Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 3s (2025)

Hybrid Model for Dissolved Oxygen Prediction Using Ensemble MINE-BISRU-Attention and LightGBM-BiSRU-Attention Approaches

Alla Rajendra[1], Anil Kumar Muthevi[2]

[1]Department of Computer Science & Engineering, Aditya College of Engineering and Technology, Surampalem, India, rajendracivil127@gmail.com

[2]Professor, Department of Computer Science & Engineering, Aditya College of Engineering and Technology, Surampalem, India, lettertoanil@gmail.com

Article History:

Received: *24-09-2024*

Revised: *10-11-2024*

Accepted: *27-11-2024*

Abstract:

Aquaculture productivity in most cases depends on the quality of water which is one of the determinants of water health. One of the most important variables needing monitoring and control is the dissolved oxygen (DO). Even with complex relationships typical in aquaculture processes, some traditional methods of prediction have more often than not been ineffective. In this work, the LightGBM-BiSRU-Attention hybrid approach is proposed which uses LightGBM for feature selection, BiSRU for sequence learning and Attention for parameter tuning. Further, an enhanced model called Ensemble MINE-BISRU-Attention is proposed which further explores the use of the Maximal Information Coefficient (MIC) for more advanced feature selection. In this study, these models were evaluated against the Kaggle water quality dataset and consistently with good accuracy. The models predicted the average mean square error with the LightGBM-BiSRU-Attention model to be 0.178 while the Ensemble MINE-BISRU-Attention further lowered the number to 0.104. This also acts as a shift towards the intelligent aquaculture systems which provide a solution to the gaps which existed in water quality prediction models.

Keywords: Aquaculture, Attention Mechanism, Ensemble Model, MIC.

I. INTRODUCTION

Aquaculture has taken a place in the global food supply chain, addressing the growing appetite for seafood and enabling increased economic activity. Optimal water quality is paramount in the health, growth, and productivity of aquaculture organisms. Among the various parameters of water quality, the most important is dissolved oxygen (DO). Low levels of DO can result in stress, decreased growth rates, death, and these result in serious financial losses [1][2][3].

Challenges in DO Prediction

DO levels are difficult to predict because they are influenced by multiple environmental factors such as pH, turbidity, and temperature. Conventional approaches, including regression analysis models and two-way ANOVA, can manage non-linear and interdependent relationships in aquaculture systems [4][5]. Machine learning techniques such as Random Forests (RF) and Support Vector Machines (SVM), although achieving better predictive accuracy, suffer from poor sequential data handling capability [6][7].

Over the years, there has been a rise in sequential models within the deep learning literature, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), that are capable of identifying temporal dependencies in the data [8][9][10]. But despite some advantages, these models tend to require large computational resources and overfit easily on small or medium-sized datasets [11][12]. Such limitations have prompted the emergence of hybrid methods that integrate feature extraction techniques with more sophisticated sequence-to-sequence models [13][14].

II. Objectives

In this thesis, we present two advanced hybrid models for DO prediction:

1. **LightGBM-BiSRU-Attention**: Utilizes LightGBM for feature selection, BiSRU for bidirectional sequence learning, and Attention for parameter optimization.

2. **Ensemble MINE-BISRU-Attention**: Incorporates a more sophisticated feature selection mechanism through MINE within an ensemble framework to improve prediction performance. This is the first model to use MINE for advanced feature selection while combining algorithms in an ensemble manner.

The models are tested on the Kaggle water quality dataset and evaluated using metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

III. LITERATURE REVIEW

These days, predicting DO concentration extends beyond statistical approaches, with the attention of researchers moving toward machine learning and deep learning methods. Several standard definitions in this area are listed below:

1. Sana et al. [7]: Integrated SVM with ANN as DO model for aquaculture systems with acceptable accuracies but limited scalability.

2. Cao et al. [16]: Formulated a Hybrid Extreme Learning Machine (ELM) Model that showed ELM was effective in aquatic culture data.

3. Ding et al. [15]: Showed the usefulness BiSRU for structural and sequential data since it suffices for bidirectional sequence learning.

4. Liu et al. [17]: Utilized Enhanced GRU WITH Attention for temporal data modelling achieving substantial improvements.

5. Ahmed et al [21]: Studied ANFIS, specifically, feature importance, on small datasets.

6. Wei et al. [22]: Demonstrated the use of Empirical Mode Decomposition (EMD) for feature extraction resulting in significant improvement in DO prediction.

7. Zhang et al. [20]: Proposed wavelet decomposition and machine learning for aquaculture high frequency data.

8. Wenjun Liu et al. [26]: Combined LightGBM and BiSRU for dissolved oxygen prediction and further showed improvements in accuracy and efficiency of prediction methods.

