INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Ensemble Machine Learning Approaches for CPUE Forecasting in Kenya's Artisanal Marine Fisheries: Application of Xgboost and Prophet with Random Forest Feature Selection

Magak C., Oscar Ngesa., Herbert Imboga

County Gevernment of Kisumu

DOI: https://doi.org/10.51584/IJRIAS.2026.11070114

Received: 10 May 2026; Accepted: 15 May 2026; Published: 08 August 2026

ABSTRACT

Catch Per Unit Effort (CPUE) remains a fundamental index of fishery productivity, supporting assessments of stock status and guiding management decisions. In many data-poor fisheries, including those in the Western Indian Ocean (WIO), CPUE forecasting is constrained by incomplete time series, high variability, and limited analytical capacity. Recent advances in machine learning (ML) offer powerful alternatives to traditional statistical models by capturing nonlinear dynamics, interactions among variables, and temporal patterns in noisy data.

This study applies an **ensemble machine learning approach** to forecast CPUE in Kenyan marine artisanal fisheries across five coastal counties. A **Random Forest (RF)** algorithm was first employed to rank predictor variables and identify the most influential features. These informed the development of county-specific forecasting models using **Extreme Gradient Boosting (XGBoost)** and **Prophet**, evaluated over 30-day holdout test periods. Data preprocessing included cubic interpolation for missing values, moderate capping of outliers, and encoding of categorical features.

Results showed that **XGBoost consistently outperformed Prophet** across all counties. Mean Absolute Error **(MAE)** ranged from **0.063 to 0.22**, Root Mean Squared Error **(RMSE)** from **0.085 to 0.29**, and **R^2 exceeded 0.996** in all counties, indicating strong predictive performance. Prophet, while capturing general seasonal cycles, exhibited wider prediction intervals and lower accuracy, particularly in highly variable catch series. Random Forest analysis revealed that **Boat Activity Coefficient (BAC)**, vessel type, and mesh size were dominant predictors of CPUE variation.

This work demonstrates the value of ensemble ML approaches for fisheries forecasting in data-poor contexts. By combining **feature selection with Random Forest** and **robust nonlinear prediction with XGBoost**, it provides a replicable framework for small-scale fisheries globally. The findings highlight both the potential of ML in operational fisheries management and its role in supporting evidence-based policy in Kenya's devolved governance structure.

Keywords: CPUE, XGBoost, Prophet, Random Forest, machine learning, artisanal fisheries, Kenya, forecasting

INTRODUCTION

Catch Per Unit Effort (CPUE) is widely used as a proxy for fish abundance and a critical measure in stock assessment (Hilborn & Walters, 1992; Maunder & Punt, 2004). In artisanal fisheries, CPUE data guide management interventions, influence licensing decisions, and provide indicators of sustainability. Yet in many developing regions, including Kenya's Western Indian Ocean (WIO) coastline, CPUE datasets are often **incomplete, noisy, and heterogeneous**, undermining their utility for robust assessments (Obura et al., 2017).

The **Kenyan marine fishery**, spanning five coastal counties (Mombasa, Kilifi, Kwale, Lamu, and Tana River), is predominantly artisanal, with over 13,000 fishers operating small vessels and traditional gears (KMFRI, 2020). It provides critical livelihoods and food security, but monitoring systems remain fragile. Challenges include

missing or inconsistent records, seasonal variability, and high gear diversity, all of which complicate CPUE estimation (Wanyonyi et al., 2018).

Traditional statistical approaches, such as linear regression or ARIMA, have limitations in these contexts. They assume linear relationships or stationarity, which are often violated in artisanal fisheries (Maunder & Punt, 2004). Moreover, such models typically perform poorly under irregular data gaps, abrupt regime shifts, or nonlinear interactions between effort and environmental factors.

