A Hybrid Approach for Predicting Dissolved Oxygen in Aquaculture to Enhance Water Quality

[1]Bhaskara Poojitha,[2]G.Praveen Babu,
[1]*MCA Student, Department of Information Technology, Jawaharlal Nehru Technological University, India*
[2]*Associate Professor of CSE, Department of Information Technology, Jawaharlal Nehru Technological University, India*

Abstract: This study proposes a hybrid model, combining Light Gradient Boosting Machine (LightGBM) and Bidirectional Simple Recurrent Unit (BiSRU), to predict dissolved oxygen (DO) levels in aquaculture environments accurately and efficiently. Initially, linear interpolation and smoothing techniques are employed to identify significant parameters, followed by LightGBM algorithm's utilization to determine the relevance of dissolved oxygen and predict its levels in intensive aquaculture settings. Furthermore, an attention mechanism is implemented to assign varying weights to BiSRU's hidden states, enhancing its predictive capabilities. The model demonstrates remarkable performance, accurately anticipating DO fluctuations over a 10-day period in just 122 seconds, with an impressive accuracy rate of 96.28%. This approach addresses the challenges faced by traditional methods in predicting DO levels due to the nonlinear, dynamic, and complex nature of aquatic environments. The significance of maintaining water quality in aquaculture for optimal productivity underscores the importance of accurate DO prediction. Overall, the proposed hybrid model offers a promising solution for enhancing DO prediction accuracy and speed, contributing to effective disease prevention and economic sustainability in aquaculture operations.

INDEX TERMS Non-linear, LightGBM, BiSRU, attention mechanism.

1. INTRODUCTION:

Aquaculture, the farming of aquatic organisms, plays a pivotal role in global food production, with China emerging as a dominant force in this sector, accounting for over 70% of the world's aquaculture output [1]. The quality of aquatic goods is intricately linked to the water quality in aquaculture environments, where dissolved oxygen (DO) serves as a critical indicator due to its fundamental role in the metabolism and survival of aquatic organisms [2], [3]. Maintaining optimal DO levels is paramount for ensuring the healthy growth of farmed fish, shrimp, and other aquatic species, as deviations can lead to stress, disease outbreaks, and mass mortality, resulting in substantial economic losses [4], [5]. Consequently, accurate prediction of DO concentrations and trends is essential for proactive water quality management and sustainable aquaculture development [6].

In recent years, significant advancements have been made in the field of dissolved oxygen prediction models, driven by extensive research and leveraging machine learning techniques. Various approaches have been explored, including integrating grey correlation degree, wavelet transformations, particle swarm optimization, deep belief networks, principal component analysis, clustering techniques, extreme learning methods, and support vector regression, among others [7]-[13]. These models aim to enhance the accuracy and efficiency of DO prediction, thereby supporting informed decision-making in aquaculture management practices.

This project focuses on developing predictive models for dissolved oxygen in intensive aquaculture settings, aiming to provide stakeholders such as fish farmers, hatcheries, and researchers with reliable tools for efficient water quality management. By harnessing the power of deep learning and analyzing large datasets, the project seeks to improve predictive accuracy, enable real-time monitoring, and facilitate proactive intervention strategies to maintain optimal conditions for aquatic organisms. Through accurate DO prediction, the project aims to enhance production efficiency, mitigate risks, and promote sustainable practices in aquaculture operations.

2. LITERATURE SURVEY

Aquaculture, particularly in China, has witnessed significant development, with the nation contributing over 70% of the global aquaculture output [1]. Water quality is crucial in aquaculture, where dissolved oxygen (DO) levels serve as a key indicator due to their influence on aquatic organisms' metabolism and survival [2]. Predicting DO concentrations accurately is essential for maintaining optimal conditions and preventing adverse effects on aquatic health and productivity [3].Hybrid machine learning techniques have also been investigated for DO prediction. Dehghani et al. utilized hybrid machine learning methods for dissolved oxygen concentration predictions in running waters [4]. Cao et al. developed a three-dimensional prediction method based on attention-GRU-GBRT for dissolved oxygen in pond culture [5]. Li et al. focused on predicting DO in fishery ponds using a gated recurrent unit (GRU) [6].