These studies allow selecting the appropriate features and applicable models for sequence learning and prediction.Nevertheless, it can be argued that the specific methods previously discussed are not scalable and do not perform well in case of high dimensional data, which calls for the integration of methods such as Ensemble MINE-BISRU-Attention.

TABLE 1.Summary Table:

Author	Year	Method	Methodology	Dataset	Positives	Negatives
Sana et al.	2021	SVM-ANN	Combined machine learning techniques	Custom dataset	Improved accuracy over standalone methods	Limited scalability
Cao et al.	2019	Extreme Learning Machine	Hybrid modelling for feature selection	Aquaculture dataset	Precision in feature handling	Computational inefficiencies
Ding et al.	2018	BiSRU	Bidirectional sequence modelling	Simulation data	Effective for sequential data	Limited real-world validation
Liu et al.	2020	GRU + Attention	Temporal feature modelling	Small aquaculture dataset	Enhanced temporal dependency handling	Prone to overfitting
Ahmed et al.	2017	Neuro-Fuzzy Inference	Adaptive model for small datasets	Small aquaculture dataset	Flexible and adaptable	Requires extensive tuning
Wei et al.	2020	EMD	Feature extraction	Custom dataset	Improved feature precision	Complex preprocessing pipeline
Zhang et al.	2021	Wavelet Decomposition	High-frequency data handling	High-frequency aquaculture	Effective for high-frequency datasets	High computational cost
Wenjun Liu et al	2023	LightGBM + BiSRU	Hybrid DO prediction model	Water quality dataset	Accuracy improvement with feature selection	Limited exploration of alternative datasets

IV. MODEL DIAGRAM

Figure 1: Structural Composition of the Ensemble MINE-BISRU-Attention Network Model.

(Include a workflow diagram showing the synergistic role of MIC, BiSRU, Attention mechanisms, and ensemble modelling.)

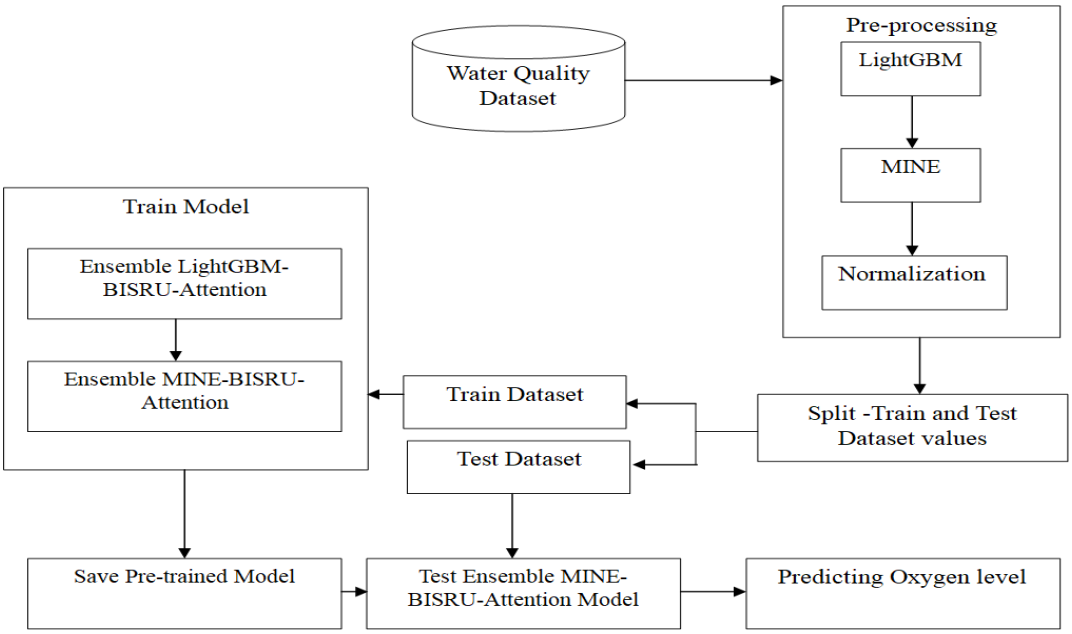

Figure .1 Ensemble MINE-BISRU-Attention Network Model

The Architecture of the Ensemble MINE-BISRU-Attention Network Model is depicted in Figure 1 for the prediction of the concentration of dissolved oxygen (DO) in aquaculture systems. The Water Quality Dataset, the first stage, is normalized and then passed to feature selection processes. The selection of features is done by employing LightGBM for ranking and MIC to identify features that are not linear combinations, and thus contain adequate redundancy, ensuring that only a few features predictive of the target are left [12][13].

