วารสารเกษตรศาสตร์และเทคโนโลยี มทร.อีสาน | Agriculture & Technology RMUTI Journal
ปีที่ 5 ฉบับที่ 3 กันยายน – ธันวาคม 2567 | Vol.5 No.3 September – December 2024

Received: May 9, 2024; Revised: September 5, 2024; Accepted: September 16, 2024

Enhancing water quality monitoring in shrimp ponds using machine learning and bio-inspired optimization

Surasit Songma[1*], Watcharakorn Netharn[1] and Rungkiat Kawpet[1]

[1]*Faculty of Science and Technology, Suan Dusit University, Bangkok.*
[*]*Corresponding Author E-mail Address: surasit_son@dusit.ac.th.*

Abstract

This work investigates how to improve water quality monitoring in shrimp ponds by combining machine learning and bio-inspired optimization techniques. Preprocessing the dataset is crucial before applying and evaluating classifiers like LR, DT, RF, SVM, KNN, NB and GBC against water quality indicators. Using criteria such as accuracy, precision, recall, F1-score and AUC, as well as computational aspects like as model size and CPU time, the study concludes that the RF model is clearly superior. It is further enhanced using approaches such as EBAO, EACO, ECBOA and EPSO, resulting in significant gains in prediction performance, particularly precision and recall. Among optimization methodologies, EACO stands out for striking a balance between performance enhancement and computing efficiency. The results highlight the importance of merging machine learning with bio-inspired algorithms in environmental monitoring, demonstrating a compelling methodology for improving water quality management in aquaculture. This complete approach not only enhances the precision of water quality assessments in shrimp farming but also establishes a precedent for future applications in environmental science and technology.

Keywords: Water quality, Shrimp ponds, Machine learning, Bio-inspired optimization techniques

Introduction

Shrimp farming is a big industry in Thailand and around the world (Laoong-U-thai et al., 2022). Thailand's dominant position in the worldwide aquaculture business, particularly in shrimp cultivation and production, can be due to its favorable geographical and meteorological conditions, as well as its sophisticated technological and regulatory framework. The country's vast coastal region, spanning 23 provinces and covering a total area of 3,151.10 kilometers, along with suitable climatic conditions, creates an ideal setting for aquaculture endeavors, particularly in the cultivation of shrimp. The favorable geographical location, combined with favorable climatic circumstances, has played a crucial role in the development of shrimp farming as a key element of Thailand's agricultural economy. This industry has made a considerable contribution to the country's gross domestic product by generating substantial export income. The international renown of Thai shrimp farming is attributed to its adoption of quality and sustainability techniques, which highlight the industry's dedication to environmental stewardship, social responsibility, and economic viability. The holistic approach to ensuring responsible farming methods is demonstrated via the application of standards such as Good/Best Aquaculture Practice (G/BAP), Code of Conduct (CoC), Aquaculture Stewardship Council (ASC), and organic aquaculture. The afore mentioned activities play a crucial role in harmonizing Thailand's shrimp farming practices with international sustainability standards, consequently bolstering the industry's worldwide competitiveness and market entry.

The goal of this work is to collect thorough data on water quality in shrimp ponds to forecast future water quality trends that. Affects the growth of shrimp. After studying relevant documents and studies, it was discovered that there is currently no standardized dataset for assessing water quality in shrimp ponds. Field survey data was received from shrimp farms. This gathered real-time data on water quality from laboratory tests performed on samples obtained from over 100 ponds. These ponds are owned by experienced shrimp farmers with 15 over years of shrimp farming expertise. The importance of water quality in shrimp farming can be stressed because it has a substantial impact on shrimp health and productivity. Farmers identified seven critical parameters to monitor: pH, alkalinity, ammonia, nitrite, calcium, magnesium, and salinity which is consistent with previous research (Ahmed et al., 2019). Each of these variables is critical to the overall health of shrimp. Traditionally, shrimp farming practices have been passed down through informal routes, frequently by word of mouth across generations. If the best advice system is not created, It will affect the growth of shrimp farming. Recognizing this gap, the researchers propose using machine learning approaches to help rookie farmers better understand and anticipate water quality in shrimp ponds. This strategy intends to dramatically minimize the learning curve for shrimp farming, allowing farmers to make more efficient and informed decisions.

Machine learning is a subset of artificial intelligence in which computers to learn from data and make predictions or judgments without explicit programming. The benefit of machine learning is learning from data and using it as a model to be used as a standard for shrimp farming data in ponds. This approach solves the problem of raising shrimp in ponds where the shrimp are not growing and promotes quality shrimp farming in ponds. Its practical uses include tailored suggestions and healthcare diagnostics (Javaid et al., 2022), financial fraud detection (Ali et al., 2022), water quality assessment (Gakii and Jepkoech, 2019; Ilić et al., 2022; Islam et al., 2022), self-driving cars, language translation and more are application of machine learning. Individuals must comprehend the principles and methodologies of machine learning before implementing it.

