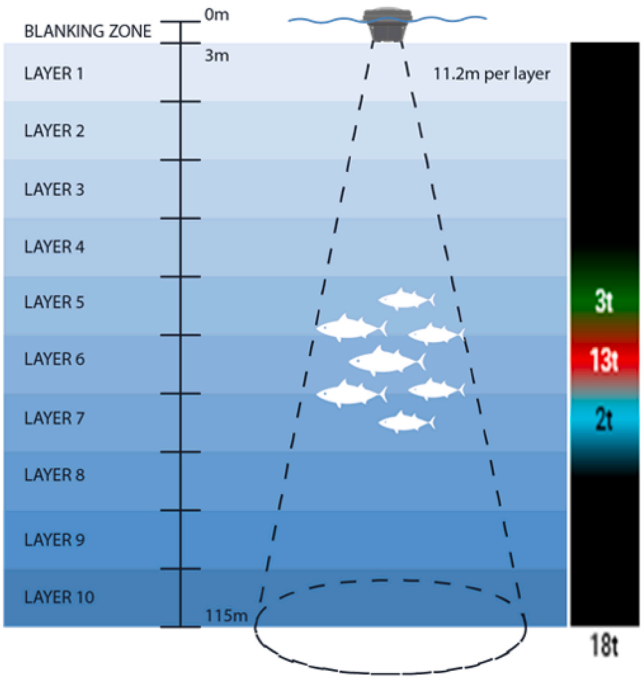


Fig. 1. Left: Depth layer configuration and set-up of the Satlink echo-sounder buoys. Right: example of the biomass estimates (in metric tons) and echogram display available to buoy users. Raw acoustic backscatter is converted into biomass estimates based on the target strength of skipjack tuna *(Katsuwonus pelamis)* using manufacturer's algorithms.

fishing crews to remotely monitor their dFADs and the biomass they aggregate in real-time, so they can target those with larger aggregated schools, thus increasing their catch while reducing searching effort (Lopez et al., 2014; Molina et al., 2003).

The widespread use of dFADs has led to large-scale changes in industrial purse-seine fishing fleets targeting tropical tunas, affecting traditional metrics such as search-time and time-at-sea, which are used to estimate Catch Per Unit Effort (CPUE) (Fonteneau et al., 2000). In this context, some authors have highlighted the need for fishery-independent abundance indices and the use of non-traditional data sources to monitor tuna stock health and the effects of fishing pressure over time (Baidai et al., 2020; Santiago et al., 2016, 2020). The echo-sounder buoys attached to dFADs across the world's oceans can be set to transmit frequent geo-referenced biomass estimates. Given the number and wide distribution of dFADs in recent years, the information provided by these echo-sounder buoys could be very valuable. Santiago et al. (2016) presented the first Buoy-Derived Abundance Index (BAI) for tropical tunas for informing stock assessments, based on the biomass estimates provided by three echo-sounder buoy brands in the Atlantic, Indian and Pacific Oceans.

However, several authors have reported substantial differences between the biomass estimates provided by echo-sounder buoys and real tuna tonnage caught by the vessels (Lopez et al., 2016; Escalle et al., 2019; Orue et al., 2019a), possibly due to the variable nature of fish aggregations under dFADs, often made up of pelagic species other than tuna (Castro et al., 2002) which do not compute in the catch data reported by vessels but would add to the biomass estimates calculated by echo-sounder buoys. Likewise, the influence of oceanic conditions on fish distribution and behavior likely drives aggregation patterns of tuna around dFADs (Lopez, 2017; Druon et al., 2017; Schaefer et al., 2007). Therefore, in order to develop a representative index of abundance from echo-sounder buoy data, it is also important to consider and understand the effect of these variables on the biomass estimates given by these echo-sounder buoys.

Although some studies have already compared biomass estimates from the buoys to catch data (Baidai et al., 2020; Lopez et al., 2016; Mannocci et al., 2021), the approach in this paper is the first to include oceanographic data as predictor variables in Machine Learning models. Likewise, others have combined oceanographic variables and catch data, without using echo-sounder buoy information (Druon et al., 2017). Lastly, some works have considered the effects of oceanographic conditions on buoy biomass estimates, without directly comparing it to catch data (Lopez, 2017; Santiago et al., 2020). In this sense, we aim to establish a well defined novel pipeline, which we have named TUN-AI, for estimating tuna biomass (in metric tons, t) under dFAD echo-sounder buoys at any given time, combining catch data, oceanographic data and echo-sounder buoy data. To achieve this, we evaluate several models at varying levels of complexity: binary classification differentiating between tuna biomass of less than 10 t (< 10 t) and greater or equal than 10 t ($\geq$10 t); three-level classification model differentiating between tuna aggregations less than 10 t (< 10 t), between 10 t and 30 t (10–30 t) or over 30 t ($\geq$30 t); and regression models aimed at estimating the exact tuna biomass (in metric tons) under the buoy. In addition, we examine the influence of different data sources, such as oceanographic data, and methods for processing echo-sounder buoy data to establish the most accurate methodology throughout.

2. Material and methods

2.1. Database description

Our study draws from three sources of information: FAD logbook data, echo-sounder buoy data, and oceanography data.

2.1.1. FAD logbook data

The first source of information corresponds to the registered interactions between fishing vessels and echo-sounder buoys, obtained from the FAD logbooks of the Spanish tropical tuna purse seine fleet operating in the Atlantic, Indian and Pacific Oceans (2018–2020, AGAC[1] ship owner's association data). This FAD logbook dataset contains almost 66,000 interactions with Satlink[2] buoys within the studied time period. Each record within the dataset can be traced to a specific buoy (using the ID and model of the buoy attached to each dFAD), and contains information about the date, time and GPS coordinates where the interaction occurred, as well as the nature of the interaction (see Ramos et al., 2017 for definitions and descriptions of each interaction).

For the purposes of the current study, only interactions registered as "Set" and "Deployment" were cross-referenced with the echo-sounder buoy dataset (see Section 2.1.2). Information from sets included the recorded catch data of skipjack tuna, yellowfin tuna and bigeye tuna, which was used as a representation of the real tuna biomass at the dFAD. This is based on the assumption that the entire tuna aggregation present at the dFAD is captured by the vessel during the set, and that total catch is accurately recorded. This is a strong assumption but inherent to this kind of large scale data. Bycatch is also recorded in the logbook, but was not considered for analysis.

It is also important to note that purse seine vessels rarely let down their nets without prior information that indicates large biomass estimates, and consequently the existence of low catches (< 10 t) is relatively uncommon (less than 8% of all the set events). Therefore, the dataset of catch events is not representative of the true data distribution, which may lead to models that overestimate the real tuna biomass and thus not widely applicable. The catch volume distribution is shown in A. To mitigate that problem we also include interactions registered as "Deployment", corresponding to new dFADs that were not previously present in the water (Ramos et al., 2017). For these interactions we assume that no tuna (0 t) was present under the buoy.

[1] Asociación de Grandes Atuneros Congeladores
[2] Satlink, Madrid, Spain, www.satlink.es

Table 1
Buoy models and characteristics; ES: Echo-sounder.

Buoy model	ES model	Frequency	Beam angle	Sampling rate
ISL+	ES12	190.5 kHz	20°	Every 15 min
SLX+	ES16	200 kHz	23°	Sunrise to sunset:every 5minSunset to sunrise: every 60 min
ISD+	ES16×2	200kHzand 38 kHznew-a	23°and 33°	Sunrise to sunset: every 5 min Sunset to sunrise: every 60 min

[a] Biomass estimates are calculated according to the acoustic response registered by the 200 kHz echo-sounder, allowing data to be comparable across all buoy models.