Furthermore, researchers have proposed hybrid neural network models for DO forecasting. Liu et al. presented a hybrid neural network model for marine dissolved oxygen concentrations time-series forecasting [7]. Ren et al. explored dissolved oxygen prediction in recirculating aquaculture systems using a deep belief network [8].

Overall, these studies demonstrate the importance of accurate dissolved oxygen prediction in aquaculture and the effectiveness of machine learning techniques in addressing this challenge. By leveraging advanced algorithms and hybrid models, researchers aim to improve predictive accuracy, enhance aquaculture management practices, and ensure the sustainability of aquatic environments.

3. METHODOLOGY

a) Proposed work:

The proposed work introduces a novel approach to predict dissolved oxygen levels in aquaculture environments using a hybrid model. This model combines LIGHTGBM for feature selection, BI-SRU for bidirectional training optimization, and Attention mechanism for parameter adjustments. By integrating these components, the model aims to enhance predictive accuracy compared to existing methods like LightGBM-LSTM and LightGBM-GRU.

Traditional methods for predicting dissolved oxygen levels, such as empirical equations, physical models, and manual monitoring, have limitations in accuracy. Despite the utilization of various algorithms like XGBOOST, CNN, LSTM, SVM, Linear Regression, and Decision Trees, the prediction accuracy remains insufficient. The proposed hybrid model addresses these limitations by leveraging advanced techniques and prioritizing minimal Mean Squared Error (MSE), Mean Absolute Error (MAE), or Root Mean Squared Error (RMSE) for improved accuracy in predicting dissolved oxygen levels.

b) System Architecture:

The system architecture for predicting dissolved oxygen levels in aquaculture environments using LightGBM involves several key steps. Firstly, the dataset is inputted, which includes parameters such as water temperature, pH, and turbidity. Next, data preprocessing techniques like cleaning, normalization, and handling missing values are applied to ensure data quality.

Following preprocessing, data visualization techniques such as histograms, scatter plots, and box plots are employed to gain insights into the distribution and relationships between variables. Feature extraction methods are then utilized to identify the most relevant features for predicting dissolved oxygen levels.

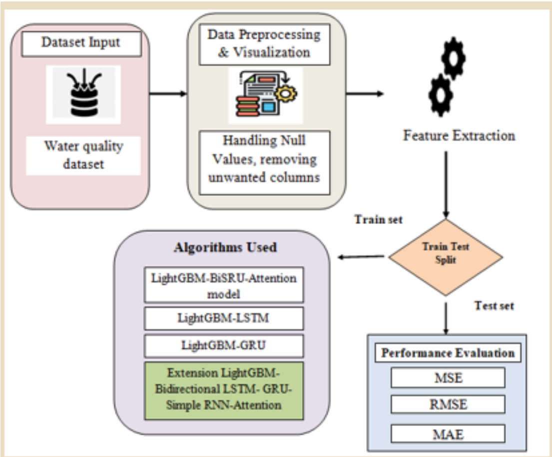

Fig 1 Proposed Architecture

The dataset is then split into training and testing sets to evaluate model performance. LightGBM, a gradient boosting framework, is employed as the primary algorithm for predicting dissolved oxygen levels. Finally, the performance of the LightGBM model is evaluated using metrics such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) to assess its accuracy and effectiveness in predicting dissolved oxygen levels accurately.

c) Dataset collection:

For this research, water quality sample data were collected from December 7, 2020, to January 18, 2021. The dataset comprised five water quality indicators and a total of 6050 observation samples.

During the data collection process, various environmental factors such as sensor aging or surface contamination may have introduced noise interference into the acquired data. Therefore, noise reduction processing was applied to the original signal of the monitored water quality data to reduce noise and recover the actual signal. In this process, abnormal data points were filled using the mean smoothing method, as illustrated in Equation 1. This ensured that the dataset used for analysis was free from noise and accurately represented the water quality parameters under consideration.

