Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Fish Disease Prediction Using Machine Learning on Water Quality Data

Rajat Kumar[1], Roshni Prasad[2]

[1] Dept. of Computer Science and Engineering Noida Institute of Engineering and Technology, Greater Noida (201310), Uttar Pradesh, India

[2] Dept. of Computer Science and Engineering Noida Institute of Engineering and Technology, Greater Noida (201310), Uttar Pradesh, India

***Corresponding Author Email**. rajat6670@gmail.com

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Fish diseases are a serious threat to the farming industry. To detect these diseases in the initial stage, a different machine learning model is proposed. This model has focused on identifying the fish diseases based on the water quality. To do this, we have used the "Aquaculture–Water Quality Dataset" data set havi15 physico-chemical parameters like PH, DO, hardness, solids, chloramines, iron, ammonia, nitrite, nitrate, phosphate, silica and the like. These attributes are linked with the existence of a disease. From the data set, we have created the Random Forest Classifier using the Python language and the Scikit library. The implementation has provided an efficiency of 89.25%. The metrics of precision, recall, and F1-score were employed to determine the efficiency. The evaluation has revealed that the proposed model has displayed the appropriate efficiency. The important feature analysis has shown that the dissolved oxygen , temperature ,PH, and ammonia are the critical factors used for the detection of the disease. Through the proposed method, a proper mechanism has been developed for detecting the fish diseases. This model can be used effectively to create proper system for the detection of fish diseases.

Keywords: Fish Disease Detection, Aquaculture, Water Quality Parameters, Machine Learning, Random Forest, Disease Classification, Environmental Monitoring, Predictive Modeling, Feature Importance

1. Introduction

A aquaculture replays a big role in the world's economy and provides a lot of our food, but it can be threatened by waterborne fish diseases that might wipe out entire stocks. Keeping an eye on water quality is really important to keep fish healthy. Things like temperature, pH, dissolved oxygen, turbidity, and ammonia levels all impact the environment where fish live. Usually, finding diseases involves waiting until about break happens or manually watching fish, which can take a lot of time and isn't very efficient. That's why people are looking for smarter systems that can predict disease risks by analyzing water quality data in real-time.

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

The attempts at using machine learning algorithms to monitor water conditions and diagnose fish diseases include the usage of SVM, Naïve Bayes, and K- Nearest Neighbours. However, most of such models were not sufficiently effective when dealing with diagnosing several diseases at once since such prediction requires the use of multiple parameters simultaneously. In addition, relatively few research papers include the creation of models that could be easily applied across various locations or even within actual fish farming facilities. The primary objective of this paper is to design a reliable machine learning model based on Random Forest to diagnose fish health problems by analysing the relevant water quality parameters. As such, a set of synthetic data, including pH, temperature, dissolved oxygen, turbidity, ammonia, nitrate, nitrite, and phosphate was developed. The data represents possible disease categories that can result from poor water quality. Thus, it may be useful in helping fish farmers to take prompt actions in order to prevent further deaths of the fish.

Organization of This Paper: Section 2 of this paper describes the methodology that we adopted including details about data cleaning and preprocessing; disease labelling based on the expert's opinion and water quality; what kind of features are important in determining whether a disease will occur or not and how the Random Forest algorithm was developed in predicting diseases. The outcome of the research including metrics of performance such as accuracy, precision, recall and confusion matrices will be presented in Section 3 of this paper. Comparison with other methods will also be done.

2. Literature review

On the other hand, Jeni Moni et al. [1] have designed a system using machine learning concepts where aquaculture conditions have been observed and fish disease is identified, leading to better health conditions in the fish population and lower financial losses. Similarly, M. Çakır et al. [2] have demonstrated that the performance of SVM and RFerns models surpasses that of standard classification methods like k-Nearest Neighbors(KNN) and Naïve Bayes (NB). Thus, their findings emphasize the usefulness of ensemble and margin maximization methods in disease detection. Taking another step ahead and shifting the focus toward the detection of water quality conditions, Al-Akhir Nayan et al. [3] applied ML-based algorithms for predicting water quality degradation, as this marks the initial indication of fish disease onset. Predictive capabilities are necessary in disease prevention measures. Moreover, Daoliang Li et al. [5] emphasize the importance of applying computer vision techniques for fish disease diagnosis by increasing accuracy.