This work entails gathering data from shrimp farms and performing laboratory checks to examine the results. The collected datasets are thoroughly cleaned and analyzed using exploratory data analysis and encoding techniques. The data is then divided into training and testing sets and input into a variety of classification models, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Gradient Boosting Classifier (GBC), to evaluate their performance. Accuracy, precision, recall, F1-Score, and Area Under Curve (AUC) are used to assess performance, in addition to model size and processing speed. The best-performing model is then chosen for optimization utilizing techniques such as Enhanced Bat Algorithm Optimization (EBAO), Enhanced Ant Colony Optimization (EACO), Enhanced Cuckoo Search-Base Optimization Algorithm (ECBOA), and Enhanced Particle Swarm Optimization (EPSO), with the goal of increasing efficiency in terms of accuracy, model size, and computation time.

The work makes several important contributions:

1. Unique dataset: The dataset utilized in this work was collected from genuine shrimp ponds by farmers with over a decade of expertise. It also includes laboratory test data, which sets it apart from past study attempts.

2. Machine learning and bio-inspired optimization Integration: In contrast to traditional methodologies, this work integrates machine learning techniques with bio-inspired optimization methods in a unified process. This integration makes the research more complete and effective.

3. Expanded performance evaluation: In addition to typical performance metrics like accuracy, the study assesses model size and processing time. These extra indicators provide useful insights into the feasibility of models for actual application.

The next sections of the paper are structured as follows: Section 2 defines the research sequence, including the research concept and methodology; Section 3 elaborates on the proposed framework and methodology; Section 4 defines the experimental setup; Section 5 presents the experimental results and associated discussions; Section 6 concludes by evaluating the model's strengths and weaknesses, and suggesting potential directions for future research.

Related work in shrimp farming has emphasized the growing worldwide and regional interest, particularly in Thailand, which has resulted in an increase in shrimp farming activities and output. Maintaining acceptable water quality is a vital element of shrimp farming, yet there is no standardized dataset for assessing it. To close this gap, researchers collected real-time water quality data from over 100 ponds, concentrating on important factors like pH, alkalinity, ammonia, nitrite, calcium, magnesium, and salinity. Given the informal nature of information flow in this business, the study suggests employing machine learning to assist rookie farmers in predicting and managing water quality changes, boosting decision-making and highlighting the potential of modern technologies in agricultural applications.

Based on document analysis and related studies. Many studies have used machine learning techniques to assess the quality of water used in shrimp, fish, and other aquaculture operations. An overview of these findings is provided in Table 1.

Table 1 Summary of studies on water quality monitoring and prediction in aquaculture

Study	Methodology/Findings
Hernández et al. (2011)	Water quality in shrimp aquaculture is evaluated using a fuzzy inference system to classify concentration levels, determining their impact on organisms. This system generates a water quality index describing ecosystem condition.
Nuanmeesri et al. (2023)	IoT devices collect data on water quality from river sensors, including salinity, pH, and temperature. This data is accurately classified using a combination of multilayer perceptron, support vector machine, and random forest algorithms. This strategy provides personalized management techniques for a variety of farm scenarios.
Somantri et al. (2018)	A water quality monitoring and aerator automation system with a timer was tested using pH and temperature sensors. Validation testing revealed proper performance with negligible mistakes across multiple scenarios. The digital data results were obtained by processing the analog sensor values.
Kajornkasirat et al. (2021)	A web and mobile application system was created for shrimp farmers and managers, using MySQL, Apache Web Server, and web technologies. Google charts and maps application programming interface assisted water quality visualization.
Ilić et al. (2022)	Using Naïve Bayes, a machine learning method, to predict water quality categories using nine characteristics. When tested using data from five locations in Vojvodina Province, Serbia, the model properly predicted 64 out of 68 incidents.
Ilić et al. (2022)	Compares the Support Vector Machine (SVM) and Artificial Neural Network (ANN) approaches for monitoring water quality. SVM obtained 90.4% accuracy, whilst ANN reached 72.1%. SVM outperformed ANN in accuracy, indicating that it is more suited for predicting water quality.

Islam et al. (2022)	A water quality prediction model based on principle component regression and gradient boosting classifier. On the Gulshan Lake dataset, it obtains 95% prediction accuracy with principle component regression and 100% classification accuracy with gradient boosting classifier.

Materials and Methods

We gained significant knowledge by analyzing data and reviewing relevant academic research, which helped shape our design strategy. This corpus of knowledge has been incorporated within a research conceptual framework, which is illustrated in Figure 1.