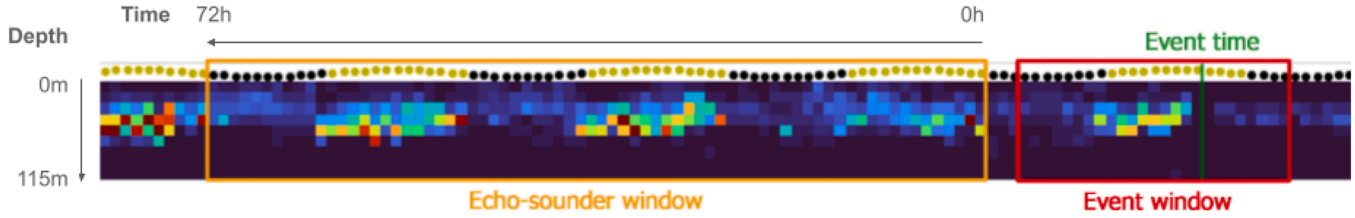


Fig. 2. Example of a selected 72 h "echo-sounder window" (yellow box) with respect to the recorded set time in the FAD logbook (green line). The day of the set (24 h) is denoted as "event window" (red box). Note that the yellow and red box do not overlap. Above the graph, the sun inclination throughout the day is represented, where day hours are depicted as yellow circles and night hours as black circles. Columns of colored squares represent the echo-sounder buoy biomass estimates for each hour, while rows are depth bins. Circadian patterns in tuna activity under the dFAD are clearly visible in this figure. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

Since the transcription to the FAD logbook is a manual process, data will sometimes contain errors. These errors include but are not limited to misreported position (either latitude or longitude), incorrect echo-sounder buoy ID, incorrect date and/or time and incorrect interaction type. To minimize this source of error, information in the FAD logbook is cross-referenced with echo-sounder data using the echo-sounder buoy ID and timestamp to check for inconsistencies (see Section 2.2).

2.1.2. Echo-sounder buoy data

The echo-sounder buoy data was collected from 15,497 Satlink buoys for which there were registered interactions in the FAD logbook data (see Section 2.1.1). This database contained over 68 million records corresponding to buoys attached to dFADs scattered over the Atlantic, Indian and Pacific Oceans. Each record is referenced to a specific buoy ID and timestamp, ranging from 2018 to 2020, and contains biomass estimates as calculated from the echo-sounder's measurements, and GPS coordinates of the buoy's last known position at the time of measuring.

For all buoys, observation range of the echo-sounder is from 3 m to 115 m depth, split into ten layers or depth bins, each with a resolution of 11.2 m (see Fig. 1). For every echo-sounder measurement, biomass estimates (in metric tons) are obtained from acoustic samples taken periodically throughout the day, and the average back-scattered acoustic response is converted into estimated tonnage, based on the target strength of skipjack tuna (see Lopez et al., 2016 for detailed explanations of the process within the buoy).

All buoy models used within the study (ISL+, SLX+ and ISD+) follow the same protocol for converting acoustic response into estimated tonnage, though there are slight differences between models, particularly in terms of sampling rate (i.e. when and how often echo-sounder measurements are taken) (Table 1). Though the echo-sounder in each buoy measures several times per hour, only that corresponding with the highest estimated tonnage each hour is transmitted by the buoy and stored in central databases, in order to reduce the total amount of information sent via satellite. If the total estimated tonnage for all measurements taken within an hour are less than 1 t, no measurement would be transmitted and thus considered a zero-reading. Likewise, total estimated tonnage for a single depth bin is capped at 63 t, due to saturation of the echo-sounder signal. Thus, the final temporal resolution of echo-sounder records contained in the current dataset is 1 h.

2.1.3. Oceanography data

Oceanographic data was downloaded from the global ocean model (products GLOBAL-ANALYSIS-FORECAST-PHY-001–024, 1/12° resolution; and GLOBAL-ANALYSIS-FORECAST-BIO-001–028, 1/4° resolution) provided by the EU Copernicus Marine Environment Monitoring Service[3] (Anon, 2018). For each position record contained in the echo-sounder buoy data (see Section 2.1.2), the following variables were downloaded: temperature (in °C), chlorophyll-a concentration (in mg/m^3), dissolved oxygen concentration (in $mmol/m^3$), salinity (in psu), thermocline depth (calculated as the depth where water temperature is 2°C lower than surface temperature, in m), current velocity (in m/s) and sea surface height anomaly or SSHa (deviation of the sea surface height from long term mean, in m). All variables, except thermocline and SSHa, were downloaded at surface level (depth = 0.494 m).

It is worth noting that all the previous variables come from oceanographic models which provide approximate values on a fixed grid. This means that they are not real observations at the exact position where the buoy is located. Nonetheless, we consider that they are sufficiently representative of oceanographic conditions for the purposes of the current study, as model accuracy has improved in recent years (see, for instance, Lellouche et al., 2018).

2.2. Data preprocessing

2.2.1. Data merging

The sets and deployments registered in the FAD logbook data were cross-referenced with each specific buoy's biomass estimates, using the buoy ID and timestamp. Oceanographic information was then collected for each position record in the echo-sounder buoy database. Since oceanographic data are available on a grid with 0.08° or 0.25° resolution, we incorporated the data from the closest point on the grid to the buoy's position. We assume oceanographic variables change on a larger spatial scale compared to the grid spacing and buoy's hourly movement, so no significant errors are incurred in this approximation.

2.2.2. Echo-sounder window

Tuna schools exhibit known circadian behaviour around dFADs: arriving at the dFAD at or near sunrise, and departing around sunset, remaining near the dFAD for several days in a row (Forget et al., 2015; Dagorn et al., 2007). To capture these patterns, we include a large enough window of echo-sounder measurements with hourly frequency as an input to the model. We tested how including time windows of

[3] http://marine.copernicus.eu/

Table 2
Number of Set and Deployment events, per ocean, per buoy model, and in total, remaining after merging echo-sounder and FAD logbook data.

	Ocean basin			Buoy model			
	Atlantic	Indian	Pacific	ISL+	SLX+	ISD+	Total
Set	1,500	2,727	974	4,877	192	132	5,201
Deployment	1,369	2,199	3,426	6,443	297	254	6,994
Total	2,869	4,926	4,400	11,320	489	386	12,195

Table 3
Grouped features used for the models. "Echo" included only data relating to echo-sounder measurements from the echo-sounder buoy database (in `blue`). "Echo + Ocean" included oceanographic data for the position and date of each record in the echo-sounder buoy database (in `green`). "All" contained further derived data from the position and time of each record in the echo-sounder buoy database (in `red`).

	Echo	Echo + Ocean	All
Biomass measurements	✓	✓	✓
Number of zero-readings	✓	✓	✓
Buoy model	✓	✓	✓
Chlorophyll-a		✓	✓
Dissolved oxygen		✓	✓
Salinity		✓	✓
Thermocline depth		✓	✓
Temperature		✓	✓
Current velocity		✓	✓
SSHa		✓	✓
Day and month			✓
Year			✓
Latitude			✓
Longitude			✓
Ocean basin			✓
Sunrise hour			✓
Sunset hour			✓

different lengths (24, 48 or 72 h) affected the model's ability to correctly estimate the daily tuna biomass. The echo-sounder window length with the best results was used in all the following analyses.

For set events, the selected window starts at sunset of the day before the event and ends 24, 48 or 72 h before that (see Fig. 2 for an example). The reason for this choice is two-fold: starting at sunset aligns all the observations with solar time regardless of time-zone, and ensures that all the echo-sounder measurements in the window will be taken before the set event, regardless of when in the day it occurred.

For deployment events, we cannot take echo-sounder measurements before the event, since the buoy is not in the water yet. For that reason, the echo-sounder window is selected after the deployment takes place, again aligned with the solar time. As mentioned in Section 2.1.1, we are assuming that no tuna is aggregated during 1–5 days after the deployment, following the conclusions derived from Orue et al. (2019b).

2.2.3. Data cleaning

As mentioned previously, data used in this study can contain errors, particularly in the case of FAD logbook data, as this information is recorded manually. To minimize any possible errors, the following conditions had to be met in order for an event (set or deployment) to be included in the final dataset:

- The buoy ID registered in the FAD logbook data must match the buoy ID in the echo-sounder database, i.e. we ensure that echo-sounder data are available for the dFAD on which the event took place. This avoids problems where the buoy ID is misreported in the FAD logbook, unless by chance the misreported ID matches the one from another buoy.
- The windows described in Section 2.2.2 from a given event cannot overlap with the window of another event. For instance, we exclude from our analysis sets that took place within a few hours of each other. This requirement is placed to ensure that during the window of echo-sounder measurements used for the estimation there is no human intervention on the dFAD.
- Following the same procedure as (Escalle et al., 2019), events with invalid positions (i.e. buoys on land) were removed from the dataset.
- Events or measurements registered at positions with less than 200 m water depth were discarded. This avoids including echo-sounder measurements that are potentially influenced by the sea-bed.
- Using the last known position of the buoys, we computed buoy speed for each position, and dropped events and measurements where buoy speed was higher than 3 knots, since the surface currents in the tropical oceans rarely exceed this speed (Orue et al., 2019a). This avoids including measurements taken on-board a vessel and not representative of a dFAD.

