Global Scientific JOURNALS

GSJ: Volume 12, Issue 5, May 2024, Online: ISSN 2320-9186
www.globalscientificjournal.com

# FISH POND WATER QUALITY ASSESSMENT MODELS USING MACHINE LEARNING ALGORITHM

Oladunjoye John Abiodun[1] Andrew Ishaku Wreford[2]

Computer Science Department, Federal University Wukari, Nigeria[1].

Computer Science Department, Federal University Wukari, Nigeria[2].

oladunjoye.abbey@yahoo.com[2].

andrew@fuwukari@edu.ng[1].

## Abstract

Traditional fish farming faces several significant challenges, including water contamination, temperature imbalances, feed management, limited land availability, and high costs. The aquaculture industry continues to face various challenges, including the need for enhanced monitoring systems, early identification of disease outbreaks, high mortality rates, and the promotion of sustainability. These issues represent ongoing concerns that require resolution and have prompted this study to conduct research on fish pond water quality management using the Woosong University fish pond dataset sourced from the Kaggle machine learning repository. The objective of this research is to develop an aquaculture solution utilizing machine learning (ML) techniques, with the aim of enhancing prawn growth and increasing productivity in pond environments. Hence, the study scrutinizes the effectiveness of some machine learning algorithms, including XGBoost, Gradient Boosting, K-Neighbors Regressor, Random Forest Regressor, and a Hybrid Ensemble Model. Evaluation metrics using some evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), R-squared (R2), and Root Mean Squared Error (RMSE) to assess the algorithms' effectiveness. The study's findings revealed that the Random Forest Regressor and the Hybrid Ensemble Model outperform other algorithms in terms of prediction accuracy, making them strong candidates for assessing water quality in fish farming.

**Key-Words:** XGB Boost (XGBOOST), Gradient Boosting (GB), K-Neighbors Regressor (KNN), Random Forest (RF)

## 1.0 Introduction

Fish farming, also known as aquaculture, has emerged as a critical solution to meet the escalating demand for seafood in a world where marine waters, both coastal and open ocean, are being harnessed to grow food (Zambrano *et al.*, 2021). This innovation-driven technology has not only fulfilled the global appetite for seafood but has also positioned itself as the future's primary method of aquatic food production.

However, the expansion of aquaculture has brought about a range of challenges and concerns. Aquaculture's impact on biodiversity, resource utilization (including land and water), emissions, and the introduction of various agents into the environment has raised important ecological questions (Yilmaz *et al.*, 2023). Greenhouse gases, leftover food waste, excrement, urine, chemotherapeutic medications, bacteria, parasites, and stray animals all fall into this category. Moreover, the transfer of eutrophicating agents, harmful chemicals, infections, diseases, and genetic material into ecosystems has raised alarms regarding its effects on wild populations and ecosystems (Feng *et al.*, 2019). These concerns extend to indirect consequences, such as habitat loss, changes in niche spaces, and disruptions in food webs.

Nonetheless, aquaculture offers a multifaceted approach. It not only serves as a means of food production but also as a tool for ecosystem restoration, stock replenishment, and the conservation of threatened and endangered species (Kaur *et al.*, 2023). This is especially evident in the distinction between marine and freshwater aquaculture, with a focus on the marine environment and estuarine creatures.

Fish farming has not only impacted the seafood industry but has also provided opportunities for current and former fishermen to complement their traditional fishing activities. For instance, farmed seafood products now contribute

significantly to the seafood market, alleviating the United States' seafood deficit and providing a domestic source of economically and ecologically sustainable seafood (Abaidoo *et al.*, 2021).

As aquaculture continues to evolve and diversify, it stands as a distinct agricultural sector characterized by its unique challenges, innovations, and contributions to global food security (Gladju *et al.*, 2022). It is a sector where scientific and technological advancements have played a pivotal role in enhancing production, improving feed conversion rates, managing diseases, and expanding species diversity.

In the pursuit of securing wholesome food sources and thus reducing poverty among coastal and rural communities, both fishing and aquaculture have pivotal roles to play (Rahman *et al.*, 2021). While aquaculture encompasses various aspects of marine life, fisheries primarily deal with the capture of wild fish or the breeding and harvesting of fish. This fundamental distinction highlights the crucial role of aquaculture in sustaining food security and economic stability. Aquaculture's profound impact on protein production cannot be overstated (Rahman *et al.*, 2021). It has become one of the fastest-growing agricultural sectors, surpassing wild fisheries production in recent years. This remarkable growth has been made possible through continuous research, technological advancements, and improvements in every facet of aquaculture. Scientific discoveries and technological innovations have led to better feed formulations, reduced disease outbreaks, and increased efficiency in aquaculture operations. However, the growing demand for seafood necessitates further expansion and innovation in aquaculture to meet the world's ever-increasing seafood needs.

