Journal of Computer Science

Research Article

Machine Learning Based Prediction of Catch Yields in Small Scale Fisheries

Epifelward Niño Olaivar Amora[1,2] and Patrick D. Cerna[1]

[1]Graduate School, State University of Northern Negros, Sagay City, Philippines
[2]College of Sciences, Bohol Island State University Candijay Campus, Philippines

Article history
Received: 11-09-2025
Revised: 03-06-2026
Accepted: 15-07-2026

Corresponding Author:
Epifelward Niño Olaivar Amora
Graduate School, State University of Northern Negros, Sagay City, Philippines
Email: ninyohamora2017@gmail.com

Abstract: Small-scale fisheries are critical for food security and livelihoods but increasingly face threats from overfishing and environmental change. This study employed machine learning to predict catch yields in Cogtong Bay, Philippines, using 3,478 records collected in 2020–2021 that included species, gear, location, season, and weather data. Models developed in Python with scikit-learn and XGBoost compared Random Forest (RF), Extreme Gradient Boosting, support vector machines, and logistic regression. Trips were classified as high- or low-yield based on total catch weight. Results indicate that RF achieved the highest performance (≈90% accuracy), correctly identifying more than 90% of high-yield trips and outperforming other models. Feature importance analysis revealed gear type and season as the strongest predictors, with location and weather exerting secondary influence. These findings demonstrate that ensemble machine learning models can capture complex fisheries dynamics and provide reliable decision support. The approach offers a scalable tool for adaptive, data-driven fisheries management and more sustainable, resilient coastal livelihoods.

Keywords: Small-Scale Fisheries, Catch Yield Prediction, Machine Learning, Random Forest, XGBoost, Cogtong Bay, Fisheries Management, Sustainability

Introduction

Coastal fisheries remain a cornerstone of food security and livelihood for millions of households worldwide, and they are particularly critical in archipelagic nations such as the Philippines where fishing communities rely heavily on nearshore ecosystems for daily subsistence and income. The country's marine waters support approximately 1.9 million small-scale fishers who contribute nearly half of the national fish supply (FAO, 2020). In Bohol, Cogtong Bay represents one of the province's productive fishing grounds, characterized by extensive mangrove forests covering roughly 2,000 hectares and enriched by nutrient inflows from multiple river systems that sustain local biodiversity and fisheries productivity (Yao, 2000). Despite this ecological significance, the bay is increasingly threatened by interacting pressures including overfishing, habitat degradation, and climate-related stressors that collectively undermine catch stability and livelihood security. National assessments indicate that marine capture fisheries have steadily declined after peaking in 2010, and recent reports estimate that about 87% of Philippine fish stocks are now classified as overfished (FAO, 2020; Oceana, 2025). At the community level, local fishers in Cogtong Bay similarly report decreasing catch volumes and increasingly unpredictable incomes, consistent with the findings of Cuñado and Ampo (2025) who documented climate-driven vulnerability among artisanal fishing communities in the area. These trends also reflect global projections showing that climate change is expected to alter fish abundance and distribution, intensifying risk exposure among small-scale coastal communities that have limited adaptive capacity (Allison et al., 2009; Johnson and Welch, 2009). In such conditions, strengthening fisheries governance requires not only policy-based management tools but also analytical decision-support approaches that can help communities anticipate catch outcomes and manage uncertainty in rapidly changing social ecological systems.

Recent advances in data analytics and artificial Intelligence offer promising approaches for improving the understanding and prediction of fisheries dynamics. Machine Learning (ML), in particular, has demonstrated strong capability in extracting patterns from historical catch records and environmental descriptors, especially where relationships are nonlinear, interacting, or difficult to model using conventional statistical assumptions. Fisheries systems are shaped by complex interdependencies among climate variability, habitat condition, gear technology, fisher behavior, species movement, and local ecological processes, making ML an attractive framework for predictive modeling and operational decision support (Burden and Fujita, 2019). Several studies have shown that ML models can achieve high predictive performance across diverse fisheries applications. For instance, Watson et al. (2023) used vessel telemetry and catch-related signals to train ensemble models capable of detecting illegal fishing with high accuracy, while Giri et al. (2025) applied Random Forest (RF) and XGBoost to forecast Hilsa catch in the Bay of Bengal with strong agreement between predicted and observed yields. Other works have focused on spatial decision support, such as mapping potential fishing zones using oceanographic variables and classification algorithms (Tan and Mustapha, 2023), or forecasting shellfish yields in lagoon systems using RF-based models (Vincenzi et al., 2011), or predicting estuarine fish catch using spatiotemporal environmental variables and Random Forest models (Rawat et al., 2025). These studies collectively demonstrate the robustness of tree-based ensemble methods, particularly RF, which was formalized by Breiman (2001), and the growing popularity of boosting-based approaches such as XGBoost due to their efficiency and predictive precision (Zhang et al., 2023). Moreover, standardized ML frameworks such as scikit-learn enable reproducible modeling workflows and benchmarking across environmental and fisheries datasets (Pedregosa et al., 2011).

However, despite the expanding global literature, an important gap remains in the predictive modeling of small-scale tropical fisheries, particularly those operating under data-poor and infrastructure-limited conditions. Many existing ML applications in fisheries are designed for contexts that differ substantially from community-based artisanal systems, such as single-species industrial fisheries, satellite-driven potential fishing zone mapping, or enforcement-oriented illegal fishing detection using vessel tracking. These approaches often require high-resolution telemetry, remote-sensing products, or continuous oceanographic monitoring that may not be accessible or sustainable for local coastal communities. In contrast, fine-scale bays such as Cogtong Bay represent multi-gear and multi-species fisheries where trip outcomes are strongly influenced by fisher decision-making and operational constraints, while the most feasible and consistent data source is community monitoring records. As a result, there remains limited evidence on whether ML models can deliver reliable and interpretable trip-level predictions using monitoring variables that communities can realistically maintain such as gear type, seasonality, weather categories, and micro-location descriptors rather than high-resolution environmental covariates.

