Machine Learning-Based Fish Species Recommendation Using Water Quality Parameters

Muhammad Owais Khan[1], Faheem Ul Haq[1], Aasif Awan[2]
[1]Robotics Team Hadaf Group of Colleges Peshawar, Pakistan
[2]Lecturer Biotechnology Hadaf College of Allied Health Sciences, Peshawar, Pakistan
*Correspondence:mowaiskhandev@gmail.com,engr.faheemulhaq@gmail.com, awan.aasif1@gmail.com

Citation | Khan. M. O, Haq. F. U, Awan. A, "Machine Learning-Based Fish Species Recommendation Using Water Quality Parameters", IJIST, Vol. 07 Special Issue. pp 110-126, May 2025 DOI:https://doi.org/10.33411/IJIST/1303
Received | April 10, 2025 **Revised |** May 03, 2025 **Accepted |** May 07, 2025 **Published |** May 09, 2025.

The integration of machine learning (ML) in aquaculture enables data-driven fish species recommendations based on water quality parameters. Traditional fish farming faces challenges like manual monitoring, inefficient species selection, and unpredictable water conditions, leading to economic losses. This paper presents a software-based fish recommendation system using ML models to analyze seven key water parameters: pH, Temperature, Turbidity, TDS, Dissolved Oxygen, Nitrate, and Ammonia. Various ML algorithms, including Random Forest, XGBoost, and SVM, were evaluated, with the optimized model achieving over 90% accuracy. A graphical user interface (GUI) allows users to input parameters and receive real-time recommendations, enhancing efficiency and sustainability in aquaculture.

Keywords: Fish Farming; Machine Learning; Water Quality Analysis; XGBoost; Smart Aquaculture.

Introduction:

Aquaculture has become an essential component of global food systems, contributing significantly to food security, nutrition, and economic development. As demand for fish continues to rise, modernizing aquaculture practices has become crucial. However, traditional fish farming remains largely dependent on manual water quality monitoring and farmer intuition, which often results in inefficient operations, inaccurate species selection, and vulnerability to environmental changes. These challenges can lead to poor yields, increased costs, and avoidable losses. Water quality parameters such as pH, temperature, total dissolved solids (TDS), turbidity, ammonia, dissolved oxygen, and nitrate directly influence fish health, growth, and survival. Monitoring these parameters manually is not only labor-intensive but also lacks the responsiveness required for real-time decision-making, especially in large-scale farming systems. To address these limitations, this study proposes a machine learning-based fish species recommendation system that predicts the most suitable species for a given aquatic environment. The objective is to support aquaculture decision-making by analyzing real-time water quality data using a variety of machine learning algorithms, including Random Forest, Decision Tree, XG-Boost, K- Nearest Neighbors, Support Vector Machine, and Logistic Regression. The system integrates preprocessing techniques such as feature scaling and dataset balancing to enhance prediction accuracy. In addition, a graphical user interface (GUI) was developed to allow farmers and aquaculture professionals to input water parameters and receive instant fish species recommendations. By automating and optimizing the species selection process, this system aims to improve the efficiency and sustainability of aquaculture operations. The novelty of this study lies in the use of additional water quality parameters not originally present in the dataset, such as TDS, dissolved oxygen, ammonia, and nitrate, generated through synthetic data. This, along with addressing class imbalance using SMOTE, improves the model's generalizability. By automating and optimizing the species selection process, this system aims to improve the efficiency and sustainability of aquaculture operations. The remainder of this paper includes a review of Literature review, a detailed methodology, results and performance analysis, discussion of findings and limitations (Section 5), and conclusions.

Literature Review:

In recent years, the integration of Internet of Things (IoT) technologies and machine learning (ML) techniques in aquaculture has garnered considerable attention. Numerous studies have focused on real-time water quality monitoring; however, the majority emphasize environmental assessment rather than intelligent fish species recommendation. This section reviews prior work on IoT-based monitoring systems and AI-driven fish species selection, identifying critical gaps that the present study aims to address.

IoT-Based Water Quality Monitoring Systems:

Several studies have proposed IoT-based solutions for continuous monitoring of aquaculture environments. These systems generally consist of sensor networks, cloud-based data storage, and remote access functionalities. However, most lack intelligent decision-making features for species recommendation. For instance, Cordova Rozas et al. presented a cloud-integrated water monitoring framework comprising five stages: data acquisition, cloud storage, database management, report generation, and prediction. Despite its effectiveness in water quality monitoring, the system does not include species-specific recommendations[1]. Gao et al. developed an IoT-enabled fish farming system focused on continuous monitoring and fish movement tracking, without offering guidance on species selection [2]. Similarly, Nagayo et al. designed a solar-powered aquaponics setup with Arduino-based temperature control but did not incorporate AI-based recommendations [3]. Pasika et al. proposed a cost-effective IoT monitoring system that measures temperature, pH, turbidity, humidity, and water level, yet lacked intelligent decision-making capabilities

[4]. Huan et al. introduced an NB-IoT-based system for real-time data acquisition to enhance aquaculture efficiency, but again, omitted species recommendation features [5].

Key Limitations in IoT-Based Systems:

Lack of intelligent decision-making: Most systems only monitor water quality without suggesting suitable fish species.

Absence of predictive analytics: Few studies offer insights into how water quality fluctuations affect fish health.

Minimal AI integration: Many systems focus solely on data collection without employing ML for species prediction.

Poor user experience: Several solutions lack user-friendly graphical interfaces tailored for farmers with limited technical expertise.

AI-Based Water Quality and Fish Species Recommendation Systems:

Though machine learning is increasingly applied in aquaculture, few studies have developed models for intelligent fish species recommendation. Most efforts are directed towards water quality assessment or predicting fish survival, without providing actionable recommendations based on real-time data. For example, Chiu et al. implemented an IoT-aided aquaculture framework using deep learning to predict growth patterns in California bass, focusing primarily on feeding behaviors rather than species suitability [6]. Niswar et al. utilized MQTT and LoRa-based sensor networks for real-time crab farming monitoring but did not include species prediction capabilities[7]. Billah et al. introduced smart instrumentation for water quality assessment, yet lacked recommendation functionality [8]. Uddin et al. developed a survival prediction model using Random Forest based on environmental parameters like pH, temperature, and turbidity. While insightful, the model did not utilize real-time data or suggest optimal fish species[9]. Abinaya et al. created an IoT-based monitoring system using Arduino and GSM modules, sending SMS alerts for water quality issues but omitting any fish recommendation logic [10]. Islam et al. developed a fish species prediction model using J48, KNN, Random Forest, and CART algorithms. Although the Random Forest model achieved 88.48% accuracy, it excluded essential parameters such as TDS, DO, nitrate, and ammonia [11]. Hemal et al. introduced Aqua Bot, an IoT-enabled water monitoring solution powered by ML, but it lacked a comprehensive species recommendation component and considered only a limited number of parameters [12].

Key Limitations in AI-Based Systems:

Limited parameter integration: Most models incorporate only 3–4 parameters, reducing predictive robustness.

Lack of real-time analytics: Few models are capable of forecasting water quality changes and their impacts.

Lack of real-time analytics: Few models are capable of forecasting water quality changes and their impacts

Limited sustainability: Despite the occasional mention of solar-powered components, their integration into intelligent systems is minimal.