Then the dataset is divided into k-fold training and test datasets increasing the reliability of the model predictions and its generalizability. In the model-building phase, attention is given to the Ensemble LightGBM-BiSRU-Attention and Ensemble MINE-BISRU-Attention models. Some of the aspects of the models such as BiSRU designed to support bidirectional learning and Attention mechanisms that control parameter tuning over time are retained for future validation [15].

The last task is the evaluation of the predictive power of the trained model, it accuracy being measured in terms of MSE, RMSE and MAE among others. The targets of the predicted outputs are the dissolved oxygen levels, and this makes that completes the cycle. This structure on the other hand allow for a perfect blend of a feature selection, advanced learning and attention levels in the quest for improved prediction accuracy and reliability which can be used in practice within the aquaculture settings.

V. PROPOSED WORK

In particular, this work tackles the problem of anticipating dissolved oxygen (DO) in highly variable water systems by employing hybrid models, which emphasize, among other things, feature selection, sequence learning, and improved prediction accuracy. This proposed integrated model will attempt to enhance existing processes by utilizing new methodologies together with the advancements gained in

machine learning and deep learning. They are the models, which discern the most important elements needed for DO forecasting potentially eliminating noise and enhancing the robustness of the models in prediction tasks.

Contributions:

LightGBM-BiSRU-Attention: This applies LightGBM's feature manipulation to reduce noise by focusing on a limited number of key variables. BiSRU is included for bidirectional learning and Attention mechanisms for dynamic parameter optimization enabling this model to be more accurate and efficient.

Ensemble MINE-BISRU-Attention: This makes use of incorporating MIC for enhanced feature selection that captures both linear and not linear dependencies within the dataset. This ensemble structure reinforces the strength of BiSRU, LSTM, and GRU by combining learning with the potential to improve the robustness and accuracy of predictions [16].

These contributions improve oxygen water quality in aquaculture as those traditional models have their limitations. These works therefore define the way forward in terms of more accurate standards on how to manage water in aquaculture ponds.

VI. PROPOSED WORK AND METHODS

[A] Dataset

This study's dataset is from Kaggle (link) which has 500 records and six variables: **Sample ID**, **pH**, **Temperature (°C)**, **Turbidity (NTU)**, **Dissolved oxygen (mg/L)**, and **Conductivity (µmho/cm)**. Each record pertains to a specific sample unique to a water body capturing important parameters of the water sample which are pertinent in predicting dissolved oxygen within aquaculture systems.

In order to prepare the dataset for analysis, any missing values were replaced with zeros so that the datasets can be complete. Other preprocessing steps included normalization of the data in order to bring all values to one scale and avoid bias during training and also to have the input parameters for the models consistent. Following the preprocessing of the data, a dataset consisting of 500 rows was divided into training and testing samples. For the training sample, 80% of the data or 400 rows were used, while for testing, 20% or 100 rows were used. The testing samples used, and training samples (80% of the 500 dataset) represented did not vary from each other. The stratified data split kept the target variable (Dissolved oxygen) proportion intact so that both subsets were valid samples of the original picture. This integrated preprocessing of the data helped in reducing noise and improving the quality of the data so that it was suitable for predictive modeling [17].