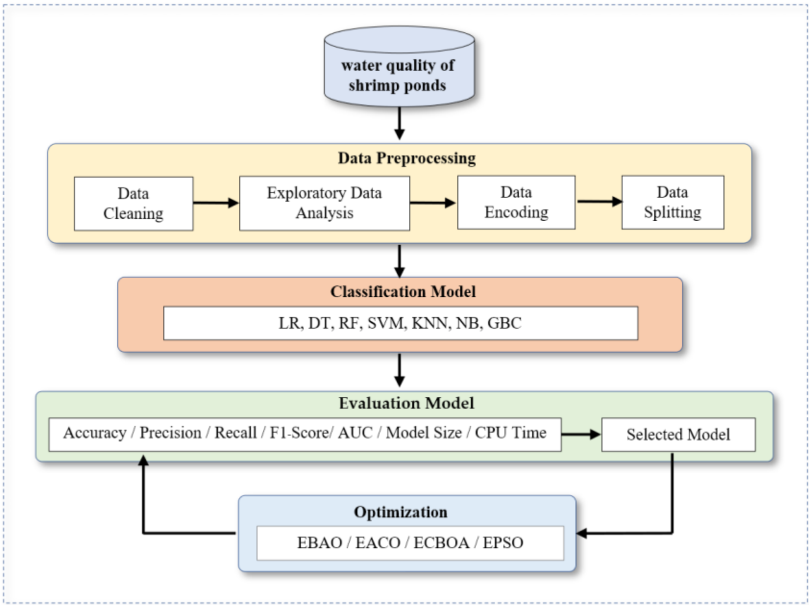

water quality of shrimp ponds
Data Preprocessing
Data Cleaning
Exploratory Data Analysis
Data Encoding
Data Splitting
Classification Model
LR, DT, RF, SVM, KNN, NB, GBC
Evaluation Model
Accuracy / Precision / Recall / F1-Score/ AUC / Model Size / CPU Time
Selected Model
Optimization
EBAO / EACO / ECBOA / EPSO

Figure 1 The proposed framework of enhancing water quality monitoring in shrimp ponds using machine learning and bio-inspired optimization

Figure 1 depicts our proposed technique, which begins by gathering data on a specific topic of concern the quality of water in shrimp farming ponds. This process includes a thorough examination of relevant facts and a review of scholarly literature.

The procedure begins with gathering and prepping a dataset on shrimp pond water quality, which includes cleaning, exploratory analysis, encoding categorical data, and separating the data for training and testing. Various classification models including LR, DT, RF, SVM, KNN, NB, and GBC are used on the preprocessed data. Each model's performance is measured using metrics such as accuracy, precision, recall, F1-Score, and Area Under the Curve (AUC), as well as model size and CPU time. The best model is then refined using bio-inspired algorithms (EBAO, EACO, ECBOA, and EPSO). The final model, based on its enhanced performance and capabilities after optimization, is projected to be extremely useful in monitoring and maintaining water quality in shrimp farming.

Experimental Setup

This work used a 64-bit Windows 11 operating system on a PC powered by an 11th Gen Intel(R) Core (TM) i7-11800H CPU @ 2.30 GHz and 32 GB of 2933 MHz DDR4 RAM. Python 3.11 was used to prepare, handle, preprocess, and analyze data, as well as train and evaluate models using certain metrics. NumPy and Pandas were utilized for efficient data handling and preprocessing, while Scikit Learn helped with model training, evaluation, and analysis of numerous performance indicators. Additionally, Matplotlib was used for data visualization.

The following sections provide further insights and details regarding the study methodology.

1. Water quality of shrimp ponds data set

In the data set presented for the water quality of shrimp ponds, our study group created a dataset that includes water quality data from over 100 shrimp ponds. These ponds, located in Suphanburi Province's Song Phi Nong District and Bang Pla Ma District, covering the period from May 2021 to June 2022. These ponds are managed by extensive shrimp farming experience, who have over 15 years of experience. The collection and subsequent lab analysis of water samples from these locations resulted in the identification of seven key water quality indicators: pH, alkalinity, ammonia, nitrite, calcium, magnesium, and salinity. These indicators are crucial in establishing the health and suitability of the water for shrimp farming, and they influence overall water quality in a variety of ways. This dataset of 1,098 items provides a solid foundation for analyzing water quality in shrimp ponds, distinguishing between good and wastewater quality (serious).

2. Data preprocessing

Data preprocessing is a critical step in data science and machine learning that prepares raw data for analysis and modeling. It modifies data to increase its quality and utility, removing missing numbers, errors, and irrelevant information. This foundational step ensures that data is in the proper format for algorithms, which has a substantial impact on model performance and accuracy, resulting in more dependable outcomes (Fan et al., 2021).

2.1 Data cleaning is the process of removing or correcting flaws, inconsistencies, and extraneous information from a dataset to improve its analytical and modeling quality. This stage may include resolving missing values, removing duplicate information, repairing errors, and standardizing data formats (Ridzuan and Wan, 2019).

2.2 Exploratory Data Analysis (EDA) is the process of evaluating a dataset to find patterns, uncover anomalies, and comprehend correlations between variables, typically through visual and statistical methods. It seeks to acquire understanding into the data's structure (Poian et al., 2023).

2.3 Data encoding converts categorical data into a numerical representation that is suitable for machine learning techniques. Depending on whether the categorical variables have a natural order, techniques such as one-hot encoding, label encoding, and ordinal encoding are applied. The choice of encoding method is critical since it has a major impact on model performance.

2.4 Data splitting divides a dataset into distinct sections for training and testing models. This step is necessary for accurately evaluating the model's performance and preventing overfitting. Typically, data is split into a training set (to learn data patterns) and a testing set (to evaluate performance on unseen data). Common split ratios are 70% for training and 30% for testing. The goal is to ensure that the model can generalize well to new data.

3. Classification model

A Classification model is a predictive model in machine learning that divides data points into specified classes depending on their characteristics. It is used in contexts with specific results, such as spam identification or disease diagnosis. Classification can be binary (two classes) or multiclass (more than two classes). The model uses a dataset with known labels to predict the class of fresh data.