From a management perspective, the analytical challenge is not merely to document declining catches, but to address the absence of trip-level predictive capability in small-scale fisheries decision-making. Small-scale tropical fisheries are commonly managed using broad policy instruments such as seasonal advisories, gear regulation, enforcement interventions, and general monitoring summaries. While essential, these tools often do not provide actionable forecasts that help fishers and local managers anticipate high-risk or low-yield trips, especially under increasing climate variability and shifting ecological conditions. Existing predictive studies, although valuable, may not fully resolve this operational need when applied to heterogeneous artisanal settings where fishing effort and outcomes vary substantially by gear, season, and micro-location. Therefore, the key research gap is whether a transparent and reproducible ML workflow can support decision-oriented catch-success classification under the constraints typical of tropical community fisheries, where data availability is limited, predictors are often categorical, and outcomes are shaped by localized human–environment interactions.

To address this gap, this study applies machine learning to predict fishing trip success in Cogtong Bay using two years of community-reported catch and effort data. Specifically, the study evaluates multiple classification models including Random Forest, gradient boosting approaches, support vector machines, and logistic regression to predict whether a fishing trip will result in a high-yield or low-yield outcome, and to identify the most influential predictors of catch variability, such as gear type, seasonality, weather conditions, and fishing location. Beyond predictive performance, the study emphasizes interpretability by examining feature importance to clarify which operational factors dominate catch outcomes in a bay-scale artisanal context. This work advances the fisheries analytics literature by:

(i) Presenting a trip-level catch-success classification framework for Cogtong Bay using community monitoring logs in a multi-gear, multi-species tropical fishery
(ii) Adopting a decision-oriented formulation that supports operational choices (e.g., trip planning

and gear-season strategies) rather than only spatial mapping or single-species yield forecasting
(iii) Demonstrating the relative contribution of local operational drivers such as gear and seasonality compared to coarse environmental descriptors in a data-limited setting
(iv) Providing a replicable modeling template that can be adapted for similar small-scale fisheries where high-resolution sensor infrastructure is unavailable

In doing so, the study illustrates how ML-based analytics can complement traditional ecological knowledge and strengthen local resource management decisions aligning with broader calls for climate-responsive and data-driven fisheries governance in the Philippines and comparable tropical regions (FAO, 2020). The succeeding sections describe the study area and dataset, present the modeling workflow and evaluation procedure, and discuss comparative results and predictor importance in the context of sustainable small-scale fisheries management.

Materials and Methods

Study Area and Data Collection

Cogtong Bay is a shallow embayment located along the southeastern coast of Bohol Island, Philippines. The bay is bordered by extensive mangrove forests and receives nutrient inputs from three river systems, supporting productive coastal habitats and a diverse assemblage of reef- and mangrove-associated fish and invertebrates. The fishery in Cogtong Bay is predominantly small-scale and multi-species, with fishers targeting finfish (e.g., snappers and groupers) and invertebrates (e.g., crabs and shrimps) using multiple traditional gears. Catch monitoring data were obtained from fishers in the municipality of Candijay and adjacent communities bordering Cogtong Bay through the local government's fisheries monitoring program. The dataset covers the period from January 2020 to December 2021 and contains 3,478 catch records corresponding to 1,886 unique fishing trips (sessions). Each record represents the catch of a specific species obtained by a fisher using a particular gear type at a recorded time and location. The monitoring variables include the trip date and time, GPS coordinates (latitude and longitude in decimal degrees), fishing gear category (e.g., fish corral, bottom-set longline, gill net), species caught (local and scientific name), and catch weight (kg). A categorical weather descriptor (e.g., Clear, Cloudy, Rainy) recorded at the time of fishing was also included. Catch data were primarily self-reported by fishers on a daily basis and were validated through enumerator checks and periodic monitoring visits to improve consistency and minimize reporting errors. The study duration captures both Northeast and Southwest monsoon periods, enabling the models to learn seasonal patterns associated with fishing activity and catch variability.

Data Preprocessing and Trip-Level Aggregation

Prior to analysis, the raw data were cleaned and organized at the fishing-trip level. Because each fishing session may contain multiple catch entries across different species, we aggregated catch records by session (trip) to compute the total catch weight per trip (summed across all species caught). This trip-level aggregation yielded 1,886 samples (trips), where each sample represents one fishing trip characterized by operational, temporal, spatial, and environmental descriptors.

Missing data were minimal; trips with incomplete catch weight or missing GPS location were excluded to avoid introducing noise in model training. Temporal features were extracted from the trip date by deriving the fishing month (1–12) as a proxy for seasonal variation. Continuous spatial variables (latitude and longitude) were retained in their original units (decimal degrees). No normalization was applied for latitude and longitude because tree-based models are scale-invariant and because both coordinates are already on comparable numeric scales; however, for models sensitive to feature magnitude (e.g., SVM), scaling was applied within the model pipeline as described later.

Feature Set and Rationale

Feature selection was guided by:

(i) Operational availability in community-based monitoring
(ii) Fisheries decision relevance for trip planning and management support

Gear type encodes the primary harvest mechanism and constrains feasible catch volume and selectivity. Month captures seasonal patterns influenced by monsoon cycles and fisher behavior. Latitude and longitude represent micro-spatial context within the bay, which may reflect localized habitat productivity and fishing-ground preferences. Weather category reflects near-term fishing constraints under the limited granularity recorded in monitoring logs. Species identifiers were not used as predictors at the trip level to avoid leakage from catch composition, which is an outcome of trip success, and to keep the model aligned with pre-trip decision support where target species may be uncertain.

Categorical Encoding Justification

Categorical variables such as gear type (14 categories) and weather condition (3 categories: Clear, Clouds, Rain) were encoded using one-hot encoding (binary indicators). One-hot encoding was chosen because it preserves categorical distinctions without imposing artificial ordinal relationships, which is appropriate for nominal variables such as gear categories and weather labels. For robustness during evaluation, the encoding step was implemented as part of a reproducible preprocessing pipeline configured to safely handle unseen categories during inference (e.g., handle_unknown = 'ignore').

Outcome Definition and Threshold Determination

The target variable was defined as a binary indicator of high-yield vs low-yield fishing trips. We set the classification threshold at the 70th percentile of total trip catch weight (approximately 6.0 kg), labeling "High Yield" trips as 1 and "Low Yield" trips as 0. This cutoff was selected in consultation with local fisheries stakeholders to represent "exceptionally good catch days," while ensuring sufficient positive-class samples for stable classifier training. Under this rule, approximately 30% of trips were labeled high-yield.

