scientific reports

Check for updates

OPEN

Machine learning for modeling North Atlantic right whale presence to support offshore wind energy development in the U.S. Mid-Atlantic

Jiaxiang Ji[1], Jeeva Ramasamy[2], Laura Nazzaro[3], Josh Kohut[3] & Ahmed Aziz Ezzat[1]✉

The Mid-Atlantic region is set to be one of the first and largest contributors to the offshore wind energy goals of the United States. Yet, the same region is home to a diverse marine ecosystem comprising important marine species such as the critically endangered North Atlantic right whale (NARW). To support the responsible development and operation of the planned offshore wind farms, there is a need for high-resolution modeling of NARW presence, i.e., at the spatial and temporal resolutions relevant to farm-level operations. Towards this, we leverage highly localized observations from nine glider deployments in the Mid-Atlantic to propose a machine learning approach for modeling NARW presence conditioned on a diverse set of glider- and satellite-based oceanic, physical, and contextual information. We find that tree- and ensemble-based models achieve the highest levels of accuracy, while maintaining a sensible balance of missed and false alarms. Interpretation of the machine-learnt features reveals interesting insights on the relative value of well-resolved satellite surface measurements to well-resolved vertical information from glider sampling in explaining the species-habitat patterns of NARWs. We then discuss the value of the models proposed herein to offshore wind developers and operators in the United States and elsewhere. Our work constitutes the first machine learning attempt to jointly leverage glider- and satellite-based information for modeling of NARWs. Data and codes for producing the results of this work have been made freely available to promote the research on this timely topic.

Keywords Marine mammals, Machine learning, Offshore wind energy

The United States (U.S.) has set an ambitious target to install 30 Gigawatts (GW) of offshore wind capacity by 2030[1]. The Northeastern U.S., with the Mid-Atlantic region at its heart, is set to be the first and largest contributor to this 30-GW-by-2030 milestone, with several utility-scale offshore wind projects that are either planned or already under construction[2]. The rapid increase in the pace and scale of the offshore wind activity therein holds great promise, both environmentally and economically[3]. Yet, the same region is home to a diverse marine ecosystem comprising several important marine species, including the critically endangered North Atlantic Right Whale (NARW), formally known as Eubalaena glacialis[4]. While there is no evidence that offshore wind farms directly contribute to marine mammal mortalities, developing and operating offshore wind farms in a highly dynamic and vibrant ocean environment could pose a number of considerable risks to NARW habitats. Noise generated during the construction and operation of offshore wind farms is a primary concern, whereas vessel traffic created by the need to service offshore wind energy assets is another risk[5]. Given their dwindling populations, effective conservation measures must be taken to ensure that offshore wind farms cause minimal to no disruption to the habitat of NARWs.

In order to inform the responsible management of offshore wind farms, a high-resolution understanding of NARW distribution is needed. Wind farm developers would highly benefit from a localized modeling of when and where it is more likely to encounter an NARW during construction and operation of their assets. There have been several efforts in the past to model how NARWs react to their local environments. Those efforts can

[1]Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ 08854, USA. [2]Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA. [3]Department of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA. ✉email: aziz.ezzat@rutgers.edu

be broadly grouped into two categories. The first category focuses on constructing statistical regression models in order to correlate historical NARW sightings to exogenous environmental covariates[6–9]. Examples include but are not limited to logistic regression models, generalized linear models, and generalized additive models. The second category develops density surface models in which distance sampling and regression models are combined to produce density estimates of NARWs that are conditioned on local environmental covariates[10–13]. Whether regression- or density-based, the vast majority of those models have traditionally relied on sighting data collected using visual surveys, often in the form of line transects, such as aerial, vessel, and more recently, drone surveys[14,15]. Those models are valuable in providing regional predictions of marine mammal density, resolving temporal and spatial scales that are typically broader than those needed to address farm-level decision making.

The last decade has seen a growing interest in leveraging passive acoustic monitoring (PAM) systems to more precisely locate marine mammals and collect localized information about their habitat at fine spatial and temporal scales[16]. A relevant example is the use of autonomous underwater gliders for marine mammal monitoring[17]. When equipped with acoustic detectors, gliders can quietly navigate in certain focus regions for fairly long periods of time to detect marine mammal presence and acquire granular habitat information[18]. This wealth of high-resolution glider data creates an opportunity for machine learning (ML) technologies to aid in high-resolution marine mammal distribution modeling, i.e., at the spatial and temporal resolutions relevant to farm-level operations. Despite some recent applications of ML for marine mammal data analysis[19–21], little attention has been devoted to developing ML methods that directly act on glider-based datasets for marine mammal predictive modeling, let alone for NARWs.

In this work, we undertake an ML approach to jointly fuse glider- and satellite-based information, for the first time, in order to predict NARW presence at granular spatial and temporal resolutions. In the context of this work, prediction refers to learning the likelihood of whale presence (or absence thereof) conditional on a set of out-of-sample environmental conditions that have not been seen by the ML model during its training stage. In doing so, we leverage highly localized observations from nine glider deployments in proximity to future offshore wind sites in the U.S. Mid-Atlantic, which comprise acoustic detections of NARWs, along with sampled profiles of oceanographic variables. Supplemented by co-located satellite information about relevant oceanographic and physical features, we train a cluster of ML classifiers to make predictions of when and where NARWs are likely to be present, conditioned on the combined set of glider- and satellite-based covariates. We find that tree- and ensemble-based models achieve the highest levels of accuracy, while maintaining a sensible balance of missed and false alarms. Interpretation of our best-performing models reveals interesting insights about the relative influence of vertically well-resolved water column observations provided by the autonomous glider missions to the spatially well-resolved surface ocean observations provided by the satellite products. We conclude by discussing the implications and utility of our models and findings to wind farm developers and operators, and highlight future research directions that can directly build on this work to unlock the promise of ML in marine mammal predictive modeling. To promote the research on this timely topic, we have made our data and codes freely available at https://github.com/Jiaxiang-J/NARW.

Results

Two datasets are merged to train the ML models in this work: a glider-based dataset denoted hereinafter by $\mathscr{D}^g$, and a satellite-based dataset denoted by $\mathscr{D}^s$. The glider-based dataset, $\mathscr{D}^g$, comprises observations from nine glider missions deployed between August 2020 and June 2022 in the south coast of New Jersey in proximity to several future offshore wind farm sites in the U.S. Mid-Atlantic. The observations include NARW detections at specific locations and time stamps, as well as vertical profiles of water temperature, salinity, oxygen concentration, and glider depth, sampled every 0.25 m in the vertical along the entire glider path. Figure 1 shows a spatial map of the NARW detections (by pooling data from all glider missions, resulting in a total of 104 detections). Co-located satellite information about four covariates, namely: frontal value, water mass, sea surface temperature, and chlorophyll, is extracted from multiple satellite products. Those constitute the variables in the dataset $\mathscr{D}^s$. The merged dataset, including co-located glider and satellite information, is denoted by $\mathscr{D}$. The full list of variables, along with a brief description for each, is presented in Table 1. More details about the data collection efforts are included in the "Materials and Methods" section.