After merging and filtering, the final dataset contained over 12,000 events (see Table 2). These occurred on 10,063 buoys, for which over 665,000 echo-sounder records were collected. It is clear from Table 2 that events were more or less evenly distributed by ocean, which made it possible to carry out further studies stratified by ocean basin. However, the situation was radically different for the buoy model, where most of the interactions occurred on ISL+ buoys, so similar studies could not be

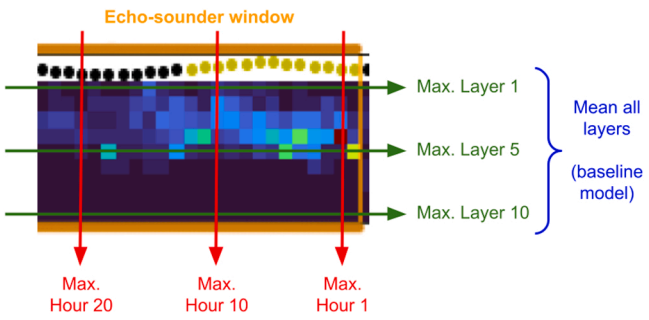


Fig. 3. Visual example of how the biomass measurements were aggregated: columns of colored squares represent the echo-sounder buoy biomass estimates for each hour, while rows are depth bins or layers. Square color represents the value of the estimated biomass. First, the maximum value for each layer is computed, resulting in a vector of size 10 (in `green`). Second, the maximum value for each hour is calculated, generating a vector of size {24, 48, 72}, depending on the size of the echo-sounder window (in `red`). Finally, the baseline model (in `blue`) was used as an input for the Machine Learning models. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

conducted with the available data.

2.3. Model selection

We tested several models varying feature sets, in order to assess the relative contribution of different features to model accuracy, as well as to evaluate the overall performance of different modelling methods.

2.3.1. Baseline model

As a baseline, we defined a model using only the biomass estimates from records contained in the echo-sounder window (see Section 2.2.2). However, as the output of the model (i.e., the overall biomass estimation) would be a single number, it was necessary to apply a set of aggregation rules on the echo-sounder window matrix. We tested the maximum, sum, mean, of the rows and columns of the echo-sounder window, and selected the one with the best performance. The details of all tested methods can be found in Appendix B. The selected baseline model was the maximum of all the hours for every layer and then the mean of all layers.

2.3.2. Feature engineering

Based on the merged dataset (see Section 2.2.1), we considered the variables included in Table 3 as features to be included in each model. The original biomass measurements form a $10 \times \{24, 48, 72\}$ matrix, depending on the echo-sounder window size. These values were not fed directly into the models; instead they were aggregated using the following rules, as shown in Fig. 3:

- By row (layer), using the maximum. This results in a vector of size 10.
- By column (hour), using the maximum. This results in a vector of size 24, 48 or 72, depending on the size of the echo-sounder window.
- By layer and by hour, computing first the maximum of the 24, 48 or 72 h and then the mean by layer. This results in a single value. The reason for using the maximum and then the mean is that it was the best performing baseline model (see Section 2.3.1).

These vectors (and baseline model) were then used directly as features for the different models. The total number of echo-sounder variables was then $\{24, 48, 72\} + 10 + 1$, depending on the size of the echo-sounder window. The previous procedure to generate feature vectors is applied to both set and deployment events.

2.3.3. Task description

Models were trained to achieve four different tasks, listed below by increasing level of complexity:

1. A binary classification task, where the target variable y (tuna biomass) could assume the values $y < 10$ t or $y \geq 10$ t.
2. A ternary classification task, where the target variable y (tuna biomass) could assume the values $y < 10$ t, $10\text{ t} \leq y < 30$ t or $y \geq 30$ t.
3. A threshold regression task, where we directly estimated the tuna biomass y, in metric tons, up to a threshold of 100 t. Estimations equal or higher than that were clipped to 100 t.
4. A regression task, where we directly estimated the tuna biomass y, in metric tons.

The thresholds to define the categories were chosen according to various criteria. In both classification tasks, the lower threshold was based on best-practice guidelines to decrease shark bycatch, which recommend avoiding sets on tuna schools less than 10 t (Dagorn et al., 2012). In the ternary classification task, the second class was further split by the median catch of the dataset (30 t). In the threshold regression task, we selected 100 t since sets above that were relatively rare (315 events, 8.1%). The full distribution of the tons of tuna captured in the set events can be seen in Appendix A.

2.3.4. Machine learning models

Following the usual approach in supervised Machine Learning, we split the dataset into training (75%, 9,152 events, 3,893 sets and 5,259 deployments) and test (25%, 3,051 events, 1,309 sets and 1,742 deployments) preserving the total class distribution. We did not stratify these splits by ocean, since the number of observations was similar (see Table 2). We considered the performance of a baseline rule-based model and four different Machine Learning models in the classification and regression tasks:

- Baseline model (see Section 2.3.1).
- Logistic Regression (LR) classifier (Cox, 1958): a linear model for the classification task.
- Elastic Net (ENet) regressor (Zou and Hastie, 2005): for the regression task, with three regularization techniques, namely $L1$ penalization, $L2$ penalization and elastic net.
- Random Forest (RF) algorithm (Breiman, 2001).
- Gradient Boosting (GB) algorithm (Friedman, 2001).
- XGBoost (XGB) algorithm (Chen and Guestrin, 2016).

For training and evaluating the models, we used the corresponding algorithms implemented in the Python `scikit-learn`(Pedregosa et al., 2011 and `XGBoost`(Chen and Guestrin, 2016) libraries. Each model was trained on three different sets of predictor variables, listed in Table 3.

2.3.5. Hyper-parameter tuning and model comparison

For each model, we performed a grid search with 5-fold cross-validation to find the optimal hyper-parameters. To select the best set of hyper-parameters for each model, we used the Area Under the Receiver Operating Characteristic Curve (ROC AUC) for the classification tasks and the Mean Absolute Error (MAE) for the regression tasks. AUC is defined by plotting the ROC curve (graphing the true positive rate against the false negative rate at several thresholds) and computing the area below the curve. MAE score is defined as the average of the absolute values of the errors when comparing the observed and the predicted values.

For each binary classification model we report the F1-score, which is the harmonic mean of precision and recall assuming than the positive class is $y \geq 10$ t. For the multi-class task, we report the averaged F1-score, weighted according to the proportion of the observations in

Table 4
Model score according to echo-sounder window size for Gradient Boosting regression and classification models.

	Classification (F1-score)		Regression (MAE)	
Hours	Binary	Three class	Standard	Threshold
24	0.911	0.811	10.16	8.70
48	0.919	0.813	10.05	8.63
72	**0.925**	**0.824**	**10.03**	**8.54**

Table 5
F1-scores for test events (classification task).

	Binary			Three class		
Models	Echo	Echo + Ocean	All	Echo	Echo + Ocean	All
Baseline	0.754	–	–	0.648	–	–
LR	0.885	0.889	0.895	0.773	0.788	0.799
RF	0.893	0.911	0.918	0.794	0.799	0.807
XGB	0.900	0.913	0.922	0.798	0.805	0.813
GB	0.907	0.924	**0.925**	0.791	0.812	**0.824**

Table 6
MAE scores for test events (regression task).

	Regression			Regression (Threshold)		
Models	Echo	Echo + Ocean	All	Echo	Echo + Ocean	All
Baseline	12.85	–	–	11.40	–	–
ENet	13.99	13.70	13.52	12.18	11.84	11.60
RF	10.74	10.30	10.20	9.42	8.93	8.84
XGB	11.37	10.86	10.76	9.60	9.13	9.02
GB	10.51	10.10	**10.03**	9.18	8.74	**8.54**

each class.

2.4. Best model performance

Finally, we performed a detailed analysis of the best models for each task. In these analyses, apart from the metrics listed above, we also computed the confusion matrix for the binary and multi-class classification problems.

In order to understand whether the models could be performing poorly in some specific subsets of the data, we also show the errors stratified by event type (set or deployment) and ocean basin. We limited these analyses to the tasks of binary classification and regression. Additional metrics were also computed: accuracy for classification and the Symmetric Mean Absolute Percentage Error (SMAPE) for regression, defined as follows,

$$\mathrm{SMAPE}(\%) = \frac{100\%}{n} * \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{|y_i| + |\hat{y}_i|}$$

where y_i is the actual value and $\hat{y}_i$ the estimated value. The main reason for reporting the SMAPE instead of the more common Mean Absolute Percentage Error (MAPE) is that the latter is undefined when $y_i = 0$, which happens with all the deployments. Besides, the SMAPE definition ranges between 0% and 100%, which makes it easier to interpret.