Recently, fish farming or aquaculture, has witnessed significant advancements driven by artificial intelligence (AI) and machine learning (ML) technologies (Hu *et al.*, 2022). These innovations have shown their potential to revolutionize fish farming practices, leading to improved efficiency, sustainability, and productivity (Kaur *et al.*, 2023). One of the critical applications of machine learning is in Water Quality Management, analyzing and monitoring water quality parameters including pH levels, temperature, oxygen levels, and turbidity (Islam *et al.*, 2023); fish health monitoring, targeted at recognizing subtle changes in fish behaviour and appearance that may indicate health issues (Yilmaz *et al.*, 2023); feed optimization, meant to optimize feed formulation by analyzing data on fish growth rates, feeding schedules, and environmental conditions (Du *et al.*, 2023); and lastly, environmental impact assessment, targeted at trailing problems regarding waste disposal and habitat alterations.

Hence, in the realm of water quality assessment, this study proposed the application of some ensemble machine learning algorithms including XGBoost, GB, KNN, and RF on the fish pond dataset sourced from the Kaggle machine learning repository.

A major contribution of this research is the integration of ensemble algorithms, harnessing their collective strengths to formulate a robust model for the evaluation of water quality in fish ponds.

## 2.0 Related Works

Xiao *et al.*, (2017) used a BP neural network with various activation functions to create machine learning and deep learning models for predicting dissolved oxygen levels in aquaculture systems. They employed 10 days of breeding data from three ponds in Beihai, Guangxi, China, for their study. The first week was dedicated to instruction, and the final three were reserved for examinations. With 5000 iterations, a learning rate of 0.01 and a goal value of 0.00000001, the neural network outperformed other standard prediction models such as curve fitting, autoregression, grey model, and support vector machines. All of the anticipated values were within a 5% range, which is acceptable for most uses. When compared to AR, GM, SVM, and CF, the neural network provided the most accurate predictions.

The purpose of Stocker *et al.*, (2022) study was to examine the potential for forecasting the amounts of Escherichia coli in agricultural pond waters. Over three years, the researchers monitored two ponds in Maryland during the irrigation season. The obtained water samples were analyzed for the presence of E. coli and another 12 indicators of water quality. Stochastic gradient boosting (SGB) machines, random forests (RF), support vector machines (SVM), and k-nearest neighbour (kNN) approaches were used to make predictions about E. coli based on the available datasets. The RMSE for predicting E. coli concentrations in both ponds using the RF model was the smallest in the majority of cases, both within individual years and between subsequent years. The calculated E. coli concentrations

(log10 CFU 100 ml1) had root mean square error (RMSE) values between 0.244 and 0.346 for Pond 1 and between 0.304 and 0.418 for Pond 2 over the years. Three-year data sets showed values of 0.334 for Pond 1 and 0.381 for Pond 2. The root mean square error (RMSE) values attained by the random forest (RF) model and those generated by other machine learning (ML) models did not differ statistically significantly ($P > 0.05$) in the vast majority of situations. The RMSE values produced from five iterations of a 10-fold cross-validation technique served as the statistical measurements for this comparison. Turbidity, dissolved organic matter content, specific conductance, chlorophyll concentration, and temperature were all found to be significant predictors of E. coli. Using 5 predictors yielded the same predictive performance as using 8 or 12 predictors in the model. This indicates that the predicted accuracy of the evaluated algorithms does not noticeably improve when new predictors are included, despite the fact that doing so requires more work and resources.

Improved water quality parameter data was used to create a novel aquaculture prediction model, which was presented in (Jiang and Yan, 2022). Using principal component analysis, the author looked at the complex relationship between dissolved oxygen and water quality. As a result, the author proposed a PCA-BP (principal component analysis backpropagation) model for predicting water quality. The PCA-BP water quality prediction model's weight and threshold parameters were optimised using a genetic algorithm. An improved PCA-BP water quality prediction model was developed after the threshold and weight of the BP neural network were determined. According to the results of several controlled experiments, the GPCA-BP model can forecast the dissolved oxygen concentration with a relative error of less than 0.76 per cent over a range of temporal and spatial water quality prediction trials. Additionally, the model's prediction accuracy is higher than that of competing models. Convergence accuracy, prediction accuracy, and the mean absolute error in performance (MAE) are all areas in which the GPCA-BP water quality prediction model excels.