Prediction results

Each glider sample is assigned a label, such that $y(\mathbf{s},t) \in \{0,1\}$ denotes whether a whale is detected at location $\mathbf{s} \in \mathbb{R}^2$ and time t. The set of co-located features is denoted by $\mathbf{x}(\mathbf{s},t) \in \mathbb{R}^p$ and includes the covariates listed in Table 1, as well as an additional categorical variable representing the season of the year in which the detection was recorded, such that $p = 9$ features. We refer to the vector formed by the values of the features and the detection label, $[\mathbf{x}(\mathbf{s},t), y(\mathbf{s},t)]^T$ as a "sample." A positive sample is one for which a whale is detected, i.e., $y(\mathbf{s},t) = 1$, whereas a negative sample characterizes whale absence, i.e., $y(\mathbf{s},t) = 0$. From there, we cast the modeling exercise as a classification task, wherein the goal is to find a mapping function (i.e., a classifier), which relates $y(\mathbf{s},t)$ to the set of features, $\mathbf{x}(\mathbf{s},t)$. In ML parlance, we seek an optimal mapping function $f(\cdot)$ such that:

$$f(\text{features}) \rightarrow \text{detection labels} \quad \text{or} \quad f(\mathbf{x}(\mathbf{s},t) \in \mathbb{R}^p) \rightarrow y(\mathbf{s},t) \in \{0,1\}. \tag{1}$$

We randomly split the merged dataset $\mathscr{D}$ into train and test subsets, and choose the latter using max-min distance sampling[23] (instead of conventional uniform sampling) to ensure that the positive samples in the test set are spatially and temporally dispersed, and that the models are evaluated on a diverse set of detections and environmental conditions. To overcome class imbalance, which is common in marine mammal surveys, the training dataset is augmented using the SMOTE method[24] to generate synthetic positive samples. Note that

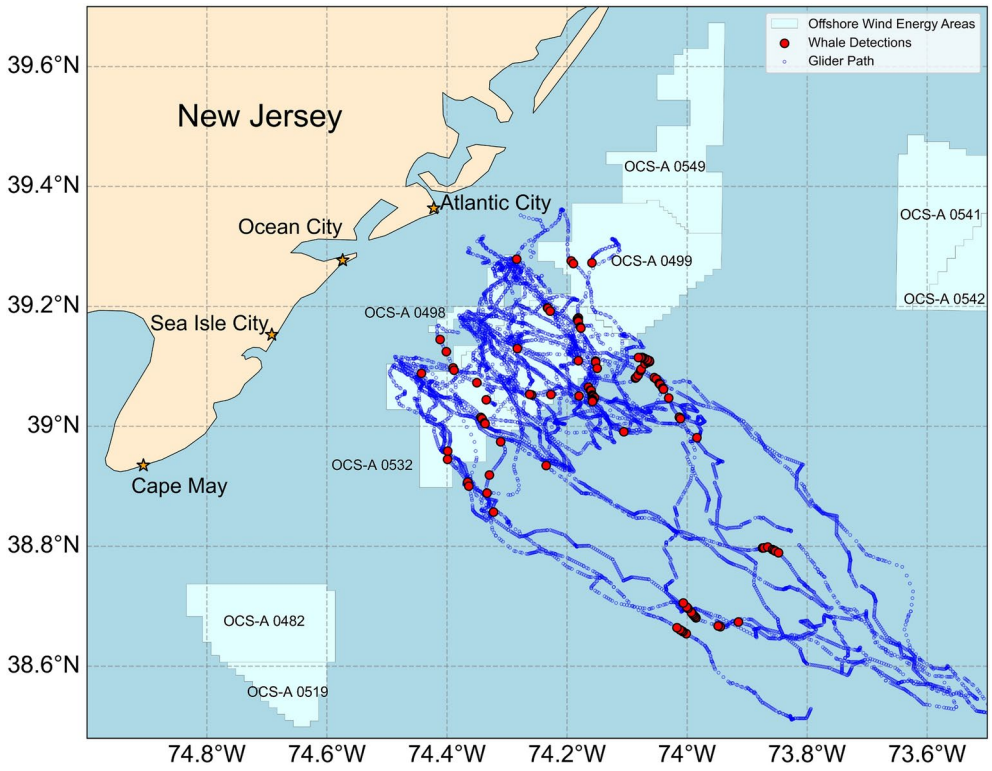

Figure 1. NARW detections (red circles) and glider paths (blue circles), on top of the offshore wind energy areas in the U.S. Mid-Atlantic (light blue polygons). The offshore wind lease areas represent their current designation as of the summer of 2024. Figure 1 is generated using the cartopy package (https://scitools.org.uk/cartopy) in Python 3.11.8.

Data Source	Variable	Description
Glider, $\mathscr{D}^g$	Temperature (°C)	Water temperature
	Salinity (psu)	Total concentration of dissolved salts
	Oxygen Concentration (*umol/L*)	Amount of dissolved oxygen in the water
	Depth (*m*)	Vertical distance from sea surface to glider's position
Satellite, $\mathscr{D}^s$	Frontal Value	Gradient strengths across water mass classifications as defined in Oliver and Irwin[22]
	Water Mass	A classification of the water column based on sea surface temperature and other hydrographic properties
	Sea Surface Temperature (°C)	Temperature of the ocean's surface layer
	Chlorophyll (mg/m^3)	Concentration of the phytoplankton pigment, chlorophyll-a

Table 1. Covariate information collected by underwater autonomous gliders (dataset $\mathscr{D}^g$) and co-located satellite imagery (dataset $\mathscr{D}^s$.).

those synthetic samples are used solely for model training, whereas the test set only contains actual samples. The whole exercise of data splitting, training, and testing is repeated 10 times using different random initiations and the overall prediction performance is reported. The combined test set formed by pooling all 10 random experiments includes a total of 8750 samples, comprising 200 positive samples (20 samples per experiment $\times 10$ experiments).

Nine prevalent classification methods are trained, namely: Logistic Regression (LR), k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Random Forest (RF), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Residual Networks (ResNet). The classifiers are chosen so as to cover a representative spectrum of ML models, from statistical to deep learning methods, and their details are deferred to the "Materials and Methods" section. We evaluate all the models under three different scenarios: (i) *Glider:* Here, we only use the covariates from the vertically resolved glider dataset $\mathscr{D}^g$ as inputs to train the ML classifiers; (ii) *Satellite:* In this scenario, we only use the covariates from the spatially resolved satellite dataset $\mathscr{D}^s$; and (iii) *Glider + Satellite:* The combined set of glider- and satellite-based covariates, $\mathscr{D}$, is used as inputs to the ML classifiers. Table 2 shows the performance of all the nine models based on three classification metrics: model accuracy (proportion of correct classifications

	Glider, $\mathscr{D}^g$			Satellite, $\mathscr{D}^s$			Glider + Satellite, $\mathscr{D}$		
	Accuracy	F1 Score	AUC	Accuracy	F1 Score	AUC	Accuracy	F1 Score	AUC
LR	0.961 ± 0.006	0.198 ± 0.068	0.625 ± 0.035	0.965 ± 0.013	0.000 ± 0.000	0.560 ± 0.036	0.972 ± 0.006	0.265 ± 0.093	0.624 ± 0.041
SVM	0.946 ± 0.004	0.216 ± 0.039	0.785 ± 0.025	0.885 ± 0.014	0.133 ± 0.020	0.733 ± 0.030	0.923 ± 0.009	0.231 ± 0.034	0.856 ± 0.027
kNN	0.950 ± 0.006	0.298 ± 0.037	0.736 ± 0.050	0.967 ± 0.005	0.431 ± 0.053	0.792 ± 0.032	0.961 ± 0.005	0.373 ± 0.039	0.791 ± 0.039
RF	0.975 ± 0.007	0.374 ± 0.071	0.778 ± 0.030	0.965 ± 0.005	0.411 ± 0.040	0.883 ± 0.017	0.975 ± 0.003	0.524 ± 0.040	0.886 ± 0.032
AdaBoost	0.972 ± 0.004	**0.377 ± 0.052**	0.815 ± 0.045	0.982 ± 0.003	0.615 ± 0.058	0.866 ± 0.028	**0.987 ± 0.002**	**0.675 ± 0.054**	**0.904 ± 0.027**
± GBoost	0.971 ± 0.005	0.359 ± 0.066	**0.823 ± 0.030**	**0.983 ± 0.003**	**0.641 ± 0.048**	**0.888 ± 0.029**	0.986 ± 0.003	0.649 ± 0.048	0.891 ± 0.028
MLP	0.969 ± 0.011	0.334 ± 0.091	0.779 ± 0.036	0.964 ± 0.006	0.358 ± 0.057	0.803 ± 0.032	0.971 ± 0.008	0.421 ± 0.080	0.863 ± 0.024
CNN	**0.977 ± 0.009**	0.344 ± 0.091	0.764 ± 0.036	0.943 ± 0.015	0.267 ± 0.064	0.801 ± 0.033	0.955 ± 0.010	0.345 ± 0.072	0.842 ± 0.025
ResNet	0.943 ± 0.002	0.223 ± 0.108	0.769 ± 0.043	0.933 ± 0.038	0.186 ± 0.065	0.735 ± 0.022	0.935 ± 0.031	0.289 ± 0.096	0.862 ± 0.031