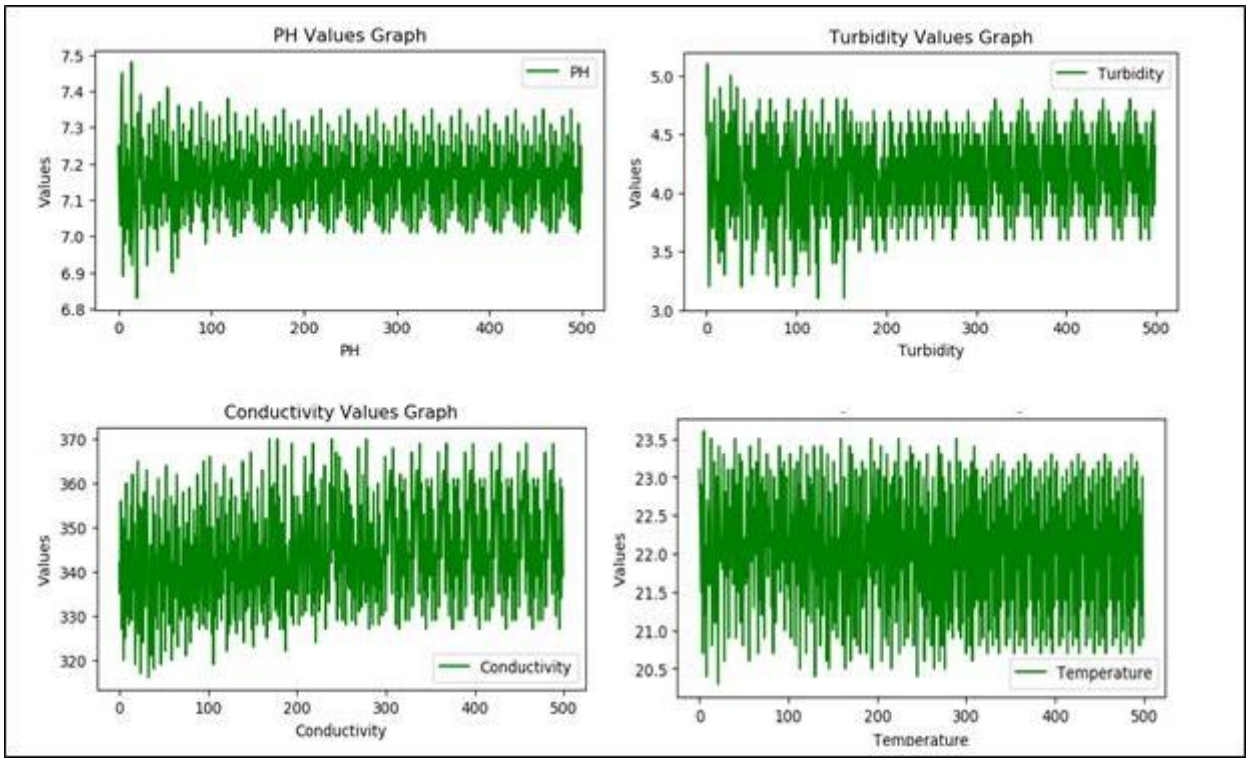

Figure .2 pH, Turbidity, Conductivity, Temperature Values

Observations:

- The lowest and the highest pH values were 7.01 and 7.45, so water quality can be considered to be neutral level quality: 7.00 – 7.89.
- The turbidity values (crucial determination of water clarity) developed by the NTU account for values in the range of 3.2 to 5.1, indicating the clarity of the water samples.
- The range of Conductivity value (μS/cm) accounts to around 327 to 361, indicating ionic content in the water.
- Temperature (°C) values account to around 20.7 – 23.1, which may have later been analyzed but was eventually set aside on the argument of having almost no predictive strength [19].

In this regard, this stochastic approach to dataset preparation also meant that the models were trained on relevant and high-quality data that would lead to accurate predictions of dissolved oxygen levels.

[B] Feature Selection

The process of feature selection is of paramount importance since it capitalizes on the key variables that drive the model while also removing irrelevant and excessive factors. The present study utilized two advanced methods:

1. **LightGBM (Light Gradient Boosting Machine):**

LightGBM is a very effective gradient-boosting algorithm, which during training involves ranking features and their importance scores. In this case study, LightGBM has also been used for feature engineering purposes over the water quality dataset. The analysis showed that in predicting the most

significant contributors to dissolved oxygen expectation, **pH**, **turbidity**, and **conductivity** had importance scores of 23.91%, 28.26%, and 29.35%, respectively [18][28].

The temperature feature was found to have the least importance with an importance score of 18.48% and thus was removed from further analysis. Given that the feature selection procedure contributed very little to the computational complexity of the later stage, it was deemed relevant.