Moreover, by incorporating advanced Deep Learning approaches, Ssekit to Isaacetal. [4][13] applied Deep Transfer Learning to increase the accuracy of disease detection. Furthermore, Explainable AI(XAI) techniques were incorporated to ensure model explainability.

In disease diagnosis experiments, specific studies were conducted by K. Sujatha et al. [9], where they used an SVM Gaussian kernel algorithm to detect Epizootic Ulcerative Syndrome (EUS) in fish with an accuracy rate of 82.75% in augmented data sets. Another study conducted by Rakesh G. et al.[12] indicated that the application of ML technology is essential in diagnosing and managing various diseases in farmed fish from small-scale and large-scale aquaculture. In relation to water quality parameters, some studies emphasized the potential role of ML in environmental monitoring. For instance, A. Zambrano et al. [6][15] applied a machine learning model called Random Forests to predict water quality with high precision based on minimal data input. Similarly, Oliver North Rogers et al.[7] reinforced the usefulness of

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

ML technology in the management of resources and regulatory compliance in aquatic environments. In another study, V. Anand et al. [8][16] demonstrated that SVM models from multispectral satellite images could be used to predict water quality parameters, confirming the viability of remote sensing in the aquaculture industry. Finally, sophisticated hybrid models were developed by V. Pe et al.[11] to enhance the performance of intelligent aquaculture systems using a deep convolutional neural network (Deep-CNN) combined with.

3. Methodology

In this paper, an attempt has been made to design a method for early disease diagnosis in fish based on machine learning using water quality parameters. This approach involves four important steps, namely data collection, preprocessing, modeling, and analysis. Here, we have used Random Forest as a classifier for diagnosing diseases based on water quality parameters.

A. Dataset: The data set used in the current analysis is named "Aquaculture – Water Quality Data Set," and is available online at Mendeley Data (Veeramsetty et al., 2024; DOI: 10.17632/y78ty2g293.1). The data set includes 15 water quality variables such as pH level, temperature, dissolved oxygen (DO), ammonia, nitrite, turbidity, biological oxygen demand (BOD), and other similar parameters measured in an aquaculture environment. In addition to the existing features, a new column was added at the end of the data set called "Predicted Disease."

B. Preprocessing: In the first stage of data preparation, we had to detect the presence of any missing or erroneous data. We concluded that the data was in its perfect form, and there was no need for data imputation. As the model is dependent on numeric values, we used the label encoding technique for the 'Predicted Disease' attribute.

The dataset was split using stratified sampling into training and testing sets with an 80:20 ratio.s

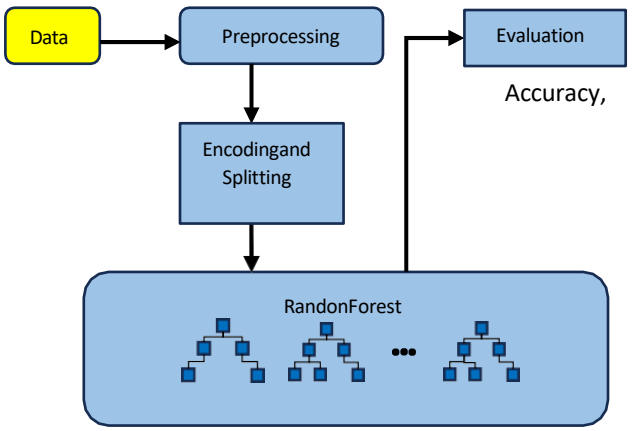

Fig 1: Sequential steps in Fish Disease Detection using Random Forest model

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

C. Model Implementation: A Random Forest classifier was implemented using Scikit-learnin Python. It was chosen forits robustness in handling high-dimensional data and non-linear relationships. The classifier was configured incorporating the following hyper parameters:

- Number of trees (n_estimators)= 100
- Maximum depth (max_depth)=10
- Criterion=Gini impurity

The Gini impurity is applied to consider the splits in every node and can be computed as:

$$G(t) = 1 - \sum_{i=1}^{n} p_i^2$$

Where pi is the proportion of samples belonging to class I in a given node t.