Table 2. Average performance of various ML classifiers, along with standard deviation, across the 10 testing experiments for the Glider scenario, $\mathscr{D}^g$ (columns 2 through 4), Satellite scenario $\mathscr{D}^s$ (columns 5 through 7), and Glider + Satellite scenario, $\mathscr{D}$ (columns 8 through 10) Bold-faced values denote best performance for each metric under each scenario.

to all classifications attempted), F1 score (harmonic mean of precision and recall), and area under the curve (AUC) of the receiver operating characteristic curve. All metrics are in the [0, 1] interval, with a score of 1.0 indicating perfect performance. The formulae to calculate those metrics using model predictions are presented in the "Materials and Methods" section. The correspondent ROC curves for the Glider + Satellite scenario are shown in Fig. 2.

The results in Table 2 indicate that all methods achieve high levels of classification accuracy (ranging from 88.5% to 98.7%). This result, albeit promising, should be interpreted with caution because the test data set is highly imbalanced. In other words, a classifier that misses most of the positive samples could still achieve decent accuracy levels. Hence, metrics like the F1 score and AUC are more realistic reflections of the model's ability to distinguish positive and negative samples. From there, we find that AdaBoost, XGBoost, and RF are consistently among the best performers across all scenarios and metrics. This suggests that tree- and ensemble-based methods appear to outperform other models in predicting NARW presence. For the Glider + Satellite scenario, AdaBoost has the best classification performance, whereas the RF is the most sensitive with 118 correctly predicted whales out of 200 positive samples in the test set. In contrast, we find that simpler statistical models such as logistic regression, significantly under-perform in terms of both the F1 and AUC scores, with only 42 correctly predicted whales out of 200 positive samples in the test set. The confusion matrices of the nine methods, under all three scenarios, are shown in Figs. S1-S3 in the Supplementary Information document appended to this manuscript.

Interestingly, we also notice that relying solely on satellite-based inputs (the satellite scenario) generally yielded better results for most models compared to only using glider-based inputs (the glider scenario). The best-performing model in the satellite scenario was XGBoost, with an average accuracy of 98.3% and an F1 score of 0.641. When only trained using the glider-based covariates (the glider scenario), the same model's performance drops to 97.1% and 0.359 in terms of accuracy and F1-score respectively. This suggests that the spatially resolved satellite-based covariates may be more useful than their vertically resolved glider-based counterparts in predicting NARW presence. Fusing both glider and satellite-based information (the Glider + Satellite scenario) almost always led to an enhancement in the predictive performance. Despite the relatively weaker performance of glider variables when used independently, the within-water column observations still provide valuable information that enhances the model's performance when combined with satellite-based information. We discuss possible explanations, as well as practical implications of those two findings in the "Discussion" section. Figure 2 confirms the findings made above, wherein among the models tested, AdaBoost, XGBoost, and RF show the highest performance with AUC values of 0.90, 0.89, and 0.88, respectively, indicating satisfactory classification performance. Logistic Regression, with an AUC of 0.62, is the weakest performer, perhaps due to its over-simplification of the complex species-habitat relationships of NARWs. The ROC curves for the glider and satellite scenarios are shown in Figs. S4 and S5 in the Supplementary Information document.

Feature importance analysis

We leverage our best-performing models to identify variables that have an influential role in driving the model predictions, and hence, can be interpreted as key NARW predictors. In this analysis, we focus on the AdaBoost and RF models which are uniquely amenable to interpretation. We carry out a feature importance analysis for those two models based on two approaches—the Gini Impurity method and Shapley's Additive Explanation (SHAP) method—and contrast the findings from both approaches. More details on those two methods are included in the "Materials and Methods" section. Figure 3a,b show the feature importance using the Gini Impurity index for RF (left) and AdaBoost (right), where longer bars indicate higher feature importance. Despite some observed differences in the order and magnitude of feature importance across both models, we find that frontal value, a satellite-based covariate, consistently ranks as the most influential feature. We also find that salinity (glider), oxygen concentration (glider), and sea-surface temperature (satellite) consistently rank among the top five influential covariates. Other oceanographic and physical features such as water mass (satellite) and

