Multivariate Analysis of Water Quality Parameters for Sustainable Prawn Farming

Vinay M T[1,2], K. T. Veeramanju[3], Sreenivasa BR[4*]

[1]*Research Scholar, Srinivas University, Mukka, Mangalore-575023*
[2]*Department of Computer Science and Engineering (Data Science), Bapuji Institute of Engineering and Technology, Davangere-577004, Karnataka, India.*
Email: vinaymt64@gmail.com
[3]*Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore-575023, Karnataka, India.*
Email: veeramanju.icis@srinivasuniversity.edu.in
[4*]*Professor, Department of Computer Science & Design, Bapuji Institute of Engineering and Technology, Davangere-577004, India.*
Email: sreenivasabr@bietdvg.edu

***Corresponding Author**: Vinay M T*
**Email: vinaymt64@gmail.com , contact number: +919353369108*

KEYWORDS
Aquaculture, Principal Component Analysis, Statistical Process Analysis, White Spot Syndrome Virus, Shrimp.

ABSTRACT
Objective: A robust framework for enhancing prawn aquaculture must be established by integrating multivariate analysis and machine learning models to evaluate and predict essential water quality parameters effectively.
Methods: The empirical water quality data taken from prawn aquaculture ponds was dimensionally reduced using Principal Component Analysis (PCA). The Successive Projections Algorithm (SPA) was used to identify key parameters, while machine learning models (XGBoost, Random Forest, and SVM) were employed to build prediction models. The models were tested for accuracy, precision, and computational efficiency.
Results: Using principal component analysis (PCA) and principal component spectral analysis (SPA), the research study was able to proficiently analyse water quality data, identify significant elements that affect prawn health, and construct and test prediction models (XGBoost, Random Forest, and SVM) that produced accurate forecasts of vital water quality indicators.
Conclusion: The integrated framework, which combines multivariate analysis and machine learning, is critical for optimising shrimp aquaculture practices. Its proactive management, continuous real-time monitoring, and precise water quality predictions enhance shrimp survival rates, mitigate disease outbreaks, and increase the sustainability of shrimp aquaculture operations.

1. Introduction

Monitoring the condition of the water is essential for shrimp farming. Excessive water quality can cause shrimp anoxia, hyperoxia, and hypoxia, so measuring dissolved oxygen is essential, according to studies [1]. The "water quality monitoring" involves gathering and evaluating water samples. We need to evaluate the water quality to determine if we are cleaning up our rivers. It shows the state of the lake, river, and stream today and historical patterns. Dissolved oxygen, pH, temperature, salinity, and nutrients are the five factors determining water quality. In aquaculture, accurate feeding and astute water quality monitoring are essential. Continuous water quality monitoring may impact the bacterial balance in aquaculture, which could lower shrimp resistance to illness [2].

In conventional aquaculture operations, skilled aquafarmers recognised and predicted farm health problems. One component is monitoring changes in water quality parameters that could affect aquaculture, such as salinity, temperature, pH, and dissolved oxygen. Lakes and fisheries must have adequate dissolved oxygen levels for shrimp farming. Besides food and fertiliser, shrimp farmers also need to supply dissolved oxygen. Sprinkle clean water from the top of the pond and take other preventative measures to keep it clean. Until the condition improves, feeding and fertilisation should be avoided. Fishing industry experts suggest oxygen-boosting drugs. The temperature and pH of the water pond must be monitored to balance hazardous and safe nitrogen molecules, such as ammonia and ammonium [3].

Aquaculture in ponds and marine areas across this low-lying nation is affected by floods, cyclones, and other natural calamities. Aquatic life may be stressed by water quality parameters that are either above or below the

optimal range, which could impact feeding, reproduction, and disease susceptibility [4]. Fishing and aquaculture are two of the most common coastal activities worldwide. Because fishing populations, which depend on fish for food and income, are susceptible to climatic variables, these activities are also becoming increasingly important in climate change. According to a recent study, bleaching and changes in organism variety and composition are just two of the detrimental effects of climate change on corals, fish populations, and aquaculture productivity [5-7].

Physicochemical parameters

Prawns' health depends on their water quality, which can affect their growth, risk of illness, and mortality rates [8]. Total dissolved solids (TDS), calcium, magnesium, nitrate, and chloride were quantified in the laboratory on the collection day following standard protocols. In contrast, pH, temperature, turbidity, dissolved oxygen, free carbon dioxide, and alkalinity were assessed on-site [9][10].

- ***Stress-Induced Susceptibility*:** Poor water quality stresses shrimps, weakening their immune systems and making them more prone to WSSV.
- ***Preventive Measures*:** Maintaining optimal water quality reduces stress, strengthens immunity, and lowers the risk of WSSV outbreaks.

By closely monitoring and managing these water quality parameters, shrimp farmers can mitigate the risk of WSSV proliferation and promote healthier shrimp populations. Table 1 demonstrates the water quality for Shrimp Farm Effluents and Coastal Waters**.**

Table 1: Summary of Water Quality Monitoring Guidelines for Shrimp Farm Effluents and Coastal Waters [12].