3. Results

3.1. Echo-sounder window selection

All GB models, regardless of task, were improved with extended echo-sounder windows. Within the classification tasks, the best overall results were achieved by the binary classification model (F1-score = 0.925, Table 4) using the 72 h echo-sounder window. Similar results are shown for the regression tasks, where models using the 72 h echo-sounder window had the lowest MAE (Table 4). Of the two regression tasks, the threshold regression performed better, with MAE almost 1.5 t lower than standard regression (Table 4).

As echo-sounder windows spanning 72 h showed the best result across all models, this was the echo-sounder window considered in the following analyses.

3.2. Classification models comparison

The performance of all the classification models tested for the 72 h echo-sounder window is shown in Table 5. The best performing model in both classification tasks was Gradient Boosting (GB). Performance also increased for every model as the number of features included in the training increased, i.e. when the models were able to learn from a larger set of features. Thus, the highest overall accuracy score was achieved by the binary classification GB model trained with all features (F1-score = 0.925, Table 5). The least accurate results were achieved by the ternary classification baseline model, which was almost 20% less accurate than the best performing model for this task, the GB model with all features. Note that there was an increase in F1-score between every ML model and the baseline.

3.3. Regression models comparison

Regarding the regression task, the results obtained by all the models trained on the different sets of predictor variables for the 72 h echo-sounder window are shown in Table 6. As in the classification tasks, the GB model also showed the overall best performance. More specifically, the threshold regression GB model was the most accurate, achieving a MAE nearly 3 t lower than the baseline model for the same task, and 1.49 t lower than the GB model for the standard regression (Table 6). It is also noteworthy that, as for the classification tasks, all models benefited from the inclusion of position and oceanography data, and were able to use this information to improve their predictions with respect to models that were only fed the echo-sounder data. Although some of these differences were small and they may be non-significant, it was clear that the Machine Learning models improved the baseline (which was in turn the best one out of several possible aggregations, B) and benefited from including all the variables.

3.4. Best models results

When analysing the confusion matrix for the test set of both classification tasks (Fig. 4), we see that the GB model in the binary classification task had a high success rate in classifying whether tuna biomass was < 10 t or ≥ 10 t, misclassifying results in only 6.03% of cases (Fig. 4). However, the ternary classification GB model found it harder to discriminate between 10 t $\leq y <$ 30 t and $y \geq 30$ t biomass estimations, having misclassified results in these two classes in 11.14% of cases (Fig. 4). Table 7 shows the results of the binary classification model by ocean basin and event type. Here, we see that the model was better at correctly estimating tuna biomass when being tested on deployments (Accuracy = 0.983) than on sets (Accuracy = 0.878), i.e. it was better at estimating when tuna biomass was < 10 t than ≥ 10 t. It is worth noting also that there was a considerable drop in accuracy in the Atlantic Ocean when compared to the rest, particularly when being tested on sets.

The MAE shown in Table 6 hides an important fact: the errors were very different for the two events included in the test data. Indeed, deployment events have by definition an observed biomass of zero: when tested over deployment events, the GB model had a MAE of 1.29 t, while the MAE for set events was 21.66 t (see Table 8). The reported overall MAE of 10.03 t is thus the weighted average of these different populations.

When looking more closely at the predictions of the best regression model, shown in Fig. 5a, it becomes apparent that for very high tuna

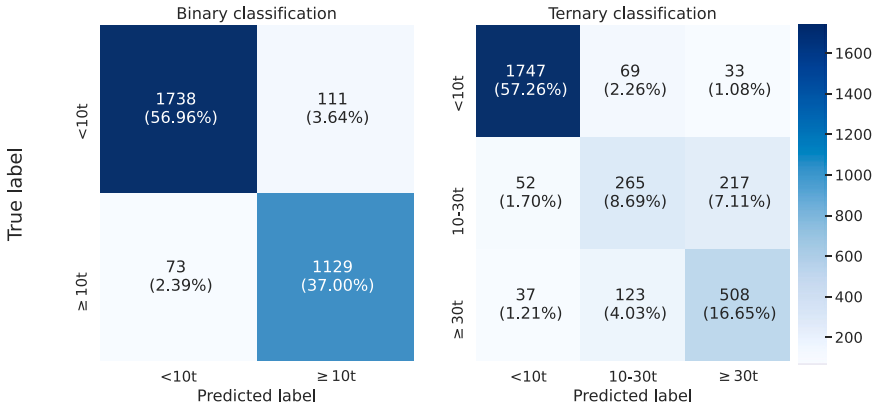


Fig. 4. Confusion matrices with the performance of the best classification models on the test set. True label refers to the category of the observed biomass, while predicted label is the category estimated by the model.

Table 7
Errors for the binary classification task and the best model (GB) by ocean and event type.

Ocean	Sets		Deployments		All	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
Atlantic	0.876	0.787	0.991	0.982	0.869	0.878
Indian	0.953	0.911	0.982	0.965	0.939	0.935
Pacific	0.964	0.930	0.997	0.994	0.953	0.980
All	0.934	0.878	0.991	0.983	0.925	0.938

Table 8
Errors for the regression task and the best model (GB) by ocean and event type.

Ocean	Sets		Deployments		All	
	MAE () t	SMAPE (%)	MAE () t	SMAPE (%)	MAE () t	SMAPE (%)
Atlantic	14.40	30.05	2.95	92.92	9.07	59.33
Indian	23.57	29.55	1.72	52.25	13.84	40.99
Pacific	27.96	28.52	0.36	32.09	6.44	31.31
All	21.66	29.51	1.29	51.15	10.03	41.86

biomass ($y \geq 100$ t) the model systematically underestimated tuna biomass. This result fits well with the improvement mentioned previously of the threshold regression task in relation to the standard regression. For this model, the MAE over set events dropped down to 18.33 t, and over deployments it decreased also to 1.18 t. However, even with this threshold the model tended to underestimate tuna biomass when observed biomass was high (Fig. 6b). Some possible factors that explain this underestimation are given in Section 4. As far as performance in each ocean basin (see Figs. 5), no large differences were apparent. The marginal distributions for observed and estimated tuna biomass on set events are depicted in Fig. 6a.

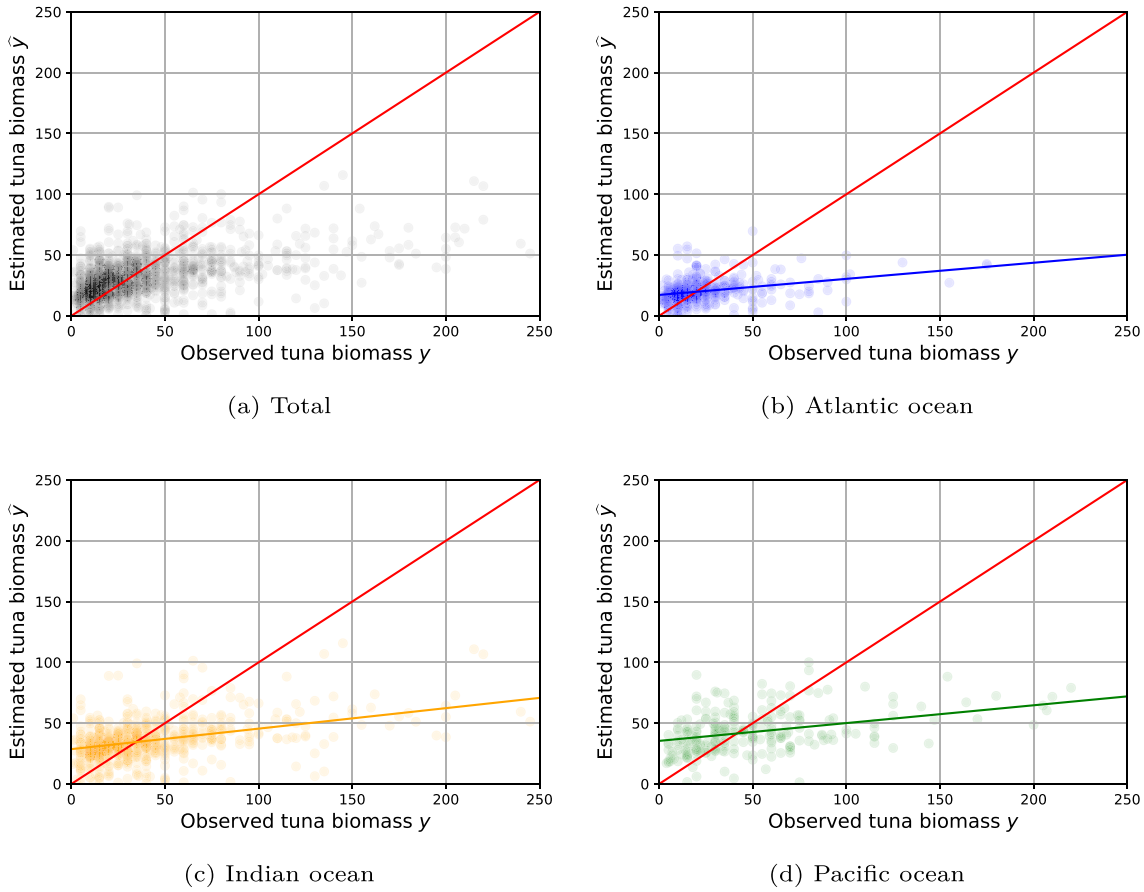
To try to understand the influence of each one of the variables, we include in C the permutation importance (Breiman, 2001). It can be seen that the top 10 variables include the baseline model (aggregated echo-sounder measurements), the position/ocean basin and some oceanographic variables such as the surface height anomaly, further evidencing that all the data sources helped the model make better estimations. Note that interpretation of feature importance must be exercised with care, since there are clearly correlations among the variables, and this has to be taken into account when interpreting the permutation importance.