3.1 Logistic Regression (LR) is a classification approach for binary tasks. It calculates the likelihood that a data point belongs to a specific class by applying a logistic (sigmoid) function to a linear combination of input features. LR learns optimal coefficients through optimization and predicts based on feature values. LR is easy, interpretable, and effective for linear or logistic associations, but it may struggle with nonlinear data (Ahmed et al., 2019; Krhoda and Owira, 2019).

3.2 Decision Tree (DT) is a versatile supervised learning technique used for classification and regression tasks. It splits the feature space recursively based on feature values, attempting to produce subsets that are as pure as feasible in terms of class labels or have the least variance. While DT are understandable and easy to envision, they are prone to overfitting. Despite this, they are still commonly used because of their simplicity, capacity to handle both numerical and categorical data, and effectiveness in capturing non-linear correlations (Ahmed et al., 2019; Gakii and Jepkoech, 2019).

3.3 Random Forest (RF) is an ensemble learning approach for classification and regression applications. It generates numerous decision trees, each trained on a different set of data and features, in order to prevent overfitting and enhance accuracy. During prediction, RF aggregates individual trees' predictions to get a final result (Lukman et al., 2023; Nuanmeesri et al., 2023). RF is a common choice in machine learning because it is adaptable, resilient, and excellent at handling a wide range of input formats.

3.4 Support Vector Machine (SVM) is a supervised learning technique used for classification and regression tasks. It determines the hyperplane that optimally separates data points from distinct classes while maximizing the margin between them. SVM can handle both linear and nonlinear data by utilizing various kernel functions. It is effective in high-dimensional spaces, resistant to overfitting, but computationally expensive. Overall, SVM is a versatile and extensively used machine learning algorithm (Guenther and Schonlau, 2016).

3.5 K-Nearest Neighbors (KNN) is a straightforward yet versatile supervised learning technique for classification and regression tasks. It predicts the class or value of a new data point based on the majority class or the average of its nearest neighbors. KNN is simple to develop and interpret, making it ideal for small datasets where interpretability is critical. However, it can be computationally expensive to predict, particularly for large datasets, and it may perform badly with high-dimensional or noisy data (Ilić et al., 2022; Uddin et al., 2022).

3.6 Naïve Bayes (NB) is a supervised machine learning method designed for classification applications. It works under the assumption of feature independence, which simplifies calculations. Despite this reduction, NB can still be effective, particularly with high-dimensional data and huge datasets. It is efficient, resistant to irrelevant features and missing data, but may suffer with feature independence breaches. NB is widely used in text classification jobs because of its simplicity and efficacy (Ilić et al., 2022)

3.7 Gradient Boosting Classifier (GBC) is an ensemble learning technique that is mostly used for classification purposes. It sequentially trains weak learners, usually decision trees, to repair errors produced by prior models. GBC uses gradient descent to minimize a loss function, allowing it to capture complicated correlations in data. It works well with diverse data types, feature interactions, and imbalanced datasets. However, GBC can be computationally expensive, necessitating hyperparameter adjustment for peak performance. Overall, it is frequently utilized due to its excellent predicted accuracy and adaptability (Islam et al., 2022).

4. Evaluation model

Evaluation accuracy is an important parameter for measuring machine learning model performance, particularly in classification tasks, because it calculates the proportion of correctly predicted instances out of all instances in the dataset, with high accuracy indicating that the model's predictions closely match the actual outcomes. A confusion matrix is a tabular tool used to measure the efficacy of a supervised learning algorithm in classification tasks. It contains real class instances in each row and predicted class instances in each column. It includes True Positives (TP) for accurate positive predictions, True Negatives (TN) for accurate negative predictions, False Positives (FP) for inaccurate positive predictions, and False Negatives (FN) for incorrect negative predictions, providing a detailed breakdown to determine the accuracy and error of the model. This work evaluates the method's performance using metrics like accuracy, precision, recall, F1-Score, and AUC classification (Nuanmeesri et al., 2022; Songma et al., 2023). It also considers model size and CPU time as supplemental evaluation criteria.

4.1 Accuracy measures the model's overall correctness and is calculated as the proportion of correct predictions to the total number of predictions made. It is suitable for classes with similar distributions. The formula for accuracy is

$$(TP+TN) / (TP+FP+FN+TN) \quad (1)$$

4.2 Precision refers to a model's ability to correctly identify only relevant instances. In other words, it measures the proportion of true positive predictions to the total number of positive predictions. The formula for precision is

$$TP / (TP+FP) \quad (2)$$

4.3 Recall, also known as sensitivity or the true positive rate, measures the percentage of true positives correctly identified by the model. The formula for recall is

$$TP / (TP+FN) \quad (3)$$

4.4 F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when precision and recall need to be balanced and when class distribution is uneven. The formula for the F1-Score is

$$2 * (Precision*Recall) / (Precision+Recall) \quad (4)$$

4.5 Area Under the Curve (AUC) statistic is used to assess the effectiveness of binary classification models. It indicates the model's ability to discriminate between positive and negative classes at all threshold values. The AUC scales from 0 to 1, with higher values indicating greater classification performance. It is a threshold-independent metric, which makes it useful for evaluating overall model performance, particularly in circumstances with imbalanced datasets or where threshold selection is not crucial. (Tohka and van Gils, 2021).