Variable	Reason for Monitoring	Guidelines for Protecting Aquatic Ecosystems
Water Temperature	Influences chemical and biological processes	Should not change by more than 2°C to prevent thermal stress on aquatic life
Dissolved Oxygen	Essential for the survival of aerobic aquatic organisms	Should be maintained at not less than 5 to 6 mg/L
pH	Affects chemical reactions and biological activities	It should be kept within the range of 6.0 to 9.0
Total Ammonia Nitrogen	Nutrient for plants but toxic at high levels; indicates pollution	Should not exceed 3 mg/L in effluents to prevent toxicity
Nitrate Nitrogen	It can be toxic to aquatic life	Should remain below 0.005 mg/L in coastal waters
Total Phosphorus	Supports plant growth; excess can cause algal blooms	Concentrations between 0.001 to 0.1 mg/L in coastal waters can lead to plankton blooms
Total Nitrogen	Nutrients contributing to eutrophication when excessive	Levels between 0.1 to 0.75 mg/L in coastal waters can cause plankton blooms; should not exceed 10 mg/L in effluents
Chlorophyll a	Indicator of phytoplankton abundance and eutrophication	Concentrations above 1 to 10 µg/L indicate eutrophication in coastal waters
Total Suspended Solids	Reflects suspended particles affecting water clarity	Should not change by more than 10% of the seasonal mean in coastal waters
Biochemical Oxygen Demand	Indicates the level of organic pollution	Should not depress dissolved oxygen levels below 5 or 6 mg/L
Salinity	High levels can lead to salinization, affecting freshwater ecosystems	Should not increase above 0.5 ppt in freshwater; no specific limit recommended for marine or brackish waters
Secchi Disk Visibility	Measures water clarity or turbidity	Should not change by more than 10% of the seasonal mean in coastal waters

These guidelines protect aquatic ecosystems and apply to the receiving water body outside the mixing zone (where effluents initially mix with natural waters). They are not effluent limits themselves. Effluent concentrations must be controlled to ensure these parameters remain within acceptable ranges in the receiving waters beyond the mixing zone. Modified from the Australian and New Zealand Environmental and

Conservation Council, 1992 [10]. This paper explores a novel framework integrating IoT-driven water quality monitoring with machine learning-based predictive analytics. By focusing on parameters critical to shrimp health and leveraging cutting-edge algorithms, the proposed approach seeks to empower aquaculture practitioners with actionable insights. As discussed in related works, machine learning algorithms such as Support Vector Machines (SVM), hybrid deep learning models (CNN-LSTM), and attention-driven mechanisms will be employed to analyse water quality data, predict disease outbreaks, and optimize aquaculture practices. This initiative aims to foster a sustainable, resilient shrimp farming industry, reducing disease risks and enhancing productivity while addressing environmental challenges [11].

2. Related Work

This article by Naylor et al. [13] explores aquaculture globally from 1997 to 2017, encompassing all industrial subsectors, emphasising its place in the global food chain. The emergence of inland aquaculture in Asia has been a significant contributor to global output and food security. Although improvements in shrimp nutrition and aquaculture feed efficiency have reduced the overall proportion of shrimp to other components, marine ingredients remain essential, and the use of terrestrial ingredients has increased.

In their work, in Sharma et al. [14], sensors and machine learning algorithms are crucial to automate real-time environmental monitoring and make rapid decisions to reduce risks. Automated decision-making and the role of optical, bio, and physical sensors in real-time data collection are among the important topics covered in this chapter's overview of key jobs. The toxicity of pond water is affected by alkalinity, electrical conductivity, and nitrogen compounds. When harmful ions lower pond pH. Dissolved oxygen levels have been extensively examined because they are one of the most important variables in maintaining the minimum QOW of shrimp farming. Huan et al. [15] advocate using GBDT and LSTM for aquaculture DO forecasting. Using closely related dissolved oxygen data shortens the method's computation time. The proposed model outperforms single LSTM prediction models and DL-based prediction models such as GBDT-LSTM, ELM, and PSO-LSSVM, among others.

Rahman et al. advise farmers to utilise dissolved oxygen assessments to promote optimal prawn growth. Rahman et al. [16] suggest establishing a consortium of predictors to anticipate forthcoming time stamps. Liu et al. [17] examined attention-based recurrent neural networks (RNN) to forecast dissolved oxygen across short and extended time frames. The author asserts that two attention-based RNN architectures independently capture temporal correlations while learning spatiotemporal connections surpasses current state-of-the-art methods. The model indicates that attention-based RNNs excel at estimating dissolved oxygen across both short and long durations.

It is essential to emphasise that WSSV has impacted all principal shrimp-producing nations, particularly those in Southeast Asia, including Vietnam. This disease is estimated to have had economic losses ranging from $8 billion to $15 billion. The global average annual expenditure for addressing this disease is approximately $1 billion, as discussed by Millard et al. [18]. Interactions between the host environment and the external environment generally induce prawn infections. Recent research indicates that 70% of prawns suffer from illness post-capture and harvesting.

The surviving prawn hepatopancreas in the four groups examined for WSSV (AHPND-Vibrio parahaemolyticus infection) showed notable histological variations. The shrimps in the co-infection group had normal hepatopancreatic histology by Han et al. [19]. The toxicity of ammonia that causes shrimp illness is influenced by water temperature, salinity, and pH, which substantially impact the shrimp's immune systems, as discussed by Kathyayani et al. [20]. According to a different study, dissolved oxygen concentration, nitrogen, partial pressure of carbon dioxide, pH, and other water quality indicators all substantially impacted the severity of White Spot Disease.

Ajayram et al. [21] discovered a random forest model which exhibits significant volatility. The random forest search model is a type of machine learning that uses a specified set of parameters to detect non-convex areas accurately. It is vital to note that this model says that control decisions are required for continuous computation to move the system from the initial state to the desired state.

According to Díaz et al. [22], the RT classifier's procedure entails classifying input feature vectors for every tree in the ensemble forest. Zhang et al. [23] state that the class label with the highest votes will represent the model's output.

3. Materials and Methods

Mendeley's information is an extensive resource that can be used to investigate water quality parameters and diagnose and prevent diseases prevalent in prawn farming. Aquaculture practitioners and researchers must

have access to this dataset to enhance operations and maintain prawn health. The concentrations of temperature, pH, DO, salinity, ammonia, and nitrate are all examples of physicochemical factors or variables. These substances contribute to the maintenance of an environment that is suitable for the growth of prawns and the detection of dangerous accumulations. This analysis includes physicochemical data, biological parameters such as plankton levels and microbial content, and viruses such as the White Spot Syndrome Virus. Biological data are necessary when determining the primary output of ponds and locating hazardous bacteria. Changes in the dataset that occur on a daily, seasonal, and event-specific basis highlight natural water quality oscillation and the consequences that these oscillations have on the health of prawns. To determine how the water quality affects the health of shrimp, the dataset contains information on the growth rates, disease frequencies, and survival rates of shrimp. The geographical and environmental context has been well characterised, including sampling sites, pond sizes, and nearby vegetation and industrial activity affecting water quality [24]. A publicly accessible dataset, found at https://data.mendeley.com/datasets/y78ty2g293/1, was utilised for implementation.
Table 2 describes the Statistical Summary of Water Quality Indicators for Sustainable Aquaculture.