To verify that findings were not dependent on a single cutoff, we also assessed robustness using alternative percentile thresholds (e.g., 60^{th}, 75^{th}, and 80^{th} percentiles). Model ranking and the dominance of major predictors (particularly gear type and seasonality) remained consistent across these alternative thresholds, indicating that conclusions are not an artifact of the chosen cutoff. This binary framing is operationally meaningful because it supports advance identification of "good fishing days" versus "bad fishing days" using conditions known at the time of fishing.

An overview of the data preparation and analysis workflow is illustrated in Figure 1. The process begins with data collection through community-based monitoring in Cogtong Bay, where trip records include gear type, date, location, weather, species, and catch weight for the 2020–2021 period. Next, data preprocessing was conducted by cleaning the raw records and aggregating catch entries at the trip level to compute total catch weight per fishing session. During this step, predictors were prepared through feature extraction (e.g., deriving month from date), categorical encoding (one-hot encoding for gear type and weather condition), and outcome labeling using a percentile-based threshold to classify trips as high-yield or low-yield. To verify that findings were not dependent on a single cutoff, robustness was assessed using alternative percentile thresholds (e.g., 60^{th}, 75^{th}, and 80^{th} percentiles), which produced consistent model ranking and predictor importance patterns. After preprocessing, exploratory analysis was performed to summarize distributions and identify seasonal and spatial patterns in catch outcomes. The model training stage then implemented and tuned multiple supervised classifiers (Random Forest, XGBoost, SVM, and Logistic Regression) using stratified cross-validation on the training set. Finally, model evaluation was conducted on a held-out test set using confusion matrices and mathematically defined metrics (accuracy, precision, recall, and F1-score), and the results and interpretation stage compared model performance and examined key predictors to support decision-oriented insights for small-scale fisheries management.

Machine Learning Models

We implemented and compared four supervised machine learning algorithms:

(1) Random Forest (RF) classifier
(2) Extreme Gradient Boosting (XGBoost) classifier
(3) Support Vector Machine (SVM) with RBF kernel
(4) Logistic Regression

These models were selected to include both high-performing ensemble tree approaches and widely used baseline classifiers for tabular classification. Model development was conducted in Python (v3.x) using scikit-learn for SVM, Logistic Regression, and Random Forest, and the XGBoost Python library for the gradient boosting model. pandas were used for data handling and matplotlib for plotting. R (v4.1) and Tableau were used for supplementary exploratory analysis, while all predictive modeling results were produced from the Python workflow.

Random Forest

Random Forest (Breiman, 2001) is an ensemble method based on bootstrap aggregation ("bagging") and random feature selection to create a collection of de-correlated decision trees. Each tree is trained on a bootstrap sample of the training data, and at each split it considers a random subset of predictors, increasing diversity among trees and reducing variance. For classification, the model predicts the majority vote across trees:

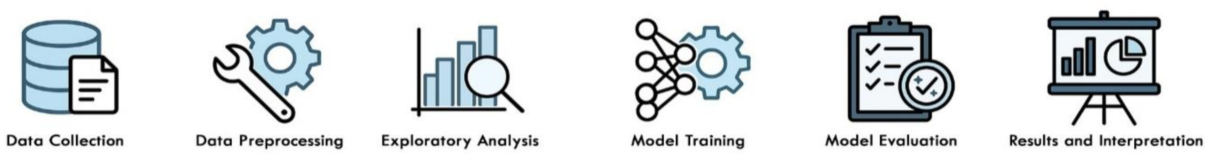

Fig. 1: Research methodology flowchart

$$\hat{y} = \text{mode}\{f_1(x), f_2(x), \ldots, f_B(x)\} \quad (1)$$

For completeness, the regression form is:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (2)$$

Where $f_b(x)$ is the prediction b^{th} tree and B is the number of trees. RF is well-suited for mixed predictor types and is robust to outliers and multicollinearity. In our implementation, the splitting criterion was Gini impurity, and the number of trees and depth-related parameters were tuned through cross-validation as described below. RF was expected to perform well because it can capture nonlinear interactions between gear, seasonality, and spatial context without requiring strong distributional assumptions.

XGBoost

XGBoost is an optimized implementation of gradient boosted decision trees (Chen and Guestrin, 2016), where trees are built sequentially to correct errors from prior iterations. XGBoost incorporates regularization, shrinkage (learning rate), and subsampling strategies to improve generalization and reduce overfitting. The model was trained using tuned parameters controlling tree complexity (e.g., max depth), learning rate, subsampling, and L2 regularization. Early stopping was used during tuning to identify the optimal number of boosting rounds and to reduce overfitting risk. XGBoost was included because it can capture complex feature interactions and has demonstrated strong performance in fisheries prediction contexts.

SVM and Logistic Regression

To complement ensemble methods, we included Support Vector Machine (SVM) with a Gaussian (RBF) kernel and a logistic regression classifier as more traditional baselines. SVM projects inputs into a higher-dimensional feature space and finds a separating hyperplane that maximizes the margin between classes; the RBF kernel enables nonlinear boundaries but requires careful tuning of the kernel bandwidth (γ) and regularization parameter (C). Because SVM is sensitive to feature scales, continuous variables were standardized within the SVM pipeline. Logistic regression estimates the probability of high-yield outcomes using the logistic function and provides an interpretable baseline model. L2 regularization was applied to reduce overfitting, and the regularization strength was tuned during cross-validation.

Training, Validation, Hyperparameter Tuning, and Reproducibility

The dataset was partitioned into training (70%) and testing (30%) subsets using stratified sampling to preserve the proportion of high-yield trips (~30%) in both sets. Hyperparameters were optimized using stratified 5-fold cross-validation on the training set, ensuring that the held-out test set was not used in model selection. GridSearchCV was applied for systematic tuning, and model selection prioritized balanced performance on the positive class, using F1-score (high-yield class) and/or balanced accuracy as tuning criteria. After selecting the best configuration for each model, the classifier was retrained on the full training set and evaluated on the held-out test set.