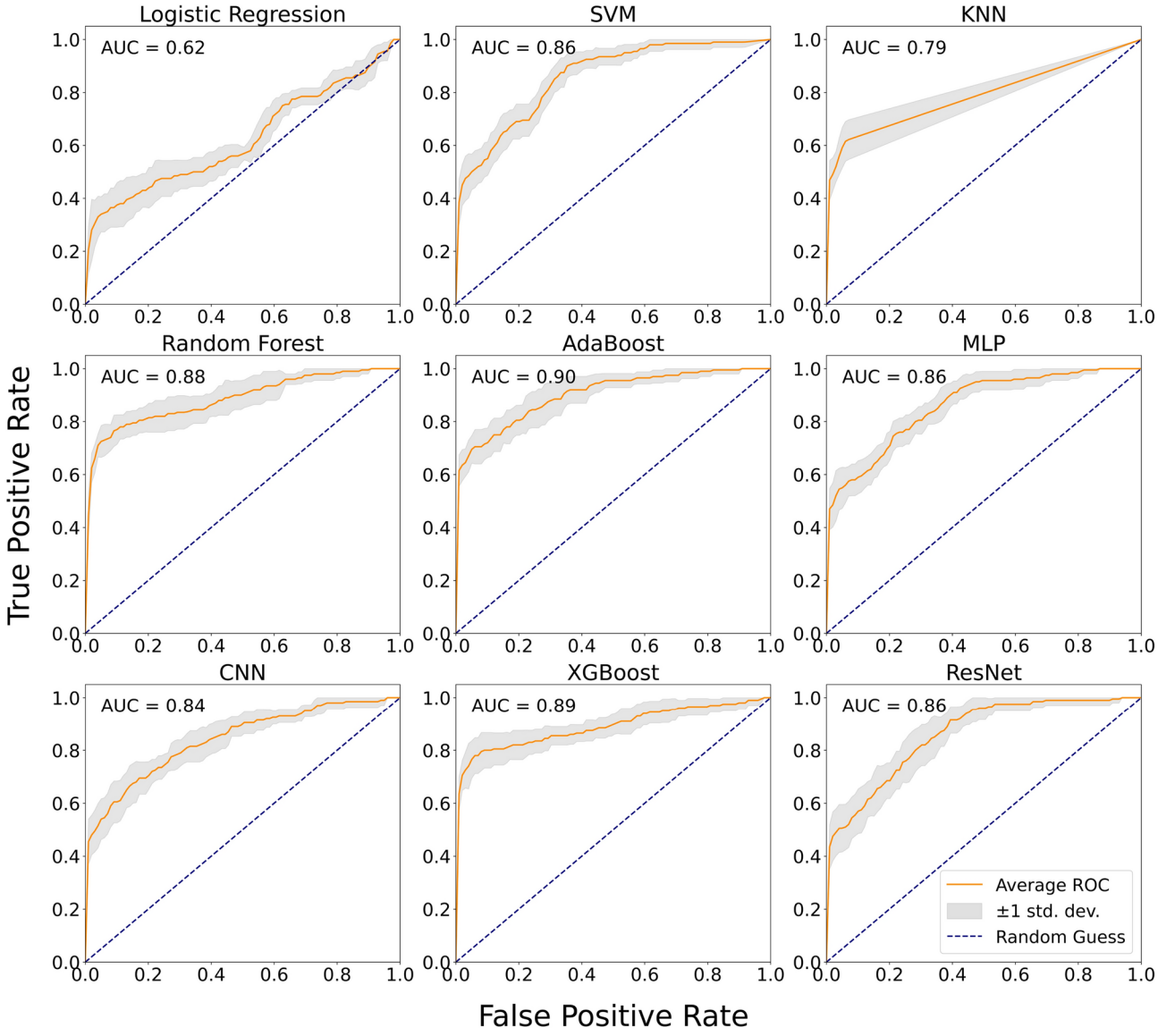

Figure 2. ROC curves of the nine classifiers for the Glider + Satellite scenario. The blue diagonal dashed line represents the performance of a random guess, whereas the orange line represents the ROC curve of the selected classifier. The shaded area around the ROC represents the standard deviation of the 10 random experiments.

chlorophyll (glider) appear to have intermediate influence, whereas glider depth appears to carry minimal importance relative to the remainder of the features.

The SHAP summary plots shown in Fig. 3c,d for RF (left) and AdaBoost (right) provide additional information. Here, a positive SHAP value (x-axis) means an increase in the likelihood of the model yielding a positive class prediction (i.e., NARW presence), and vice-versa. Consistent with the Gini impurity analysis, frontal value (satellite) stands out as the most significant feature for both RF (left) and AdaBoost (right). Smaller frontal values (blue-colored points) appear to coincide with negative SHAP values, whereas most of the red-colored points (higher frontal values) correspond to positive SHAP values. This means that the model suggests that NARWs generally show a preference for higher frontal values. In contrast, salinity (glider) appears to have an opposite effect, with the model predicting NARWs to prefer environments with lower salinity levels. Our survey of the literature reveals little in-depth investigation on the influence of oceanic fronts on NARW habitats. Yet, frontal information has been mentioned in some prior studies as a proxy for oceanographic processes that could influence prey distribution[18,25,26]. Our ML-learnt results support those findings suggesting that locations and times experiencing stronger fonts may be associated with increased foraging opportunities, and hence higher likelihood of NARW presence. Meanwhile, chlorophyll, salinity, and sea-surface temperature have been consistently invoked as covariates in NARW habitat models[7,13,27].

The seasonal covariate appears to have intermediate predictive power throughout the feature importance analysis. It is regarded as an influential feature in Fig. 3b (Gini Impurity analysis) but not as powerful in the

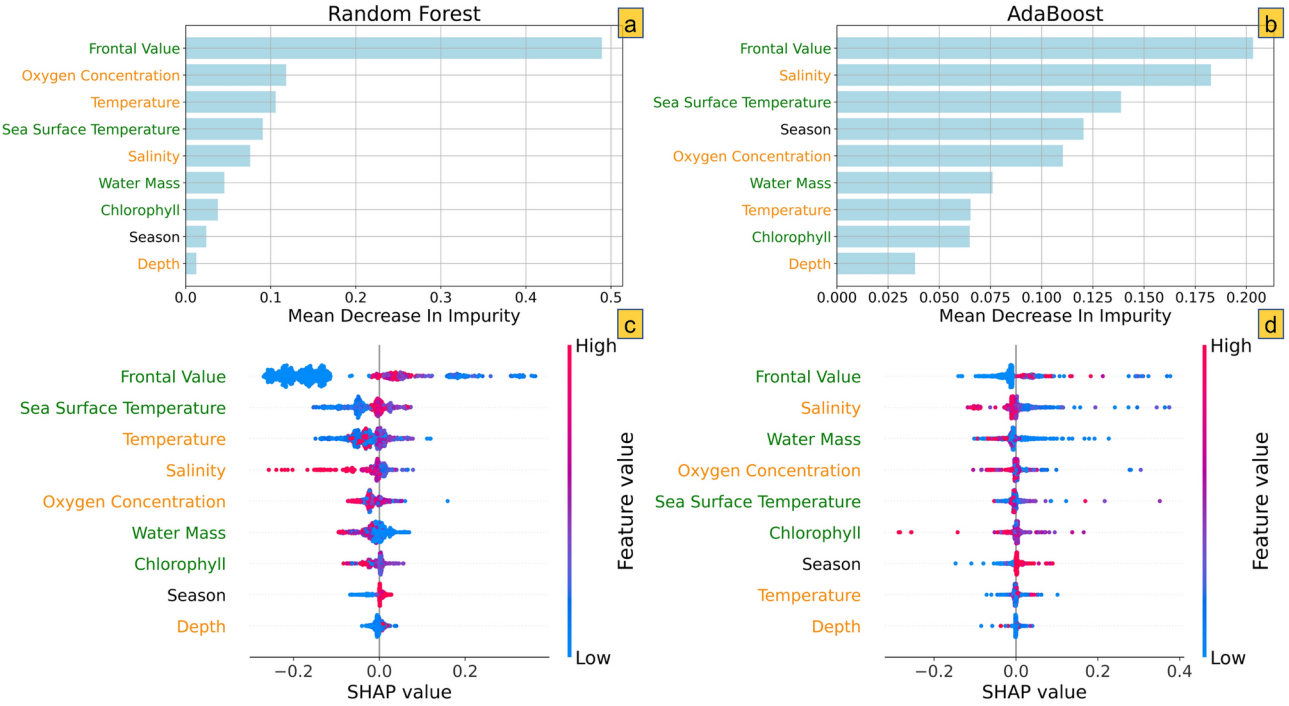

Figure 3. Feature importance for Random Forest (left) and AdaBoost (right). Panels (**a**,**b**) are based on Gini Impurity, whereas Panels (**c**,**d**) are based on the SHAP method. Features are in descending order of importance. Glider- and satellite-based covariates are color-coded in orange, and green, respectively.

SHAP analysis (Fig. 3d). Despite this, we conjecture that seasonality plays an influential role in driving ML predictions. This can be explained in light of the uneven distribution of NARW detections across the four seasons, which is likely attributed to the seasonal migratory patterns of NARWs. Specifically, the winter season had the highest number of NARW detections, followed by Fall, Spring, and then Summer, which had no detections (See Table S1 in the supplementary Information material). This aligns to some extent with prior survey efforts in this region[28]. Disaggregating the best-performing ML model's predictions by season shows significantly higher predictive performance in the Fall and Winter seasons, followed at some distance by the Spring season. The model consistently (and correctly) predicted no whales in the summer season. The season-specific ROC curves are shown in Fig. S6 of the Supplementary Information document. Given the significant oceanographic variability between seasons in the study site, this suggests that the ML models more effectively learn the oceanographic features describing NARW presence during seasons when whales are more available for detection. Additional subset selection experiments, presented in Fig. S7 of the Supplementary Information document, show that dropping the seasonal covariate from the best-performing ML model (while assuming all other covariates are fixed) results in a noticeable reduction in predictive performance, further confirming the importance of considering seasonality effects when modeling NARW habitat.