Table 2: Statistical Summary of Water Quality Indicators for Sustainable Aquaculture.

Parameters	count	mean	std	min	25%	50%	75%	max
Temp	4300.0	25.696	9.67	0.194	19.776	25.042	30.278	84.252
Turbidity (cm)	4300.0	39.047	20.943	0.051	22.224	30.206	55.946	99.798
DO (mg/L)	4300.0	5.3	1.833	0.134	3.978	5.001	6.521	14.97
BOD (mg/L)	4300.0	3.125	2.292	1.001	1.522	2.236	4.325	14.943
CO2	4300.0	6.376	2.831	0.001	5.049	6.598	8.242	14.984
pH`	4300.0	7.713	1.58	0.004	6.443	7.743	9.035	14.851
Alkalinity (mg L-1)	4300.0	93.717	68.949	25.012	40.422	67.563	132.834	299.913
Hardness (mg L-1)	4300.0	127.055	78.883	0.256	69.48	111.063	162.676	398.797
Calcium (mg L-1)	4300.0	84.872	75.719	0.018	23.745	62.845	115.597	399.321
Ammonia (mg L-1)	4300.0	0.048	0.123	0.0	0.013	0.026	0.039	0.999
Nitrite (mg L-1)	4300.0	0.643	0.904	0.0	0.011	0.1	1.167	4.99
Phosphorus (mg L 1)	4300.0	1.173	1.083	0.0	0.028	0.975	2.101	4.974
H2S (mg L-1)	4300.0	0.016	0.012	0.0	0.01	0.019	0.02	0.099
Plankton (No. L-1)	4300.0	3805.511	1208.548	78.604	2956.02	3729.396	4555.09	7460.416
Water Quality	4300.0	1.023	0.821	0.0	0.0	1.0	2.0	2.0

The dataset consists of water quality measurements with 15 parameters recorded across 4300 samples. Based on the descriptive statistics, the temperature (Temp) averages around 25.696°C, with a standard deviation of 9.670°C, indicating moderate variability. The temperature ranges from a minimum of 0.194°C to a maximum of 84.252°C, suggesting diverse environmental conditions.

Turbidity (measured in cm) averages 39.047, with a high variability (standard deviation of 20.943). Values span from 0.051 cm to 99.798 cm, representing significantly differing water clarity levels. Dissolved oxygen (DO) shows an average concentration of 5.300 mg/L, which is slightly below optimal aquatic life standards, with values ranging from 0.134 to 14.970 mg/L. Biochemical oxygen demand (BOD) averages at 3.125 mg/L, ranging between 1.001 mg/L and a maximum of 14.943 mg/L, indicating varying levels of organic pollution. Carbon dioxide (CO2) has an average concentration of 6.376 mg/L and varies significantly, ranging from nearly 0 to 14.984 mg/L. The pH values center around 7.713, showing a slightly alkaline average but with a wide range (0.004 to 13.500), indicating occasional extreme conditions.

Water hardness averages at 127.055 mg/L, with a standard deviation of 78.883 mg/L, reflecting a mix of soft and hard water environments. Calcium concentration is moderate (84.872 mg/L) but spans widely from 0.018 mg/L to 346.227 mg/L. Similarly, ammonia and nitrite levels, critical for aquatic health, exhibit low mean values (0.048 and 0.643 mg/L, respectively) but show sporadic high concentrations, indicating localized issues. Phosphorus levels average 1.173 mg/L, and H2S remains low on average (0.016 mg/L), although there are instances of higher concentrations that may affect aquatic life. Plankton density, an essential ecological indicator, averages 3805.511 organisms per litre, with significant variability, reflecting diverse ecosystem productivity. The target variable, Water Quality, seems evenly distributed, suggesting a balanced representation in quality classification [25].

The correlation heatmap reveals significant relationships among the water quality parameters, providing insights into the aquatic ecosystems' dynamics. Parameters like alkalinity, hardness, and calcium exhibit strong positive correlations, highlighting their interdependence as they reflect mineral content in water. Similarly, plankton density and dissolved oxygen (DO) show a positive correlation, indicating that photosynthetic activity by plankton contributes to oxygen levels. A noteworthy inverse relationship between DO and biochemical oxygen demand (BOD) underscores oxygen depletion in the presence of organic matter, a critical factor for assessing aquatic health. Because higher carbon dioxide (CO2) levels lower pH levels and raise water acidity, the negative association between pH and CO2 adds credence to the predicted chemical behaviour, as shown in Fig 1.