Non-intuitive interfaces: Many AI-based models are not paired with graphical user interfaces, limiting usability for non-technical users.

Identified Research Gaps:

From the reviewed literature, several key gaps emerge:

Insufficient AI-driven decision support: Existing systems focus on monitoring rather than intelligent species selection.

Limited environmental parameter scope: Models often neglect critical factors like nitrate, ammonia, TDS, and DO.

Lack of real-time predictive capability: There is a noticeable absence of models that integrate real-time data streams for forecasting species suitability.

Sustainability concerns: Few studies incorporate renewable energy solutions for long-term operation.

Usability limitations: Many systems do not offer interactive GUIs, hindering accessibility for fish farmers.

Research Contribution:

To address these gaps, this study proposes a machine learning-based fish species recommendation system that: Utilizes seven key water quality parameters (pH, Temperature, Turbidity, TDS, DO, Nitrate, Ammonia), Implements advanced ML algorithms (Random Forest, XG- Boost, SVM, KNN, etc.). Achieves over 90% accuracy in predicting optimal fish species, Offers an intuitive graphical user interface (GUI) for easy user interaction and decision-making.

Methodology:

This section outlines the methodology employed in the development of the machine learning-based fish species recommendation system. It covers the dataset, preprocessing techniques, selection of machine learning models, evaluation metrics, and the development of a user-friendly graphical interface.

Water Quality Parameters:

To ensure accurate and context-specific fish species recommendations, the system utilizes seven essential water quality parameters: temperature, turbidity, pH, total dissolved solids (TDS), dissolved oxygen (DO), ammonia, and nitrate. These parameters are widely recognized in aquaculture research for their significant influence on fish health, growth, and survival.

Each fish species thrives within specific environmental conditions. Therefore, deviations from optimal ranges can result in stress or mortality. Our recommendation system evaluates these parameters to identify the most suitable fish species for a given aquatic environment Table 1 summarizes the standard reference ranges for each water quality parameter, which serve as the baseline for model training and decision-making.

By incorporating a comprehensive set of water quality indicators, the system aims to offer a more holistic and intelligent recommendation approach, moving beyond simple monitoring to enable informed aquaculture management decisions.

Table 1. Optimal ranges of water quality parameters

Water Quality Parameter	Value
Temperature	25◦C–32 ◦C or >20 ◦C
pH	6.5–8.5
Turbidity	30–80 cm
Dissolved Oxygen (DO)	>5 mg/L
Total Dissolved Solids (TDS)	400 mg/L
Nitrate	0–100
Ammonia	0–0.2

In this section, we present the methods employed in developing the machine learning-based fish species recommendation system. It includes details about the dataset, preprocessing techniques, machine learning models, performance evaluation metrics, and the development of the graphical user interface (GUI).

We have taken into consideration seven key water quality parameters to determine the suitability of water for different fish species. These parameters are temperature, turbidity, pH, total dissolved solids (TDS), dissolved oxygen (DO), ammonia, and nitrate. Based on these parameters, our system provides fish species recommendations that are best suited for the given water conditions. The reference values for these parameters are presented in Table 1. Any significant deviation from the optimal ranges can negatively affect fish health and survival, making it essential to select the right species for specific water conditions.

Each of these parameters is essential in evaluating water suitability for fish farming:

- Temperature affects fish metabolism, growth, and oxygen availability. Extreme temperatures can cause stress and even lead to fish mortality.
- pH influences water acidity or alkalinity, affecting fish health. Most fish species thrive within a pH range of 6.5 to 8.5, while extreme values can be harmful or fatal.
- Turbidity indicates water clarity. High turbidity can reduce light penetration, disrupt photosynthesis, and create unfavorable conditions for fish.
- Dissolved Oxygen (DO) is critical for fish survival, with levels dropping below 5 mg/L potentially leading to stress and reduced growth.
- TDS (Total Dissolved Solids) represents minerals, salts, and organic matter in water. High TDS levels can disrupt fish osmoregulation.
- Nitrate and Ammonia are nitrogen-based compounds. High nitrate levels (>100 mg/L) can lead to excessive algal growth, while ammonia is toxic to fish even at low concentrations (>0.2 mg/L).

Since each of these parameters plays a critical role in fish health and survival, feature selection was not applied before model training. All seven parameters were deemed essential for accurately assessing water quality and determining the suitability of water for various fish species. Omitting any of these parameters could lead to a loss of important information, potentially compromising the accuracy of the species recommendations.

Data Collection:

This study utilizes the Real-Time Pond Water Dataset for Fish Farming, sourced from Kaggle [13], originally collected by the Faculty of Fisheries at the University of Dhaka, Bangladesh. The dataset consists of 591 samples with four primary features: temperature, turbidity, pH, and fish species. The independent variables pH (91 unique values), temperature (51 unique values), and turbidity (108 unique values) reflect diverse water quality conditions. The dependent variable includes 11 fish species such as Katla, Song, Prawn, Rui, Koi, Pangas, Tilapia, Silver Carp, Karpio, Magur, and Shrimp.

To extend the dataset and enable more accurate fish species predictions, additional water quality parameters Total Dissolved Solids (TDS), Dissolved Oxygen (DO), Ammonia, and Nitrate were incorporated through synthetic data generation. These parameters were not originally present in the dataset but are widely recognized as essential indicators for aquaculture suitability[14][15]. The absence of these features could limit the model's ability to generalize across realistic aquaculture scenarios. Synthetic values were generated based on standard ranges reported in aquaculture literature. For instance, DO values were sampled between 4 and 10 mg/L, which supports optimal respiration for most freshwater fish species. TDS values were generated between 200 and 500 mg/L, ammonia between 0 and 1.5 mg/L, and nitrate between 0 and 50 mg/L—ranges consistent with those used in fish farming management guidelines [14]. A practical example: for a given row with pH 7.8, temperature 29°C, and turbidity 7 NTU (favoring Tilapia), a synthetic DO value of 6.9 mg/L and TDS value of 350 mg/L were generated using a uniform distribution in Python. These values were generated using *numpy.random.uniform()* to reflect real-world variability while ensuring ecological validity. By enhancing the dataset with realistic synthetic data, the feature set was expanded, resulting in improved model learning and robustness. The distribution of samples by fish species is provided in Table 2.

To further improve the dataset and address the issue of limited sample size and class imbalance, we employed the Synthetic Minority Over-sampling Technique (SMOTE) to generate additional synthetic samples. This process increased the total number of samples from 591 to 1,320, which helped balance the dataset and improve model generalization.

Although the system enables real-time predictions through manual user input of

water parameters, it does not integrate automatic real-time data acquisition from IoT sensors or cloud-based sources. The current approach relies on users manually entering measured parameters via the GUI for fish species recommendation. Future enhancements could focus on connecting the system to real-time water monitoring devices for fully automated operation.

Table 2 presents the distribution of fish species in the dataset, revealing a noticeable class imbalance. Tilapia (129), Rui (99), and Pangas (78) are the most represented species, making them the majority classes. In contrast, species like Koi (15), Prawn (14), and Magur (11) have significantly fewer samples, making them minority classes. This imbalance can lead to biased model predictions, where the classifier favors majority classes while misclassifying underrepresented species. To address this issue and improve prediction accuracy for all fish species, techniques such as SMOTE are essential for balancing the dataset.