Discussion

In this study, we have developed a cluster of ML models to jointly leverage, for the first time, vertically resolved glider- and spatially resolved satellite-based information for modeling the presence of NARWs near the offshore wind energy areas in the U.S. Mid-Atlantic. We find that tree- and ensemble-based models, like random forests and AdaBoost, achieve the highest levels of accuracy, while maintaining a sensible balance of false and missed alarms. Neighborhood-based methods like kNN and SVM, as well as deep-learning-based models (e.g., CNN, ResNet, MLP) appear to yield intermediate predictive performance, whereas simpler models like logistic regression fall short, especially when it comes to correctly predicting positive samples. When using both glider- and satellite-based covariates, AdaBoost is found to be the best-performing model in terms of predictive performance, whereas RF is the most sensitive with 118 correctly predicted whales out of 200 positive samples. While this corresponds to a true positive rate of 59.0%, we are inclined to look at this result favorably, especially considering the absence of an established ML baseline for this problem, and the difficulty of predicting NARWs at such high granularity with fairly limited data. We consider this work as a stepping stone for subsequent works in which ML engineers and practitioners can work hand-in-hand with ecologists and oceanographers to develop more advanced ML approaches that may achieve even greater heights compared to the one proposed in this work. To enable this and accelerate the research on this timely topic, we have made our data and codes freely available at https://github.com/Jiaxiang-J/NARW.

Although we can confidently claim that tree-based models appear to be the best-performers for predicting NARW presence as evidenced by our quantitative experiments, this finding may be partly due to the way we

decided to cast the modeling problem as a classification task. To be more specific, our findings align with the mounting evidence in the ML literature suggesting that tree- and ensemble-based models often perform very well in learning from low-dimensional, tabular datasets, not to mention their scalability and interpretable nature. However, our raw data are by no means low-dimensional. Raw satellite information comes in the form of evolving image-based inputs, whereas raw glider inputs are high-dimensional spatial-temporal vertical profile data. Our choice to select single points in space and time "compresses" the problem in a low-dimensional manifold, merely for simplification. We expect deep-learning-based approaches to be strong contenders if richer information from satellites and/or gliders are used, but this comes with significant challenges related to lack of data coverage (especially in satellite imagery), scalability in terms of computational requirements needed to train the models and make timely predictions, and most importantly, interpretability, which is a crucial aspect in ecological applications. This is an ongoing area of research.

Interestingly, our results suggest that spatially resolved satellite-based covariates play a greater role than vertically resolved glider-based information in predicting NARW presence. This is evidenced by how our ML models perform significantly better under the satellite scenario versus the glider scenario (recall the results in Table 2). Two likely explanations are possible. First, satellite-based covariates appear to provide unique and rich information about spatial variability in oceanographic processes as proxies for both NARW behavior and its prey concentration. Second, this can be also linked to the surface feeding behavior of NARWs, which may be better captured using satellite surface measurements. From a practical viewpoint, this is a very promising result, since a wind farm developer or operator can run and update the predictive models using publicly available satellite information, without the need for continuous glider monitoring of their study region, which is both impractical and expensive. In that case, the decision-maker, be it the farm developer or operator, would only need to deploy initial glider campaigns for collecting historical detection data. Once this is complete and enough data has been gathered, the ML models can be trained and run in real-time solely off of satellite imagery inputs. Nevertheless, we find that using the combination of glider- and satellite-based information almost always yields the best performance. Albeit not highly useful on their own, glider-based covariates appear to complement the information provided by satellite-based inputs. This synergy between highly localized information sampled by gliders and macro-scale information obtained from satellite imagery appears to furnish two complementary "world views" that maximize the ability of ML models to capture the complex, multi-scale species-habitat patterns of NARWs, leading to improved predictions and inference, and further underscoring the importance of a holistic, multi-modal approach to NARW predictive modeling.

The utility of our proposed ML approach stems from its potential value to farm-level planning and decision-making. Predicting NARW presence at finer spatial and temporal resolutions—unlocked through an ML approach —can enable operators and developers of future offshore wind farms to make responsible and effective decisions that are highly localized in space and time. As a case in point, ML-enabled granular predictions of NARW encounter risk can be used by offshore wind developers during construction operations to cease or delay high-intensity activities such as pile driving. Another relevant application is to inform dynamic and highly localized vessel collision risk models for managing the projected increase in vessel traffic needed to reliably service offshore wind farms. In the long-run, this could have considerable economic and environmental implications for offshore wind farm construction, operations, and maintenance planning[29,30]. Such fine resolutions may be difficult to attain using coarser-scale species distribution models. Beyond their practical utility, our ML models can offer timely insights about the key drivers of NARW presence. For example, interpretation of our best-performing models suggest that whales tend to be located in waters with higher frontal values and lower salinity levels.

The results presented in this work thus far utilize the most recently available covariate information in order to predict NARW presence. However, if ML-based models are able to "forecast" the likelihood of whale presence for future horizons, then such look-ahead predictions would be highly valuable to inform operational decision-making by offshore wind stakeholders and other ocean users. To investigate this, we carry out an additional analysis where we use lagged (instead of current) covariate information and re-evaluate the predictive performance of the AdaBoost model. The results, shown in Table S3 of the Supplementary Information document, suggest that decent levels of predictive power are still maintained up to 48 hours ahead. Understandably, the predictive performance degrades gradually as the forecast horizon extends (up to 11.56% reduction in F-1 scores for 48-hour ahead versus same-day prediction). Although this result attests to the potential merit of ML in look-ahead prediction, two key re-modeling efforts are needed in order to develop a full-fledged ML-based NARW forecasting tool. First, an embedded forecasting system is needed for the oceanographic covariates. This entails ensuring that the propagation of forecast errors in those covariates does not severely affect the downstream NARW prediction task. Second, there is a need to redesign the experimental ML training and evaluation setups. In a forecasting setup, the training set would only consist of information that occurred prior to the glider observations. Thus, data from gliders deployed in later years (e.g., 2022) cannot be used to make predictions of whale presence in earlier years (e.g., 2021). We believe that this is an important future research direction, since such forecasting tool, if effective, can be of immense value to offshore wind planning and operations.

Despite the value of this study, we would also like to point out the limitations of our models and conclusions. The whale presence data collected by gliders are samples of where whales have been detected during select times and study regions, and may invoke elements of sampling, spatial, and temporal biases. Even with their immense potential, ML models are fairly less developed compared to traditional statistical models when it comes to formally accounting for sampling biases. In glider-based surveys, one such bias is related to detection availability—gliders only detect whale presence when whales vocalize, and hence, a negative detection is not equivalent to whale absence. Although recent validation studies have shown that Slocum gliders have fairly low missed occurrence rates[31], those biases can potentially impact the ML model estimates of false and true negatives, but would have minimal influence on true and false positives. We therefore envision this work to be

a stepping stone for subsequent efforts to design ecologically-informed ML models that can recognize the bias types that have long been acknowledged in the ecological modeling literature. For example, ML approaches can be embedded within classical spatial models to carry out specific learning tasks on complex datasets that are difficult to process or "mine" using traditional statistical methods. Fusing those two distinct modeling paradigms in some form of an ensemble prediction is also an interesting avenue to pursue.

This work can be extended in several ways. The first direction is to use richer and larger information inputs. For example, satellite maps can be used instead of being confined to select single gridded values. Similarly, spatio-temporal glider profiles can be leveraged instead of single points in space and time to account for derived glider gradients along the profile, including mixed layer presence, and depth. Importantly, many of the covariates included in the model, such as frontal values, serve as proxies for prey concentration, indirectly informing NARW preferences. Thus, incorporating relevant prey covariates to inform the ML-based models is an important research direction to pursue. For example, recent studies have explored the utility of satellite-based visual spectra to detect copepod density, which could offer direct insights into prey availability[32,33]. Collectively, augmenting the input data to our model brings about two advantages: First, it can provide richer information about the broader spatial and temporal variability in oceanic features that are relevant to predict NARW presence. Second, it may provide a more robust approach against abnormal values that may mislead ML model training. However, a major challenge with satellite images in particular is the weak spatio-temporal coverage, requiring advanced image processing and raising the question of the optimal window size to "zoom" in as input to the ML models. Smaller windows would most likely be relevant to a particular NARW detection, but run into the risk of weaker coverage, and vice-versa. Finally, another interesting area of research pertains to the use of ML models to guide future glider deployments. It is clear that our models would highly benefit from more positive samples due to the highly imbalanced nature of the problem. Leveraging the ML models to decide on where and when to deploy gliders in order to balance exploration and exploitation efforts is a relevant question to pursue.