	Sample ID	pH	Temperature (°C)	Turbidity (NTU)	Dissolved Oxygen (mg/L)	Conductivity (µS/cm)
0	1	7.25	23.1	4.5	7.8	342
1	2	7.11	22.3	5.1	6.2	335
2	3	7.03	21.5	3.9	8.3	356
3	4	7.38	22.9	3.2	9.5	327
4	5	7.45	20.7	3.8	8.1	352
...	...	...	...	...	...	...
495	496	7.01	20.8	4.6	7.1	327
496	497	7.31	22.5	3.8	9.4	361
497	498	7.02	21.2	4.7	7.5	334
498	499	7.25	23.0	3.9	8.7	359
499	500	7.12	20.9	4.4	8.2	339

Fig 2 Data Set

d) DATA PROCESSING

In the data processing phase of the study, unwanted columns from the dataset were identified and removed to streamline the data for further analysis. This step involved carefully examining each column to determine its relevance to the prediction of dissolved oxygen levels in intensive aquaculture settings. Columns that were deemed irrelevant or redundant for the prediction task were excluded from the dataset to reduce noise and focus only on the essential features.

e) Feature Selection

In the feature selection phase of the study, a rigorous process was employed to identify the most relevant features that significantly influence the prediction of dissolved oxygen levels in intensive aquaculture environments. This involved analyzing the correlation between each feature and the target variable, as well as assessing the importance of each feature in the context of the prediction task.

Various techniques such as correlation analysis, feature importance ranking, and domain knowledge were utilized to prioritize features that exhibit strong predictive power and are meaningful in the context of aquaculture water quality management. Features that were found to have low correlation with the target variable or were deemed irrelevant to the prediction task were excluded from the modeling process.

f) TRAINING AND TESTING

In the training and testing phase, the dataset was randomly divided into two subsets: a training set and a testing set. The training set, comprising 80% of the total data, was used to train the hybrid model on historical observations of dissolved oxygen levels and corresponding feature values. This process involved optimizing the model parameters and learning the underlying patterns in the data.

Following training, the testing set, which accounted for the remaining 20% of the data, was utilized to evaluate the performance of the trained model. The testing set served as an independent dataset to assess the model's generalization ability and its capability to accurately predict dissolved oxygen levels in unseen data.

By splitting the dataset into training and testing subsets, the hybrid model's performance could be objectively evaluated, providing insights into its effectiveness in predicting dissolved oxygen levels and ensuring efficient water quality management in intensive aquaculture systems.

g) ALGORITHMS:

Light GBM-LSTM

Light GBM-LSTM is a hybrid model that combines the Light GBM algorithm for feature selection and the Long Short-Term Memory (LSTM) [17]neural network for sequential data processing. It leverages the strengths of both algorithms to enhance prediction accuracy. In the project, Light GBM-LSTM is utilized to predict dissolved oxygen levels in intensive aquaculture systems. The Light GBM algorithm selects relevant features from the dataset, while the LSTM[17] network learns temporal dependencies in the data. By integrating these two techniques, the model can effectively capture complex patterns in dissolved oxygen levels over time, contributing to efficient water quality management in aquaculture operations.

Light GBM-GRU

LightGBM-GRU is a hybrid model combining LightGBM for feature selection and Gated Recurrent Unit (GRU) [9] for sequence modeling. It selects relevant features using LightGBM and employs GRU to capture temporal dependencies in sequential data. In the project, LightGBM-GRU[9] is applied to predict dissolved oxygen levels in intensive aquaculture systems. LightGBM assists in identifying important features from the dataset, while GRU processes the sequential nature of the data. By integrating these two techniques, the model effectively captures temporal patterns in dissolved oxygen levels, contributing to improved predictions and efficient water quality management in aquaculture settings.

Light GBM-BISRU-Attention

Light GBM-BISRU-Attention is a hybrid model designed for predicting dissolved oxygen levels in intensive aquaculture environments. It combines Light GBM for feature selection, Bidirectional Simple Recurrent Unit (BISRU) for bidirectional training optimization, and Attention mechanism for parameter adjustments. Light GBM identifies relevant features, BISRU[25] optimizes bidirectional training to capture temporal dependencies, and Attention mechanism adjusts parameters dynamically. This model is utilized in the project to enhance the accuracy of dissolved oxygen predictions by effectively leveraging feature selection, bidirectional training, and parameter adjustments, thus facilitating efficient water quality management in aquaculture systems.