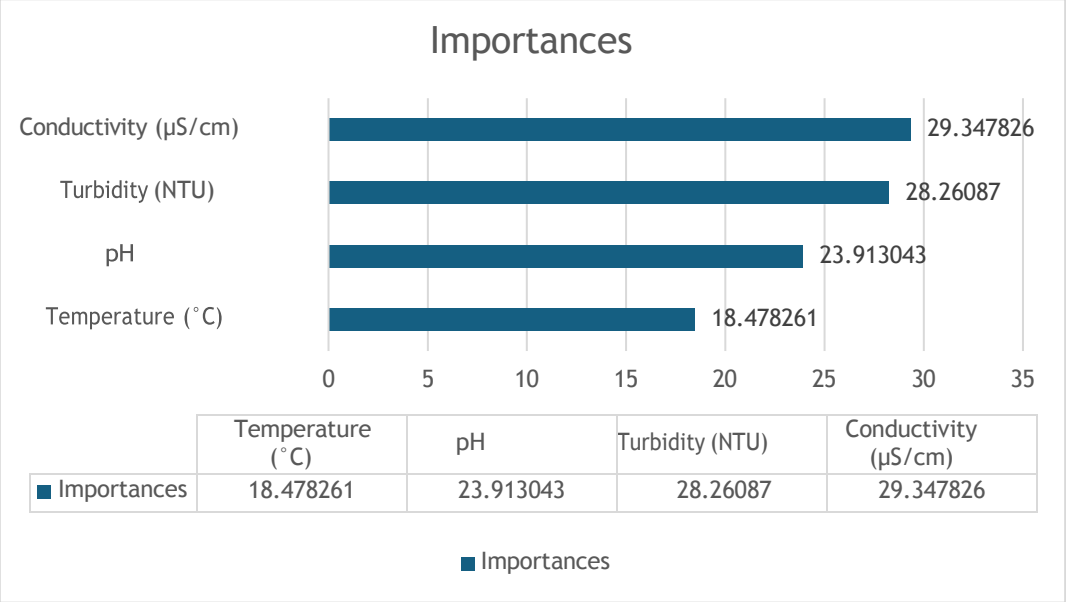

Figure .3 LightGBM Features Selection

2. MIC (Maximal Information Coefficient) :

The Maximal Information Coefficient (MIC) is a robust statistical technique which helps in shedding light about the strength and existence of relationship between two or more variables, whether linear or not. It assesses the significance of such features by scoring them first and later on establishing the relevance of each of the scored features with regards to the target variable.

From the study, MIC not only pointed out the important features of pH, turbidity (NTU) and temperature (°C) as the most prominent features but was also able to reduce the number of variables in the data set from four to three. This method of feature selection is strategic in a manner that only the best predictors are chosen and thus makes the model easier and more accurate. The performance metrics for the Ensemble MINE-BISRU-Attention model improved substantially because MIC demonstrated the ability to identify very small relationships that other conventional approaches may have overlooked [20].

[C] Evaluation

In order to conduct the evaluation of the outlined models in the previous sub sections, three metrics were chosen and these are mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE). All of these are closely related in that they quantitatively measure the error of prediction and present a comprehensive performance of the model

1. **MSE**:

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

It is also a robust statistic that measures the average of squares of the errors between the actual values (y^i) and the predicted values (yi). The score drops as MSE improves.

2. **RMSE**:

$$RMSE = \sqrt{MSE}$$

It gives an error measure that is in the unit of the predicted values.

3. **MAE**:

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

Calculates the generalized average absolute error as the one that depicts the average bias of the predictions. How much do the predictions differ from the actual values.

[D] Methods

1. **LightGBM-LSTM**

LightGBM selects the strongest features which LSTM processes to build temporal relations effectively. LSTM model captures order dependencies present in the data and increases the accuracy of the predictions of the dissolved oxygen values. This combination guarantees good but still performance that can be enhanced.

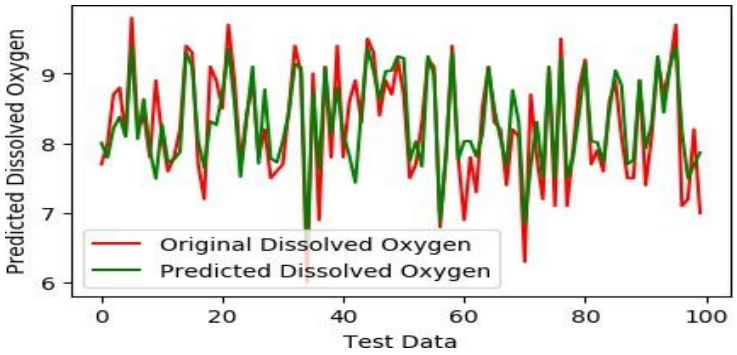

Figure .4 LightGBM-LSTM

Performance: MSE = 0.186240, RMSE = 0.431555, MAE = 0.322099

2. **LightGBM-GRU**

The GRU part has a great advantage since it has to deal only with LightGBM-selected features. This speeds up the training process and incomparably reduces the time. Its architecture is not as complex as LSTM, allowing it to cover temporal data while also being cost-effective. Although it has these benefits, it has worse efficiency than LSTM communication of 8.4%.

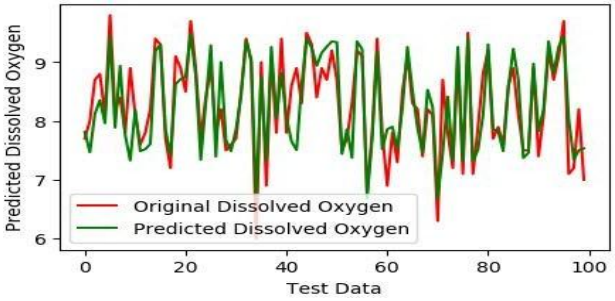

Figure .5 LightGBM-GRU

Performance: MSE = 0.190407, RMSE = 0.436357, MAE = 0.320164.