4.6 Model size is the amount of memory or storage needed to store a trained machine learning model. It includes all parameters, coefficients, and data structures connected with the model. Model complexity, feature count, model type, and format all have an impact on model size. Understanding model size is critical for deployment and scalability, as it influences resource use and inference speed. Optimizing model size while preserving performance is critical in machine learning model building (Songma et al., 2023).

4.7 CPU Time is the time it takes a CPU to execute a task, which is critical for determining computational efficiency. Lower CPU Time signifies faster processing, vital for optimizing performance in real-time systems and large-scale data processing (Songma et al., 2023).

5. Optimization

5.1 Bat Algorithm Optimization (BAO) is a bio-inspired computer algorithm developed by Xin-She Yang in 2010 based on bat echolocation behavior. This algorithm simulates how bats use echolocation at changing frequencies, loudness, and pulse rates to navigate and locate prey in the dark. In BAO, virtual bats fly through the solution space, using their echolocation abilities to evaluate the quality of solutions and thereby guiding the search to the best answer. The program modifies the flight behavior of the simulated bats based on the proximity of their present position to the target, efficiently balancing exploration and exploitation in the solution space. This makes BAO particularly useful for tackling optimization problems in a variety of domains, including engineering design and data science, by identifying the best parameters or configurations that maximize or reduce a given objective function (Agarwal and Kumar, 2022).

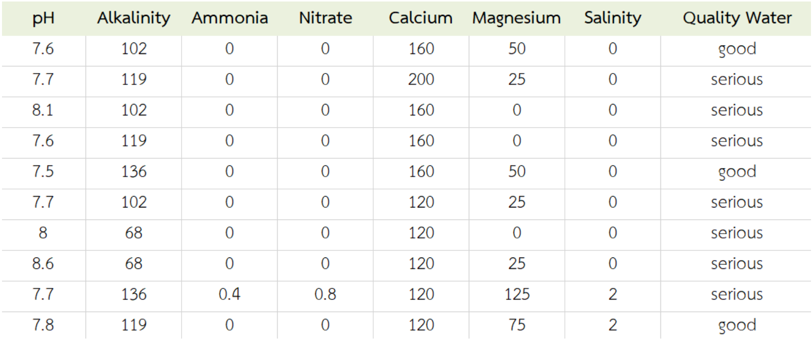
5.2 Ant Colony Optimization (ACO) is a probabilistic method for addressing computational problems that can be simplified to finding good paths via graphs. ACO utilizes an approach inspired by ant foraging behavior, in which ants deposit pheromones on the ground to identify some favored paths that other members of the colony should follow. ACO algorithms, first introduced by Marco Dorigo in the early 1990s (Dorigo and Blum, 2005; Azargashb et al., 2021), model ant behavior, including pheromone deposition and evaporation, to discover optimal solutions. Each ant symbolizes a potential solution to the problem, and the intensity of the pheromone trail on paths affects the likelihood that succeeding ants will follow those paths. Over time, the algorithm converges on the best option as the pheromone trails leading to ideal solutions strengthen, guiding more ants along these paths. ACO has been successfully used to a variety of issues, including routing, scheduling, and assignment, demonstrating its versatility and effectiveness in identifying optimal pathways in complex search spaces (Wang and Han, 2021).

5.3 Cuckoo Search-Base Optimization Algorithm (CBOA) is a nature-inspired metaheuristic method introduced by Xin-She Yang and Suash Deb in 2009 (Yang and Deb, 2009). Drawing from the obligate brood parasitism of certain cuckoo species, where eggs are laid in other birds' nests, the algorithm simulates this process by representing solutions as cuckoo eggs in randomly chosen nests. Through iterations involving randomization and selection, solutions are refined based on an objective function, with the best nests passing down to subsequent generations. By incorporating Lévy flights, which mimic cuckoos' flight behavior during nest hunting, CBOA efficiently balances exploration and exploitation, making it adept at navigating complex multidimensional spaces. This unique blend of natural inspiration and optimization techniques enables CBOA to effectively tackle a diverse range of optimization problems across engineering, design, and data science domains.

5.4 Particle Swarm Optimization (PSO) is a computational method for optimizing a problem by iteratively improving a candidate solution based on a given quality measure. Kennedy and Eberhart created it in 1995 as a way to imitate the social behavior of birds in a flock or fish schools. In PSO, each particle in the swarm is a possible solution to the optimization issue. These particles explore the solution space by modifying their position in response to their own experience and the achievements of surrounding particles, similar to how natural swarms move to ideal feeding sites or traverse situations. The approach is distinguished by the sharing of information among particles, which drive toward their individual best-known places in the solution space as well as the overall best position discovered by the swarm. This methodology enables PSO to efficiently search for global optima over a large range of functions and problems, making it especially valuable in domains such as engineering optimization, where older approaches may struggle due to the complexity of the solution space (Gad, 2022; Shami et al., 2022).