Light GBM- Bidirectional LSTM- GRU-Simple RNN-Attention

The LightGBM-Bidirectional LSTM-GRU-Simple RNN-Attention algorithm is a comprehensive hybrid model developed to predict dissolved oxygen levels in intensive aquaculture. It integrates LightGBM for feature selection, Bidirectional LSTM and GRU layers for capturing temporal dependencies bidirectionally, Simple RNN for sequential modeling, and Attention mechanism for dynamic parameter adjustment. This model offers enhanced accuracy in dissolved oxygen prediction by leveraging feature selection, bidirectional and sequential modeling, and dynamic parameter adjustments. It is applied in the project to improve the understanding of dissolved oxygen dynamics in aquaculture systems, facilitating more effective water quality management for optimized production outcomes.

4. EXPERIMENTAL RESULTS

MSE

Mean Squared Error (MSE) is a statistical metric used to measure the average squared difference between the predicted values and the actual values in a dataset. It is calculated by taking the average of the squared differences between each predicted value and its corresponding actual value. MSE provides a quantitative measure of the overall accuracy of a predictive model, where smaller values indicate better performance. It is widely used in regression analysis and machine learning to assess the goodness-of-fit of a model and to compare the performance of different models based on their prediction accuracy.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

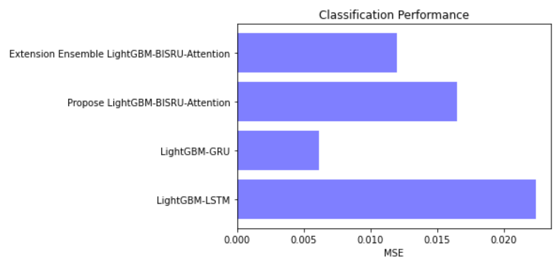

Fig 3 COMPARISON GRAPHS

RMSE

Root Mean Squared Error (RMSE) is a statistical measure of the average magnitude of the differences between predicted values and observed values in a dataset. It is calculated by taking the square root of the mean of the squared differences between predicted and actual values. RMSE provides a measure of the accuracy of a predictive model, similar to MSE, but it is more interpretable as it is in the same unit as the target variable. RMSE is commonly used in regression analysis and machine learning to evaluate the performance of models, where lower values indicate better predictive accuracy.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

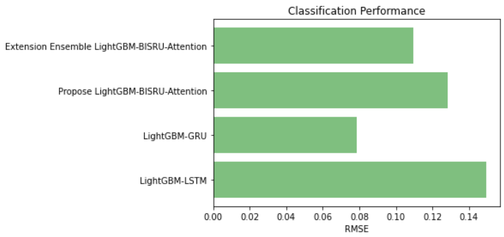

FIG 4 COMPARISON GRAPHS

MAE

Mean Absolute Error (MAE) is a metric used to measure the average magnitude of errors between predicted and actual values in a dataset. It is calculated by taking the average of the absolute differences between predicted and observed values. MAE provides a straightforward measure of the accuracy of a predictive model, representing the average absolute deviation of predictions from the true values. Unlike other error metrics such as MSE or RMSE, MAE is less sensitive to outliers since it does not square the errors. It is commonly used in regression analysis and machine learning to assess model performance, where lower MAE values indicate better predictive accuracy.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

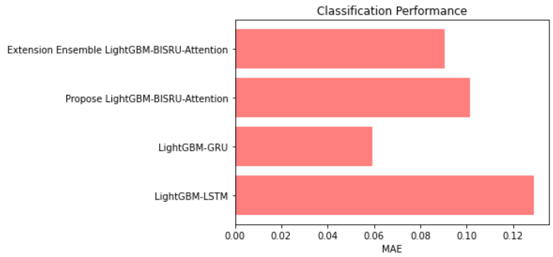

Fig 5 COMPARISON GRAPHS

	ML Model	MSE	RMSE	MAE
0	LightGBM-LSTM	0.022421	0.149737	0.129270
1	LightGBM-GRU	0.006148	0.078412	0.059561
2	Propose LightGBM-BISRU-Attention	0.016512	0.128499	0.101702
3	**Extension LightGBM- Bidirectional LSTM- GRU-Simple RNN-Attention**	**0.011986**	**0.109481**	**0.090655**