4. Discussion

The purpose of this paper is to present a new pipeline for estimating tuna biomass aggregated at dFADs, named TUN-AI. The pipeline uses echo-sounder buoy, oceanographic and FAD logbook data to train multiple Machine Learning models that solve different tasks relevant to fisheries operations. To find the most accurate methodology, we tested the performance of classification and regression methods, as well as the relative impact of including different data sources on model performance. The approach used in the current study differs from previous work in several ways. Although the methodology in Baidai et al. (2020) is similar to ours, they only tackle the classification problem, and thus they are not able to directly estimate the metric tons of tuna under the dFAD. They also have a smaller sample size in terms of sets (albeit similar) that covers only the Atlantic and Indian oceans. Finally, we have also tested several models for each task in order to find the one with the best overall performance. Model performance of the two studies are hard to compare directly, since the models have been trained on different datasets. It is worth noting that other studies that address the regression problem, like Orue et al. (2019a); Lopez et al. (2016), cannot be directly compared with this study for a number of reasons: First, their sample size is orders of magnitude smaller (21 and 138 sets, respectively). Second, they only have data from a single ocean (Atlantic and Indian, respectively). Finally, they perform a statistical model fit, while our study involves a full ML approach with train-test split and a much larger dataset. This means that TUN-AI is expected to have the reported performance on new, unseen data, while there is no guarantee that the models in Orue et al. (2019a) and Lopez et al. (2016) will generalize as well, as they use the same dataset for model fit and error evaluation.

In addition, the assumptions and data-processing methods applied in other work may not be directly comparable to the process described here. For example, Orue et al. (2019a); Lopez et al. (2016) assume that tunas only occupy layers deeper than 25 m, thus omitting biomass estimates from shallower layers in their analyses. In our case, all layers were considered, as skipjack tuna are known to prefer warmer surface waters in areas where the thermocline is shallow (Andrade, 2003). In fact, later studies using the same approach as Lopez et al. (2016) did not achieve significant improvements on biomass estimates (Orue et al., 2019b). When developing tuna presence/absence and classification models, Baidai et al. (2020) also chose to consider all layers in their analyses, which used data from a different brand of echo-sounder buoys in the Atlantic and Indian oceans, but did not consider oceanographic parameters in their models.

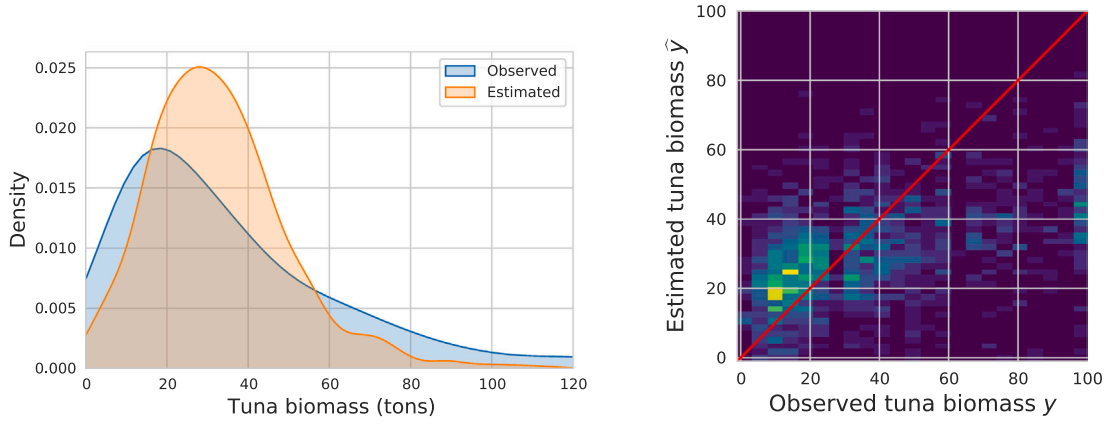
Our analysis also evaluated the impact of oceanographic conditions and position-derived variables on model performance. Across all tasks and models, the inclusion of additional features clearly improved when compared to the model that only used echo-sounder data. This



(a) Total (b) Atlantic ocean

(c) Indian ocean (d) Pacific ocean

Fig. 5. Scatter plot of the observed tuna biomass against estimated biomass in set events by ocean. Red line is $\hat{y} = y$, shown for reference.



(a) Density distributions (set events) of observed and estimated tuna biomass for the standard regression task.

(b) Two-dimensional histogram of observed and estimated tuna biomass for the regression threshold task plot.

Fig. 6. Error distribution of the two regression tasks.

highlights the importance of enriching biomass estimates with contextual information when using data from echo-sounder buoys attached to dFADs. Although at first glance this would prove laborious, the current pipe-line draws from automated processes for extracting the oceanographic data and relating it to the other available datasets, thus the added complexity translated to only a few minutes of additional computation time on standard equipment. Given the improvement in model accuracy when including this information, and the potential applications of having accurate methods for estimating tuna biomass at dFADs, we consider that it is worthwhile to use all available information. Previous studies have investigated the relationship between tropical tuna distribution and oceanographic conditions, both through catch data from observer logbooks and from dFAD data. For instance, in the Atlantic and Indian oceans skipjack tuna has been known to aggregate around upwelling systems and productive features where feeding habitat is favorable, and variables such as sea surface temperature or SSH have been shown to have a significant relation with tuna distribution (Druon et al., 2017; Lopez, 2017). In addition, Spanish fishers using

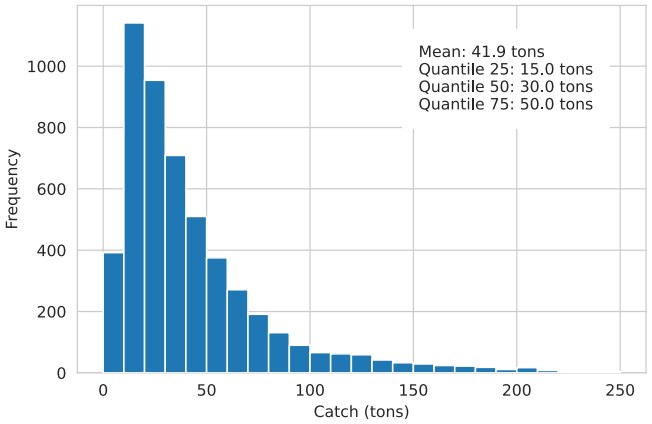


Fig. A.7. Distribution of tons of tuna caught, from a total of 5202 sets.

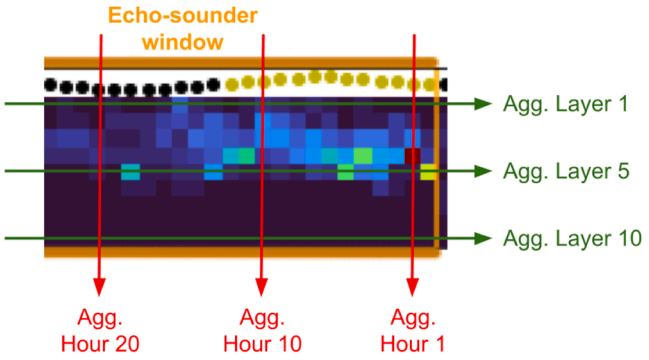


Fig. B.8. Visual example of how the biomass measurements are aggregated.

Table B.9
Different aggregations of data for the regression baseline model.