Table 2. Distribution of fish species in the dataset

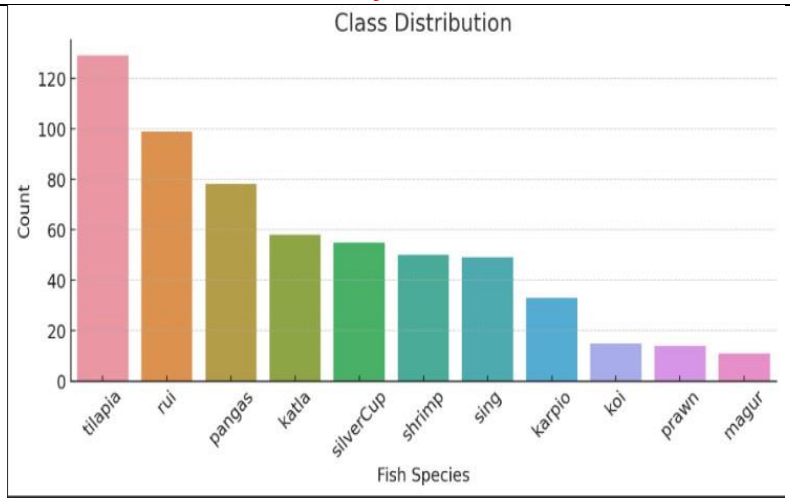
Class Label	Number
Tilapia	129
Rui	99
Pangas	78
Katla	58
Silver Cup	55
Shrimp	50
Sing	49
Karpio	33
Koi	15
Prawn	14
Magur	11

Table 2 presents the distribution of fish species in the dataset, revealing a noticeable class imbalance. Tilapia (129), Rui (99), and Pangas (78) are the most represented species, making them the majority classes. In contrast, species like Koi (15), Prawn (14), and Magur (11) have significantly fewer samples, making them minority classes. This imbalance can lead to biased model predictions, where the classifier favors majority classes while misclassifying underrepresented species. To address this issue and improve prediction accuracy for all fish species, techniques such as SMOTE are essential for balancing the dataset.

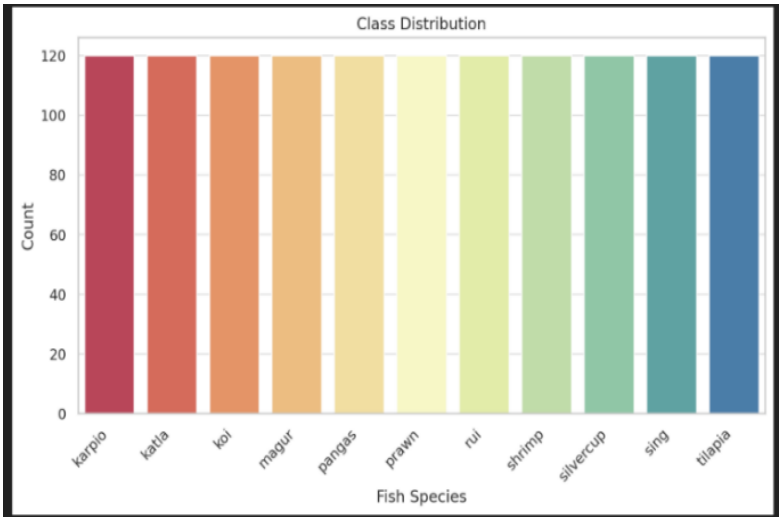
Data Preprocessing:

To ensure the reliability of model predictions, it is essential to preprocess the dataset by handling missing values, addressing class imbalance, and normalizing data. The following section details these preprocessing techniques and their impact on model performance.

Handling Missing Data: Any missing values in the dataset were addressed using mean or median imputation to maintain data integrity. Splitting the Dataset: The dataset was divided into two sets: 80% for training data and 20% for testing data, allowing an effective evaluation of the model's predictive performance. Balancing the Dataset: The dataset was checked for class imbalance, and SMOTE (Synthetic Minority Oversampling Technique) was applied. This method generates synthetic samples for underrepresented classes, improving model performance on imbalanced data. These preprocessing steps ensure that the dataset is well-prepared for model training, minimizing biases and improving predictive accuracy. By handling missing values, we maintain data consistency, while proper dataset splitting allows the model to generalize effectively. Additionally, balancing the dataset using SMOTE helps address the issue of class imbalance, ensuring that minority classes are adequately represented in training. These refinements collectively enhance the robustness of the machine-learning model, leading to more reliable fish species recommendations.



a. Before SMOTE



b. After SMOTE

Figure 1. Distribution of fish species before and after applying SMOTE.

Figure 1 illustrates the distribution of fish species before and after applying the Synthetic Minority Over-sampling Technique (SMOTE). In Figure 1(a), the dataset is imbalanced, with certain fish species, such as Tilapia and Rui, having significantly more samples compared to others like Prawn and Magur. This imbalance can negatively impact the performance of machine learning models by causing them to favor majority classes. In Figure 1(b), after applying SMOTE, the dataset is balanced, meaning all fish species have an equal number of samples. SMOTE achieves this by generating synthetic data for the underrepresented classes, improving model performance by reducing bias and enhancing generalization.

A.A. Implemented Machine Learning Algorithm This study utilizes six machine-learning models to predict the most suitable fish species based on various water quality parameters. The implemented models include Random Forest (RF), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression.

1. Random Forest (RF): Configured with 100 decision trees and a random state of 42. Hyperparameters were manually tuned by iteratively adjusting parameters based on validation performance.

2. Extreme Gradient Boosting (XGBoost): Configured with optimized hyper-parameters using GridSearchCV, employing a learning rate of 0.1, max depth of 6, and 100 estimators.

3. K-Nearest Neighbors (KNN): The optimal k- k-value was determined as seven after testing multiple values.
4. Support Vector Machine (SVM): Configured with a regularization parameter (C) of 10 and a gamma value of 0.1. These values were manually selected after trial-and-error testing to improve model performance, with a random state of 42 for reproducibility.
5. Decision Tree (DT): Configured with a maximum depth of 5 and a random state of 42. Hyperparameters were selected manually through iterative testing on validation data.
6. Logistic Regression (LR): Configured with L2 regularization (Ridge), a learning rate of 0.1, and a random state of 42. Hyperparameters were manually adjusted based on validation performance.

These models were chosen for their effectiveness in accurately predicting fish species based on water quality parameters, and they were fine-tuned using hyperparameter optimization techniques to improve predictive accuracy. The selection of the final model was determined by comparing performance metrics, including accuracy, precision, recall, and F1 score.