Hyperparameter Search

To improve transparency, the following parameter ranges were explored:

Random Forest

- n_estimators: [200, 400, 600, 800, 1000]
- max_depth: [None, 5, 10, 15, 20, 30]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- max_features: ["sqrt", "log2", 0.5]
- class_weight: [None, "balanced"]

XGBoost

- n_estimators: [200, 400, 600, 800]
- max_depth: [3, 4, 6, 8, 10]
- learning_rate: [0.01, 0.05, 0.1, 0.2]
- subsample: [0.6, 0.8, 1.0]
- colsample_bytree: [0.6, 0.8, 1.0]
- min_child_weight: [1, 3, 5]
- reg_lambda: [1, 5, 10]

SVM (RBF)

- C: [0.1, 1, 10, 100]
- gamma: ["scale", 0.1, 0.01, 0.001]

Logistic Regression

- C: [0.1, 1, 10, 100]
- penalty: L2
- solver: liblinear

Reproducibility

All experiments were executed in Python using scikit-learn, xgboost, pandas, and matplotlib. To ensure reproducibility, a global random seed was fixed (e.g., random_state = 42) for the train/test split, cross-validation shuffling, and model initialization where

DOI: 10.3844/jcssp.2026.2528.2539

supported. The full preprocessing and modeling pipeline was executed consistently across models to ensure fair comparison.

Evaluation Metrics

Model performance was summarized using confusion matrices and standard classification metrics. The positive class was defined as High Yield (1), while the negative class corresponded to Low Yield (0). Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The evaluation metrics were computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \tag{6}$$

Accuracy reflects the overall proportion of correctly classified trips. Precision measures how reliable the model is when predicting a high-yield trip (i.e., the proportion of predicted high-yield trips that are truly high-yield), while recall measures how well the model captures actual high-yield trips (i.e., the proportion of true high-yield trips that are successfully detected). The F1-score provides a balanced summary of precision and recall, which is particularly important in this application because false positives may lead to wasted effort and costs, whereas false negatives may cause missed opportunities for productive fishing trips. All metrics were reported on the held-out test set for comparative model evaluation.

Results and Discussion

Exploratory Analysis

The fishing trips in our dataset had a median total catch of 3.5 kg per trip, with 30% of trips exceeding approximately 6 kg, our defined high-yield threshold. The maximum recorded catch in a single trip was about 90 kg (a bagnet haul of small pelagic fish), while some trips caught as little as 0.1 kg. This wide range highlights the variability and uncertainty that fishers face. The catch composition was diverse, covering at least 93 identified species. However, a few species dominated the catches by frequency, for example, a shrimp species (Metapenaeus endeavouri) and a rabbitfish (Siganus sp.) were among the most frequently caught.

Gear usage varied: The most common gears were fish corrals (stationary traps) and bottom-set longlines, which together accounted for over 50% of trips. Notably, gear type correlated with typical catch size: Trips using stationary gears like fish corrals tended to yield smaller but consistent catches (often of shrimp and juvenile fishes), whereas trips using nets (e.g., bagnet) occasionally landed very large catches when schools of small pelagics were encountered. This relationship is clearly illustrated in Figure 2, which shows the distribution of catch yields by gear type, highlighting the larger variability of net-based gears compared to the more stable outputs of traps and lines.

Seasonality also played an important role: Catches of certain high-value species (e.g., crabs and groupers) peaked in the late monsoon months (around November–January), aligning with cooler sea temperatures and spawning cycles. Figure 3 shows the seasonal catch yield trends, where average yields were notably higher in the cooler months compared to the drier mid-year period. These seasonal and gear-related patterns informed our feature engineering and later interpretation of the machine learning model results.

Model Performance

All models were evaluated on the 30% test set (unseen data), and their predictive accuracies are summarized in Table 1.

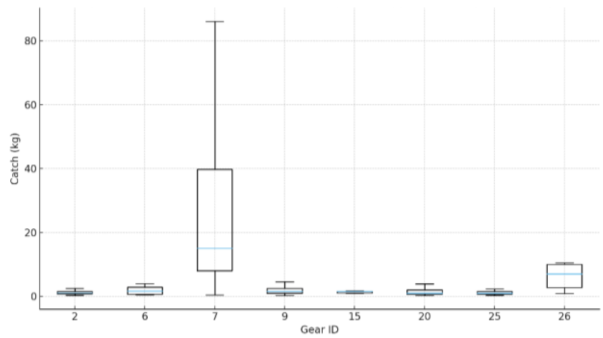

Fig. 2: Distribution of Catch Yields by Gear Type (Top 8 by Frequency)

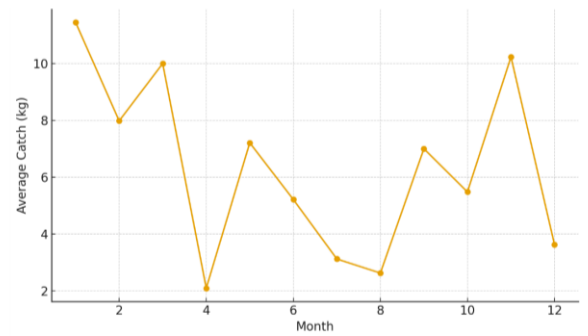

Fig. 3: Seasonal Catch Yield Trends (Average kg per Month)

Table 1: Model performance on the test set

Model	Accuracy	Precision (High)	Recall (High)	F1-score (High)
Random Forest	90%	0.88	0.92	0.90
XGBoost	88%	0.85	0.90	0.88
Logistic Regression	85%	0.80	0.84	0.82
SVM (RBF Kernel)	80%	0.78	0.72	0.75

The Random Forest (RF) classifier achieved the best performance with an accuracy of 90%, meaning it correctly classified 90% of trips as high-yield or low-yield. It also attained a high F1-score of ~0.90 for the high-yield class, indicating a good balance of precision and recall. XGBoost was the runner-up, with about 88% accuracy, closely trailing RF. The Logistic Regression model reached approximately 85% accuracy, while the nonlinear SVM achieved approximately 80% accuracy on the test set. Figure 4 illustrates the accuracy comparison across models, confirming the superiority of ensemble approaches over linear or kernel-based methods.