First aggregation direction	First aggregation function	Second aggregation function	MAE
layer	max	mean	13.86
layer	mean	sum	13.98
hour	mean	max	15.64
hour	max	mean	15.69
layer	max	max	16.47
layer	mean	max	16.64
layer	mean	mean	17.59
hour	sum	max	28.42
layer	sum	mean	37.67
layer	max	sum	59.91
layer	sum	max	143.78
hour	max	sum	243.02
layer	sum	sum	483.00

echo-sounder buoys on dFADs consider that the oceanographic context of the dFAD, and the characteristics of each ocean influence the accuracy of biomass estimates provided by buoys (Lopez and Scott, 2014).

It is worth noting that the oceanographic variables included in the current study were at surface level only (except for thermocline depth and SSHa). However, given the fact that tuna distribution within the water column is largely temperature dependent (Aoki et al., 2020; Hino et al., 2019; Tanabe et al., 2017) it is likely that models would further improve when considering variables depth-wise. Models could be also enriched by considering dFAD soak time, which has been relevant in previous research, or presence/absence of bycatch species and other species of tuna (Orue et al., 2019b; Lopez, 2017; Forget et al., 2015). Indeed, the presence of small schooling bycatch species such as oceanic triggerfish (*Canthidermis maculata*), which has been found at dFADs

Table C.10
Top ten most important features for the GB model in each of the 4 tasks. The variables are coloured depending on the feature group they belong. `blue` corresponds to echo-sounder features, `green` refers to oceanografic variables and `red` identifies geographical coordinates (or features derived from them).

	Classification		Regression	
Rank	Binary	Ternary	Standard	Threshold
1	Max.L5	Max.L5	Baseline	Baseline
2	Longitude	Longitude	Max.L5	Max.L5
3	Ocean	Ocean	Longitude	Longitude
4	N_Zero	Max.L2	Latitude	Max.L2
5	Latitude	Max.H10	Max.H32	SSHa.D0
6	Max.L7	SSHa.D0	Max.L2	Latitude
7	Max.L1	Latitude	SSHa.D0	Max.L6
8	Max.L6	SSHa.D3	Max.H35	SSHa.D1
9	Baseline	02.D1	SSHa.D1	SSHa.D3
10	Max.L2	02.D0	SSHa.D3	Max.H10

Table D.11
Grid of hyper-parameters used in the Random Forest models.

Parameter	Classification	Regression
n estimators	[200, 500, 1000]	[100, 200, 500]
max samples	[None, 0.8]	[None, 0.8]
max depth	[None, 2, 4]	[None, 4, 8]
min samples split	[2, 8, 32]	[2, 8, 32]
min samples leaf	[1, 4, 16]	[1, 4, 16]
max features	[None, sqrt, log2]	[None, sqrt, log2]

Table D.12
Grid of hyper-parameters used in the Gradient Boosting models.

Parameter	Classification	Regression
n estimators	[50, 100, 200]	[400]
learning rate	[0.01, 0.1, 0.2]	[0.01, 0.1, 0.2]
max depth	[None, 3, 6]	[None, 3, 6]
min samples split	[2, 4, 8]	[2, 4, 8]
min samples leaf	[1, 2, 4]	[1, 2, 4]
max features	[None, sqrt, log2]	[None, sqrt, log2]

(Forget et al., 2015), could be affecting the biomass estimates provided by the echo-sounder buoys. We see some evidence of this when examining the accuracy of our models in the binary classification task (see Table 7): the model performed worse when being tested on sets. This is likely due to cases where species other than tuna were contributing to the echo-sounder signal, so the buoy's biomass estimates were high even though real catch of tuna at the dFAD was low. This was reflected particularly in the Atlantic Ocean, where accuracy was lowest and bycatch is higher than in other oceans (Restrepo et al., 2017). Future studies could include the bycatch information recorded in the FAD logbooks to account for this effect. Looking more closely into the binary classification model, we can see that the worst performance comes from trying to distinguish sets that catch less than 10 t from the rest. This makes sense, since the fact that a fishing vessel decided to set on a specific FAD is probably already a good indicator of the echo-sounder measurements showing a strong signal, which probably correspond to other fish species. These are clearly the hardest observations to distinguish.

In the current study, species composition of the catch data was not considered. As the echo-sounder buoys used in this study calculate biomass estimates based on the target strength of skipjack tuna, it is likely that the presence of other tuna species such as bigeye, which has a lower target strength (Boyra et al., 2018, 2019) and thus stronger acoustic response, would impact biomass estimates from the

Table D.13
Grid of hyper-parameters used in the XGBoost models.

Parameter	Classification	Regression
n estimators	[50]	[50, 100, 200]
learning rate	[0.2]	[0.01, 0.1, 0.2]
max depth	[2,4]	[2, 4, 6]
subsample	[1.0]	[0.7, 1.0]
colsample bytree	[1.0]	[0.5, 1.0]

Table D.14
Best hyper-parameters for the Random Forest models trained on all features.

	Classification		Regression	
Parameter	Binary	Three class	Standard	Threshold
n estimators	1000	500	100	200
max samples	None	None	None	None
max depth	None	None	None	None
min samples split	2	8	8	2
min samples leaf	1	1	4	4
max features	sqrt	sqrt	None	None

Table D.15
Best hyper-parameters for the Gradient Boosting models trained on all features.

	Classification		Regression	
Parameter	Binary	Three class	Standard	Threshold
n estimators	200	200	400	400
learning rate	0.2	0.1	0.01	0.01
max depth	6	6	None	None
min samples split	8	2	2	4
min samples leaf	4	2	4	8
max features	log2	log2	auto	auto

Table D.16
Best hyper-parameters for the XGBoost models trained on all features.

	Classification		Regression	
Parameter	Binary	Three class	Standard	Threshold
n estimators	50	50	200	100
learning rate	0.2	0.2	0.01	0.1
max depth	4	4	6	6
subsample	1.0	1.0	0.7	1.0
colsample bytree	1.0	1.0	0.5	0.5

echo-sounder buoys, contributing to errors within the models used to estimate aggregation size. Most traditional echo-sounder buoys do not currently differentiate between species when giving biomass estimates, though recent buoy models, such as the ISD+ buoys included in the study, provide a daily estimate of species composition together with biomass estimates. Although previous studies have highlighted the importance of considering species composition when estimating biomass (Moreno et al., 2019; Santiago et al., 2016), the information from these buoys has not yet been applied, and should be considered in future studies. Likewise, this study only used echo-sounder information from one buoy manufacturer, although vessels may use buoys from up to four different brands. The echo-sounders and buoys from each manufacturer vary on a number of levels: beam angle, sampling method, echo-sounder frequency, etc.; so the same machine learning models used here cannot be applied directly. Nonetheless, future studies could explore the application of similar ML models to the biomass estimates provided by other buoy brands in order to establish manufacturer-specific echo-sounder signal processing pipelines.

In the case of classification models, the confusion matrices in Fig. 4 showed that most cases where the model misclassified the tuna aggregation size were when biomass estimates were medium ($10\ t \leq y < 30\ t$) or high ($y \geq 30\ t$). On the other hand, when examining the regression models we found that estimated tuna biomass tended to be lower than observed tuna biomass as the latter increased (Fig. 6). This could be due to various factors: firstly, catches over 100 t were relatively rare (in our data, 315 events, 8.1%) and thus the model did not have sufficient examples to properly learn from them; secondly, buoys are only able to estimate the biomass of tuna within the echo-sounder beam, and in tuna aggregations over 100 t it is unlikely that the entire school is under the buoy at the same time. Furthermore, it is possible that with large schools of tuna the echo-sounder signal saturates, and the biomass estimates provided by the echo-sounder buoy become an underestimation. To resolve this issue, it could be interesting to apply specialist models which could be adjusted according to when aggregations are predicted to be small or large. It is also worth noting that fishermen do not choose on which buoys they set at random, but based on the biomass estimation provided to them, and thus could be biased towards buoys with higher biomass estimations. This could be a further reason why our ML models underestimated the observed tuna biomass when its values were above 30 t in the case of the ternary classification, or 100 t in the case of the regression tasks. Future studies exploring the reasons behind fishermen's decisions to visit a buoy could provide further insight into this point. This tendency to underestimate should also be taken into account when using information derived from echo-sounder buoys for stock assessments (Santiago et al., 2016), although consistent underestimation should have no effect on patterns present in the temporal series.