Nyumba ya Mungu Dam's fishing potential has dropped by 95% between 1972 and 2018, therefore researchers (Mangi *et al.*, 2023) looked into the link between water quality and fish productivity to figure out why. Using standard procedures, we analysed the temperature, pH, dissolved oxygen (DO), turbidity, total nitrogen, total phosphorus, chlorophyll, and depth of the water at the fishing net location. Over a year, from January to December of this year, fish biomass/productivity was analysed by looking at catches per unit of effort. Pearson's correlation analysis showed a significant positive relationship between turbidity ($r = 0.461$, p 0.01) and TP ($r = 0.405$, p 0.01) and fish catch per unit effort. The results of the stepwise multiple regression model indicated that turbidity, dissolved oxygen, and the depth at which the fishing net was placed were statistically significant predictors of fish catch per unit effort. Fish catch per unit effort was also shown to be modified by 24% once turbidity and dissolved oxygen were incorporated into the model. There was a 28.9% difference in fish catch per unit effort, but that was only after factoring in turbidity, dissolved oxygen, and the depth of the fishing net site.

## 3.0 Methodology

A three-phase methodological approach was utilized in the development of pond machine learning models. In the initial stage, the dataset was accessed using the Panda's library. This library provides a range of methods for reading datasets in various file formats, such as comma-separated values (CSV), which is the format employed in the dataset utilized for this study. The second phase of the study emphasized data preprocessing, encompassing many activities like the elimination of extraneous features, standardization, and encoding of the dataset's attributes. The final step of our methodology consisted of inputting the processed features, which were filtered, scaled, and encoded, into the designated machine learning algorithms, namely K-Nearest Neighbors (KNN), XGBoost, Random Forest, and Gradient Boosting. Before inputting the data, we partitioned the dataset into a training set comprising 70% of the data and a test set including the remaining 30%. The training dataset was employed to train the machine learning models, whereas the test dataset was utilized to assess the models' correctness. During the third phase, some performance evaluation metrics were implemented to ascertain the models that exhibited superior performance. Figure 1 depicts the sequential methodological approach as previously explained.