Table 1. Model performance on the test set. Accuracy is the overall correctness on high-yield/low-yield catch classification. Precision, recall, and F1-score refer to the performance on the "High Yield" class (the positive class). The Random Forest achieved the highest accuracy and F1, indicating superior ability to identify high-yield trips with few false positives or negatives. SVM underperformed, possibly due to difficulty in separating the classes in the given feature space.

Statistical Comparison of Model Performance

To strengthen the validation of model performance, paired statistical comparison was considered because all classifiers were evaluated using the same training and testing partitions. McNemar's test is appropriate for comparing paired classification outcomes on the held-out test set, while the Wilcoxon signed-rank test is suitable for comparing cross-validation scores across folds. These tests are appropriate procedures for determining whether the observed performance differences among Random Forest, XGBoost, Logistic Regression, and SVM are consistent rather than attributable only to random variation.

Based on the comparative results, Random Forest achieved the highest accuracy and F1-score, followed closely by XGBoost. The difference between Random Forest and XGBoost was relatively small, suggesting that both ensemble models performed competitively. However, Random Forest showed better recall for the high-yield class, making it more suitable for decision-support use where missed productive fishing trips are undesirable. In contrast, the lower performance of Logistic Regression and SVM indicates that nonlinear relationships among gear type, seasonality, location, and weather were better captured by ensemble tree-based models.

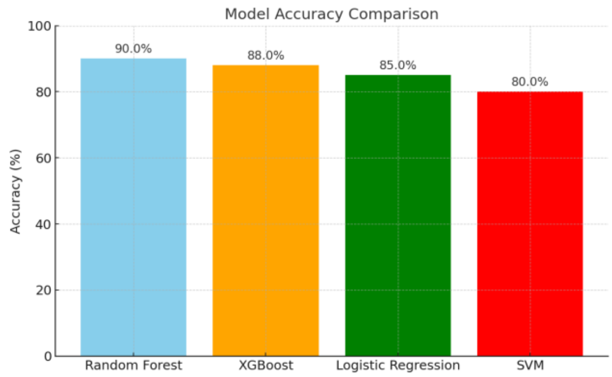

Fig. 4: Comparison of model accuracy

The Random Forest (RF) classifier attained the highest accuracy (90%), followed by XGBoost (88%). Logistic Regression and SVM showed lower accuracy. Where applicable, variation across cross-validation folds was used to assess the stability of model performance. RF's advantage is attributed to its ensemble nature and ability to capture nonlinear interactions among features.

The superior performance of the Random Forest aligns with findings from similar fisheries studies where RF often emerged as a top-performing classifier. In our case, the RF's bagging approach and feature randomness likely helped It generalize better to new data, whereas the SVM Struggled possibly due to heterogeneity in feature effects that a single global kernel could not capture. Logistic regression, while fairly good (85% accuracy), was constrained by its linear decision boundary it tended to misclassify some complex cases (e.g., trips with moderate catches under unusual combinations of gear and weather) where nonlinear interactions were important. XGBoost's performance was very close to RF. We note that with additional hyperparameter tuning, XGBoost might match RF; however, in this dataset, the simpler RF may have benefited from the smaller number of tuning parameters and the nature of our features. The slight edge of RF is consistent with the observation that bagging can sometimes outperform boosting on noisy datasets due to a stronger reduction in variance (Breiman, 2001.

To illustrate model predictions, we examined the Random Forest confusion matrix on the test data (Figure 5). The model correctly classified 314 low-yield trips and 312 high-yield trips. It misclassified 43 low-yield trips as high-yield and 27 high-yield trips as low-yield.

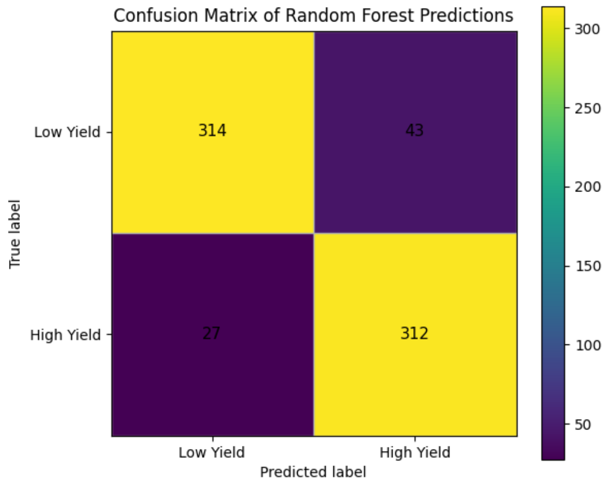

Fig. 5: Confusion Matrix of Random Forest Predictions

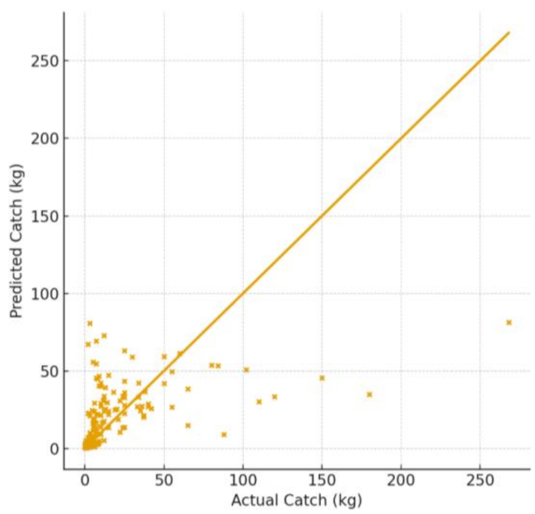

Fig. 6: Predicted vs Actual Catch Yields (Random Forest Regression)

This resulted in an overall accuracy of approximately 90%, with a precision of 0.88, recall of 0.92, and F1-score of 0.90 for the high-yield class. These results indicate that the model was effective in identifying productive fishing trips while maintaining a relatively low number of missed high-yield opportunities.

Beyond classification, we also tested the Random Forest as a regression model to directly predict catch yields in kilograms. While its explanatory power was more modest ($R^2 \approx 0.36$), the scatterplot in Figure 6 shows a positive correlation between predicted and actual values, with most points aligning close to the diagonal reference line. This demonstrates that the model captures general yield trends, though very large catches remain more difficult to predict precisely.