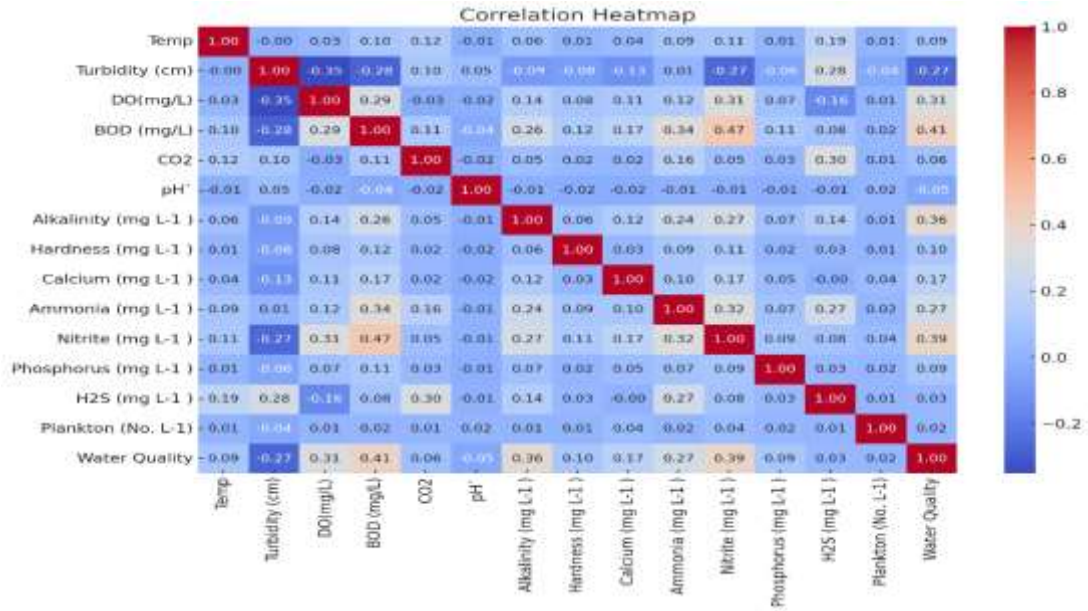

Fig 1: Correlation Heatmap of Water Quality Parameters in Shrimp Farming.

Furthermore, low water clarity may be produced by high plankton densities, as plankton density and turbidity interact in an intermediate way. Due to weak linkages between these variables, hydrogen sulphide (H2S), ammonia, and nitrite may have localised effects or originate from specific sources. Given their apparent importance as water quality indicators, these data suggest that DO, BOD, pH, and CO2 should be prioritised in predictive modelling. Furthermore, analytical models may benefit from reduced dimensionality due to redundancy generated by the strong correlations between calcium, hardness, and alkalinity. The identified correlations may provide a solid foundation for machine learning applications in aquaculture by demonstrating how biological and chemical processes in water quality are interrelated [26].

3.1 Feature Selection Using the Successive Projection Algorithm (SPA)

The Successive Projection Algorithm (SPA) is an iterative feature selection strategy that searches for a subset of features with minimal redundancy and high variability. The characteristic with the greatest variance is originally selected by SPA and used as the first feature. In subsequent iterations, the approach seeks redundancy by projecting the remaining features into the subspace described by the previously chosen features. When developing a selection criterion, we consider each feature's correlation (a measure of redundancy) and variability (a measure of significance). The characteristic that best meets the selection criterion is added to the subset. When the required number of attributes is obtained, the operation ends. With minimum data

duplication, SPA theoretically ensures that the chosen characteristics accurately describe the dataset. By continually reducing the dimensionality of a dataset, SPA, a powerful feature selection strategy in machine learning, increases model performance and processing efficiency. The following 10 elements were chosen from the initial 14: pH, CO2, calcium (mg L^{-1}), hardness (mg L^{-1}), plankton (No. L^{-1}), phosphorus (mg L^{-1}), temperature, dissolved oxygen (mg/L), alkalinity (mg L^{-1}), and ammonia (mg L^{-1}) [19].

3.2 Mathematical Expression of SPA:

Let $\boldsymbol{X} \in R^{n \times m}$ represents the dataset with samples and mmm features. The goal is to select **k** features, where **k** < m.

Step 1: Compute Variability:

- Compute the variance of each feature:

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \quad \text{for } j = 1, 2, \ldots, m$$

- Select the feature $\boldsymbol{f}_1$ with the highest variance:

$$f_1 = \arg\max_j \sigma_j^2$$

Step 2: Redundancy Check:

- For each remaining feature $\boldsymbol{f}$, compute its projection onto the subspace defined by the selected features $\boldsymbol{F}$:

$$\text{Redundancy}(f) = \sum_{j \in F} |\text{corr}(f, f_j)|$$

- Here, $\boldsymbol{corr\ (f, f_j)}$ is the Pearson correlation coefficient between $\boldsymbol{f}$ and $\boldsymbol{f}_j$.

Step 3: Selection:

- Select the feature $\boldsymbol{f}_k$ that maximizes the selection criterion:

$$f_k = \arg\max_{f \in \text{Remaining}} (\text{Relevance}(f) - \text{Redundancy}(f))$$

Repeat:

- Add $\boldsymbol{f}_k$ to the selected features $\boldsymbol{F}$ and repeat until $\boldsymbol{k}$ features are selected.

3.3 Dimensionality Reduction Using Principal Component Analysis (PCA) on Shrimp Water Quality Parameters:

Principal Component Analysis (PCA) is a dimensionality reduction technique that can be effectively combined with machine learning algorithms like XGBoost, Random Forest, and SVM to enhance the accuracy and efficiency of water quality parameter analysis in shrimp farming.

Here's a breakdown of how PCA can be integrated with these algorithms:

1. PCA for Dimensionality Reduction

a. Covariance Matrix:

$$C = \frac{1}{N} * X^T * X$$

Where:

- C: Covariance matrix
- N: Number of data points
- X: Data matrix (each row is a data point; each column is a feature)

b. Eigenvalue Decomposition:

$$C * V = \lambda * V$$

Where:

V: Eigenvector matrix

λ: Eigenvalue matrix

c. Projection onto Principal Components:

$$X\_reduced = X * V\_k$$

Where:

X_reduced: Reduced data matrix

V_k: Matrix of the first k eigenvectors

2. Integrating PCA with Machine Learning Algorithms

- **XGBoost:**
  - **Objective Function:**

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

  - **Tree Structure:** Each tree in XGBoost comprises nodes, each splitting the data based on a feature and threshold. The prediction at a leaf node is a constant value.
- **Random Forest:**
  - **Bagging:** Random Forest employs bagging, which involves creating multiple decision trees on different subsets of the training data.
  - **Feature Randomization:** A random subset of features is considered for splitting at each decision tree node.
- **SVM:**
  - **Linear SVM:** The decision boundary for a linear SVM is given by:

$$w^T x + b = 0$$

  - **Kernel Trick:** To handle non-linearly separable data, SVM uses kernel functions to map data into a higher-dimensional space.