Results and Discussions

In the first phase, the water quality dataset for shrimp ponds is imported and prepared for future study. This process begins with data cleaning, which entails finding and correcting duplicate entries, mistakes, and missing values. The dataset for shrimp pond water quality has 1,098 items and 9 attributes, including ID, pH, alkalinity, ammonia, nitrate, calcium, magnesium, salinity, and water quality. Show in Table 2, and exam data water quality dataset show in Figure 2.



pH	Alkalinity	Ammonia	Nitrate	Calcium	Magnesium	Salinity	Quality Water
7.6	102	0	0	160	50	0	good
7.7	119	0	0	200	25	0	serious
8.1	102	0	0	160	0	0	serious
7.6	119	0	0	160	0	0	serious
7.5	136	0	0	160	50	0	good
7.7	102	0	0	120	25	0	serious
8	68	0	0	120	0	0	serious
8.6	68	0	0	120	25	0	serious
7.7	136	0.4	0.8	120	125	2	serious
7.8	119	0	0	120	75	2	good

Figure 2 Exam water quality dataset

Table 2 Water quality parameters description

Attributes	Description
ID	Assigns a unique identifier to each water sample.
pH	Measures the water's acidity or alkalinity.
Alkalinity	Assesses a water body's ability to neutralize acid, indicating its PH stability.
Ammonia	Indicates potential contamination from various sources in water.
Nitrate	Chemicals that may become dangerous pollutants in water.
Calcium	Indicates water hardness and affects taste, pipe scaling, and soap effectiveness.
Magnesium	Adds to water hardness.
Salinity	Influences taste, health, and agriculture through compounds like sodium chloride
Quality Water	Categorizes water into different levels (e.g., "good," "serious") based on measured characteristics

The Exploratory data analysis of 1,098 water samples revealed that the water is slightly basic, with an average pH of 7.82 and moderate alkalinity levels. Major pollutants like ammonia and nitrate are found in low amounts, indicating a low pollution concern. Calcium and magnesium levels suggest that the water is usually soft to moderately hard, whereas salinity, particularly sodium chloride, is low and has no effect on flavor or health. Overall, the water samples are of adequate quality for a variety of purposes, as shown in Table 3.

Table 3 A summary of water quality parameters

Parameter	Mean	Min	Max
pH	7.82	7.3	9.0
Alkalinity	113.50	36	175
Ammonia	0.02	0.0	1.8
Nitrate	0.01	0.0	0.8
Calcium	132.86	40	200
Magnesium	39.03	0.0	125
Salinity	0.18	0.0	2.0

The dataset is complete, with no missing values in any column, making it suitable for detailed study. It has a "quality water" column that categorizes water quality as "good" or "serious." With an average score of 0.67, most water samples are likely categorized as "Good" under the grading criteria. This dataset contains a complete overview of numerous water quality parameters, demonstrating that the samples are generally safe for human consumption and environmental health. Indicators analyzed, such as low levels of pollutants like ammonia and nitrates, point to effective management for both human usage and ecological balance. The qualitative water quality labels are encoded as numerical codes ('0' for 'Good' and '1' for 'Serious'), simplifying the data for further analysis and modeling.

Data splitting is the process of dividing a dataset into training and testing sets to evaluate machine learning models. A standard strategy for a 1,098-sample dataset is to allocate 70% (768 samples) for training and 30% (330 samples) for testing. Stratified splitting should be employed to maintain the category ratio ('Good' and 'Serious') across these sets, ensuring that the models appropriately represent real-world data distributions.

Table 4 Quality water distributes training and testing datasets.

Quality Water	Training		Testing	
	Amount of Data	Ratio (%)	Amount of Data	Ratio (%)
Good	257	33.46	104	31.52
Serious	511	66.54	226	68.48
Total	768	100.00	330	100.00

After complete data processing, the dataset is evaluated using machine learning algorithms created during the preprocessing steps. There are distinctions when examining the utility of classifiers such as LR, DT, RF, SVM, KNN, NB, and GBC utilizing measures such as AUC, F1-Score, recall, precision, and accuracy. RF and GBC came out as the top performers, having the highest marks across all metrics. Specifically, RF achieved near-perfect accuracy, precision, recall, and F1-Score, as well as an impressive AUC score of 0.993, exhibiting remarkable performance and identifying it as the top model for the task. Table 5-6 show summarizes the results.

Table 5 Performance metrics comparison of machine learning classifiers

Classifiers	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.948	0.969	0.956	0.962	0.983
DT	0.985	0.996	0.982	0.989	0.986
RF	0.997	0.996	1.000	0.998	0.993
SVM	0.948	0.982	0.942	0.962	0.987
KNN	0.927	0.981	0.912	0.945	0.971
NB	0.627	0.990	0.460	0.628	0.980
GBC	0.997	0.996	0.995	0.997	0.990

Table 6 A comparison of CPU time and model size for machine learning classifiers

Classifiers	CPU Time (s)	Model Size (MB)
LR	0.017	0.50
DT	0.008	1.00
RF	0.213	1.20
SVM	0.097	2.50
KNN	0.027	4.50
NB	0.005	0.20
GBC	0.125	2.00

Following the study of Tables 5–6, the researcher identified RF as the best-performing model and chose to improve its efficiency using EBAO, EACO, ECBOA, and EPSO. The goal of this optimization work was to fine-tune the n_estimators parameter while leaving other essential parameters constant, such as max_depth=None, min_samples_split=2, and min_samples_leaf=1. The results of these optimization processes are extensively reported in Table 7-8, which demonstrates the efforts to determine the most effective n_estimators number for the Optimized RF.