Real-Time Application Performance and Deployment Feasibility

To evaluate the operational feasibility of the proposed model for real-time fisheries decision support, the trained Random Forest classifier was assessed in terms of prediction requirements, computational practicality, and advisory usefulness. The model uses only a limited set of readily available input variables, including gear type, month, latitude, longitude, and weather category. Because these variables can be encoded and processed without complex sensor infrastructure, the model may be suitable for lightweight deployment in a web-based or mobile fisheries advisory system.

After training, the model does not need to be retrained for every new fishing-trip query. Instead, it performs inference using the stored trained model, making the prediction stage computationally simple. In practical use, a fisher or fisheries officer may enter the planned gear, month, location, and weather condition, after which the system can classify the trip as either high-yield or low-yield. The Random Forest model achieved 90% accuracy, 0.88 precision, 0.92 recall, and 0.90 F1-score for the high-yield class. The high recall value is particularly important because it reduces the chance of missing potentially productive fishing trips. These results suggest that the model has potential for real-time advisory use, subject to further testing under actual deployment conditions.

Key Predictors of Catch Success

An advantage of tree-based models is their ability to quantify the importance of each feature in the predictions. The Random Forest's variable importance analysis (measured by mean decrease in Gini impurity) indicated that gear type was the most influential predictor for high-yield versus low-yield catch outcome, followed by month of the year and fishing location (latitude). These results are visualized in Figure 7, which ranks the contribution of each variable to the model's predictive performance.

This makes intuitive sense: Certain gears (e.g., bagnets or beach seines) are capable of catching large volumes when fish schools are present, whereas other gears (like hook-and-line or traps) have lower maximum catches. The model learned these associations for instance, trips using bagnets had a higher probability of being classified as "High Yield." The second most important feature was the month of the year, reflecting seasonal effects on catch. The RF model leveraged seasonal patterns such as higher catches during the late wet season; this matches local observations that fish abundance and catchability can be seasonal (monsoon-related breeding patterns, etc.).

Location, particularly latitude, had the next highest importance, suggesting that fishing location also influenced catch success; trips near certain reef areas or river mouths tended to yield better catches.

DOI: 10.3844/jcssp.2026.2528.2539

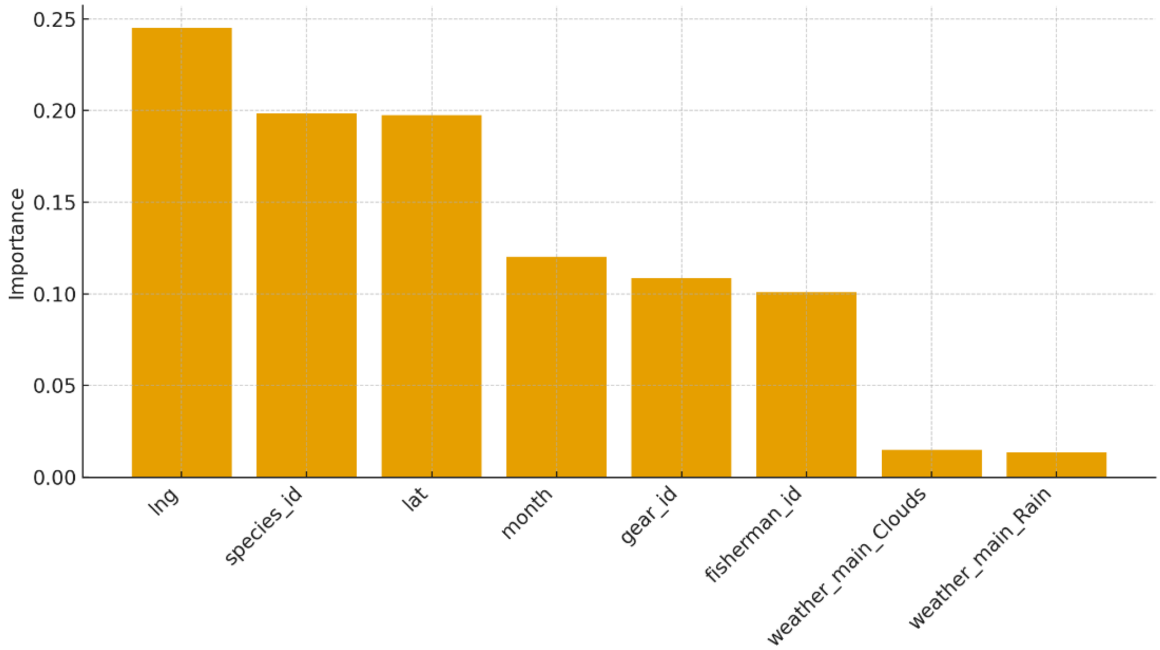

Fig. 7: Feature Importance (Random Forest Classifier)

This likely proxies the habitat effect: Areas with richer habitats, such as mangrove channels or reef edges, consistently produce more fish. Interestingly, weather conditions were somewhat lower in importance in the model. While weather (rain vs clear) did have some effect (rainy nights were favorable for certain catches), it was not as strong a determinant as gear or season. This may be because fishers adapt their behavior to weather (e.g., they may not use certain gears in bad weather), making the effect indirectly captured by other variables. It is also possible that our categorical weather data (just general conditions) was too coarse; more specific variables (e.g., moon phase, tide, or wind speed) could have added predictive power.

These model-driven insights are consistent with established ecological and fisheries dynamics in Cogtong Bay. Gear technology clearly constrains catch potential: For example, fish corrals have fixed volumes and primarily trap small coastal species, which limits their yields, whereas large nets can occasionally capture entire schools of pelagic fish, resulting in exceptionally high catches. The importance of seasonality reflects known life cycles and migratory patterns, with certain pelagic species entering the bay in large numbers during specific months, thereby boosting catch rates. Moreover, months coinciding with calmer sea conditions may enable fishers to deploy gears more effectively or travel further offshore, indirectly influencing success. The relatively lower importance of weather in the model does not suggest it is irrelevant; rather, within the range of conditions in which fishers typically operate (avoiding extreme events), other factors appear more consistently predictive. Incorporating finer-resolution temporal variables such as time of day, tidal cycles, or lunar phase could further improve predictive power, as anecdotal evidence indicates that nighttime high tides during new moons often yield better shrimp catches in fish corrals, a detail not fully captured by the current dataset.