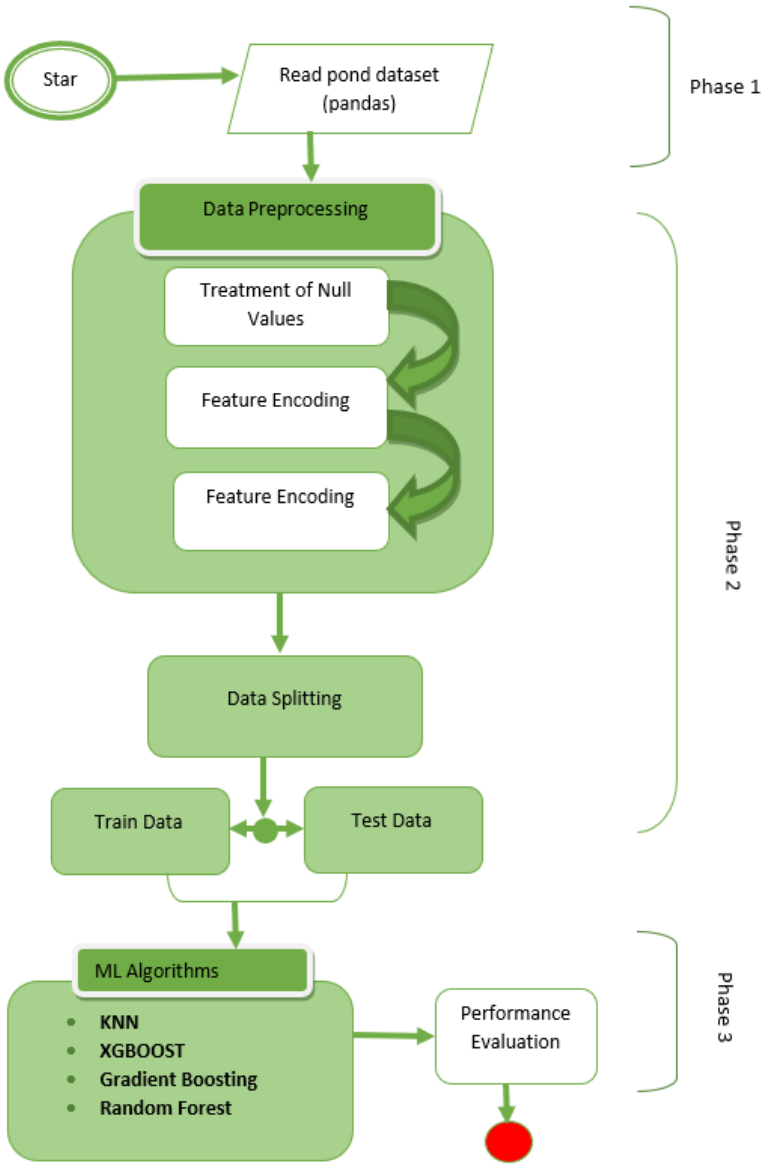


**Figure 1:** Methodological Framework

## 3.1 Dataset Description

The dataset utilized is the Woosong University fish pond dataset. Islam *et al.*, (2021) acquired the dataset in real time by utilizing a specifically constructed Internet of Things (IoT) framework for monitoring aquatic environments. This framework involved the use of an Arduino and sensors. During the process of data collecting, the researchers employed three distinct sensors, namely pH, temperature, and turbidity sensors, to monitor and assess the water quality of a total of five ponds. The dataset consists of 591 rows and 4 columns. The factors under consideration include pH, temperature, turbidity, and the presence of fish. The target variable in this study is fish, whereas the remaining variables are considered independent variables. There exists a total of 11 separate fish classifications, 86 distinct values for pH, 46 distinct values for temperature, and 85 distinct values for Turbidity.

## 3.2 Data Normalization

To enhance the efficacy of a proposed machine learning regression model, the introduction of the normalization technique was suggested. The objective of normalization in this context is to standardize the values of numerical columns within the dataset while preserving the relative disparities in the ranges of values. This objective is accomplished by normalizing each feature variable, denoted as $V_{ij}$, to the interval (0, 1).

$$\frac{V^{ij} - min^{j}(V^{ij})}{max^{j}(V^{ij}) - min^{j}(V^{ij})} \dots\dots\dots\dots\dots 1$$

## 3.3 Gradient Boosting

Gradient boosting regression tree algorithms utilize an ensemble learning technique to create robust predictive models (Singh *et al.*, 2021). They achieve this by combining multiple individual regression trees, which are often considered weak learners. These weak learners are characterized by having high bias, low variance, and limited performance on their own. The primary goal of the algorithm is to reduce the errors made by these weak models, improving their predictive power. Boosting algorithms typically consist of three key components: an additive model, the weak learners, and a loss function. This algorithm can effectively capture nonlinear relationships in data, making it suitable for tasks like modelling wind power curves. It employs various differentiable loss functions and continually learns during the iterative process, adjusting its predictions based on input features. Gradient boosting machines (GBM) work by recognizing the weaknesses of these weak models through gradient information. This is achieved through an iterative process where the algorithm aims to combine base learners to minimize prediction errors. Decision trees are integrated into the model in an additive manner, and the loss function is reduced using gradient descent, allowing the algorithm to progressively improve its predictions. The GBT (gradient boosting tree) $F_n(x_t)$ can be defined as the summation of n regression trees.

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \dots\dots 2$$

Where every $F_i(x_t)$ is a decision tree. The ensemble of a tree is constructed sequentially by estimating the new decision tree $F_{n+1}(x_t)$ with the help of the following equation:

$$argmin \sum_{t} L\big(y_t . F_n(x_t) + F_{n+1}(x_t)\big) \dots\dots 3$$

Where $L$. Is differentiable for loss-function $L(.)$. The optimization is solved by the steepest descent method.

## 3.4 K-Nearest Neighbors (KNN)

The KNN regression model is a straightforward approach that relies on the similarity between data points to make predictions for new observations (Sumayli, 2023). When using KNN regression, the model determines how far off a new observation is from each of the data points in the training dataset. Euclidean distance, the most widely used distance measure, is determined using the equation.

$$d\big(x_i, x_j\big) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \dots\dots\dots 4$$

Input feature count is denoted by $p$, and the value of feature $k$ for observation $i$ is denoted by $x_{ik}$, and feature $k$ for observation $j$ is denoted by $x_{jk}$. The KNN method takes the estimated distances and chooses the k closest neighbours. Next, we use the average (or median) of the values of the target variable among these K nearest neighbours to get a prediction for the new observation. The ease with which KNN regression can be understood is one of its primary benefits. However, finding an appropriate number for $k$ is essential, as doing so may lead to either over-fitting or under-fitting if the value is too little or too big. The choice of 5 neighbours reflects this.