FIG 6 PERFORMANCE EVALUATION TABLE

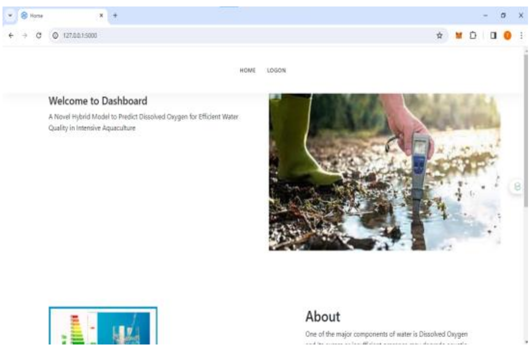

FIG 7 HOME PAGE

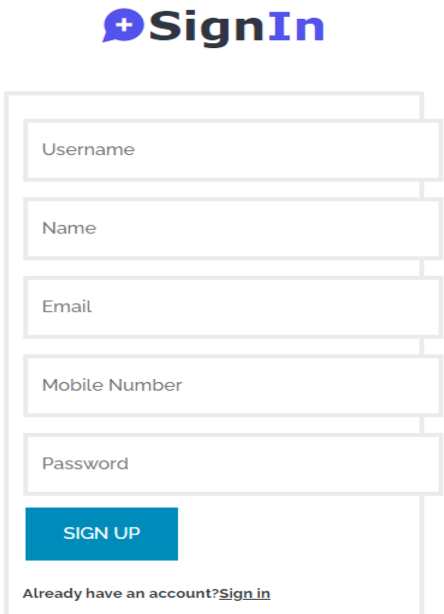

FIG 8 Sign Up

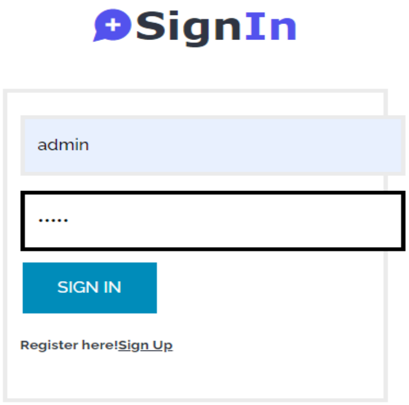

Fig 9 SIGN IN

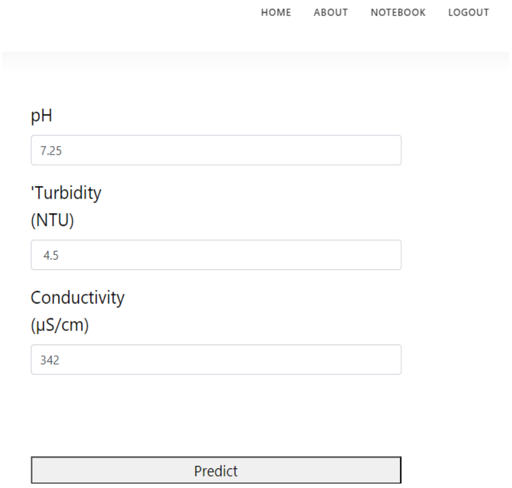

Fig 10 Upload Input Data

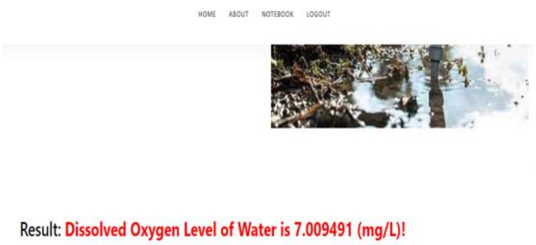

Fig 11 Predicted Result

5. CONCLUSION

In conclusion, the proposed LightGBM-BISRU-Attention model presents a significant advancement in predicting water quality in intensive aquaculture settings. By integrating LightGBM for feature selection, Bidirectional Simple RNN for training optimization, and Attention for parameter updates, the model achieves notable improvements in dissolved oxygen prediction accuracy. Comparative analysis against variants like LightGBM+LSTM and LightGBM+GRU underscores the superiority of the hybrid approach in both accuracy and efficiency. Moreover, the Ensemble model, combining various algorithms, further enhances prediction accuracy, demonstrating the efficacy of ensemble techniques. The integration of Flask with SQLite offers a user-friendly interface for model testing, facilitating practical usability for aquaculture industries, environmental agencies, and researchers. Overall, accurate dissolved oxygen prediction supports sustainable aquaculture practices and environmental conservation efforts, benefiting communities reliant on aquatic food exports.

6. FUTURE SCOPE

In the realm of predicting dissolved oxygen for efficient water quality management in intensive aquaculture, there exist several avenues for future exploration and enhancement. Firstly, the integration of additional advanced machine learning techniques such as deep learning architectures like transformers or attention mechanisms could further refine prediction accuracy and efficiency. Secondly, incorporating real-time sensor data and IoT devices can enable continuous monitoring and adaptive management of aquaculture systems, enhancing responsiveness to dynamic environmental conditions. Additionally, expanding the scope of the model to encompass predictions for other critical water quality parameters beyond dissolved oxygen, such as pH, ammonia levels, and temperature, would provide a comprehensive tool for holistic aquaculture management. Furthermore, collaboration with industry stakeholders and environmental agencies can facilitate the development of tailored solutions and implementation strategies, ensuring the practical applicability and scalability of the model in diverse aquaculture settings. Overall, continued research and innovation hold the potential to advance the efficacy and sustainability of intensive aquaculture practices.