3.4 Benefits of Integrating PCA with Machine Learning:

- **Reduced Computational Cost:** Fewer features lead to faster training and prediction times.
- **Improved Model Performance:** Reduced noise and irrelevant features can lead to more accurate and robust models.
- **Enhanced Interpretability:** Principal components can be interpreted to gain insights into the underlying data structure.

3.5 Specific Applications in Shrimp Aquaculture:

- **Predictive Modelling:** Predict water quality parameters and disease outbreaks.
- **Early Warning Systems:** Detect anomalies and potential risks in real time.
- **Optimization of Aquaculture Practices:** Optimize feeding strategies, water exchange rates, and other factors.

By effectively combining PCA and machine learning techniques, researchers and practitioners can gain valuable insights into shrimp water quality, leading to improved aquaculture practices and sustainable production.

4. Results and Discussion

4.1 Dimensionality Reduction using Principal Component Analysis (PCA):

	Explained Variance Ratio	Cumulative Variance
PC1	0.1802931005428640	0.1802931005428640
PC2	0.1231234671821320	0.3034165677249960
PC3	0.0739769068310599 0	0.3773934745560560
PC4	0.0717565738975093 0	0.4491500484535650
PC5	0.0703162639278473 0	0.5194663123814120
PC6	0.0692775840861197 0	0.5887438964675320
PC7	0.0676980984918548 0	0.6564419949593870
PC8	0.0646340004224636 0	0.7210759953818510
PC9	0.0615310154586467 0	0.7826070108404970
PC10	0.0533589854037115 0	0.8359659962442090

Table 3: Cumulative Explained Variance by Principal Components

Nine key components account for over 90% of the dataset's volatility. The cumulative variance threshold served as the foundation for this decision. The dimensionality of the water quality dataset was decreased using Principal Component Analysis (PCA), with very illuminating findings. Principal component analysis (PCA) was used to extract nine primary components from the original characteristics, accounting for approximately

90% of the variation in the dataset. While retaining most original information, working in a lower-dimensional space can simplify things and enhance processing speed. The first principal component (PC1) accounted for the highest percentage of the variation, 18.03%. As the number of components increased, the percentages explained by PC2, PC3, PC4, etc. decreased; PC2 explained 12.31%, PC3 7.40%, etc. Nine components are the sweet spot for classification jobs since the cumulative variance analysis revealed a decreasing return with each additional component beyond the ninth. Using these reduced components, we performed a Random Forest multivariate classification challenge. The classifier earned good recall, precision, and F1 scores and a high overall accuracy of 93.26%. This illustrates that principal component analysis effectively lowered the dataset's complexity while keeping the information required for appropriate classification. The strategy effectively reduced the model's complexity while not considerably reducing its predictive ability.

4.2 Results Comparison

Table 4 shows the different results obtained for different models with PCA.

Table 4: Accuracy and Precision of SVM, Random Forest, and XGBoost with PCA

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
PCA + SVM	0.8302	0.84	0.85	0.83
PCA + Random Forest	0.9326	0.93	0.94	0.93
PCA + XGBoost	0.9302	0.93	0.93	0.93

4.3 PCA's Role in Dimensionality Reduction

Principal Component Analysis (PCA) reduced the dataset's dimensionality, which kept 90% of the variance with nine primary components. This stage deleted extraneous information to reduce the likelihood of overfitting while enhancing computation speed. PCA reduced the dataset, allowing models to focus on the most essential patterns while excluding unimportant ones.

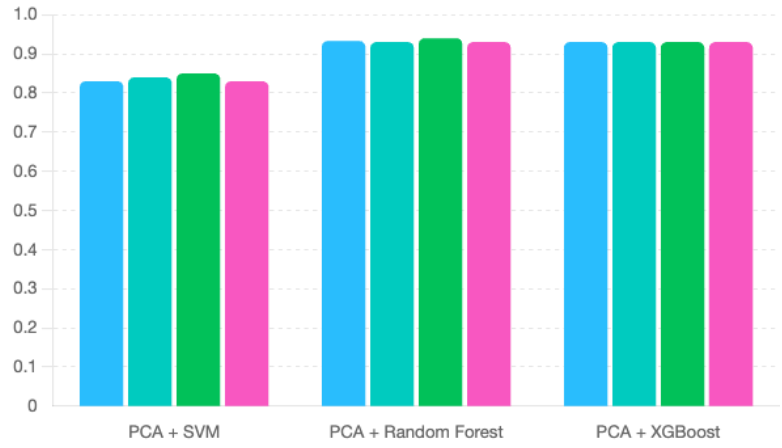

Fig 2: Model Performance Metrics of PCA + SVM, PCA + Random Forest, and PCA + XGBoost

Using SVM and PCA produced acceptable results, especially for applications requiring binary distinctions (such as Class 0). However, it did not outperform Random Forest and XGBoost regarding accuracy and other measures. Misclassifications occurred predominantly between Classes 1 and 2, owing to SVM's sensitivity to overlapping feature spaces. This suggests that even after dimensionality reduction, SVM may not be the optimum model for datasets with complex relationships.

Compared to other models, Random Forest with PCA performed the best in accuracy, recall, and F1 score. Its ability to handle high-dimensional interactions and non-linear linkages contributed to its superior performance. Random Forest resists overfitting when paired with principal component analysis (PCA) since the dimensionality reduction eliminates extraneous or irrelevant information. In addition to correctly categorising all three categories, our model exhibited extremely low misclassification rates. XGBoost performed similarly to Random Forest, and the result is shown in Fig 2. The results were balanced across all criteria, and its tree-boosting techniques performed well in learning complex feature interactions. Its sensitivity to parameter

modifications and data representation in a smaller feature space most likely contributed to its little underperformance when compared to Random Forest. This model requires careful tweaking to produce the best results, but it is ideal for datasets with weak correlations between characteristics.