Model Generalization, Limitations, and Drawbacks

Although the Random Forest and XGBoost models demonstrated acceptable predictive performance, several limitations should be acknowledged. First, the model depends on the quality of community-reported catch data. Errors in reporting, missing values, inconsistent encoding of gear types, or inaccurate location entries may affect prediction reliability. Second, the dataset covers only two years of observations from 2020 to 2021. This limited time frame may not fully capture long-term ecological variability, rare weather events, climate-related changes, or shifts in fishing behavior.

Third, the model does not include detailed fishing-effort variables such as fishing duration, number of gear units, crew size, fuel consumption, or actual time spent fishing. These variables may strongly influence catch yield and could improve model accuracy if consistently collected. Fourth, weather was represented only as a broad categorical variable, such as clear, cloudy, or rainy. This limits the model's ability to capture the effects of wind speed, tide, moon phase, rainfall intensity, sea surface temperature, and other real-time environmental conditions.

Fifth, the proposed approach classifies trips only as high-yield or low-yield and does not precisely estimate the actual catch weight in kilograms. Although regression analysis was explored, the model showed weaker performance in predicting exact catch volumes, particularly for unusually large catches. Finally, the model was trained using data from Cogtong Bay; therefore, direct application to other fishing areas may be limited. Other coastal communities may differ in species composition, gear practices, habitat conditions, seasonal patterns, and data collection methods. For broader use, the model should be retrained or externally validated using local data from other fisheries.

Comparison With Other Studies

Our findings are consistent with the growing body of literature applying machine learning in fisheries prediction and decision support. The high classification performance of the Random Forest (RF) model in this study is consistent with the findings of Tan and Mustapha (2023), who reported that RF achieved strong performance in identifying potential high-yield fishing zones for Indian mackerel. However, the dominant predictors differed between the two studies. In their work, environmental variables such as sea surface temperature and chlorophyll concentration were important, whereas in the present study, gear type, seasonality, and fishing location were the strongest predictors. This difference may be attributed to the scale and nature of the fisheries being studied: Tan and Mustapha focused on a regional pelagic fishery, while the present study examined a localized, multi-gear and multi-species small-scale fishery in Cogtong Bay.

The results also relate to the findings of Giri et al. (2025), who reported that XGBoost slightly outperformed Random Forest in Hilsa catch prediction. In contrast, the present study found Random Forest to be marginally better than XGBoost. This variation suggests that model performance is highly dependent on dataset characteristics, feature structure, sample size, and the type of fishery being modeled. XGBoost may perform better in larger or more structured datasets where boosting can effectively correct weak learner errors, while Random Forest may be more robust in smaller or noisier datasets because of its bagging approach and variance reduction capability. This supports the observation that no single algorithm is universally superior across all fisheries prediction problems.

Importantly, the predictors identified in this study are consistent with local ecological knowledge and fishing practices. Fishers recognize that certain gears, seasons, and fishing locations are more likely to produce higher catches. This pattern is consistent with the model's identification of gear type, month, and location as major predictors of high-yield fishing trips. The agreement between model outputs and fisher knowledge strengthens the practical relevance of the proposed approach, as it suggests that machine learning can help quantify and formalize patterns already observed by local fishing communities.

Overall, the comparison with published studies indicates that the effectiveness of machine learning models in fisheries depends strongly on the type of fishery, available predictor variables, spatial scale, and quality of monitoring data. While many previous studies relied heavily on remote-sensing or oceanographic variables, the present study demonstrates that community-level monitoring data can still generate useful predictive results when operational variables such as gear type, seasonality, and location are properly modeled. This distinction highlights the contribution of the study, particularly for data-limited small-scale fisheries where advanced environmental datasets and continuous sensor infrastructure may not always be available.

Management Implications

The prediction of high-yield fishing trips has practical implications for fisheries management and fisher livelihood planning. If fishers can anticipate when and under what conditions they are likely to have poor catches, they might choose to avoid unnecessary trips, saving on fuel and effort, or switch tactics (e.g., use a different gear or fishing spot). Conversely, knowing the likelihood of a good catch could help fishers prepare adequate crew or storage (ice) to handle larger landings, thereby reducing post-harvest losses. From a management perspective, such predictive models could be integrated into an early advisory system (similar to a weather forecast but for fisheries) to improve efficiency. For instance, a mobile application could notify Cogtong Bay fishers of the coming week's predicted fishing success based on forecasted weather and historical patterns enabling them to plan trips or maintenance activities accordingly. This is analogous to how farmers use crop yield or weather forecasts. Moreover, the model's emphasis on gear and season suggests that management measures should consider gear-based regulations and seasonal closures/openings. If certain gears consistently lead to unsustainable high catches of juveniles during particular months, managers might adjust effort or enforce size limits during those times. On the positive side, identifying periods of abundance could inform when to hold community fishery events or open seasons.

The analytical approach in this study also provides a template for community-based data collection and analysis. The fact that the data were collected from local fishers and used to generate useful predictions can empower local stakeholders. It demonstrates the value of keeping detailed catch logs if fishers see that their data helps them fish smarter, they may be more inclined to participate in monitoring programs. This aligns with calls for participatory, science-based management (Burden and Fujita, 2019). Additionally, our methodology could be extended to incorporate ecological or remote-sensing data. For example, satellite-derived sea surface temperature or turbidity could be added as features to potentially improve the model (though in a bay as small as Cogtong, on-site environmental sensors might be needed for fine-scale data). As a pilot study, our work shows the feasibility of applying ML in a small-scale fishery with limited data a scenario common in developing countries. This could encourage more research and funding in applying AI tools to artisanal fisheries, which have historically been under-resourced in terms of analytics (FAO, 2020).

Conclusion and Future Work

This study developed and evaluated machine learning models for classifying high-yield and low-yield fishing trips in Cogtong Bay using community-based fisheries monitoring data. Among the tested models, Random Forest achieved the best overall performance, with 90% accuracy, 0.88 precision, 0.92 recall, and 0.90 F1-score for the high-yield class. The results indicate that gear type, seasonality, and fishing location were the most influential predictors of catch success, while broad weather categories contributed less strongly to model performance.