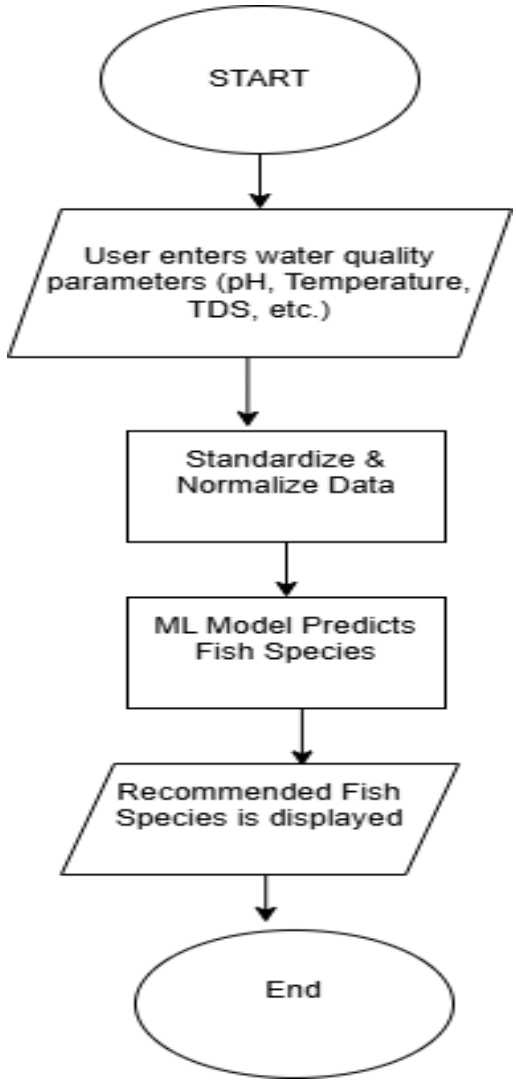


Figure 2. Flowchart of ML Model.

Figure 2 (Flow Diagram) illustrates a machine learning-based fish species recommendation system developed to assist in aquaculture. The process starts with the user entering key water quality factors such as pH, temperature, turbidity, total dissolved solids (TDS), dissolved oxygen, nitrate, and ammonia. These values undergo standardization and

normalization to maintain consistency and improve model accuracy. After preprocessing, the data is provided to a trained ML model, which analyzes the parameters and predicts the most appropriate fish species for aquaculture. The system processes this information and displays the recommended fish species to the user through a user-friendly interface. By leveraging machine learning, this system enhances decision-making in aquaculture, reduces reliance on manual monitoring, improves efficiency, and promotes sustainable fish farming practices.

Performance Evaluation:

The trained models were evaluated using the following metrics:

Accuracy: The percentage of correctly classified instances.

Precision – The ratio of correctly predicted positive instances to the total predicted positive instances.

Recall: The ratio of correctly predicted positive instances to total actual positive instances.

F1-Score: The harmonic mean of precision and recall, offering a balanced assessment of model performance.

Among all models, XGBoost achieved the highest accuracy, exceeding 90%, and was selected for final deployment.

GUI Design:

A simple GUI-based application was developed to take seven water quality parameters, i.e. pH, Temperature, Turbidity, TDS, Dissolved Oxygen, Nitrate, and Ammonia, as input. The trained machine-learning model processes these inputs and provides a fish species recommendation. The interface ensures easy data entry and quick predictions for users.

Results:

Evaluation Metrics for Model Performance:

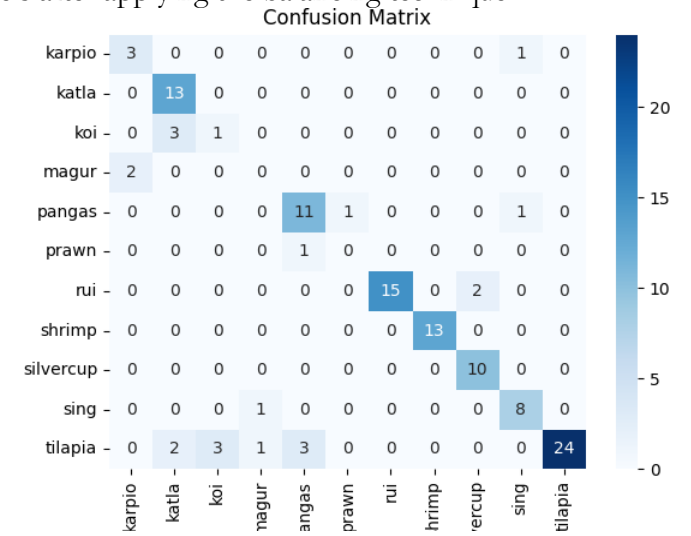
To assess the performance of the classification models, several key evaluation metrics were employed. One of the most fundamental tools in classification tasks is the confusion matrix, which provides a comparison between predicted and actual values. The confusion matrix consists of four essential components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP represents the number of positive instances correctly identified, whereas FP denotes negative instances incorrectly classified as positive. Similarly, TN refers to correctly identified negative instances, while FN represents positive instances that were misclassified as negative. From this matrix, multiple performance metrics are derived to measure model effectiveness.

- **Accuracy** determines the percentage of correctly classified instances in the dataset.
- **Precision** evaluates the proportion of correctly predicted fish species among all positive predictions.
- **Recall (Sensitivity)** measures the model's ability to correctly identify fish species when they are present.
- **F1-Score** represents the harmonic mean of Precision and Recall, providing a balance between the two.
- **Matthews Correlation Coefficient (MCC)** assesses overall classification quality by incorporating all confusion matrix components.
- **ROC-AUC** visualizes the trade-off between TPR and FPR across thresholds, using a one-vs-all strategy for multi-class classification. A higher AUC score indicates better model performance.

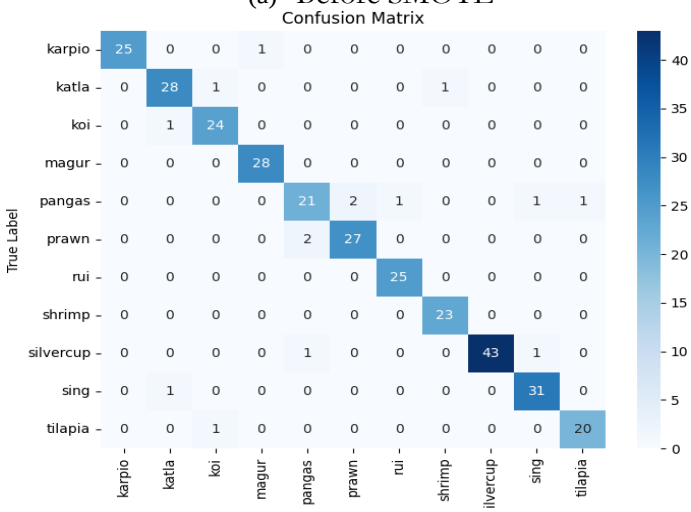
Machine Learning Model Result Comparison:

We applied six ML algorithms to predict suitable fish for individual ponds based on several parameters, including accuracy, precision, recall, F1-score, and the ROC curve. We first

evaluated the performance of these models without balancing the dataset. Afterward, we also evaluated the models after applying the balancing technique.



(a) Before SMOTE



(b) After SMOTE

Figure 3. Confusion matrix of ensemble model before and after SMOTE.

Figure 3(a), Before SMOTE, highlights the class imbalance in the dataset. The model performs well for majority classes like Tilapia (24), Rui (15), and Shrimp (13), but struggles with minority classes such as Prawn and Magur, leading to frequent misclassifications. Some classes, like Karpio and Koi, also show lower prediction counts. This imbalance affects the overall performance and justifies the need for SMOTE to enhance prediction accuracy for underrepresented fish species.