7. REFERENCES

[1] F. Hu, ‘‘Development of fisheries in China,’’ Reprod. Breeding, vol. 1, no. 1, pp. 64–79, 2021.
[2] S. Ayesha Jasmin, P. Ramesh, and M. Tanveer, ‘‘An intelligent framework for prediction and forecasting of dissolved oxygen level and biofloc

amount in a shrimp culture system using machine learning techniques,'' Expert Syst. Appl., vol. 199, Aug. 2022, Art. no. 117160.

[3] M. H. Ahmed and L.-S. Lin, ''Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique,'' J. Hydrol., vol. 597, Jun. 2021, Art. no. 126213.

[4] R. Dehghani, H. Torabi Poudeh, and Z. Izadi, ''Dissolved oxygen concentration predictions for running waters with using hybrid machine learning techniques,'' Model. Earth Syst. Environ., vol. 8, no. 2, pp. 2599–2613, Jun. 2022.

[5] X. Cao, N. Ren, G. Tian, Y. Fan, and Q. Duan, ''A three-dimensional prediction method of dissolved oxygen in pond culture based on attention-GRU-GBRT,'' Comput. Electron. Agricult., vol. 181, Feb. 2021, Art. no. 105955.

[6] W. Li, H. Wu, N. Zhu, Y. Jiang, J. Tan, and Y. Guo, ''Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU),'' Inf. Process. Agricult., vol. 8, no. 1, pp. 185–193, Mar. 2021.

[7] H.Liu, R. Yang, Z. Duan, and H. Wu, ''A hybrid neural network model for marine dissolved oxygen concentrations time-series forecasting based on multi-factor analysis and a multi-model ensemble,'' Engineering, vol. 7, no. 12, pp. 1751–1765, Dec. 2021.

[8] Q. Ren, X. Wang, W. Li, Y. Wei, and D. An, ''Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network,'' Aquacultural Eng., vol. 90, Aug. 2020, Art. no. 102085.

[9] X. Cao, Y. Liu, J. Wang, C. Liu, and Q. Duan, ''Prediction of dissolved oxygen in pond culture water based on K-means clustering and gated recurrent unit neural network,'' Aquacultural Eng., vol. 91, Nov. 2020, Art. no. 102122.

[10] P. Shi, G. Li, Y. Yuan, G. Huang, and L. Kuang, ''Prediction of dissolved oxygen content in aquaculture using clustering-based softplus extreme learning machine,'' Comput. Electron. Agricult., vol. 157, pp. 329–338, Feb. 2019.