The pelagic and migratory nature of tuna make it a challenging species to study using traditional methods, as only a small part of this species' habitat can be observed in real time at any given moment. However, dFADs equipped with high-tech echo-sounder buoys, as those used in the current study, can be used as floating open-ocean sampling stations, gathering constant and current information from various sensors. As shown here, and as remarked by other authors (Orue et al., 2019a), although the information provided by the echo-sounder alone is valuable, it still requires extensive cleaning and filtering prior to use. These initial protocols can avoid errors due to measurements taken by the buoys on board or on land, but the ML techniques used here go a step further in processing the echo-sounder signal to correctly estimate tuna biomass beneath any given echo-sounder buoy. This way, the echo-sounder signals can provide insight into the whereabouts and behaviour of tuna around dFADs, at a fraction of the cost of what scientific expeditions with the same scope could achieve. As highlighted by previous authors, this type of data could be used for fishery-independent abundance indices, improving knowledge on species distribution or better understanding the factors driving aggregation and disaggregation processes of tuna at dFADs (Santiago et al., 2016; Lopez et al., 2016; Moreno et al., 2019). The current study represents an important step in this direction, being the first to successfully evaluate the performance of numerous ML models, following correct ML methodology using large amounts of data to both train and test each model. This has allowed Tun-AI to tackle the most tasks of various complexities, including directly estimating the amount of tuna aggregated to the dFAD, achieving high degrees of accuracy. As evidenced here, when the massive data provided by echo-sounder buoys attached to dFADs is further enriched with remote-sensing data on conditions across oceans, and trained with reliable ground-truthed data, ML proves a powerful tool for extracting otherwise hidden patterns in these datasets, potentially furthering knowledge on pelagic species.

CRediT authorship contribution statement

Daniel Precioso: Data Curation, Software, Methodology, Writing – original draft, Writing – review & editing, Visualization, Formal analysis. **Manuel Navarro-Garcia:** Data Curation, Software, Methodology, Writing – original draft, Writing – Review & Editing, Visualization, Formal analysis. **Kathryn Gavira-O'Neill:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Validation,

Supervision, Project administration. **Alberto Torres-Barrán:** Data Curation, Software, Methodology, Writing – original draft, Writing – review & editing, Project administration, Formal analysis. **David Gordo:** Conceptualization, Methodology, Resources. **Víctor Gallego:** Conceptualization, Methodology, Resources. **David Gómez-Ullate:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study has been conducted using E.U. Copernicus Marine Service Information. We also thank AGAC for providing the logbook data used in the analysis and the helpful comments about the manuscript. The authors would also like to thank Carlos Roa for rendering available the Satlink echosounder dataset. The research of DGU has been supported in part by the Spanish MICINN under grants PGC2018-096504-B-C33 and RTI2018-100754-B-I00, the European Union under the 2014-2020 ERDF Operational Programme and the Department of Economy, Knowledge, Business and University of the Regional Government of Andalusia (project FEDER-UCA18-108393). The research of Manuel Navarro-García has been financed by the research project IND2020/TIC-17526 (Comunidad de Madrid). The research of Alberto Torres has been financed in part by a Torres Quevedo grant PTQ2019-010642 from Agencia Estatal de Investigación (Spain). The research of Daniel Precioso has been financed by an Industrial PhD grant from the University of Cádiz.

Appendix A. Catch volume distribution

When defining the categories for the classification task, the thresholds were chosen according to various criteria. In both binary and ternary tasks, the lower threshold was based on best-practice guidelines to decrease shark bycatch, which recommend avoiding sets on tuna schools less than 10 t (Dagorn et al., 2012). In the ternary classification task, we used the median catch (30 t) of the dataset. The full distribution of the tons of tuna is shown in Fig. A.7.

B. Baseline model

The raw echo-sounder measurements provided by dFAD buoys are already converted into biomass estimates in metric tons. However, to provide a single estimation, we still need to transform the $72 \times 10 = 720$ matrix into a single value. These values can be aggregated in two directions (see Fig. B.8): by rows (layers) or by columns (hours), and also using different aggregation rules. In this work we have decided to test all possible combinations of aggregation rules (mean, maximum and sum) and aggregation directions (by hour and by depth bin or layer), selecting the one that obtains the lowest Mean Absolute Error (MAE), that is, the one that is able to best estimate the tons of tuna captured under the dFAD.

It is important to note that some combinations lead necessarily to the same MAE, and in these cases only one of them is reported. For instance, if the aggregation function by layers and hours is the same, then the order in which these aggregations are performed is irrelevant. The same applies to aggregations that contain only the sum and the mean, since they are both linear functions.

The results for all possible combinations can be found in Table B.9. As an example of how to read this table, consider the best performing aggregation: layer, max, mean. The final predictions for this aggregation are obtained by first taking the **maximum** value of the 72 h for every **layer**. This results in 10 numbers, one per layer. Then, we further aggregate these 10 numbers by computing the **average**, resulting in a single number.

The baseline for the classification models are directly computed from the previous estimate, just by checking whether the output is above or below the given thresholds.

C. Feature importance

A challenging task in ML projects is their interpretability: to understand which features the model considers most relevant in its calculations. There exist several approaches to assess feature importance, and in this work we employ permutation importance (Breiman, 2001), which evaluates the importance of a given feature as the drop in model efficiency when the values of that column are shuffled randomly in the training set.

In Table C.10 we rank the 10 most important features for each task for the GB model. We next explain the meaning of each feature: .

- `Baseline` is the tons of tuna estimated by the baseline model, which aggregates the values of the 72×10 echo-sounder window.
- `N_Zero` is the number of zero-readings in the echo-sounder window, ranging from 0 to 72. The buoy does not send an hourly biomass estimate to the satellite if the total biomass estimation is below 1 t.
- `Max.LY` is obtained by computing the maximum of the 72 values for each layer `Y`, where layer 1 is the one closest to the surface.
- `Max.HX` is obtained by computing the maximum of the 10 values for each hour `X`, where 0 is the one closest to the event.
- `O2. DX` and `SSHa.DX` are the dissolved oxygen and SSHa, respectively, for each day `X`, where 0 is the closest to the event.
- `Latitude` and `Longitude` are the buoy coordinates closest to the event.
- `Ocean` is a categorical variable that indicates the ocean basin where the event took place.

Interpretation of feature importance must be exercised with care, since there are clearly correlations among the variables, and this has to be taken into account when interpreting permutation importance. However, we may briefly discuss the appearance of some of them in Table C.10, though further studies would be necessary to verify their validity. For instance, the feature `Baseline` emerges as the most important explanatory variable in both regression tasks, which is reasonable since it is itself a first-stage biomass estimation. Conversely, `N_Zero` appears as a relevant predictor only in the binary classification task, since a reading of < 1 t likely implies that there is no tuna under the dFAD. Moreover, it is remarkable the geographical coordinates `Latitude` and `Longitude` appear as meaningful covariates for every task, highlighting the importance of considering this information source in the models. We also observe that `Max.L5` appears consistently as one of most important features for every task, which could be evidence of the prevalent depth at which tuna are present. Finally, some oceanographic variables such as `SSHa` also appear as significant features in the three most complex tasks, highlighting the fact that they are helpful to the models for estimating the tons of tuna more accurately.

D. Hyper-parameter search

A grid search with cross-validation has been conducted to find the best hyper-parameters for each model by maximizing the ROC AUC. The hyper-parameter grid for all the models using the full set of features in each of the four tasks is shown in Tables D.11, D.12 and D.13. For Logistic Regression and Elastic Net, a grid search was conducted using the standard classes `LogisticRegressionCV` and `ElasticNetCV`, with a `l1_ratio` grid of [0.0, 0.2, 0.4, 0.6, 0.8, 1.0] and [0.1, 0.5, 0.9, 1.0] respectively. The final values selected by cross-validation were 1 and 0.9. The name of the hyper-parameters coincide with the names that

they have in `scikit-learn` library (Pedregosa et al., 2011). All the parameters not shown here were set to their default value (see the documentation for more details). Finally, the set of optimal hyper-parameters for each of the models is shown in Tables D.14, D.15 and D.16.

References

Andrade, H.A., 2003. The relationship between the skipjack tuna (Katsuwonus pelamis) fishery and seasonal temperature variability in the south-western Atlantic. Fish. Oceanogr. 12 (1), 10–18.

Anon. Global Monitoring and Forecasting Center 2018. Operational Mercator global ocean analysis and forecast system, E.U. Copernicus Marine Service Information. ⟨https://resources.marine.copernicus.eu⟩ (Accessed 15th January 2021).

Aoki, Y., Aoki, A., Ohta, I., Kitagawa, T., 2020. Physiological and behavioural thermoregulation of juvenile yellowfin tuna Thunnus albacares in subtropical waters. Mar. Biol. 167 (6), 71.