Figure 3(b), After SMOTE, demonstrates a significant improvement in classification performance across all fish species. Unlike the "Before SMOTE" matrix, where minority classes were often misclassified, this matrix shows more balanced predictions, indicating the effectiveness of SMOTE in addressing class imbalance.

Key improvements:

- Better classification of previously underrepresented species (e.g., Magur, Prawn, and Pangas now have strong diagonal values).
- Fewer misclassifications across all classes, indicate that the model has learned better decision boundaries.

- Higher overall accuracy, as false positives and false negatives have been reduced compared to the previous model.

Table 3. ML algorithm performance evaluation before smote.

Without Smote					
ALGORITHM	**ACC**	**PRE**	**REC**	**F1**	**MCC**
XG	0.79	0.83	0.8	0.8	0.77
RF	0.77	0.78	0.77	0.77	0.74
DT	0.75	0.79	0.76	0.76	0.73
SVM	0.59	0.61	0.6	0.58	0.54
KNN	0.49	0.53	0.5	0.5	0.42
LR	0.52	0.5	0.53	0.51	0.45

Table 3 presents the performance comparison of various machine-learning algorithms without applying SMOTE. The results indicate that **XGBoost (XG)** achieved the highest performance, with 79% accuracy, 83% precision, 80% recall, 80% F1-score, and 77% MCC, making it the best-performing model. **Random Forest (RF)** followed closely, obtaining 77% accuracy, 78% precision, 77% recall, 77% F1-score, and 74% MCC. The **Decision Tree (DT)** model performed moderately well, achieving 75% accuracy, whereas **SVM, KNN**, and **Logistic Regression (LR)** demonstrated lower performance levels, with **LR** showing the weakest results (52% accuracy, 50% precision, 53% recall, 51% F1-score, and 45% MCC). These results suggest that ensemble models like **XGBoost** and **Random Forest** outperform other models in this scenario.

Before applying SMOTE, the results from Table 3 demonstrate that XGBoost outperforms all other models, achieving the highest accuracy and balanced performance across metrics such as precision, recall, and F1-score. Random Forest follows closely, offering strong performance as well. However, models like SVM, KNN, and Logistic Regression lag, with Logistic Regression showing particularly poor results. These findings suggest that more complex models like XGBoost and Random Forest are better suited for the classification task, while simpler models struggle to handle the complexities of the dataset.

Table 4. ML algorithms performance evaluation after smote.

After Applying Smote					
Algorithm	**ACC**	**PRE**	**REC**	**F1**	**MCC**
XG	0.968	0.968	0.968	0.968	0.968
RF	0.966	0.970	0.970	0.970	0.960
DT	0.902	0.900	0.900	0.900	0.890
SVM	0.842	0.850	0.840	0.840	0.830
KNN	0.840	0.845	0.840	0.840	0.830
LR	0.644	0.650	0.650	0.620	0.610

Table 4 presents the performance metrics of various machine learning models after applying SMOTE, the XGBoost (XG) model achieved the highest performance, with 96.8% accuracy, 96.8% precision, 96.8% recall, a 96.8% F1-score, and a 96.8% MCC. The Random Forest (RF) model followed closely, achieving 96.6% accuracy, 97% precision, 97% recall, a 97% F1-score, and a 96% MCC. Among other models, the Decision Tree (DT) achieved 90.2% accuracy, while the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) performed similarly, with 84.2% and 84% accuracy, respectively. Logistic Regression (LR) exhibited the lowest performance, with 64% accuracy, 65% precision, 65% recall, 62% F1-score, and 61% MCC.

The ROC curve was also utilized to compare model performance, applying the One-vs-Rest method for multi-class classification. The XGBoost model demonstrated the best differentiation between classes, confirming its effectiveness in fish species classification. These

results indicate that XGBoost is the most effective model for fish species classification after SMOTE, followed by Random Forest, while Logistic Regression remains the least effective for handling this classification problem.

Figure 4 (a: ROC Curve for XGBoost) illustrates the performance of the XGBoost model before applying SMOTE. The ROC curve represents the trade-off between the true positive rate and the false positive rate for each fish species classification. While the model performs well for certain species, achieving AUC values close to 1.0, other species exhibit lower AUC scores, indicating difficulties in accurate classification. This discrepancy arises due to the imbalanced nature of the dataset, where some fish species have significantly fewer samples compared to others. The model struggles to learn distinctive patterns for these minority classes, leading to misclassifications. This demonstrates the limitations of training on an imbalanced dataset, as the model tends to favor the majority classes, reducing its overall reliability in classifying underrepresented species. The impact of this imbalance will be analyzed further in Figure 4 after the application of SMOTE to assess improvements in classification performance.

Figure 5 (a: ROC Curve for XGBoost) displays the ROC curve of the XGBoost model after applying SMOTE to balance the dataset. Compared to Figure 4, the ROC curves show a noticeable improvement in classification performance across different fish species. The AUC values for previously underrepresented classes have increased, indicating that the model is now better at distinguishing between species. This improvement results from SMOTE generating synthetic samples for minority classes, allowing the model to learn more patterns that are representative. By addressing class imbalance, SMOTE reduces the model's tendency to favor majority classes, leading to a more balanced classification performance.

The enhanced ROC curve demonstrates the effectiveness of oversampling in improving model generalization, ensuring more reliable predictions across all fish species. The comparative analysis of ROC curves for various models further highlights the effectiveness of SMOTE in improving classification performance. While XGBoost and Random Forest show significant enhancements in their AUC scores after addressing the class imbalance, models like Decision Tree and SVM exhibit moderate improvements. However, Logistic Regression still struggles to differentiate between fish species due to its linear nature, reinforcing its limitations in handling complex patterns These findings emphasize the importance of selecting robust machine learning models and applying appropriate preprocessing techniques to ensure accurate and reliable fish species classification. The following figures illustrate the ROC curves of different models, providing a visual representation of their classification capabilities before and after the SMOTE application.

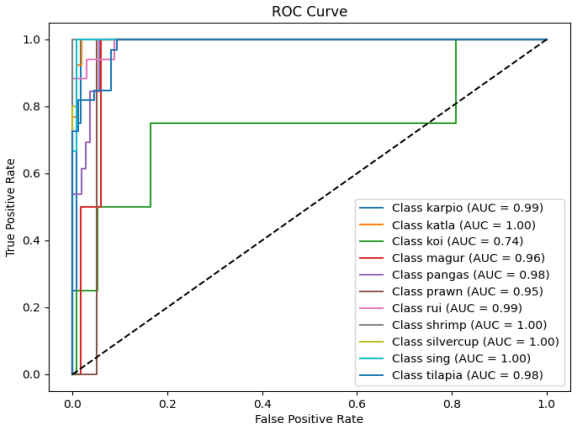
The results shown in Figure 4 and Figure 5 emphasize the pivotal role that data preprocessing plays in improving model performance. The marked improvement in AUC scores after applying SMOTE not only validates the effectiveness of oversampling techniques but also underscores the importance of handling class imbalances in machine learning tasks. The enhanced ROC curve for XGBoost after the SMOTE application confirms that this method can mitigate the risks of overfitting to majority classes, resulting in better generalization across all species. The improvement in AUC for minority classes is particularly notable, as it highlights how SMOTE's synthetic samples allow the model to capture a broader range of patterns and distinguish more accurately between different species.