In recent years, **machine learning (ML)** has emerged as a transformative approach in fisheries science. ML models are particularly adept at handling nonlinearities, high-dimensional feature sets, and complex time dependencies (Zhu et al., 2019; Choi et al., 2021). Algorithms such as Random Forest (RF), XGBoost, and neural networks have been successfully applied in predicting fish distributions, estimating biomass, and modeling catch dynamics (Zhou et al., 2022; Mbaru & Obura, 2020). In East Africa, applications remain limited but growing, with Kenyan studies exploring ML for small-scale fisheries catch estimation (Mbaru et al., 2021).

This study builds on these advances by applying a **hybrid ensemble ML approach** to CPUE forecasting in Kenya. Specifically, it integrates:

1. **Random Forest feature importance** to identify dominant predictors of CPUE.
2. **Prophet**, a time-series model capable of handling seasonality and missing data.
3. **XGBoost**, a gradient boosting algorithm optimized for structured prediction tasks.

The novelty of this study lies in its **ensemble design**: combining RF feature selection with county-specific forecasting using Prophet and XGBoost, and systematically comparing their performance. This contributes to both **fisheries science** (by refining CPUE modeling approaches in artisanal systems) and **policy relevance** (by supporting Kenya's devolved fisheries management).

The objectives of this study were:

- To preprocess and structure CPUE datasets from five Kenyan coastal counties.
- To identify the most influential features using Random Forest.
- To develop and evaluate county-specific CPUE forecasts using Prophet and XGBoost.
- To compare model performance and draw implications for fisheries management.

MATERIALS AND METHODS

Study Area and Data

The study analyzed daily CPUE data spanning two years (2021–2023) from five Kenyan coastal counties. Data were collected through **routine catch assessments** conducted by county fisheries officers and the Kenya Marine and Fisheries Research Institute (KMFRI). Variables included:

- Date (year, month, week), County
- Vessel propulsion mode (motorized, non-motorized, foot-fisher, horsepower)
- Gear characteristics (gear type, mesh/hook size an number)
- Boat Activity Coefficient (BAC; fraction of active fishing days)
- Catch per unit effort (kg/fisher/day)

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Data Preprocessing

- **Missing data:** Imputed using **cubic interpolation**, which preserves smooth seasonal trends for vessel hp,
- **Outlier handling:** Extreme CPUE values were capped at the 95th percentile to moderate the influence of anomalous observations, for total catches and number of crew.
- **Categorical encoding:** Counties, propulsion modes and vessel types were one-hot encoded for compatibility with ML models.
- **Target variable:** CPUE served as the dependent or target variable.

Random Forest Feature Selection

Random Forest (Breiman, 2001) was applied to identify predictors with the strongest influence on CPUE. Feature importance was ranked using mean decrease in impurity. Across counties, BAC, vessel type, and mesh size consistently ranked highest, followed by gear type and temporal variables. These findings guided the selection of input features for forecasting models.

Forecasting Models

Prophet

Prophet (Taylor & Letham, 2018) is a time-series forecasting model developed by Facebook, designed to handle missing data and strong seasonality. Model parameters used in this study are summarized below:

Prophet Model Parameter Configurations

Parameter	Value	Description
Growth Function	Additive	Linear growth with changepoint flexibility.
Seasonality Mode	Additive	Seasonality effects added to trend.
Weekly Seasonality	TRUE	Captured intra-week fishing activity patterns.
Yearly Seasonality	TRUE	Accounted for ecological/seasonal cycles.
Daily Seasonality	FALSE	Not included, since data was aggregated beyond daily cycles.
Fourier Order	10	Flexibility in capturing seasonal fluctuations.
Changepoint Prior Scale	0.05	Controlled flexibility of trend changes.
Seasonality Prior Scale	10.0 (default)	Strength of seasonal component regularization.
Holidays/Events	None	Not specified due to lack of structured event data.
Missing Data Handling	Cubic Interpolation	Ensured continuity in CPUE, BAC, and vessel attributes.
Outlier Handling	Moderate capping	Prevented distortion from extreme CPUE peaks.
Evaluation Metrics	MAE, RMSE, MAPE, R^2	Used for model performance assessment.