The model was trained on the training subset and evaluated on the test set. After training, feature importance scores were extracted to identify which water quality parameters most influenced predictions. These were visualized in Fig1, showing that Dissolved Oxygen, Temperature, pH, and Ammonia were among the most significant features.

The importance of each feature j was calculated as the average reduction in impurity over all the trees in the forest:

$$FI_j = \sum_{t \in T} \frac{N_t}{N} . \Delta i_t(j)$$

Where$FI_j$is the value of importance of feature j ,T is the collection of all decision trees, N_t is the number of in stances at node t, N is the over all number of instances, and $\Delta it(j)$ is the reduction in impurity at node t due to splitting on feature j.

D. Evaluation Metrics: To evaluate the performance of the and rom Forest model in classifying fish diseases based on water quality parameters, several standard evaluation metrics were utilized, covering metrics such as accuracy, confusion matrix, precision, recall, and F1-score.

These measures are calculated as follows:

- Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision:

$$Precision = \frac{TP}{TP + FP}$$

- Recall:

$$Recall = \frac{TP}{TP + FN}$$

- F1 Score:

$$F_1 = 2. \frac{Presion. Recall}{Precision + Recall}$$

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Furthermore feature importance was analyzed to determine which parameters most influenced the prediction outcomes.

4. Results and discussion

The Random Forest Model achieved an accuracy rate of 89.25% on the test data, indicating that the model was able to predict the disease status of most samples, thereby showing the ability to detect diseases early in aquaculture farms.

Precision, Recall, and F1-Score

In order to evaluate the performance of the proposed model across various disease classes, precision, recall, and F1-score were calculated and presented in Table 1.

These parameters provide an insight into the sensitivity and specificity of the model for each category.

Table1: Precision, Recall, and F1-Score per Class Feature Importance

Disease	Precision	Recall	F1-Score
Healthy	0.91	0.88	0.89
Acid Stress	0.85	0.86	0.85
Fungal Infection	0.84	0.82	0.83
Gill Damage	0.87	0.89	0.88
Bacterial Infection	0.81	0.79	0.80
Low Oxygen Stress	0.83	0.84	0.83
Nitrite Poisoning	0.79	0.76	0.77

Figure Y demonstrates the significance of various water quality parameters in predicting diseases. The following parameters have been found to be very important for the model, including Dissolved Oxygen (D.O.), Temperature, pH, and Ammonia.

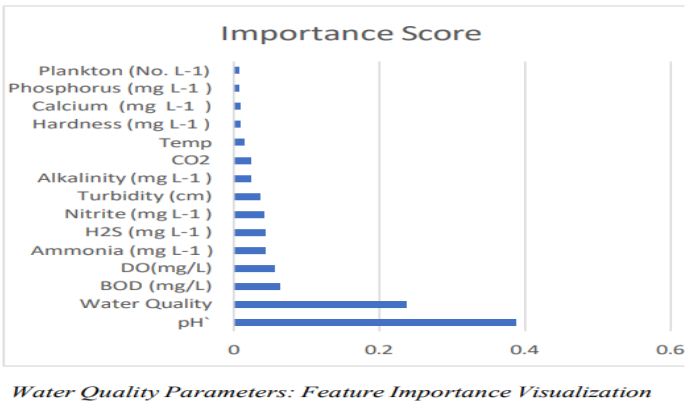

Fig 2. Water Quality Parameters: Feature Importance Visualization

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

5. Conclusion

This successful completion of the project enabled machine learning algorithms to predict fish diseases from the perspective of water quality, thus giving aquaculture systems a valuable means of prediction and prevention. The application of Random Forest algorithm together with the corresponding environmental data enabled the system to detect any disease and analyze its causes, and therefore intelligent aquaculture control is now within reach.