In contrast, models like Logistic Regression, which rely on linear decision boundaries, continue to struggle with the complexities of multi-class classification, particularly in imbalanced datasets. The results indicate that Logistic Regression is not well-suited for problems involving intricate, non-linear patterns, such as those found in fish species classification. Although Logistic Regression performs reasonably in simpler scenarios, its limitations become evident when applied to more complex, imbalanced datasets, as seen in

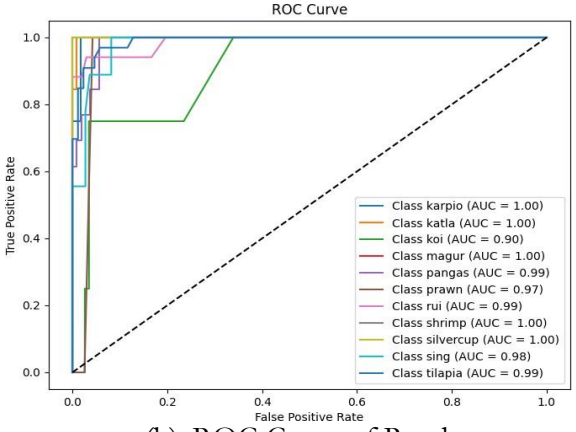
the lower AUC scores for various fish species.

This comparative analysis reveals a critical insight into model selection: while SMOTE can significantly improve performance, it cannot fully compensate for the limitations of less complex models like Logistic Regression. Thus, selecting an appropriate model, such as XGBoost or Random Forest, becomes crucial when addressing classification problems involving imbalanced and complex datasets. These models, coupled with techniques like SMOTE, offer the best approach to ensuring reliable predictions across all classes, particularly in challenging scenarios such as fish species classification.

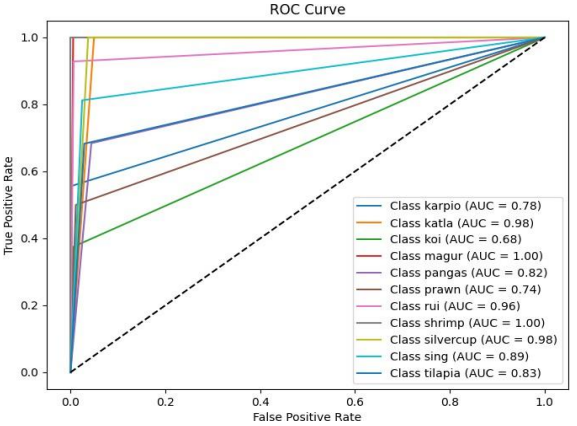
The following figures further demonstrate the impact of various preprocessing techniques and models on classification performance. Through a detailed analysis of ROC curves, the results advocate for the importance of both the right choice of model and preprocessing method in enhancing classification reliability, especially in multi-class classification tasks.



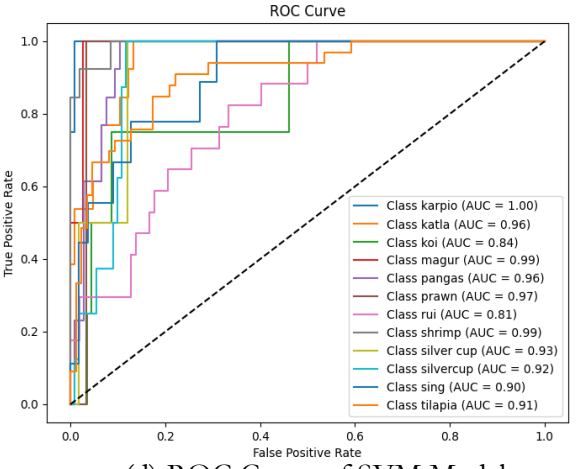
(a) ROC Curve of XGBoost Model



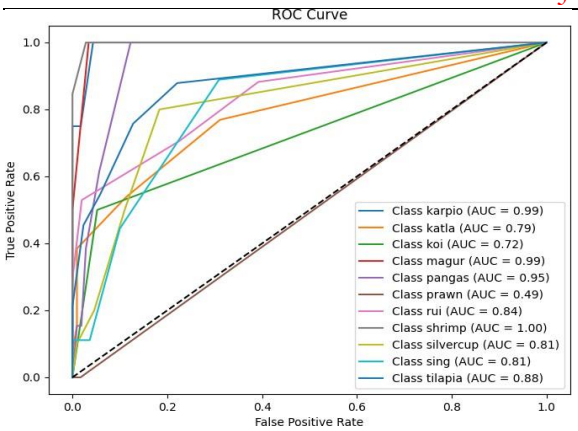
(b) ROC Curve of Random Forest Model



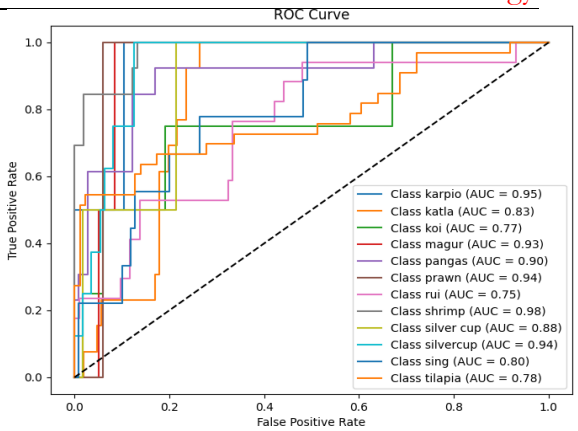
(c) ROC Curve of Decision Tree Model



(d) ROC Curve of SVM Model

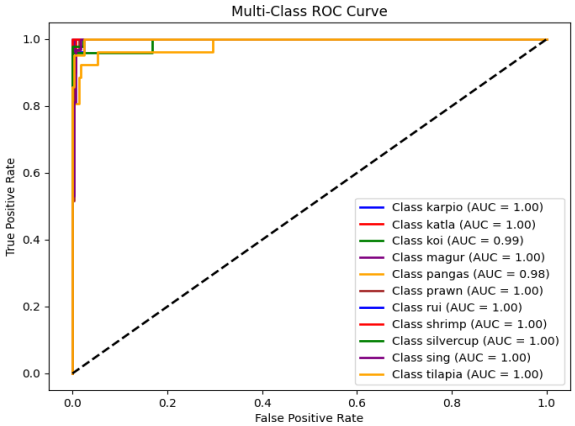


(e) ROC Curve of KNN Model

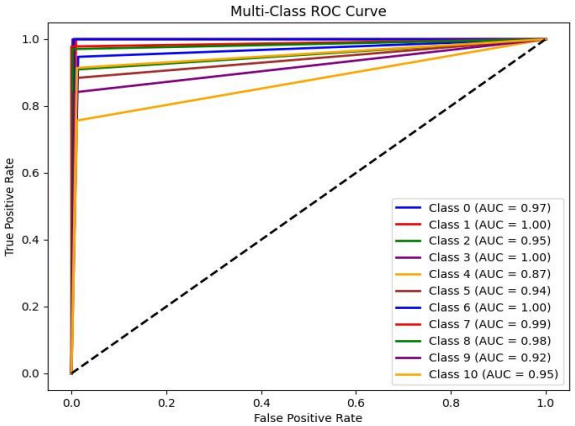


(f) ROC Curve for the LR Model

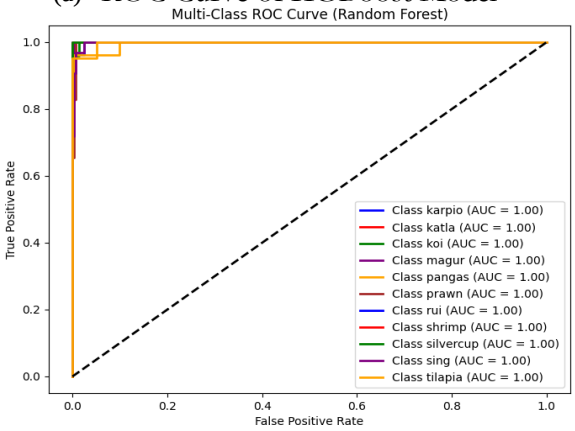
Figure 4 ROC Curves of Different ML Algorithms Before Applying SMOTE.