XGBoost

XGBoost (Chen & Guestrin, 2016) is a gradient boosting algorithm optimized for tabular datasets. The following configuration was applied:

XGBoost Model Parameter Configurations		
Parameter	**Value**	**Description**
Objective	reg:squarederror	Minimized mean squared error for continuous targets (CPUE).
Booster	gbtree	Gradient-boosted decision trees.
Learning Rate (eta)	0.1	Shrinked feature weights to prevent overfitting.
Max Depth	6	Maximum depth of each decision tree.
Subsample	0.8	Fraction of training samples per boosting round.
Colsample_bytree	0.8	Fraction of features used per tree.
Estimators (nrounds)	500	Number of boosting iterations.
Min_child_weight	1	Minimum sum of instance weight for a child node.
Gamma	0	Minimum loss reduction for further partitioning.
L1 Regularization (α)	0	No L1 penalty (lasso).
L2 Regularization (λ)	1	Ridge penalty to reduce overfitting.
Evaluation Metric	MAE, RMSE, MAPE, R^2	Metrics used for assessing forecast accuracy.

Model Evaluation

Both models were trained on 80% of the dataset and tested on the final 20% with 30-day holdout periods. Performance metrics included:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- R^2 Score

RESULTS

Random Forest Feature Importance

RF analysis consistently highlighted **BAC, vessel type, and mesh size** as the strongest predictors across all counties. This justified their inclusion as external regressors in forecasting models. Temporal features (month, week) and county identifiers showed moderate influence, while gear type contributed marginally.

Model Performance Comparison

XGBoost outperformed Prophet in all counties (Table 1).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VII July 2026

Table 1. Model performance across counties.

Model Performance Comparison					
County	Model	MAE	RMSE	MAPE (%)	R^2
0	Prophet	0.7761	1.0769	54.91	-0.5783
	XGBoost	0.0841	0.1228	1.52	0.9988
1	Prophet	0.6607	0.9316	55.14	-0.4925
	XGBoost	0.0635	0.0851	4.61	0.9973
2	Prophet	0.7879	1.0754	93.44	-0.6148
	XGBoost	0.1524	0.2431	5.37	0.9962
3	Prophet	0.6497	0.8748	42.12	-0.6649
	XGBoost	0.0633	0.102	1.68	0.998
4	Prophet	0.9094	1.2327	53.74	-0.7604
	XGBoost	0.2198	0.2858	5.27	0.9991

XGBoost achieved **near-perfect fit ($R^2 > 0.996$)**, while Prophet often underestimated CPUE peaks and overestimated troughs.

County-Specific Observations

- **County 0:** CPUE predictions were stable; XGBoost accurately tracked fluctuations, Prophet over-smoothed.
- **County 1:** Strong seasonal cycles captured, but Prophet produced wider uncertainty intervals.
- **County 2:** Prophet failed to capture sharp spikes, whereas XGBoost aligned closely with observed values.
- **County 3:** XGBoost's fine-grained predictions reduced MAE to 0.063, while Prophet missed several peaks.
- **County 4:** Most variable series; Prophet performed poorest (negative R^2), XGBoost retained high accuracy ($R^2 = 0.999$).

DISCUSSION

This study confirms the superiority of **ensemble ML approaches** over classical time-series models in data-poor fisheries. The integration of **Random Forest feature selection and XGBoost forecasting** proved particularly powerful, yielding high-precision CPUE predictions across diverse counties.

Comparison with Previous Studies

Globally, ML applications in fisheries have shown promise. For example, Choi et al. (2021) applied gradient boosting for stock abundance prediction, while Zhou et al. (2022) highlighted ensemble models in marine resource management. In East Africa, Mbaru & Obura (2020) demonstrated the feasibility of ML in coral reef fisheries monitoring, and Mbaru et al. (2021) extended this to artisanal catch prediction in Kenya. Our findings align with these trends, reinforcing the value of ML for complex, noisy datasets.

Role of Random Forest in Feature Selection

Random Forest was critical in isolating key drivers of CPUE, especially VESSEL_LEN, NUM_CREW, VESSEL_HP, BAC, NUM_GEAR, MESH_SIZE, , NUM_HOOKS_PER_LINE, propulsion mode and mesh size. This step improved model interpretability, prevented overfitting, and provided fisheries managers with tangible variables linked to catch efficiency.