Baidai, Y., Dagorn, L., Amande, M.J., Gaertner, D., Capello, M., 2020. Machine learning for characterizing tropical tuna aggregations under Drifting Fish Aggregating Devices (DFADs) from commercial echosounder buoys data. Fish. Res. 229, 105613.

Boyra, G., Moreno, G., Sobradillo, B., Pérez-Arjona, I., Sancristobal, I., Demer, D.A., 2018. Target strength of skipjack tuna (Katsuwanus pelamis) associated with fish aggregating devices (FADs). ICES J. Mar. Sci. 75 (5), 1790–1802.

Boyra, G., Moreno, G., Orue, B., Sobradillo, B., Sancristobal, I., 2019. In situ target strength of bigeye tuna (Thunnus obesus) associated with fish aggregating devices. ICES J. Mar. Sci. 76 (7), 2446–2458.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Castro, J.J., Santiago, J.A.J., Santana-Ortega, A.T., 2002. A general theory on fish aggregation to floating objects: alternative to the meeting point hypothesis. Rev. Fish. Biol. Fish. 11 (3), 24.

Anon. ISSF 2021, Status of the World Fisheries for Tuna. Mar 2021. ISSF Technical Report 2021–10, March 2021(March): 1–120.

Chen, T. , Guestrin, C. , {XGBoost}: A Scalable Tree Boosting System.In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 785–794, New York, NY, USA. ACM. Boosting System Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., KDD '16 2016 785 794.(New York, NY, USA. ACM).

Cox, D.R., 1958. The regression analysis of binary sequences. J. R. Stat. Soc.: Ser. B Methodol. 20 (2), 215–232.

Dagorn, L., Holland, K.N., Itano, D.G., 2007. Behavior of yellowfin (Thunnus albacares) and bigeye (T. obesus) tuna in a network of fish aggregating devices (FADs). Mar. Biol. 151 (2), 595–606.

Dagorn, L., Filmalter, J.D., Forget, F., Amandè, M.J., Hall, M.A., Williams, P., Murua, H., Ariz, J., Chavance, P., Bez, N., 2012. Targeting bigger schools can reduce ecosystem impacts of fisheries. Can. J. Fish. Aquat. Sci. 69 (9), 1463–1467.

Davies, T.K., Mees, C.C., Milner-Gulland, E.J., 2014. The past, present and future use of drifting fish aggregating devices (FADs) in the Indian Ocean. Mar. Policy 45, 163–170.

Druon, J.N., Chassot, E., Murua, H., Lopez, J., 2017. Skipjack tuna availability for purse seine fisheries is driven by suitable feeding habitat dynamics in the Atlantic and Indian Oceans. Front. Mar. Sci. 4, 315 (OCT):

Escalle, L., Heuvel, B.V., Clarke, R., Brouwer, S., Pilling, G., Lauriane Escalle, Heuvel, B. V., Clarke, R., Brouwer, S., Pilling, G., 2019. Report on preliminary analyses of FAD acoustic data. West. Cent. Pac. Fish. Comm. 53 (9), 17.

Fonteneau, A., Pallarés, P., Pianet, R., 2000. A worldwide review of purse seine fisheries on FADs. Reg. Synth. 21.

Forget, F.G., Capello, M., Filmalter, J.D., Govinden, R., Soria, M., Cowley, P.D., Dagorn, L., 2015. Behaviour and vulnerability of target and non-target species at drifting fish aggregating devices (FADs) in the tropical tuna purse seine fishery determined by acoustic telemetry. Can. J. Fish. Aquat. Sci. 72 (9), 1398–1405.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Hino, H., Kitagawa, T., Matsumoto, T., Aoki, Y., Kimura, S., 2019. Changes to vertical thermoregulatory movements of juvenile bigeye tuna (Thunnus obesus) in the northwestern Pacific Ocean with time of day, seasonal ocean vertical thermal structure, and body size. Fish. Oceanogr. 28 (4), 359–371.

Lellouche, J.M., Greiner, E., LeGalloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C.E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., LeTraon, P.Y., 2018. Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1g 12° high-resolution system. Ocean Sci. 14 (5), 1093–1126.

Lopez, J., 2017. Environmental preferences of tuna and non-tuna species associated with drifting fish aggregating devices (DFADs) in the Atlantic Ocean, ascertained through fishers' echo-sounder buoys. Deep Sea Res. II 12.

Lopez, J., Scott, G.P., 2014. The use of FADs in tuna fisheries. Eur. Union 1, 70.

Lopez, J., Moreno, G., Sancristobal, I., Murua, J., 2014. Evolution and current state of the technology of echo-sounder buoys used by Spanish tropical tuna purse seiners in the Atlantic, Indian and Pacific Oceans. Fish. Res. 155, 127–137.

Lopez, J., Moreno, G., Boyra, G., Dagorn, L., 2016. A model based on data from echosounder buoys to estimate biomass of fish species associated with fish aggregating devices. Fish. Bull. 114 (2), 166–178.

Mannocci, L., Baidai, Y., Forget, F., Tolotti, M.T., Dagorn, L., Capello, M., 2021. Machine learning to detect bycatch risk: Novel application to echosounder buoys data in tuna purse seine fisheries. Biol. Conserv. 255, 109004.

Maufroy, A., Chassot, E., Joo, R., Kaplan, D.M., 2015. Large-scale examination of spatio-temporal patterns of drifting fish aggregating devices (dFADs) from tropical tuna fisheries of the Indian and Atlantic Oceans. PLOS ONE 10 (5), e0128023.

Molina, D.D., Pallares, P., Areso, J.J., Ariz, J., 2003. Statistics of the purse seine spanish fleet in the Indian Ocean (1984-2002). IOTC Proc. 6 (6), 115–128.

Moreno, G., Boyra, G., Sancristobal, I., Itano, D., Restrepo, V., 2019. Towards acoustic discrimination of tropical tuna associated with Fish Aggregating Devices. PLOS ONE 14 (6), e0216353.

Orue, B., Lopez, J., Moreno, G., Santiago, J., Boyra, G., Uranga, J., Murua, H., 2019a. From fisheries to scientific data: a protocol to process information from fishers' echo-sounder buoys. Fish. Res. 215, 38–43.

Orue, B., Lopez, J., Moreno, G., Santiago, J., Soto, M., Murua, H., 2019b. Aggregation process of drifting fish aggregating devices (DFADs) in the Western Indian Ocean: who arrives first, tuna or non-tuna species? PLOS ONE 14 (1), e0210435.

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., 2011. Scikit-learn: in Python. Python 6.

Ramos, M.L. , Báez, J.C. , Grande, M. , Herrera, M.A. , López, J. , Justel, A. , Pascual, P.J. , Soto, M. , Murua, H. Muniategi, A. , Abasca, F.J. , Spanish FADs logbook: solving past issues, responding to new global requirements.1st Ad-Hoc IOTC Working Group on FADs, 2017(April): 1–24.

Restrepo, V., Dagorn, L., Itano, D., Justel-Rubio, A., Forget, F., Moreno, G. , 2017 , A Summary of Bycatch Issues and ISSF Mitigation Initiatives To Date in Purse Seine Fisheries, with emphasis on FADs. ISSF Technical Report 2017–06, ISSF (November2017): 1–40.

Santiago, J. , Lopez, J. , Moreno, G. , Murua, H. , Quincoces, I. , Soto, M. , Towards a Tropical Tuna Buoy-Derived Abundance Index (TT-BAI).Collective Volume of Scientific Papers ICCAT, 72: 714–724.

Santiago, J. , Uranga, J. , Quincoces, I. , Orue, B. , Grande, M. , Murua, H. , Merino, G. , A Novel Index of Abundance of Juvenile Yellowfin Tuna in the Atlantic Ocean Derived from Echosounder Buoys.Collective Volume of Scientific Papers ICCAT, 76: 321–343.

Schaefer, K.M., Fuller, D.W., Block, B.A., 2007. Movements, behavior, and habitat utilization of yellowfin tuna (Thunnus albacares) in the northeastern Pacific Ocean, ascertained through archival tag data. Mar. Biol. 152 (3), 503–525.

Tanabe, T., Kiyofuji, H., Shimizu, Y., Ogura, M., 2017. Vertical distribution of juvenile skipjack tuna Katsuwonus pelamis in the tropical western Pacific ocean. Jpn. Agric. Res. Q. 51 (2), 181–189.

Wain, G., Guéry, L., Kaplan, D.M., Gaertner, D., O'Driscoll, R., 2021. Quantifying the increase in fishing efficiency due to the use of drifting FADs equipped with echosounders in tropical tuna purse seine fisheries. ICES J. Mar. Sci. 78 (1), 235–245.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.