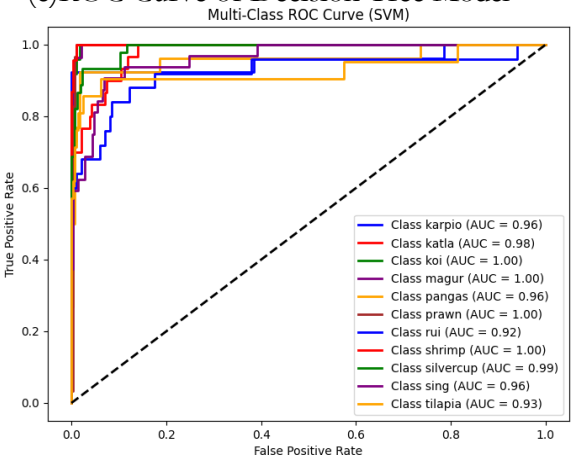
(a) ROC Curve of XGBoost Model



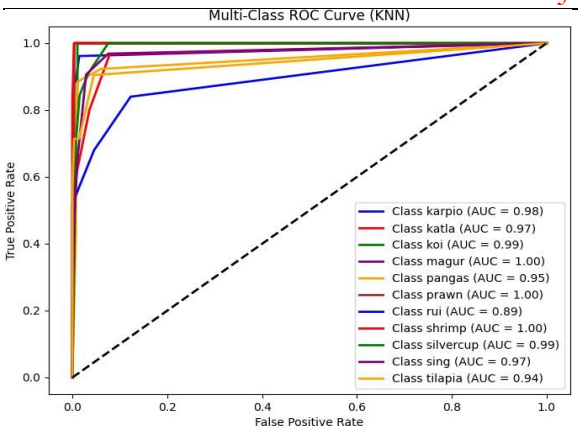
(c)ROC Curve of Decision Tree Model



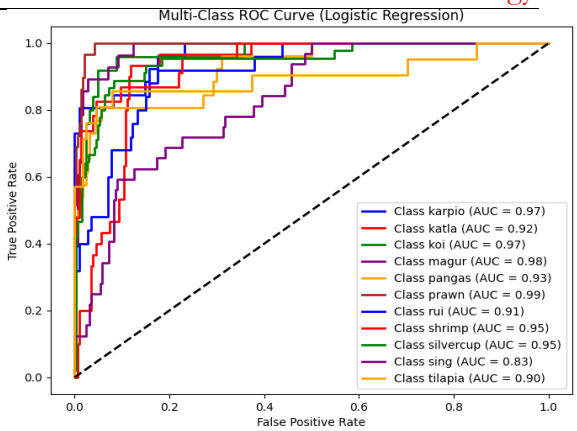
(b) ROC Curve of LR Model



(d) ROC Curve of SVM Model



(e) ROC Curve of KNN Model



(f) ROC Curve of LR Model

Figure 5. ROC Curves of Different ML Algorithms After Applying SMOTE.

Fish Recommendation System GUI:

The Graphical User Interface (GUI) of the fish species recommendation system offers an intuitive and user-friendly platform for inputting water quality parameters and obtaining classification results. Designed with ease of use in mind, the GUI enables users to input values such as pH, temperature, turbidity, TDS, dissolved oxygen, nitrate, and ammonia. These inputs are then processed using the developed machine-learning model. Upon submission, the system predicts the most suitable fish species based on the provided parameters and displays the results in a clear and readable format. Figure 6 illustrates the GUI, showcasing its layout and functionality, which ensures seamless user interaction.

The GUI is built using Python and Tkinter, providing a simple yet effective framework for developing desktop applications. The machine learning model is integrated through scikit-learn, enabling real-time classification of water quality data. Pandas is utilized for efficient data manipulation, while Matplotlib is employed for visualizing the prediction results. To enhance usability, PyInstaller is used to convert the Python script into a standalone executable (.exe) file. This ensures that users can run the application without needing to install Python or any dependencies locally, offering a smooth and hassle-free experience.

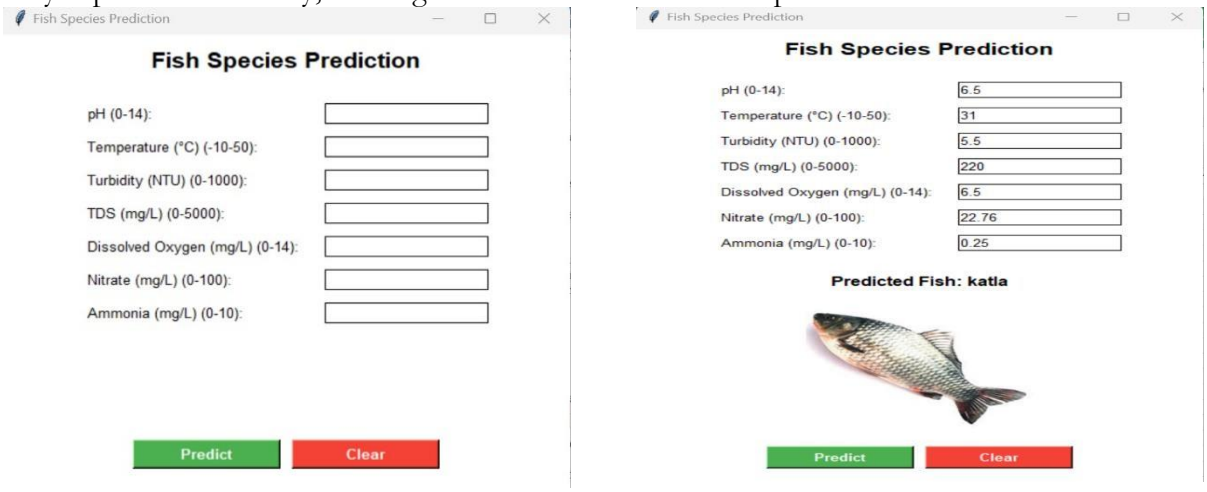


Figure 6. GUI Of Fish Recommender System.

Discussion:

This project represents a novel approach by integrating machine learning (ML) with real-time water quality monitoring to enhance traditional fish farming practices. While previous studies primarily focus on water quality monitoring, this research innovates by introducing an AI-powered fish recommendation system that aids farmers in making more informed decisions about the most suitable fish species for their farming conditions.