## 3.5 XGBoost

XGBoost regression, a term denoting extreme gradient boost regression, represents an iterative decision tree algorithm with the primary objective of enhancing model performance through the utilization of residuals (Huang *et al.*, 2022). This algorithm boasts several key attributes such as its support for parallel computing, enabling the simultaneous utilization of all available computer cores. This parallel processing capability significantly enhances the algorithm's efficiency, making it well-suited for tasks requiring rapid model development and evaluation. Moreover, XGBoost incorporates internal regularization techniques, a critical feature that helps prevent overfitting. Imposing constraints on the model during training ensures that the resulting model maintains robustness and generalizability, even when dealing with complex datasets. XGBoost also has built-in cross-validation and missing value-handling mechanisms. These features simplify the model validation process and facilitate the management of missing data, streamlining the overall workflow of model development. To achieve all these, XGBoost uses a loss function mathematically defined as follows:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + C \ldots\ldots 5$$

Then, Taylor's second-order expansion of the objective function is performed:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_t f_t(x_i) + \frac{1}{2} h_j f_t^2(x_i)\right) + \Omega(f_t) + C \ldots\ldots 6$$

Finally, the evaluation function of the tree structure is obtained. The smaller the value is, the smaller the error is:

$$L^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \Upsilon^T \ldots\ldots\ldots 7$$

## 3.6 Random Forest

In ensemble learning, the RF regression algorithm pools data from several different regression trees. An informal definition of a regression tree in this setting is "a set of conditions or rules organized hierarchically and applied systematically from the root to the leaves" (Zhou, 2016). Multiple bootstrap samples, which are subsets picked at random with replacements from the original training dataset, are generated to begin the RF procedure. A regression tree is built for each of these bootstrap samples individually. During the process of building each tree, a random subset of the total set of input variables is selected and used to create binary divisions at each node. Selecting the input variable with the smallest Gini Index serves as the condition for splitting the regression tree.

$$IG\left(t_{X(x_i)}\right) = 1 - \sum_{j=1}^{m} f(t_{X(x_i)}, j)^2 \ldots\ldots\ldots 8$$

where $f(t_{X(x_i)}, j)^2$ represents the fraction of samples where $x_i$ is a value from branch *j* of node *t*. An observation's projected value is obtained by taking the mean across all trees. The number of regression trees (ntree; default value is 500 trees) and the number of input variables per node (mtry; default value is 1/3 of the total number of variables) are the two parameters that need to be optimised in the RF.

## 3.7 Hybridization Approach

Stacking refers to an ensemble technique wherein a meta-learner algorithm is used with one or more base-level classifiers. The initial dataset serves as the input for multiple separate models in the stacking technique. The metaclassifier is subsequently employed to assess the input, output, and weights of each model. The models that demonstrate superior performance are selected, and the remaining models are excluded. The stacking technique utilizes a metaclassifier to combine many basic classifiers that have been trained using distinct learning methods on a unified dataset. The model's predictions are combined with the inputs from each subsequent layer to produce a fresh set of predictions.

## 3.8 Evaluation metrics

The performance of the model was evaluated using the below indicators:

**Mean Square Error (MSE):** serves as a metric for evaluating the proximity of a regression line to the adapted dataset. It essentially represents a measure of risk, corresponding to the anticipated value of the squared error loss. To compute the MSE, one takes the average, or mean, of the squared errors between the data points and a related function. A higher MSE signifies that the data points are widely spread around their central point (mean), while a lower MSE indicates the opposite. A smaller MSE is preferable, as it suggests that the data points are closely clustered around their central mean.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(F_d(i) - F(i))^2 \dots\dots\dots 9$$

**Root Mean Squared Error (MSE):** It is the average of the squared difference between the predicted and actual value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i^f - y_i^{ob})^2}{n}} \dots\dots\dots 10$$

Where $y_i^f$ is the $i^{th}$ forecasted data, $y_i^{ob}$ is the $i^{th}$ observed data, and n is the amount of data.

**MAPE (Mean Absolute Percentage Error):** provides a metric for assessing the typical deviation from a predicted value to the actual value. Simply said, it provides a numerical value for the typical percentage by which a model's predictions differ from the true values.

$$MAPE = \frac{1}{n}\sum\frac{|o_t - p_t|}{|o_t|} * 100 \dots\dots 11$$

Where n indicates the sample size, $o_t$ indicates the actual data value, $p_t$ indicates the forecasted data value

**$R^2$-Score:** When examining the efficacy of a machine learning model that uses regression, the R2 score is crucial. It measures how well the model's predictions account for the diversity in a given dataset. It's a way to evaluate how far the model's predictions deviate from the observed data.