Table 7 Performance metrics for optimized RF using Bio-inspired algorithms

Classifiers	Accuracy	Precision	Recall	F1-Score	AUC
Before	0.997	0.996	1.000	0.998	0.993
EBAO	0.998	0.997	1.000	0.998	0.985
EACO	0.999	0.998	1.000	0.999	0.990
ECBOA	0.998	0.997	1.000	0.998	0.985
EPSO	0.999	0.998	1.000	0.999	0.990

Table 8 A comparison of CPU time and model size for machine learning classifiers

Classifiers	CPU Time (s)	Model Size (MB)
Before	0.213	1.20
EBAO	0.220	1.30
EACO	0.219	1.40
ECBOA	0.218	1.35
EPSO	0.230	1.45

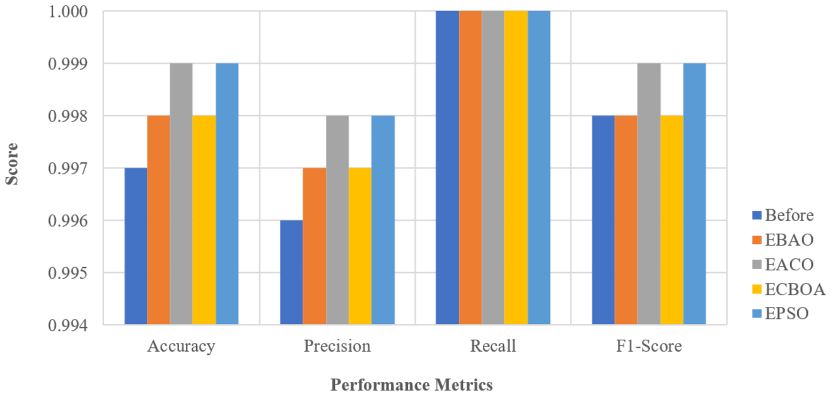


Figure 3 Performance of optimized RF with bio-inspired algorithm enhancements

Table 7-8 and Figure 3 show how bio-inspired algorithms such as EBAO, EACO, EBOA, and EPSO improve the performance metrics of an optimized RF model. Initially, performance benchmarks like as accuracy, precision, recall, F1-Score, AUC, CPU Time, and Model size are established as a baseline ("Before") for comparison. Consequently, each optimization strategy improves across all measures. EACO and EPSO stand out as top performers, with the highest results in accuracy, precision, F1-Score, and AUC while preserving flawless recall. Among these methods, EACO marginally outperforms EPSO because to its low increase in CPU time and model size, demonstrating an optimal balance of improved performance and resource efficiency. The parameter configuration with n_estimators=160, max_depth=None, min_samples_split=2, and min_samples_leaf=1 is the most successful, demonstrating EACO's ability to determine optimal settings for the RF model, resulting in peak prediction performance. Figure 3 shows how the researcher visually portrays these data using a bar chart.

In the discussion of experimental findings, it is clear that bio-inspired algorithms, specifically EBAO, EACO, ECBOA, and EPSO, improve the performance of an optimized RF model highlighting the RF model's predictive powers. EACO and EPSO stand out as front-runners, with significant improvements in key performance indicators while preserving perfect recall. The minor preference for EACO, which is due to its low impact on CPU time and model size, emphasizes the significance of establishing a compromise between better model performance and resource efficiency. demonstrating EACO's success in directing the RF model to peak performance. This thorough optimization strategy not only demonstrates the promise of bio-inspired algorithms for complementing machine learning models, but also provides the framework for future research efforts to improve model performance.

Conclusion

This work unambiguously shows that bio-inspired enhanced algorithms such as EBAO, EACO, ECBOA, and EPSO can significantly improve the performance of an enhanced RF model when refined. Through painstaking parameter modifications, these methods have delivered considerable increases in key measures like as accuracy, precision, recall, F1-score, and AUC, demonstrating the promise of nature-inspired strategies in improving the efficacy of machine learning models. Furthermore, the study emphasizes the need of achieving a balance between increased performance and resource efficiency with EACO standing out for its low impact on CPU time and model size. Overall, the findings support the use of bio-inspired techniques to drive advances in machine learning model performance.

This work demonstrates the efficiency of bio-inspired algorithms in improving an enhanced RF model, although it focuses on specific parameters and a single model type, limiting generalizability. It fails to thoroughly investigate computational costs, scalability, comparisons with alternative optimization methodologies, and the impact on model interpretability. These observed limits provide recommendations for further research.

In the future, the research will result in the creation of an advanced model capable of producing web and mobile applications suited to shrimp producers. These applications will increase usability and simplify extended testing, particularly for shrimp industry entrepreneurs. This program will include new research aimed at gathering more information about the underlying causes of common problems, as well as developing effective tactics or solutions to address them. The ultimate goal is to assure continuous optimization and maximum profit from the application's utilization.