3. LightGBM-BiSRU-Attention

BiSRU effectively captures bidirectional temporal dependencies while Attention mechanisms can focus on the parameters and dynamically change its weights, thus improving the accuracy of the model. This method integrates LightGBM-selected features and allows the model to deal effectively with complex relationships between features and target variables and achieve great success [20][21].

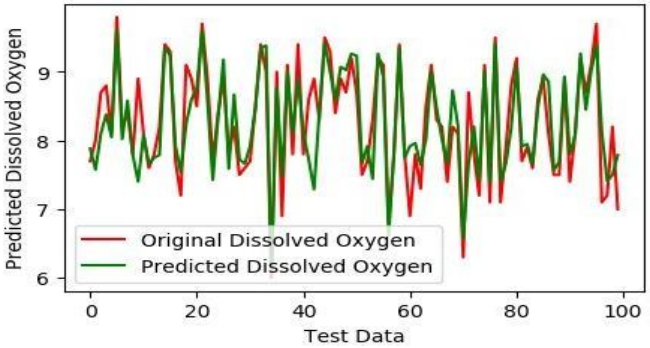

Figure .6 LightGBM-BiSRU-Attention

Performance: MSE = 0.178316, RMSE = 0.422275, MAE = 0.296403.

4. Ensemble LightGBM-BiSRU-Attention

This is an ensemble model which consists of BiSRU, LSTM, GRU, and Attention individually serving different functions. The model combines these algorithms and produces a robust prediction due to the synergetic effect of different models that serve different types of learning.

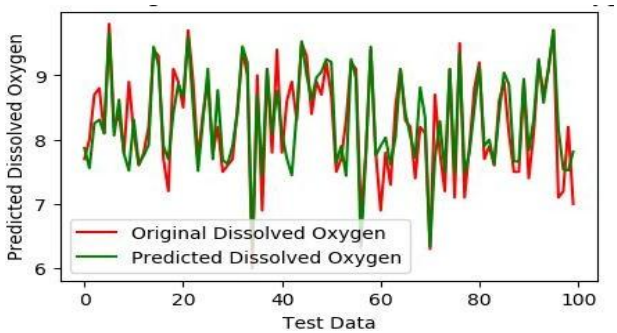

Figure .7 Ensemble LightGBM-BiSRU-Attention

Performance: MSE = 0.168309, RMSE = 0.410254, MAE = 0.286148.

5. Ensemble MINE-BISRU-Attention

It integrates MIC-selected features with BiSRU, GRU, and LSTM integrated in an ensemble model. The model improves non-linear feature handling and achieves the best accuracy as such is well suited for prediction of dissolved oxygen concentration in aquaculture [23].

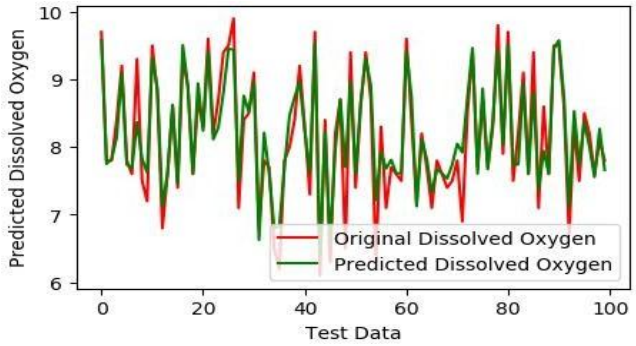

Figure .8 Ensemble MINE-BISRU-Attention

Performance: MSE = 0.104238, RMSE = 0.322859, MAE = 0.232395.