Reference

[1] Moni, J., Jacob, P., Sudeesh, S., Nair, M., George, M., & Thomas, M. (2024). A smart aquaculture monitoring system with automated fish disease identification. 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST), 1–6. https://doi.org/10.1109/ICTEST60614.2024.10576108

[2] Çakır, M., Yilmaz, M., Oral, M., Kazanci, H., & Oral, O. (2023). Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. Journal of King Saud University – Science. https://doi.org/10.1016/j.jksus.2023.102754

[3] Nayan, A., Mozumder, A., Saha, J., Mahmud, K., & Azad, A. (2021). Early detection of fish diseases by analyzing water quality using machine learning algorithm. arXiv.https://arxiv.org/abs/2102.09390

[4] Isaac, S., Daniel, O., Grivin, M., Kyagaba, J., Lule, E., & Marvin, G. (2024). Explainable machine vision techniques for fish disease detection with deep transfer learning. 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), 1218–1227. https://doi.org/10.1109/ICESC60852.2024.10689742

[5] Li, D., Li, X., Wang, Q., & Hao, Y. (2022). Advanced techniques for the intelligent diagnosis of fish diseases: A review. Animals, 12.https://doi.org/10.3390/ani12212938

[6] Zambrano, A., Giraldo, L., Quimbayo, J., Medina, B., & Castillo, E. (2021). Machine learning for manually measured water quality prediction in fish farming. PLOS ONE, 16. https://doi.org/10.1371/journal.pone.0256380

[7] Rogers, O., & S, A. (2024). Water quality prediction with machine learning algorithms. EPRA International Journal of Multidisciplinary Research (IJMR). https://doi.org/10.36713/epra16318

[8] Anand, V., Oinam, B., & Wieprecht, S. (2024). Machine learning approach for water quality predictions based on multispectral satellite imageries. Ecological Informatics, 84, 102868. https://doi.org/10.1016/j.ecoinf.2024.102868

[9] Sujatha, K., & Mounika, P. (2023). Evaluation of ML models for detection and prediction of fish diseases: A case study on epizootic ulcerative syndrome. 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 1–7. https://doi.org/10.1109/ICEEICT56924.2023.10156914

[10] M, A., Dinesh, M., Lakshmipriya, C., Sharmila, V., Muthuram, A., & R, S. (2023). Water quality prediction using machine learning: A comparative study. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 348–353. https://doi.org/10.1109/ICAISS58487.2023.10250743

[11] P, V., K, S., D, B., & Reshma, R. (2023). Predicting and analyzing water quality using machine learning for smart aquaculture. 2023 International Conference on Sustainable

Received: 2026-01-10
Accepted: 2026-03-15
Published Online: 2026-03-30
DOI: 10.70454/JRIST.020105

Journal of Recent Innovation in Science and Technology
E-ISSN: 3117-3926, P- 3117-535X

Computing and Data Communication Systems (ICSCDS), 354–359. https://doi.org/10.1109/ICSCDS56580.2023.10104677

[12] Gr, R., Raghavendra, C., Rohit, S., Shetty, P., & M. P., S. (2023). An overview of machine learning techniques for identification of diseases in aquaculture. 2023 4th International Conference for Emerging Technology (INCET), 1–5. https://doi.org/10.1109/INCET57972.2023.10170241

[13] R. K. Singh, I. Siraj, A. Kumar, D. Singh, R. Mandal, and P. Kumar, "Fir robot: A federated learning approach to information retrieval in robotic edge devices," in 2025 2nd International Conference On Multidisci plinary Research and Innovations in Engineering (MRIE), Gurugram, India, 2025, pp. 733–737.

[14] Xu, P., Liu, X., Liu, J., Cai, M., Zhou, Y., Hu, S., & Chen, M. (2024). Survey on machine vision-based intelligent water quality monitoring techniques in water treatment plants: Fish activity behavior recognition-based schemes and applications. Demonstratio Mathematica. https://doi.org/10.1515/dema-2024-0010

[15] R. K. Singh, V. Kochher, H. Mehta, S. Gupta, P. Kumar, and L. Verma, "Optimizing security in high-speed networking environments: An inte grated framework using aes, mpls, and ids for enhanced data protection and performance," in 2025 International Conference on Electronics, AI and Computing (EAIC), Jalandhar, India, 2025, pp. 1–6.

[16] P. K. Sharma et al., "Hybrid machine learning system for recognizing vehicle number plates in hazy environments is utilized for safety and security at tourist destinations," in ICT Analysis and Applications, ser. Lecture Notes in Networks and Systems, S. Fong, N. Dey, and A. Joshi, Eds. Cham: Springer, 2026, vol. 1653. [Online]. Available: https://doi.org/10.1007/978-3-032-06694-7 3