Materials and methods

Data description and preparation

The glider dataset, $\mathscr{D}^g$, used in this research comes from nine glider deployments[34], spanning from August 2020 to June 2022, off the south coast of New Jersey between $38.5°N$ to $39.5°N$ and $74.5°W$ to $73.5°W$. The NARW detections were based on an autonomous glider mounted hydrophone. The WHOI-developed digital acoustic monitoring (DMON) instrument running the low-frequency detection and classification system[35] was integrated into the Slocum glider a decade ago[36] and has been used on over 90 glider missions in U.S., Canadian, and Chilean waters to date. The accuracy of the DMON for baleen whale detection from gliders and buoys is well characterized[37,38]. Each DMON-equipped glider mission monitored multiple marine mammal species in near real time, including fin, sei, humpback, blue and NARW. During a mission, real-time detections are telemetered back to shore for verification by a trained analyst. The detections relayed only the position of the platform when a sound is detected, not the position of the sound source[31]. Thus, whale detection location was based on the glider position at the time of detection. A total of 104 NARW recordings, which are shown in Fig. 1, have been detected. Table 3 shows the dates of each trip, together with the number of detections in each. Only five deployments contain confirmed NARW detections. For each mission, the glider simultaneously collected profiles of oceanographic variables at 0.25m vertical resolution, including water temperature, oxygen concentration, salinity, and glider depth. Figure 4b shows an example of the profiles generated by the gliders at different depths and time stamps.

We leverage the spatial and temporal information from the glider data to extract co-located information from satellite imagery. Satellite data used in this research come from multiple sources, and include information about key oceanic variables, including sea surface temperature (SST) and chlorophyll. Derived satellite products based on spatial observations of SST and ocean color include water mass type and frontal value[22,39]. The frontal value reflects the magnitude of difference between adjacent water masses in SST and ocean color variable space. Due to cloud effects, the satellite imagery may often have weak spatial coverage. To address the data coverage issues, we combine multiple satellite source data to have better coverage. The three satellite sources we use are "NOAA/NESDIS/STAR GHRSST GOES16 SST Daily Composite SST" (GOES)[40], "VIIRS Suomi NPP 1-Day 750 m Composite Northwest Atlantic" (VIIRS)[41], and "MODIS Aqua 3-Day 1 km Composite Northwest Atlantic"

Trips	Number of detections
2020-07-29 to 2020-08-26	0
2020-10-03 to 2020-11-05	0
2020-11-19 to 2020-12-21	47
2021-02-08 to 2021-03-08	21
2021-11-20 to 2021-12-17	3
2022-01-13 to 2022-02-11	17
2022-02-15 to 2022-03-16	16
2022-03-30 to 2022-04-22	0
2022-05-20 to 2022-06-10	0
Total	104

Table 3. Glider mission information and number of detections per mission.

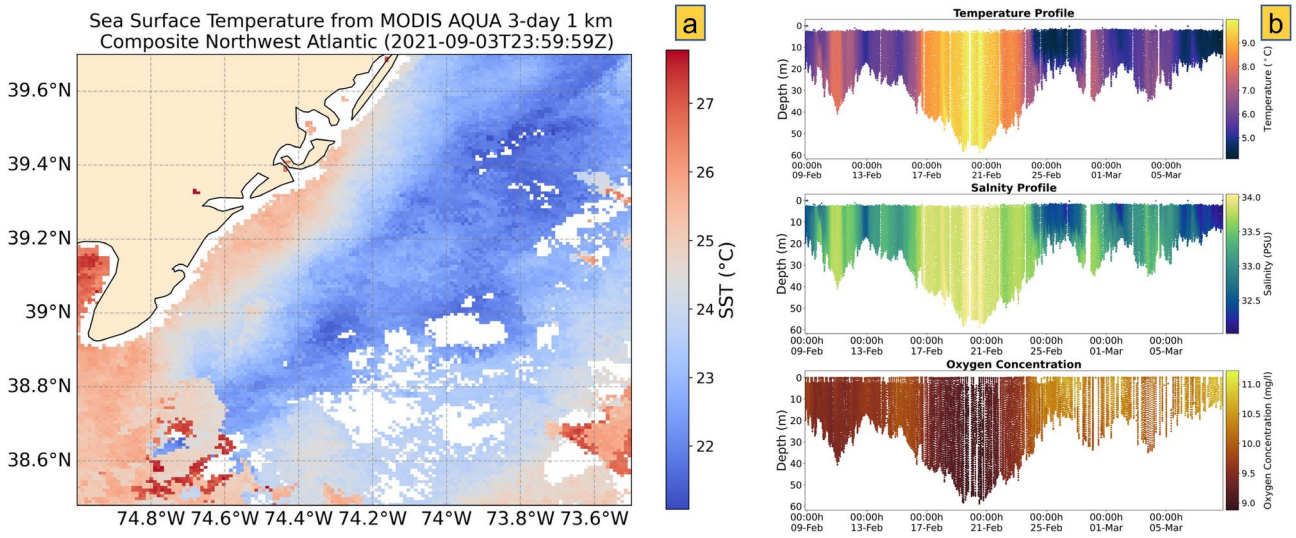

Figure 4. Sample visualization of satellite and glider data. Panel (**a**) presents sea surface temperature measured by MODIS at 2021-09-03. Panel (**b**) presents glider variables collected from 08 Feb 2021 to 10 Mar 2021. (**a**) is generated using the cartopy package (https://scitools.org.uk/cartopy) in Python 3.11.8. (**b**) is generated using functionalities in the cmocean package (https://matplotlib.org/cmocean/) in Python 3.11.8.

(MODIS)[42]. Since GOES data is likely to be less affected by clouds due to its geostationary nature, it is almost always our first choice. If GOES data at a particular location and time is unavailable, we turn to data from VIIRS as it offers higher spatial resolution compared to MODIS. MODIS data is the last resort for sea surface temperature/chlorophyll, as it has lower spatial resolution compared to GOES and VIIRS. Since water mass and frontal value are only available from MODIS, we solely rely on MODIS data for these variables. Figure 4a shows an example of sea surface temperature from MODIS[42].

Even when using multiple satellite data sources, the problem of data missingness is not fully resolved. To address this issue, the kriging method is used to interpolate data for continuous variables[43], whereas random forests are used for categorical variables[44]. Once the dataset is interpolated and merged, we divide it into training and testing sets using an 0.8:0.2 ratio, which is a common choice in ML practice[44]. However, we enforce our training data to include 80% of the NARW detections. Furthermore, observing that whale detections tend to cluster, we use the max-min distance sampling (instead of conventional uniform sampling) to ensure that the detections in the test set are spatially and temporally dispersed and that the models are tested on a diverse set of environmental conditions and detections. Due to the severe imbalance in our datasets, the Synthetic Minority Over-sampling Technique (SMOTE) is utilized to achieve a more balanced class distribution, and we set the target majority-to-minority ratio to 0.45[24]. Using the SMOTE algorithm, new synthetic samples are generated by learning features of the actual minority samples (here, the positive samples that constitute whale presence). These synthetic samples retained the characteristic distribution of the actual positive samples data to a large extent but also had a certain degree of diversity typical of real-world datasets. Figures S8 and S9 in the Supplementary Information document present a comparative analysis of the actual versus synthetic datasets confirming the effectiveness of SMOTE in reproducing the key features of actual positive samples. Consequently, the inflated training dataset consists of 4959 samples, comprising 3420 negative samples and 1539 positive samples. This whole exercise is repeated 10 times (using different random initializations) and the overall predictive performance, across the 10 random experiments, is reported. Note that the SMOTE-generated synthetic samples are used for model training only. Test sets only contained actual samples and did not contain any synthetic samples. In that way, the combined set formed by pooling all 10 test sets had a total of 8550 negative samples, and 200 positive samples (20 samples per experiment $\times 10$ experiments).