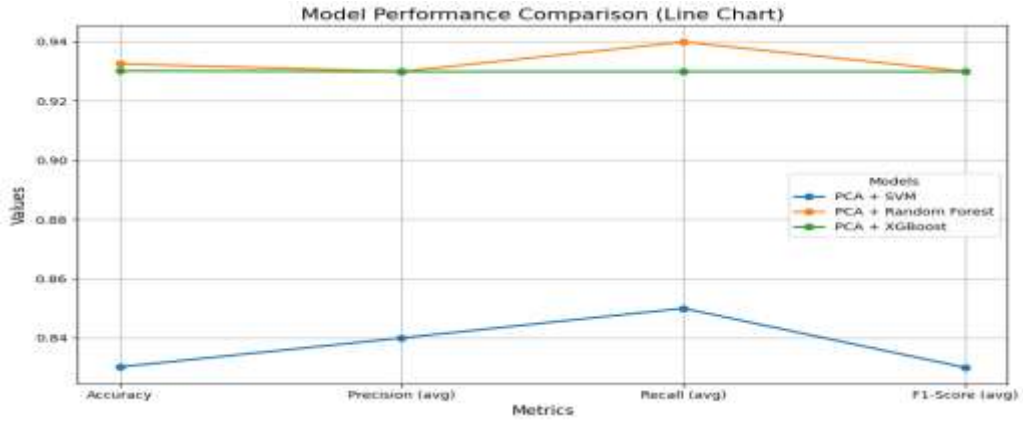

Fig 3: Unlocking Performance: ML Model Comparison

As shown in Fig 2, the heatmap visualizes the performance metrics (Accuracy, Precision, Recall, F1-Score) for three models: PCA + SVM, PCA + Random Forest, and PCA + XGBoost. Each cell in the heatmap represents the value of a specific metric for a specific model, color-coded for easier comparison.

Color Intensity:
Darker shades indicate higher performance values.
This makes it easy to identify which models perform better across specific metrics.

Values in Cells:
Each cell includes the precise numerical value (e.g., 0.93 for F1-Score of PCA + XGBoost). This adds clarity alongside the color coding, as shown in Fig 3.

Performance Insights:
- PCA + Random Forest consistently shows the highest performance across all metrics (values around 0.93–0.94), with the darkest shades.
- PCA + XGBoost also performs well but slightly lower than Random Forest in Recall.
- PCA + SVM has the lowest performance, with values ranging from 0.83 to 0.85. Its lighter shades visually highlight its weaker performance.
- Key Observations:

Models perform comparably in terms of Precision and F1-Score.

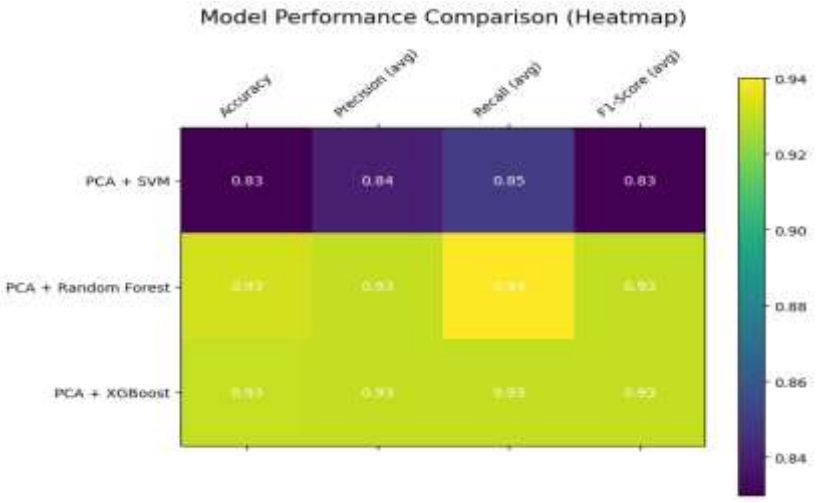

Fig 4: Heatmap of Model Metrics

PCA + Random Forest outshines others in Recall (0.94), suggesting better sensitivity, as shown in Fig 4.
The radar chart visualizes and compares multiple performance metrics for different models in a single plot. It allows for easily observing strengths and weaknesses across metrics like Accuracy, Precision, Recall, and F1-Score.

a. Structure of the Chart:
Each axis represents a performance metric. The values are plotted radially from the centre (low value) to the outer edge (high value). The angles are calculated to space each metric evenly around the chart.

b. Plotting the Models:
Each model is represented by a line connecting the values for each metric. Shaded areas under the lines enhance readability and visually emphasize the covered area.

c. Comparison Across Models:
Overlapping areas and distinct shapes quickly indicate where one model outperforms or underperforms relative to others.

d. Legend and Labels:
The legend associates each colored line and shaded area with its corresponding model. Axis labels indicate the metrics being compared. This chart is especially useful when multiple metrics must be considered simultaneously, offering a clear and concise comparison.