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y_i})^2}{\sum(y_i - \overline{y_i})^2} \dots\dots\dots 12$$

## 4.0 Experimental setup

In this investigation, we used a 64-bit Windows OS on a computer with an Intel(R) Corel Trade Mark (TM) i5-2560QM CPU @2.40GHZ and 8.00 GB of RAM (Random Access Memory) to conduct experiments and make predictions on pond water quality. The Anaconda environment with the Python 3.11 software development kit was used to put the program code into action. Sklearn, Pandas, Matplotlib, Seaborn, and NumPy were used as their respective application programming interfaces.

### 4.1 Result Presentation

Table 1 presents a comparative analysis of different machine learning algorithms, including XGB Boost, Gradient Boosting, K-Neighbors Regressor, Random Forest Regressor, and a Hybrid Ensemble Model, based on various evaluation metrics: MSE, MAPE, R2, and RMSE. The MSE measures the average squared difference between the actual and predicted values. A lower MSE indicates a better fit to the data. Among the algorithms, Gradient Boosting, Random Forest Regressor, and the Hybrid algorithm achieved the lowest MSE values, with 0.012, 0.0033,

and 0.0031 respectively. This suggests that these models provide more accurate predictions compared to others. The MAPE calculates the percentage difference between actual and predicted values. A smaller MAPE indicates better predictive accuracy. In this analysis, the Random Forest Regressor, and the Hybrid algorithm exhibit the lowest MAPE at 0.0014, indicating that it provides the most accurate percentage-wise predictions. The R-squared (R2) metric measures how well the model explains the variance in the data. An R2 value of 0.99 is achieved by all the algorithms, indicating that they are excellent at explaining the variance in the target variable. This suggests a strong correlation between the predicted and actual values for all models. The RMSE is a measure of the standard deviation of prediction errors. Smaller RMSE values indicate more accurate models. Both the Random Forest Regressor and the Hybrid Ensemble Model have the lowest RMSE values, with 0.058 and 0.055, respectively, implying that they provide the best overall prediction accuracy.

**Table 1: Result Presentation**

| Algorithm | MSE | MAPE | R2 | RMSE |
|---|---|---|---|---|
| XGB Boost | 0.14 | 0.16 | 0.99 | 0.37 |
| Gradient Boosting | 0.012 | 0.16 | 0.99 | 0.11 |
| K-Neighbors Regressor | 0.14 | 0.048 | 0.99 | 0.38 |
| Random Forest Regressor | 0.0033 | 0.0014 | 0.99 | 0.058 |
| Hybrid Ensemble Model | 0.0031 | 0.0014 | 0.99 | 0.055 |

In Figure 2, a visual representation is provided to showcase the performance of different individual models based on four distinct evaluation metrics: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), R2 Score, and Root Mean Squared Error (RMSE). The y-axis on the graph represents the values of these evaluation metrics, measured in decimal numbers, reflecting the quantitative assessment of each model's performance. On the x-axis, each model is represented, and the length of the bars extending from each model denotes the corresponding scores achieved for the mentioned metrics. For the MSE, MAPE, and RMSE a short bar represent better performance whereas for the R2 Score, the lengthy bars depicts better performances.