Classification methods

A total of nine prevalent ML models have been trained in this work. All data preprocessing and ML models were implemented in Python. Specifically, scikit-learn was used for implementing most classification models, and imbalanced-learn[45] for SMOTE implementation. Deep learning models were implemented in TensorFlow and PyTorch, whereas feature importance analysis was performed using the SHAP package[46]. A brief description of those methods is provided below.

Statistical and neighborhood-based approaches

LR serves as a baseline model in our analysis, providing a probabilistic approach for binary classification tasks based on a parametrized logistic function. kNN makes predictions using consensus voting of the k nearest neighbors to the target prediction location. Here, we find that $k = 5$ yields the best performance. SVM is a kernel-based separating hyperplane approach which finds the optimal decision boundary that separates two

(or more) classes. For SVM, we use the RBF kernel, which is a popular choice in ML, and set the cost penalty parameter to 5.00.

Tree- and ensemble-based methods
RF is an ensemble tree-based approach that combines multiple decision trees to make a consensus-based prediction. For RF, we use a total of 100 decision trees with a maximum depth of 5. AdaBoost is an ensemble tree-based approach that sequentially adjusts the weights of incorrectly classified instances, thus focusing more on the difficult cases in subsequent trees. We use a total of 100 decision trees when training AdaBoost. XGBoost is a boosting approach that combines multiple weak learners (here, decision trees) for improved classification performance[47]. For XGBoost, the number of boosting rounds is set to 100, the maximum depth of the tree is 6, and the learning rate is set at 0.3.

Deep learning methods
MLP is a (typically shallow) artificial neural network architecture. Here, we use two hidden layers consisting of 128/64 neurons with ReLU activation, one dropout layer in between with a dropout rate of 0.5, and a dense layer of 1 neuron with sigmoid activation to produce the final output. CNN is a deep learning architecture which utilizes convolutional layers to automatically learn spatial (or cross-correlation) hierarchies within the input space. For CNN, we use a convolution layer with 64 filters and the kernel size is set to 3. After max pooling and flattening, we use a dense layer with 50/1 neurons to obtain the output. ResNet is a deep learning approach that is typically designed for sequential data and is characterized by skip connections that help overcome the vanishing gradient problem. Herein, we adopt a version of ResNet that is adapted for tabular data[48]. For ResNet, use 5 epochs with batch size 50 and the learning rate is set at 0.001.

Evaluation metrics

To comprehensively evaluate the performance of our models, we employed several metrics: Accuracy, F1 score, and AUC scores. For a binary classification problem, four components affect those metrics: True positives (TP) which denote the correctly predicted positive observations (i.e., NARW presence); True negatives (TN) denoting correctly predicted negative observations (i.e., NARW absence); False positives (FP) indicating incorrectly predicted positive observations (i.e., false alarms); and false negatives (FNs) denoting incorrectly predicted negative observations (i.e., missed alarms). Accuracy is the proportion of correctly predicted test samples out of the total test samples, and is defined as $\frac{TP+TN}{TP+TN+FP+FN}$. For imbalanced datasets, accuracy as a single metric may be misleading, and hence, additional metrics such as F1 scores and AUC can provide important indications about the model performance. F1 score is the harmonic mean of precision and recall, and is defined as $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, such that precision is defined as the fraction of TP to all predicted positive samples, i.e., $\frac{TP}{TP+FP}$, whereas recall is defined as the fraction of TP to all positive samples, i.e., $\frac{TP}{TP+FN}$. Thus, the F1 score provides a balance between precision and recall. ROC curves plot the so-called false positive rate (FPR) against the true positive rate (TPR) at various classification threshold values. TPR is the same as recall (or sensitivity), whereas TNR defined as $\frac{TN}{TN+FP}$, also known as specificity, measures the proportion of actual negatives that are correctly identified by the model. The AUC provides a single measure of overall performance based on ROC curves, with values ranging from 0 to 1 and a higher AUC indicates a better-performing model.

Feature importance metrics

Two feature importance measures are used in this work, namely: the Gini index and SHAP values in our analysis. For a given dataset S and classes $C_1, C_2, \cdots, C_k$, the Gini index is defined as $Gini(S) = 1 - \sum_{i=1}^{k} p_i^2$, where p_i is the proportion of class C_i in the dataset S. For tree-based models, each node in the tree splits the data into two subsets, S_1 and S_2, based on certain values for a feature. Then, the correspondent Gini index for this split is the weighted average of the Gini indices of these subsets, $Gini_{split}(S) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$. A higher reduction of impurity means that the feature can split the node better, thus indicating higher feature importance. SHAP (Shapley Additive exPlanations), on the other hand, quantifies the contribution of each feature in a predictive model to the overall prediction[49]. It is rooted in cooperative game theory, where it is used to determine the fair distribution of payoffs among players. Each feature of the dataset is treated as a "player" and the "payoff" is the improvement in the prediction error contributed by including the feature in the model. The Shapley value for feature i in a model f is defined as $\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!} \left(f(S \cup \{i\}) - f(S)\right)$, where N is the set of all features. S is a subset of features excluding i. $f(S)$ is the prediction model evaluated with features in S. The formula calculates the average marginal contribution of feature i across all possible combinations of other features in the model. This value helps in understanding how the presence or absence of a feature affects the prediction outcome.

Data availability

Data and codes to reproduce the results in this research are freely available at https://github.com/Jiaxiang-J/NARW.

Received: 5 August 2024; Accepted: 14 November 2024
Published online: 25 November 2024