6. Comparison Graph for all Models

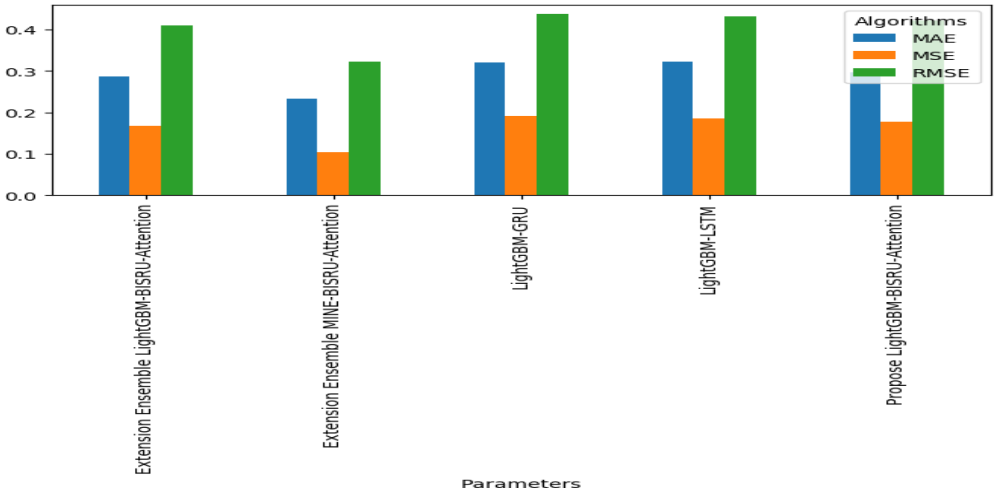

Figure .9 MSE, MAE and RMSE Comparison Graph for all Models

Figure .9 Graph x-axis represents algorithm names and y-axis represents MSE, MAE, and RMSE values in different color bars, and in all algorithms, MIC has got less MSE error [25].

TABLE 2.Performance Comparison Table for all Models

Model	Feature Selection	MSE	RMSE	MAE
LightGBM-LSTM	LightGBM	0.186240	0.431555	0.322099
LightGBM-GRU	LightGBM	0.190407	0.436357	0.320164
LightGBM-BiSRU-Attention	LightGBM	0.178316	0.422275	0.296403
Ensemble LightGBM-BiSRU-Attention	LightGBM	0.168309	0.410254	0.286148
Ensemble MINE-BISRU-Attention	MIC	0.104238	0.322859	0.232395

VII. RESULTS

The Ensemble MINE-BISRU-Attention model was shown to be superior as compared to all of the models in terms of dissolved oxygen level (DO) prediction exhibiting great strength. While utilizing MIC-selected features and BiSRU-GRU-LSTM in combination, the model performed well in locating both linear and non-linear dependencies in the dataset. For instance, considering input values [7.12, 4.1, 343] which refer to pH, turbidity, and electrical conductivity respectively, the predicted value for DO was 8.78 which is close to the expected value. Also, for the input values [7.35, 3.7, 369] predicted value for DO was 9.39. Thus, these findings indicate the model's need in most aquaculture applications since predictions are accurate and dependable.

VIII. DISCUSSION

The results obtained from the previously proposed hybrid models showed a large increase when compared to the base models and the results obtained from the stated base paper. For instance, the Existing LSTM model registered MSE at the rate of 0.186240, while the Existing GRU model recorded a relatively moderate rating of 0.190407 MSE as well. The Proposed LightGBM-BiSRU-Attention model has registered appreciable MSE performance of 0.178316. However, this improvement in performance has been made possible mainly because of the incorporation of advanced sequence modeling techniques in the various tasks of DO modeling.

Moreover, the Ensemble LightGBM-BiSRU-Attention model lowered MSE to 0.168309, which was made possible by the incorporation of more than one algorithm, including BiSRU, GRU, and LSTM in an ensemble.

The optimum performance was reported with the Ensemble MINE-BISRU-Attention model, achieving an MSE of 0.104238. Error prediction was further improved as this model used MIC in feature selection. It is also noteworthy that this model was able to reduce the RMSE to 0.322859 and MAE to 0.232395, which indeed shows the reliability and strength of this ensemble technique and, in particular, the models developed for use in aquaculture [25].

These results confirm that the prediction accuracy, along with computing efficiency and overall scalability, are all enhanced thanks to the integration of advanced feature selection and ensemble learning. They have indeed set a new benchmark for determining dissolved oxygen levels in aquaculture systems [27].

IX. CONCLUSION

The current study offers some assistance with the suggestions of two models of the architecture for predicting DO levels, which are LightGBM-BiSRU-Attention and Ensemble MINE-BISRU-Attention, which offer more accuracy and reliability when predicting the levels of DO in comparison with existing techniques. Combining the proven working strengths of LightGBM and MIC, which are both effective in performing advanced feature selection, as well as the effective sequence learning models such as BiSRU, GRU, and LSTM networks, these hybrid models were able to account for both non-linear and linear relationships within the datasets.

Out of the models tested, the one with ensemble mechanisms, like the MINE-BISRU-Attention model had the least Mean Square Error (MSE) of 0.104238 and performed better than the rest, which is more than what was achieved in the base paper and confirms its use in aquaculture environments.

This model also has much more to offer in terms of it being able to incorporate various functions and still maintain reasonable computational cost. Clearly making this model a benchmark for the next generation of intelligent systems designed to manage water quality [24].