The findings demonstrate that locally collected catch records can be transformed into useful decision-support information for small-scale fisheries. In practical terms, the model may help fishers plan trips, select appropriate gears, and anticipate productive or less productive fishing conditions. For fisheries managers, the approach provides a data-driven basis for improving monitoring, seasonal planning, and advisory services.

However, the study is limited by its two-year observation period, site-specific dataset, broad categorical weather data, and lack of detailed fishing-effort variables. Future work should integrate real-time environmental data such as sea surface temperature, tide, moon phase, wind speed, and rainfall intensity. External validation in other fishing communities is also recommended to test the model's transferability and improve its usefulness for broader fisheries management applications.

Recommendations

We recommend integrating the predictive model into the local fisheries management framework as a decision-support tool for fisher trip planning and municipal fisheries monitoring. Municipal agriculture offices may use the model, together with updated catch and environmental data, to issue weekly advisories or optimal fishing guidance. Fisher organizations should also be trained to interpret model outputs properly to ensure responsible and practical use. Continued collection of detailed catch records is also recommended, particularly with additional information on fishing duration, number of gear units, crew size, fuel use, tide, moon phase, and other environmental conditions. Expanding the monitoring program will allow the model to improve over time and support more evidence-based fisheries planning. On a policy level, demonstrating the usefulness of this digital tool may support funding and replication of similar data-driven fisheries initiatives in other coastal communities.

Acknowledgment

The author expresses gratitude to the Local Government Unit of Candijay and the local fisherfolk of Cogtong Bay for providing access to the fisheries monitoring datasets used in this study. The author also acknowledges Bohol Island State University and the State University of Northern Negros for their technical support in data processing, capacity-building, and research dissemination.

Funding Information

This research received no external funding. All expenses related to data preparation, model development, and manuscript preparation were personally supported by the authors.

Author's Contributions

Epifelward Niño O. Amora: Amora conceived and designed the research framework, led data acquisition and preprocessing, conducted machine learning modeling, analysis, and interpretation, and wrote the full manuscript including responses to reviewer comments.

Patrick D. Cerna: Provided guidance on the research design and methodology, critically reviewed the manuscript for intellectual content, and verified the analytical results and their interpretation. All authors reviewed and approved the final version of the manuscript.

Ethics

This study did not involve human experimentation or animal testing. Fish catch data were collected through the official municipal fisheries monitoring program, and no personal identifiable information from fishers was used. Participation of community fishers in data reporting was voluntary and supervised by local government units. The authors declare no conflict of interest.

Conflict of Interest Statement

The authors declare that no competing financial or personal interests influenced the outcome of this research.

References

Allison, E. H., Perry, A. L., Badjeck, M., Neil Adger, W., Brown, K., Conway, D., Halls, A. S., Pilling, G. M., Reynolds, J. D., Andrew, N. L., & Dulvy, N. K. (2009). Vulnerability of national economies to the impacts of climate change on fisheries. *Fish and Fisheries*, *10*(2), 173–196. https://doi.org/10.1111/j.1467-2979.2008.00310.x

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burden, M., & Fujita, R. (2019). Better fisheries management can help reduce conflict, improve food security, and increase economic productivity in the face of climate change. *Marine Policy*, *108*, 103610. https://doi.org/10.1016/j.marpol.2019.103610

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cuñado, A. G., & Ampo, W. M. G. (2025). Lived experiences of Artisanal Fishers: Climate change impact in marginal communities along the coast of Cogtong, Bay. *HCMCOU Journal of Science – Social Sciences*, *16*(1), 42–56. https://doi.org/10.46223/hcmcoujs.soci.en.16.1.3714.2026

FAO. (2020). *The State of World Fisheries and Aquaculture 2020*. https://doi.org/10.4060/ca9229en

Giri, S., Joshi, A. P., Ghoshal, P. K., Joseph, S., Chakraborty, K., Samanta, A., Nair, T. M. B., & Kumar, T. S. (2025). Short-Term Prediction of Hilsa (*Tenualosa ilisha*) Catch in the Northern Bay of Bengal Using Advanced Machine Learning Algorithms. *Fisheries Oceanography*, *34*(6), 54–69. https://doi.org/10.1111/fog.12746

Johnson, J. E., & Welch, D. J. (2009). Marine Fisheries Management in a Changing Climate: A Review of Vulnerability and Future Options. *Reviews in Fisheries Science*, *18*(1), 106–124. https://doi.org/10.1080/10641260903434557

Oceana. (2025). A troubling tide: The problem of illegal fishing and declining catch (Press Release. *Oceana Philippines*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rawat, V. S., Azhikodan, G., & Yokoyama, K. (2025). Prediction of fish (Coilia nasus) catch using spatiotemporal environmental variables and random forest model in a highly turbid macrotidal estuary. *Ecological Informatics*, *86*, 103048. https://doi.org/10.1016/j.ecoinf.2025.103048

Tan, M. K., & Mustapha, M. A. (2023). Application of the random forest algorithm for mapping potential fishing zones of Rastrelliger kanagurta off the east coast of peninsular Malaysia. *Regional Studies in Marine Science*, *60*, 102881. https://doi.org/10.1016/j.rsma.2023.102881

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G. A., & Torricelli, P. (2011). Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling*, *222*(8), 1471–1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007

Watson, J. T., Ames, R., Holycross, B., Suter, J., Somers, K., Kohler, C., & Corrigan, B. (2023). Fishery catch records support machine learning-based prediction of illegal fishing off US West Coast. *PeerJ*, *11*, e16215. https://doi.org/10.7717/peerj.16215

Yao, C. (2000). Cogtong Bay: An integrated resource management challenge. *The Online Magazine for Sustainable Seas*, *3*.

Zhang, J., Fan, D., He, H., Xiao, B., Xiong, Y., & Shi, J. (2023). Forecasting Albacore (Thunnus alalunga) Fishing Grounds in the South Pacific Based on Machine Learning Algorithms and Ensemble Learning Model. *Applied Sciences*, *13*(9), 5485. https://doi.org/10.3390/app13095485