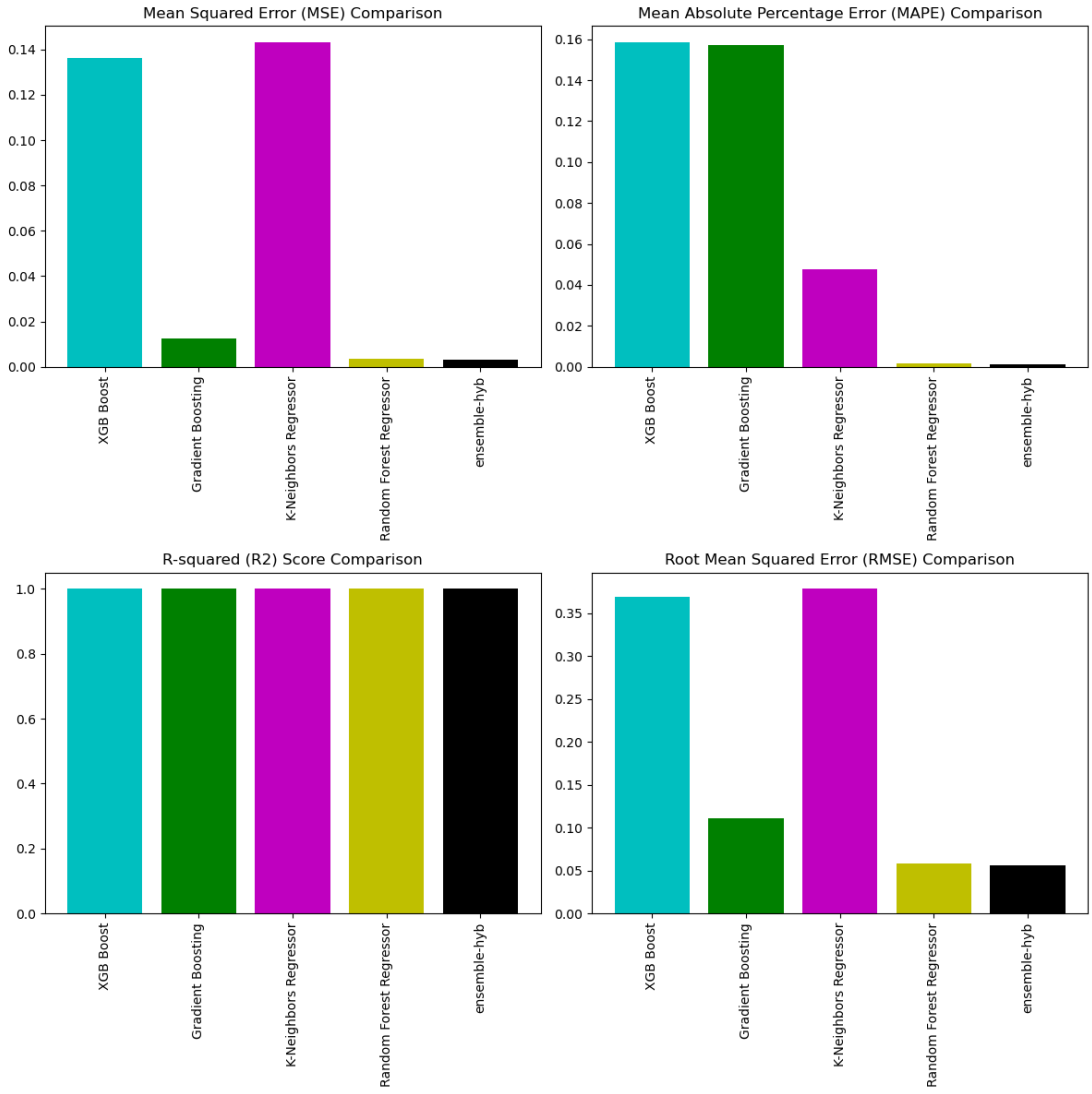


**Figure 2:** Result Comparison Graph

## 5.0 Conclusion

This study has comprehensively investigated machine learning approaches for smart fishing farming and has thus applied the usage of some ensemble machine learning approaches on the fish pond dataset sourced from the Kaggle machine learning repository. The study has applied the viabilities of four machine learning algorithms including the Random Forest, KNN, Gradient Boosting, and XGBoost algorithm and has additionally hybridised the algorithm using the stacking approach. The results indicate that Gradient Boosting, Random Forest Regressor, and the Hybrid Ensemble Model outperformed the others, with the lowest MSE values of 0.012, 0.0033, and 0.0031, respectively, signifying their superior predictive accuracy. Furthermore, the Random Forest Regressor and the Hybrid algorithm exhibited the lowest MAPE at 0.0014, underscoring their precision in percentage-wise predictions. All models achieved a high R2 value of 0.99, showcasing their excellence in explaining variance, and both the Random Forest Regressor and Hybrid Ensemble Model displayed the lowest RMSE values of 0.058 and 0.055, confirming their superior overall prediction accuracy. Conclusively, these findings highlight the effectiveness of Random Forest and the Hybrid Ensemble Model in fish pond water quality assessment based on the utilized performance evaluation metrics. An extension of the study can include the integration of IoT devices to enhance real-time monitoring of

water quality in the pond while strategically placing sensors measuring parameters like temperature, pH, and dissolved oxygen.

## Availability of Dataset

https://www.kaggle.com/datasets/monirmukul/realtime-pond-water-dataset-for-fish-farming

## Competing Interest

The authors declare that they have no competing interests.