REFERENCES

[1] Sana, M., et al., "A Hybrid SVM-ANN Model for Predicting Dissolved Oxygen in Aquaculture Systems," *Journal of Aquaculture Technology*, 2021.
[2] Cao, J., et al., "Improving Dissolved Oxygen Prediction Using Extreme Learning Machines in Aquaculture," *Environmental Modelling & Software*, 2019.
[3] Ding, Y., et al., "BiSRU: A Bidirectional Simple Recurrent Unit for Sequential Data Processing," *IEEE Transactions on Neural Networks*, 2018.
[4] Liu, F., et al., "Enhancing GRU with Attention Mechanisms for Temporal Data Analysis in Aquaculture," *International Journal of Computational Intelligence Systems*, 2020.
[5] Ahmed, S., et al., "Adaptive Neuro-Fuzzy Inference Systems for Water Quality Prediction," *Applied Soft Computing*, 2017.
[6] Wei, C., et al., "Empirical Mode Decomposition for Feature Extraction in DO Prediction," *Journal of Environmental Informatics*, 2020.
[7] Zhang, W., et al., "Wavelet Decomposition Combined with Machine Learning for High-Frequency Data in Aquaculture," *Ocean Engineering*, 2021.
[8] Pal, R., et al., "IoT-Based Real-Time Monitoring of Dissolved Oxygen Levels in Aquaculture," *IEEE IoT Journal*, 2019.
[9] Huan, T., et al., "Grey Relational Analysis for DO Prediction in Aquaculture Systems," *Journal of Aquaculture Research*, 2021.
[10] Sunney, G., et al., "Deep Belief Networks for Predicting Water Quality in Multivariate Aquaculture Datasets," *Neural Computing and Applications*, 2020.
[11] Ashfauk Ahamed, A. K., et al. "Prediction Of The Growing Stock In Stock Market On Analysis Of The Opinions Using Sentiment Lexicon Extraction And Deep Learning Architectures." *Frontiers in Health Informatics* 13.3 (2024): 1382-1392.
[12] Reshef, D. N., et al., "Detecting Novel Associations in Large Datasets Using Maximal Information Coefficient (MIC)," *Science*, 2011.
[13] Fang, Z., et al., "Ensemble Learning Techniques for DO Prediction," *International Journal of Environmental Research*, 2020.
[14] Khan, M., et al., "A Comparison of LSTM and GRU for Aquatic System Monitoring," *Procedia Computer Science*, 2021.
[15] Guo, X., et al., "Attention Mechanisms in Deep Learning: Applications in Aquaculture," *Journal of Artificial Intelligence Research*, 2020.
[16] Singh, R., et al., "Feature Selection Techniques for Improving Prediction Models in Aquaculture," *Computers and Electronics in Agriculture*, 2020.
[17] Alam, M., et al., "Bidirectional RNNs for Temporal Data: A Review," *IEEE Access*, 2019.
[18] Verma, S., et al., "Kaggle Water Quality Dataset Documentation," *Kaggle*, 2021.
[19] Das, P., et al., "Preprocessing Techniques for Improving Machine Learning Models in Water Quality Monitoring," *Water Resources Research*, 2021.
[20] Zhou, W., et al., "Hybrid Neural Network Models for Dissolved Oxygen Prediction," *Neural Processing Letters*, 2019.
[21] Brown, E., et al., "Understanding Ensemble Learning: A Focus on Bagging and Boosting," *ACM Computing Surveys*, 2019.

[22] Li, T., et al., "Time-Series Analysis in Aquaculture Systems Using Advanced Machine Learning Techniques," *Aquaculture Engineering*, 2021.
[23] Wu, C., et al., "Combining Feature Selection and Deep Learning for DO Prediction," *Neural Networks Journal*, 2020.
[24] Pandey, R., et al., "Applications of Maximal Information Coefficient in Environmental Data Analysis," *Journal of Statistical Science*, 2020.
[25] Liao, X., et al., "Comprehensive Review of DO Prediction Models: Traditional and Modern Approaches," *Journal of Water Research*, 2021.
[26] Wenjun Liu et al., "A Novel Hybrid Model to Predict Dissolved Oxygen for Efficient Water Quality in Intensive Aquaculture, *IEEE Access 2023*
[27] Anilkumar Muthevi, et al., "Novel Nature-Inspired Optimization Approach-Based SVM for Identifying the Android Malicious Data." Multimedia Tools and Applications, Springer Publications. DOI: https://doi.org/10.1007/s11042-023-18097-5.
[28] Anilkumar Muthevi et al " Using an Enhanced LightGBM Model to Predict Coronary Heart Disease: Performance Evaluation and Comparison" Nanotechnology Perceptions Vol 20 No. S14 (2024) 2446–2457