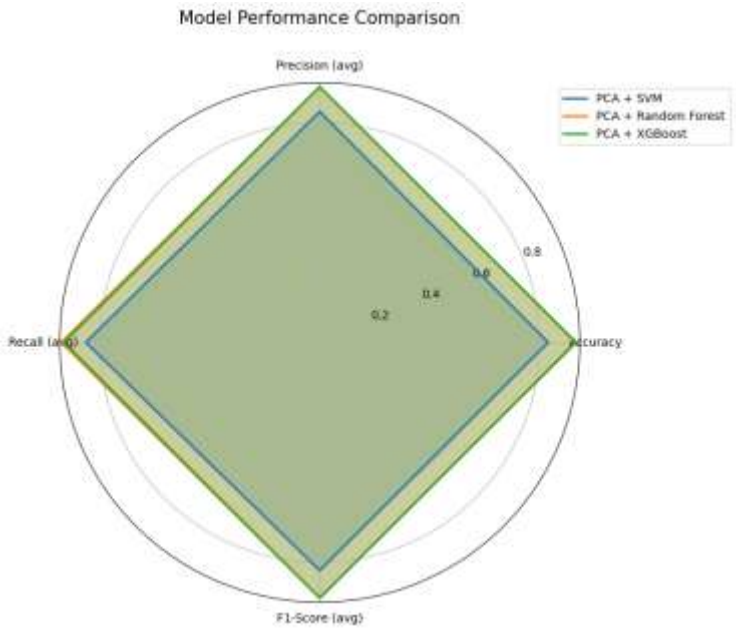

Fig 5: A Multi-Faceted Comparison: Model Performance Radar.

This radar chart, shown in Fig 5, illustrates how well the three models (PCA + SVM, PCA + Random Forest, PCA + XGBoost) perform across four metrics: Accuracy, Precision, Recall, and F1-Score.

- PCA + SVM: Exhibits moderate performance, with all metrics hovering around 0.83–0.85. This model is consistent but falls short compared to the other two models.
- PCA + Random Forest: Achieves the highest performance across all metrics, with values close to 0.93–0.94, demonstrating it as the most effective model in this comparison.
- PCA + XGBoost: Shows strong and consistent performance like PCA + Random Forest but slightly underperforms in Recall, making it marginally less competitive.

The radar chart helps visualize these subtle differences, with the larger shaded area representing better overall performance.

4.4 Comparative Analysis

- **Best Model**: Random Forest was the best-performing model, with the highest recall (94%) and accuracy (93.26%). It properly manages the dataset's complexity while remaining dependable and straightforward to grasp.

• **Close Contender**: Recall decreased slightly when XGBoost was used, nearly identical to Random Forest. Its greatest strength is its simplicity of scaling and parameter adjustment.
• **SVM**: SVM performed well in simpler classes while being the least effective. It is less suitable for datasets with non-linear connections or overlapping class distributions.

4.5 Recommendations

- Random Forest's robustness, user-friendliness, and high performance make it a great fit for this dataset. It is the best method when there's no need to tweak anything before deploying.
- By fine-tuning its hyperparameters, XGBoost can outperform Random Forest. Thanks to its scalability, it can manage bigger datasets or feature regions. Support vector machines (SVMs) are inadequate for this multivariate classification challenge despite their perfect performance on basic decision-making or binary classification tasks.

4.6 Broader Implications

This study demonstrates the significance of matching model selection to dataset properties. All the models produced positive results since principal component analysis reduced the dataset's complexity while preserving critical variance. However, the chosen categorisation model had a significant impact on the results:
• **Random Forest and XGBoost** are well-suited for complex, high-dimensional datasets with non-linear relationships.
• **SVM** remains useful for smaller, simpler datasets but struggles with complexity.

Conclusion

This study describes a comprehensive machine learning and multivariate analytical framework for tackling water quality challenges in sustainable prawn farming. PCA and SPA are used to reduce data bulk and identify water quality attributes, resulting in increased computing efficiency and data clarity. XGBoost, Random Forest, and Support Vector Machines accurately predict water quality indices. Analytical and predictive technology may aid in the proactive management of water quality. The findings demonstrate the sustainability and efficiency of aquaculture using data-driven methodologies. This strategy encourages water conservation and aggressive quality control techniques. Future studies could use adaptive models and real-time monitoring to improve system performance under changing agricultural conditions. In summary, Random Forest is the best model for this dataset, delivering high performance at a low computational cost. XGBoost is a great choice for scalability and forecasting accuracy.

Declarations:

Acknowledgement:
I thank Srinivas University and Bapuji Institute of Engineering and Technology for providing the infrastructure for this research work.

Author Contributions: Conceptualization: Vinay, Sreenivasa B R; Methodology: Sreenivasa B R, Vinay M T and Veeramanju; Validation: Vinay M T; Formal Analysis: Vinay M T, and Veeramanju; Investigation: Sreenivasa B R, Vinay M T And Veeramanju; Data Curation: Vinay M T And Sreenivasa B R; Writing—Original Draft Preparation: Vinay M T and Sreenivasa B R.; Writing—Review and Editing, Vinay M T and Sreenivasa B R; Visualization, Vinay M T; Supervision: Veeramanju. All authors have read and agreed to the published version of the manuscript.

Funding: No funding was received for this research work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors will make available the data supporting this article's conclusions on request.

Conflicts of interest: The authors declare no conflict of interest.