Unlike black-box ML approaches, Random Forest feature importance bridges **predictive accuracy and policy relevance**, enabling targeted interventions such as regulating mesh sizes or monitoring vessel effort in terms of hp, vessel length and number of crew.

Policy Implications

Kenya's devolved fisheries governance requires county-specific management strategies. Accurate short-term forecasts of CPUE can support:

- **Licensing and effort controls** based on anticipated resource pressure.
- **Revenue forecasting** for county governments reliant on fish levies.
- **Early warning for stock depletion** in artisanal fisheries.

By demonstrating a scalable ML framework, this study supports Kenya's broader blue economy agenda and aligns with FAO recommendations on data innovation in fisheries (FAO, 2020).

Limitations and Future Directions

The findings highlight several methodological limitations that warrant consideration. Prophet demonstrated difficulty in capturing sharp, high-variance fluctuations, indicating that its strength lies more in identifying underlying trends rather than delivering precise short-term forecasts. In contrast, XGBoost achieved superior predictive accuracy but at the expense of interpretability, a drawback relative to regression-based approaches that provide clearer insights into feature contributions.

Future research should address these gaps by;

- Investigating advanced deep learning models such as NeuralProphet and Long Short-Term Memory (LSTM) networks to improve long-term CPUE forecasting.
- Incorporating environmental covariates (e.g., sea surface temperature and chlorophyll concentration) as external regressors to enhance predictive performance.
- Expanding ensemble modeling frameworks to cover broader Western Indian Ocean (WIO) datasets, thereby supporting transboundary fisheries stock assessment and management.

CONCLUSION

This study demonstrates the effectiveness of **ensemble machine learning approaches** in forecasting CPUE for Kenya's artisanal marine fisheries. By leveraging Random Forest for feature selection and XGBoost for nonlinear forecasting, it achieved unprecedented accuracy across all five coastal counties. Prophet provided complementary insights on seasonality but was less suited for short-term predictions.

The methodological framework is **scalable, replicable, and policy-relevant**. It addresses key challenges in data-poor fisheries, offering both predictive precision and managerial interpretability. Moving forward, integrating additional ML models and environmental predictors will further enhance forecasting capacity, strengthening the scientific foundations of fisheries governance in Kenya and the wider Western Indian Ocean region.

REFERENCES

1. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
3. Choi, J., Park, H., & Lee, J. (2021). Machine learning applications for predicting fish stock abundance. Ecological Informatics, 63, 101312.
4. FAO (2020). The State of World Fisheries and Aquaculture 2020. Rome.
5. Hilborn, R., & Walters, C. (1992). Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty. Springer.
6. KMFRI (2020). Status of the Kenya Marine Fisheries 2019/2020 Report.
7. Mbaru, E. K., & Obura, D. (2020). Machine learning approaches to coral reef fisheries monitoring. Marine Ecology Progress Series, 634, 1–15.
8. Mbaru, E. K., Ochiewo, J., & Obura, D. (2021). Predicting artisanal fisheries catches in Kenya using machine learning. African Journal of Marine Science, 43(2), 203–214.
9. Maunder, M. N., & Punt, A. E. (2004). Standardizing catch and effort data: A review of recent approaches. Fisheries Research, 70(2–3), 141–159.
10. Obura, D. O., et al. (2017). Coral reef fisheries management in Kenya: Successes, challenges, and policy directions. Marine Policy, 81, 81–92.
11. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.
12. Wanyonyi, I., Mangi, S. C., & Catchpole, T. L. (2018). Socioeconomic impacts of marine artisanal fisheries in Kenya. Ocean & Coastal Management, 163, 372–381.
13. Zhou, C., Sun, S., & Xu, Y. (2022). Ensemble learning for marine resource prediction: A review and case study. Ecological Modelling, 464, 109819.
14. Zhu, G., Zhang, H., & Wang, S. (2019). Machine learning in fisheries science: Advances and applications. Frontiers in Marine Science, 6, 662.