The system is built around an architecture that seamlessly combines data collection, processing, and analysis. This results in an optimized fish species selection based on seven crucial water quality parameters: pH, temperature, turbidity, total dissolved solids (TDS), dissolved oxygen, nitrate, and ammonia. A key feature of this study is its comprehensive approach, as it considers a broader range of water quality parameters than many previous studies, which typically focus only on pH and temperature. Parameters like turbidity, TDS, and ammonia, which are critical for fish health and farming success, are often overlooked in existing literature. By incorporating these additional parameters, this research enhances the accuracy and practicality of the fish species recommendation system. Moreover, Table 5 provides a comparative analysis between this study and prior works. While many studies have focused on Internet of Things (IoT)-based water quality monitoring systems, very few have combined machine learning with fish species recommendation. Even among those that do use ML, most rely on a limited set of parameters. In contrast, this research applies an optimized machine learning algorithm to a well-structured, multi-parameter dataset, resulting in a fish prediction accuracy exceeding 90%. This highlights the project's contribution to advancing the use of ML in aquaculture, particularly in integrating multiple water quality parameters for more accurate species selection.

The integration of machine learning with water quality monitoring in this manner offers a promising tool for farmers, allowing them to optimize fish farming practices based on comprehensive environmental data.

Conclusion:

This research introduces a machine learning-based fish species recommendation system aimed at improving decision-making in aquaculture. By incorporating seven key water quality parameters: pH, temperature, turbidity, total dissolved solids (TDS), dissolved oxygen, nitrate, and ammonia, the model predicts the most suitable fish species for given water conditions, enhancing fish farming efficiency and sustainability.

The study evaluated several machines learning algorithms, including XGBoost, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR). Among these, XGBoost achieved the highest performance with an accuracy of 96.8% after the application of SMOTE (Synthetic Minority Over-sampling Technique), followed closely by Random Forest with 96.6%. Decision Tree and SVM also demonstrated strong results, with accuracies of 90.2% and 84.2%, respectively. However, KNN and Logistic Regression performed less effectively, recording accuracies of 84% and 64%, respectively.

The use of SMOTE significantly enhanced classification accuracy, particularly for underrepresented fish species, by addressing dataset imbalance. This improvement ensures more reliable predictions across all species, making the model more robust in real-world applications. Furthermore, a graphical user interface (GUI) was developed to make the system user-friendly and accessible. The GUI allows fish farmers and aquaculture professionals to input water quality parameters and receive real-time recommendations for the most appropriate fish species.

The integration of this machine-learning model into practical aquaculture operations has the potential to significantly enhance fish farming practices. By improving decision-making, it can help optimize yields, reduce costs, and promote more sustainable aquaculture practices, ultimately contributing to the advancement of the industry.

Table 5. Comparison of ml model performance with previous studies.

Reference	Models Evaluated	Best Performing Model	Data Preprocessing	Accuracy

Islam et al.	J48,J48, KNN, RF,NB, CART	RF	None	88.48%
Hemal et al.	XGBoost, RF, DT, KNN, LR, SVM	RF	SMOTE	94%
Proposed Model	XGBoost, RF, DT, KNN, SVM	XGBoost	Feature Scaling, SMOTE	96.8%

Table 5 compares the performance of different models from previous studies with the proposed model. Islam et al. achieved 88.48% accuracy using Random Forest (RF) without any preprocessing, while Hemal et al. improved performance to 94% by applying SMOTE with RF. The proposed model outperformed both studies, achieving 96.8% accuracy using XGBoost with feature scaling and SMOTE, demonstrating the effectiveness of data preprocessing and advanced ensemble learning in improving classification performance.

References:

[1] C. Cordova-Rozas, M. Orellana, and F. Sepúlveda, "Cloud- based water monitoring system: A framework for real-time aquaculture management," *J. Aquac. Eng*, vol. 75, no. 2, pp. 102–115, 2021.

[2] Y. C. H. Gao, L. Zhang, "An IoT-based fish farming [1] and tracking system: Enhancing aquaculture productivity through real-time monitoring," *Sensors*, vol. 19, no. 8, pp. 1895–1910, 2020.

[3] T. F. M. Nagayo, K. Yamamoto, "Design and implementation of a solar-powered aquaponics system for sustainable fish farming," *Renew. Energy Sustain. Dev*, vol. 10, no. 3, pp. 55–67, 2019.

[4] R. Pasika and S. Gandla, "IoT-based real-time water quality monitoring system for aquaculture: A low-cost approach," *IEEE Access*, vol. 8, pp. 160566–160578, 2020.

[5] M. L. X. Huan, Z. Wang, "NB-IoT-enabled real-time water moni- toring for smart aquaculture," *IEEE Internet Things J*, vol. 7, no. 5, pp. 4123–4135, 2021.

[6] H. S. Y. Chiu, P. Lin, "Deep learning-based predictive analysis of fish growth in IoT aquaculture systems," *Comput. Electron. Agric*, vol. 178, p. 105780, 2020.

[7] M. Y. A. Niswar, R. Abdul, "Crab farming with IoT monitoring using MQTT and LoRa-based sensor networks," *J. Aquac. Res*, vol. 44, no. 3, pp. 299–312, 2019.

[8] M. M. Billah, Z. M. Yusof, K. Kadir, A. M. M. Ali, and I. Ahmad, "Quality Maintenance of Fish Farm: Development of Real-time Water Quality Monitoring System," *2019 IEEE 6th Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2019*, Aug. 2019, doi: 10.1109/ICSIMA47653.2019.9057294.

[9] A. I. M. S. Uddin, M. S. Hossain, M. M. Hossain, "Fish survival prediction in an aquatic environment using random forest model," *J.Aquat. Environ. Predict*, vol. 35, no. 4, pp. 215–230, 2021.

[10] T. Abinaya, J. Ishwarya, and M. Maheswari, "A Novel Methodology for Monitoring and Controlling of Water Quality in Aquaculture using Internet of Things (IoT)," *2019 Int. Conf. Comput. Commun. Informatics, ICCCI 2019*, Jan. 2019, doi: 10.1109/ICCCI.2019.8821988.

[11] T. J. S. M. Islam, S. Rahman, "ML-based fish species predic- tion using ensemble learning techniques," *J. Appl. Sci. Comput*, vol. 27, no. 1, pp. 109–124, 2022.

[12] R. V. A. Hemal, P. K. Singh, "AquaBot: An AI- driven smart pond water quality monitoring system," *Comput. Water Environ. Eng*, vol. 11, pp. 30–46, 2021.

[13] M. Monir, "Realtime pond water dataset for fish farming," *Kaggle*, 2023, [Online]. Available: https://www.kaggle.com/datasets/monirmukul/realtime-pond-water-dataset-for-fish-farming

[14] C. E. Boyd, "Water quality for pond aquaculture," *Auburn Univ.*, 1998, [Online]. Available: https://aurora.auburn.edu/handle/11200/49690

[15] A. Bhatnagar and P. Devi, "Water quality guidelines for the management of pond fish culture," *Int. J. Environ. Sci*, vol. 3, no. 6, pp. 980– 2009, 2013.



Copyright © by authors and 50Sea. This work is licensed under Creative Commons Attribution 4.0 International License.