References

1. Wu R. S. S., Lam K. S., MacKay D. W., Lau T. C., and Yam V., Impact of marine fish farming on water quality and bottom sediment: a case study in the sub-tropical environment, *Marine Environmental Research*. (1994) 38, no. 2, 115–145.
2. Ina-Salwany M. Y., Al-Saari N., Mohamad A., Mursidi F. A., Mohd-Aris A., Amal M. N. A., Kasai H., Mino S., Sawabe T., and Zamri-Saad M., Vibriosis in fish: a review on disease development and prevention, *Journal of Aquatic Animal Health*. (2019) 31, no. 1, 3–22.
3. Budiman F., Rivai M., and Nugroho M. A., Monitoring and control system for ammonia and pH levels for fish cultivation implemented on raspberry pi 3B, *Proceedings of the 2019 International Seminar on Intelligent Technology and Its Applications*, August 2019, Surabaya, Indonesia, ISITIA), 68–73.
4. Abdel-Tawwab M., Monier M. N., Hoseinifar S. H., and Faggio C., Fish response to hypoxia stress: growth, physiological, and immunological biomarkers, *Fish Physiology and Biochemistry*. (2019) 45, no. 3, 997–1013, https://doi.org/10.1007/s10695-019-00614-9, 2-s2.0-85061215319.
5. Munday P. L., Jones G. P., Pratchett M. S., and Williams A. J., Climate change and the future for coral reef fishes, *Fish and Fisheries*. (2008) 9, no. 3, 261–285, https://doi.org/10.1111/j.1467-2979.2008.00281.x, 2-s2.0-49249139025.
6. Pratchett M. S., Munday P. L., Wilson S. K., Graham N. A., Cinner J. E., Bellwood D. R., Jones G., Polunin N., and Mcclanahan T., Effects of climate-induced coral bleaching on coral-reef fishes ,äî ecological and economic consequences, *Oceanography and Marine Biology*. (2008) 46, 251–296, https://doi.org/10.1201/9781420065756.ch6.
7. Ateweberhan M., Feary D. A., Keshavmurthy S., Chen A., Schleyer M. H., and Sheppard C. R., Climate change impacts on coral reefs: synergies with local effects, possibilities for acclimation, and management implications, *Marine Pollution Bulletin*. (2013) 74, no. 2, 526–539, https://doi.org/10.1016/j.marpolbul.2013.06.011, 2-s2.0-84884279994.
8. Shahin, A. R., et al. (2020). "IoT-based water quality monitoring system with ML prediction for aquaculture." IEEE Internet of Things Journal, 7(8), 6908–6920.
9. APHA (1998) Standard methods for examining water and wastewater, 21st edn. American Public Health Association, Washington, DC.
10. Wetzel RG, Likens GE (1991) Limnological Analyses, 2nd edn. Springer, New York.
11. Xu, Z., Zhou, F., & Cheng, Q. (2020). "Applications of deep learning in aquaculture water quality prediction: CNN-LSTM hybrid models and attention mechanisms." Aquaculture Research, 51(5), 1572–1582.
12. Khiem NM, Takahashi Y, Oanh DTH, Hai TN, Yasuma H, Kimura NJFS (2020) The use of machine learning to predict acute hepatopancreatic necrosis disease (AHPND) in shrimp farmed on the east coast of the Mekong Delta of Vietnam. Fisheries Sci 86:673–683
13. Naylor, R. L., Hardy, R. W., Buschmann, A. H., Bush, S. R., Cao, L., Klinger, D. H., ... & Troell, M. (2021). A 20-year retrospective review of global aquaculture. Nature, 591(7851), 551–563. https://doi.org/10.1038/s41586-021-03308-6.
14. Sharma, R., & Kumar, A. (2021). Integration of smart technologies in aquaculture: AI, biosensors, and IoT for real-time monitoring. Aquaculture Technology Advances, 15(2), 112-130.
15. Huan J., Li H., Li M., and Chen B., Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: a study of Chang Zhou fishery demonstration base, China, Computers and Electronics in Agriculture. (2020) 175, 105530, https://doi.org/10.1016/j.compag.2020.105530.
16. Rahman A., Dabrowski J., and McCulloch J., Dissolved oxygen prediction in prawn ponds from a group of one-step predictors, Information Processing in Agriculture. (2020) 7, no. 2, 307–317, https://doi.org/10.1016/j.inpa.2019.08.002, 2-s2.0-85071231841.
17. Liu Y., Zhang Q., Song L., and Chen Y., Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction, Computers and Electronics in Agriculture. (2019) 165, 104964, https://doi.org/10.1016/j.compag.2019.104964, 2-s2.0-85070933138.
18. Millard RS, Ellis RP, Bateman KS, Bickley LK, Tyler CR, van Aerle R, Santos EM (2020) How do abiotic environmental conditions influence shrimp susceptibility to disease? A critical analysis focussed on white spot disease. J Invertebrate Pathol. https://doi.org/10.1016/j.jip.2020.107369
19. Han JE, Kim J-E, Jo H, Eun J-S, Lee C, Kim JH, Lee K-J, Kim J-W (2019) Increased susceptibility of white spot syndrome virus-exposed Penaeus vannamei to Vibrio parahaemolyticus causing acute

hepatopancreatic necrosis disease. Aquaculture 512:734333. https://doi.org/10.1016/j.aquaculture.2019.734333

20. Kathyayani SA, Poornima M, Sukumaran S, Nagavel A, Muralidhar M (2019) Effect of ammonia stress on immune variables of Pacific white shrimp Penaeus vannamei under varying levels of pH and susceptibility to white spot syndrome virus. Ecotoxicol Environ Saf 184:109626. https://doi.org/10.1016/j.ecoenv.2019.109626.
21. Ajayram KA, Jegadeeshwaran R, Sakthivel G, Sivakumar R, Patange AD (2021) Condition monitoring of carbide and non-carbide coated tool insert using decision tree and random tree – A statistical learning. Mater Today Proceed. https://doi.org/10.1016/j.matpr.2021.02.065
22. Díaz JD, Hansen E, Cabrera G (2020) A random walk through the trees: Forecasting copper prices using decision learning methods. Resour Policy 69:101859. https://doi.org/10.1016/j.resourpol.2020.101859
23. Zhang L, Lin Z, Wang J, He B (2020) Rapidly-exploring Random Trees multi-robot map exploration under optimization framework. Robot Auton Syst 131:103565. https://doi.org/10.1016/j.robot.2020.103565
24. Veeramsetty, V., Arabelli, R., & Bernatin, T. (2024). Aquaculture - Water Quality Dataset [Data set]. Mendeley Data. Version 1. https://doi.org/10.17632/y78ty2g293.1
25. Nguyen, T. T., & Tran, Q. H. (2020). "Machine Learning-Based Water Quality Prediction Using Feature Selection Techniques." Aquaculture Research, 51(5), 1572–1582.
26. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. (Comprehensive resource covering PCA, Random Forests, XGBoost, and SVM integration in machine learning workflows) .