References

Agarwal T. and Kumar V. (2022). A systematic review on bat algorithm: Theoretical foundation, variants, and applications. Archives of Computational Methods in Engineering. 29(5): 2707-2736.

Ahmed U., Mumtaz R., Anwar H., Shah A.A., Irfan R. and Garcí-Nieto J. (2019). Efficient water quality prediction using supervised machine learning. Water. 11: 2210. https://doi:10.3390/w11112210.

Ali A., Razak A.S., Othman S.H., Eisa, T.A.E., Al-Dhaqm A., Nasser M., Elhassan T., Elshafie H. and Saif A. (2022). Financial fraud detection based on machine learning: A Systematic Literature Review. Applied Sciences. 12: 9637. https://doi.org/10.3390/app12199637.

Azargashb L.S., Hashemy S.S.M. and Roozbahani A. (2021). Minimization of operational and seepage losses in agricultural water distribution systems using the ant colony optimization. Water Resources Management. 35(3): 827-846.

Dorigo M. and Blum C. (2005). Ant colony optimization theory: A survey. Theoretical Computer Science. 344(2–3): 243-278.

Fan C., Chen M., Wang X., Wang J. and Huang B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in Energy Research. 9: 652801. https://doi.org/10.3389/fenrg.2021.652801.

Gad A.G. (2022). Particle swarm optimization algorithm and its applications: A systematic review. Archives of Computational Methods in Engineering. 29(5): 2531-2561.

Gakii C. and Jepkoech J. (2019). A Classification model for water quality analysis using decision tree. European Journal of Computer Science and Information Technology. 7(3): 1-8.

Guenther N. and Schonlau M. (2016). Support vector machines. The Stata Journal. 16(4): 917-937.

Hernández J.J.C., Fernández L.P.S. and Pogrebnyak O. (2011). Assessment and prediction of water quality in shrimp culture using signal processing techniques. Aquaculture International. 19(6): 1083-1104.

Ilić M., Srdjević Z. and Srdjević B. (2022). Water quality prediction based on Naïve Bayes algorithm. Water Science and Technology. 85(4): 1027-1039.

Islam K.M.S., Islam N., Uddin J., Islam S. and Nasir M.K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences. 34(8): 4773-4781.

Javaid M., Haleem A., Pratap S.R., Suman R. and Rab S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks. 3: 58-73.

Kajornkasirat S., Ruangsri J., Sumat C. and Intaramontri P. (2021). Online analytics for shrimp farm management to control water quality parameters and growth performance. Sustainability. 13: 5839 https://doi.org/10.3390/su13115839.

Krhoda G.O. and Owira A.M. (2019). Groundwater quality prediction using logistic regression model for Garissa County. Africa Journal of Physical Sciences. 3: 13-27.

Laoong-U-thai Y., Wanna W. and Kaikaew A. (2022). DNA binding activity of marine shrimp LvProfilin. Applied Science and Engineering Progress. 15(3): 5522. https://doi.org/10.14416/j.asep.2021.10.002

Lukman H., Na S. and Liris M. (2023). Combining generative model and random forest to predict shrimp disease occurrence. Proceedings of the International Conference on Fisheries and Aquaculture. 10(1): 35-44.

Nuanmeesri S., Poomhiran L., Kadmateekarun P. and Chopvitayakun S. (2023). Improving the water quality classification model for various farms using features based on artificial neural network. TEM Journal. 12(4): 2144-2156.

Poian D.V., Theiling B., Clough L., McKinney B., Major J., Chen J. and Hörst S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. Frontiers in Astronomy and Space Sciences. 10: 1134141. https://doi.org/10.3389/fspas.2023.1134141.

Ridzuan F. and Wan Z.W.M.N. (2019). A review on data cleansing methods for big data. Procedia Computer Science. 161: 731-738.

Shami T.M., El-Saleh A.A., Alswaitti M., Al-Tashi Q., Summakieh M.A. and Mirjalili S. (2022). Particle swarm optimization: A comprehensive survey. IEEE Access. 10: 10031-10061.

Somantri M., Herawati V.E., Sofwan A., Abdurrasyiid H. and Arfan M. (2018). Design of water quality control for shrimp pond using senso-cloud integration. In Proceeding of 2018 5th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE). Semarang, Indonesia. 331-335

Songma S., Sathuphan T. and Pamutha T. (2023). Optimizing intrusion detection systems in three phases on the CSE-CIC-IDS-2018 Dataset. Computers. 12(12): 245 https://doi.org/10.3390/computers12120245.

Tohka J. and van Gils M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. Computers in Biology and Medicine. 132: 104324. https://doi.org/10.1016/j.compbiomed.2021.104324

Uddin S., Haque I., Lu H., Moni M.A. and Gide E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports. 12: 6256. https://doi.org/10.1038/s41598-022-10358-x.

Wang Y. and Han Z. (2021). Ant colony optimization for traveling salesman problem based on parameters optimization. Applied Soft Computing. 107: 107439. https://doi.org/10.1016/j.asoc.2021.107439.

Yang X.S. and Deb S. (2009). Cuckoo search via Lévy flights. In 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC). Coimbatore. India. 210-214.