References

1. Office of the Press Secretary. Fact sheet: Biden administration jumpstarts offshore wind energy projects to create jobs (2021). https://www.whitehouse.gov/briefing-room/statements-releases.
2. Bureau of Ocean Energy Management. Lease and Grant Information (2024). https://www.boem.gov/renewable-energy/lease-and-grant-information.
3. U.S. Department of Energy. DOE releases strategy to accelerate and expand domestic offshore wind deployment (2023). https://www.energy.gov/articles/doe-releases-strategy-accelerate-and-expand-domestic-offshore-wind-deployment.
4. Pettis, H., Pace III, R. & Hamilton, P. North Atlantic Right Whale consortium 2020 annual report card. Tech. Rep., NOAA Washington, DC, USA (2021).
5. New York Energy Research & Development Authority. NYSERDA Master Plan 2.0 (2023). https://www.nyserda.ny.gov/All-Programs/Offshore-Wind/About-Offshore-Wind/Master-Plan.
6. Best, B. D. et al. Online cetacean habitat modeling system for the US east coast and Gulf of Mexico. *Endangered Spec. Res.* **18**, 1–15 (2012).
7. Pendleton, D. E. et al. Weekly predictions of North Atlantic right whale Eubalaena glacialis habitat reveal influence of prey abundance and seasonality of habitat preferences. *Endangered Spec. Res.* **18**, 147–161 (2012).
8. Moses, E. & Finn, J. T. Using geographic information systems to predict North Atlantic right whale (Eubalaena glacialis) habitat. *J. Northwest Atl. Fish. Sci.* **22**, 37–46 (1997).
9. Monsarrat, S. et al. A spatially explicit estimate of the prewhaling abundance of the endangered North Atlantic right whale. *Conserv. Biol.* **30**, 783–791 (2016).
10. Miller, D. L., Burt, M. L., Rexstad, E. A. & Thomas, L. Spatial models for distance sampling data: Recent developments and future directions. *Methods Ecol. Evol.* **4**, 1001–1010 (2013).
11. Hedley, S. L. & Buckland, S. T. Spatial models for line transect sampling. *J. Agric. Biol. Environ. Stat.* **9**, 181–199 (2004).
12. Roberts, J. J. et al. Habitat-based cetacean density models for the US Atlantic and Gulf of Mexico. *Sci. Rep.* **6**, 22615 (2016).
13. Roberts, J. J. et al. North Atlantic right whale density surface model for the US Atlantic evaluated with passive acoustic monitoring. *Mar. Ecol. Prog. Ser.* **732**, 167–192 (2024).
14. Davis, G., Tennant, S. & Van Parijs, S. Upcalling behaviour and patterns in North Atlantic right whales, implications for monitoring protocols during wind energy development. *ICES J. Mar. Sci.* fsad174 (2023).
15. Department of the Navy. Marine species monitoring for the U.S. Navy's Atlantic fleet training and testing (AFTT) – 2022 annual report. Annual Report, U.S. Fleet Forces Command, Norfolk, Virginia (2023).
16. Fucile, P. D., Singer, R. C., Baumgartner, M. & Ball, K. A self contained recorder for acoustic observations from AUV's. In *OCEANS 2006*, 1–4 (IEEE, 2006).
17. Schofield, O. et al. Slocum gliders: Robust and ready. *J. Field Robot.* **24**, 473–485 (2007).
18. Dreyfust, C. *et al.* Aligning the seasonal migration of North Atlantic right whales with oceanic features. In *OCEANS 2022, Hampton Roads*, 1–9 (IEEE, 2022).
19. Rubbens, P. et al. Machine learning in marine ecology: An overview of techniques and applications. *ICES J. Mar. Sci.* **80**, 1829–1853 (2023).
20. Shiu, Y. et al. Deep neural networks for automated detection of marine mammal species. *Sci. Rep.* **10**, 607 (2020).
21. Bach, N. H., Vu, L. H., Nguyen, V. D. & Pham, D. P. Classifying marine mammals signal using cubic splines interpolation combining with triple loss variational auto-encoder. *Sci. Rep.* **13**, 19984 (2023).
22. Oliver, M. J. & Irwin, A. J. Objective global ocean biogeographic provinces. *Geophys. Res. Lett.***35** (2008).
23. Gramacy, R. B. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences* (Chapman and Hall/CRC, 2020).
24. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
25. Baumgartner, M. F., Cole, T. V., Clapham, P. J. & Mate, B. R. North atlantic right whale habitat in the lower bay of fundy and on the SW Scotian shelf during 1999–2001. *Mar. Ecol. Prog. Ser.* **264**, 137–154 (2003).
26. Sorochan, K., Plourde, S., Baumgartner, M. & Johnson, C. Availability, supply, and aggregation of prey (Calanus spp.) in foraging areas of the north atlantic right whale (eubalaena glacialis). *ICES J. Mar. Sci.* **78**, 3498–3520 (2021).
27. Garrison, C. A. K. L., Rene Baumstark, L.I.W.-G. & Hines, E. Application of a habitat model to define calving habitat of the north atlantic right whale in the southeastern united states. *Endangered Spec. Res.* **18**, 73–87 (2012).
28. Whitt, A. D., Dudzinski, K. & Laliberté, J. R. North atlantic right whale distribution and seasonal occurrence in nearshore waters off New Jersey, USA, and implications for management. *Endangered Spec. Res.* **20**, 59–69 (2013).
29. Papadopoulos, P., Coit, D. W. & Ezzat, A. A. Seizing opportunity: Maintenance optimization in offshore wind farms considering accessibility, production, and crew dispatch. *IEEE Trans. Sustain. Energy* **13**, 111–121 (2021).
30. Silber, G., Dangerfield, A., Smith, J., Reeb, D. & Levenson, J. Offshore wind energy development and north atlantic right whales. *Sterling (VA): US Department of the Interior, Bureau of Ocean Energy Management* (2023).
31. Johnson, H. D., Taggart, C. T., Newhall, A. E., Lin, Y.-T. & Baumgartner, M. F. Acoustic detection range of right whale upcalls identified in near-real time from a moored buoy and a slocum glider. *J. Acoust. Soc. Am.* **151**, 2558–2575 (2022).
32. McCarry, C. L., Basedow, S. L., Davies, E. J. & McKee, D. Estimating surface concentrations of calanus finmarchicus using standardised satellite-derived enhanced RGB imagery. *Remote Sens.* **15**, 2987 (2023).
33. Basedow, S. L. et al. Remote sensing of zooplankton swarms. *Sci. Rep.* **9**, 686 (2019).
34. ERDDAP. Glider delayed science profile. http://slocum-data.marine.rutgers.edu/erddap/tabledap. (Accessed: 2024-04-24).
35. Baumgartner, M. F. & Mussoline, S. E. A generalized baleen whale call detection and classification system. *J. Acoust. Soc. Am.* **129**, 2889–2902 (2011).
36. Baumgartner, M. F. et al. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. *J. Acoust. Soc. Am.* **134**, 1814–1823 (2013).
37. Baumgartner, M. F. et al. Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: System description and evaluation. *Methods Ecol. Evol.* **10**, 1476–1489 (2019).
38. Baumgartner, M. F. et al. Slocum gliders provide accurate near real-time estimates of baleen whale presence from human-reviewed passive acoustic detection information. *Front. Mar. Sci.* **7**, 100 (2020).
39. Oliver, M. J. *et al.* Bioinformatic approaches for objective detection of water masses on continental shelves. *J. Geophys. Res.: Oceans***109** (2004).
40. NOAA/NESDIS/STAR. NOAA/NESDIS/STAR GHRSST GOES16 SST Daily Composite. Available online at http://basin.ceoe.udel.edu/erddap/griddap/daily_composite_JPL_SST.html (n.d). (Accessed: 2024-04-24).
41. ERDDAP. VIIRS Suomi NPP 1-Day 750 m Composite Northwest Atlantic. Available online at http://basin.ceoe.udel.edu/erddap/griddap/VIIRS_NWATL.html (n.d). (Accessed: 2024-04-24).
42. ERDDAP. MODIS Aqua 3-Day 1 km Composite Northwest Atlantic. Available online at http://basin.ceoe.udel.edu/erddap/griddap/MODIS_AQUA_3_day.html (n.d). (Accessed: 2024-04-24).
43. Williams, C. K. & Rasmussen, C. E. *Gaussian Processes for Machine Learning* Vol. 2 (MIT press Cambridge, 2006).
44. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Vol. 2 (Springer, 2009).

45. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
46. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30, 4765–4774 (Curran Associates, Inc., 2017).
47. 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*, 785–794 (2016).
48. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural. Inf. Process. Syst.* **34**, 18932–18943 (2021).
49. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process Syst.***30** (2017).

Author contributions

J.J. conducted the experiments, analyzed the results, and contributed to the conceptualization of the methodology, as well as reporting the outcomes of the study. J.R. and L. Nazzaro helped with the data extraction, processing, and interpretation. J.K. and A.A.E. contributed to the conceptualization of the methodology, interpretation and analysis of results, as well as acquiring funding needed for this study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80084-z.

Correspondence and requests for materials should be addressed to A.A.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.