## Authors' Contributions

The manuscript was written by Oladunjoye John Abiodun, while the code was written by Andrew Ishaku Wreford. It should be noted that both authors contributed substantial contributions to the study.

## Reference

Abaidoo, E., Melstrom, M., & Malone, T. (2021). The Growth of Imports in US Seafood Markets. *Choices*, *36*(4), 1-10.

Du, Z., Cui, M., Wang, Q., Liu, X., Xu, X., Bai, Z., ... & Li, D. (2023). Feeding intensity assessment of aquaculture fish using Mel Spectrogram and deep learning algorithms. *Aquacultural Engineering*, 102345.

Feng, Z., Zhang, T., Li, Y., He, X., Wang, R., Xu, J., & Gao, G. (2019). The accumulation of microplastics in fish from an important fish farm and mariculture area, Haizhou Bay, China. *Science of the Total Environment*, *696*, 133948.

Gladju, J., Kamalam, B. S., & Kanagaraj, A. (2022). Applications of data mining and machine learning framework in aquaculture and fisheries: A review. *Smart Agricultural Technology*, *2*, 100061.

Hu, W. C., Chen, L. B., Huang, B. K., & Lin, H. M. (2022). A computer vision-based intelligent fish feeding system using deep learning techniques for aquaculture. *IEEE Sensors Journal*, *22*(7), 7185-7194.

Huang, Y., Chen, C., & Miao, Y. (2022). Prediction Model of Bone Marrow Infiltration in Patients with Malignant Lymphoma Based on Logistic Regression and XGBoost Algorithm. *Computational and Mathematical Methods in Medicine*, *2022*.

Islam, M. M., Kashem, M. A., Alyami, S. A., & Moni, M. A. (2023). Monitoring water quality metrics of ponds with IoT sensors and machine learning to predict fish species survival. *Microprocessors and Microsystems*, 104930.

Islam, M.M., Mohammed, A.K., and Jia, U (2021). Fish survival prediction in an aquatic environment using random forest model. *Int J Artif Intell*, ISSN 2252.8938 (2021): 8938.

Jiang, Y., & Yan, F. (2022). Aquaculture Prediction Model Based on Improved Water Quality Parameter Data Prediction Algorithm under the Background of Big Data. *Journal of Applied Mathematics*, *2022*.

Kaur, G., Adhikari, N., Krishnapriya, S., Wawale, S. G., Malik, R. Q., Zamani, A. S., ... & Osei-Owusu, J. (2023). Recent Advancements in Deep Learning Frameworks for Precision Fish Farming Opportunities, Challenges, and Applications. *Journal of Food Quality*, *2023*.

Mangi, H. O., Onywere, S. M., & Kitur, E. C. (2023). Fish productivity response to water quality variations: A case study of nyumba ya mungu dam, in pangani water basin, Tanzania. *International Journal of Ecology*, *2023*.

Rahman, L. F., Marufuzzaman, M., Alam, L., Bari, M. A., Sumaila, U. R., & Sidek, L. M. (2021). Developing an ensembled machine learning prediction model for marine fish and aquaculture production. *Sustainability*, *13*(16), 9124.

Singh, U., Rizwan, M., Alaraj, M., & Alsaidan, I. (2021). A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. *Energies*, *14*(16), 5196.

Stocker, M. D., Pachepsky, Y. A., & Hill, R. L. (2022). Prediction of E. coli concentrations in agricultural pond waters: application and comparison of machine learning algorithms. *Frontiers in Artificial Intelligence*, *4*, 768650.

Sumayli, A. (2023). Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. *Arabian Journal of Chemistry*, *16*(7), 104833.

Xiao, Z., Peng, L., Chen, Y., Liu, H., Wang, J., & Nie, Y. (2017). The dissolved oxygen prediction method is based on a neural network. *Complexity*, *2017*.

Yilmaz, M., Çakir, M., Oral, M. A., Kazanci, H. Ö., & Oral, O. (2023). Evaluation of disease outbreak in terms of physico-chemical characteristics and heavy metal load of water in a fish farm with machine learning techniques. *Saudi Journal of Biological Sciences*, *30*(4), 103625.

Zambrano, A. F., Giraldo, L. F., Quimbayo, J., Medina, B., & Castillo, E. (2021). Machine learning for manually-measured water quality prediction in fish farming. *Plos one*, *16*(8), e0256380.

Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, *4*(3), 212-219.