Predicting fish growth using Internet of Things and Artificial Intelligence in Aquaculture systems

Prepared By

Name: Ahmed Mohammed Ahmed Alnahari

ID No.: 251003035

Submitted in fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: **Dr. Hager Saleh**

Collage of Informatics, Midocean University, Comoros

Email Address : 251003035@st.midocean.edu.km

Acknowledgements

I would like to express my deepest gratitude to everyone who has been part of my graduation Thesis journey.

First and foremost, my heartfelt thanks go to my supervisor, **Dr Hager Saleh**, for their invaluable guidance, patience, and encouragement throughout every stage of this work. Your insights and constructive feedback have been instrumental in shaping the quality of this Thesis.

I am also sincerely grateful to my family and friends for their unwavering support, understanding, and motivation, especially during the most demanding moments. Your belief in me has been my greatest source of strength.

Finally, I extend my appreciation **to Midocean University / Science in Artificial Intelligence Department** for providing the resources, environment, and opportunities that made this project possible.

This achievement is not mine alone - it is the result of the collective effort, kindness, and encouragement of all those mentioned above.

Abstract

This thesis presents an integrated data-driven framework that combines Internet of Things (IoT) technologies with advanced machine learning (ML) techniques to enhance monitoring, prediction, and decision-making in aquaculture systems. The study focuses on developing a smart IoT-based fishpond monitoring model that continuously measures critical water quality parameters—such as temperature, pH, dissolved oxygen, turbidity, ammonia, and nitrate—through distributed sensors connected to an ESP-32 microcontroller. The collected data were processed, normalized, and stored in a cloud database to form a reliable dataset for predictive analysis.

Several supervised regression algorithms were trained and evaluated to predict two biological growth indicators: fish length and fish weight. The models included Decision Tree, Random Forest, Support Vector Regression (SVR), Lasso, Ridge, and XGBoost. Model performance was assessed using standard statistical metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Among all tested models, ensemble-based approaches—particularly Random Forest and XGBoost—achieved the highest accuracy, with R^2 values exceeding 0.99 for both target variables, demonstrating their superior capability in handling nonlinear relationships within IoT-derived aquaculture data.

Beyond experimentation, the research included a comprehensive literature review of fifteen recent studies (2022–2025) on IoT-enabled aquaculture analytics, revealing that hybrid deep learning and ensemble techniques consistently outperform traditional models. The findings emphasize the importance of integrating IoT sensing with predictive intelligence to support early warnings, enhance fish survival rates, and improve operational efficiency.

Overall, the proposed framework contributes to the growing field of smart aquaculture by offering a robust methodological approach that bridges real-time IoT sensing and predictive modeling. The study concludes that machine learning–driven IoT systems can transform aquaculture management into a proactive, sustainable, and data-centric practice, while future work should explore larger datasets, multimodal sensing, and autonomous control mechanisms for fully intelligent aquaculture environments.

Table of Contents

List of Abbreviations viii
List of Figures ix
List of Tables x

Chapter 1: Introduction 1
1.1 Problem Definition 2
1.2 Thesis Objectives 3
1.3 Thesis Contribution 5
1.4 Thesis Organization 6

Chapter 2: Overview 8
2.1 Introduction 8
2.2 Machine Learning (ML) 8
2.3 Deep Learning (DL) 9
2.4 Relevance of AI Concepts to Fishpond Water Quality Prediction 9
2.5 Some of Models (Regression Models) 10

Chapter 3: Related Work 24
3.1 Introduction 24
3.2 Literature Review 25
3.4 Comparison Table 40

Chapter 4: Chapter 4 Proposed Approach (Methodology)..... 42
4.1 Dataset Description..... 42
4.2 Data Pre-processing ……..…….. 43
4.3 Splitting the dataset …… 44
4.4 Models (Regression Models) …… 44

Chapter 5: Experimental Results and Discussion ……. 47
5.1 Introduction …….. 47
5.2 Experimental Setup …….. 47
5.3 Training Parameters ……. 48
5.4 Results and Discussion: Predicting Fish Length …….. 48
5.5 Results and Discussion: Predicting Fish Weight ……. 50
5.6 Performance analysis and overview …… 52
5.7 Comparative Analysis……...…… 55

Chapter 6: Conclusion and Future Work 56
6.1 Conclusion …… 56
6.2 Important Features …… 57
6.3 Future Work …… 59

References …… 61

List of Figures

- Figure 1 Decision Tree Regression (DT) ……..…..… 10
- Figure 2 Random Forest Regressor (RF) ….……..…..… 12
- Figure 3 Support Vector Regression (SVR) ……..……..….14
- Figure 4 Support Vector Machine (SVM) …….……15
- Figure 5 Lasso Regression (LR) …….……...…….17
- Figure 6 Ridge Regression …….……..….19
- Figure 7 XGBoost Regression ……..……..……21
- Figure 8 XGBoost Regression ……..……..22
- Figure 9 Tree Algorithm …….……23
- Figure 10 Methodology overview …….42
- Figure 11 all algorithm Result Fish Length R^2 …….49
- Figure 12 All algorithm Result Fish weight ……..51

List of Abbreviations

- Acc.: Accuracy
- ANN: Artificial Neural Network
- ML: Machine Learning
- DL: Deep Learning
- XAI: Explainable Artificial Intelligence
- SVM: Support Vector Machine
- RF: Random Forest
- DT: Decision Tree
- LSTM: Long Short-Term Memory
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- LR: Logistic Regression
- KNN: K-Nearest Neighbors
- MAE: Mean Absolute Error
- MSE: Mean Squared Error
- R^2: R-Squared
- LDA: Linear Discriminant Analysis
- LLM: Large Language Model
- NLP: Natural Language Processing
- EDM: Educational Data Mining

List of Tables

- Table 1 Comparison of previous studies …….. 40

- Table 2 X: Performance Comparison of Regression Models for Fish Length Prediction …… 48

- Table 3 X: Performance Comparison of Regression Models for Fish Weight Prediction …….50

- Table 4 X: Comparative Performance of Regression Models for Fish Length and Weight Prediction ……..55

Chapter 1: Introduction

Aquaculture has emerged as one of the fastest-growing food production sectors, driven by the increasing global demand for fish and aquatic products. Fishponds contribute significantly to freshwater aquaculture production; however, their sustainability and productivity strongly depend on maintaining optimal water quality conditions. Even minor fluctuations in parameters such as temperature, pH, dissolved oxygen (DO), and turbidity can trigger stress, disease outbreaks, or mass mortality in fish populations. Traditional water monitoring methods, which rely on manual sampling and laboratory analysis, are time-consuming, intermittent, and unable to capture real-time variations in dynamic aquatic environments.

The integration of the Internet of Things (IoT) into aquaculture introduces a transformative approach to environmental monitoring by enabling continuous sensing, automated data collection, and remote supervision of fishponds. Through distributed sensors and wireless communication, IoT systems generate large volumes of time-series water quality data, providing a rich foundation for predictive analytics and intelligent decision-making. When coupled with machine learning (ML) models—such as Random Forest (RF), XGBoost, Support Vector Regression (SVR) , K-Nearest Neighbors, Linear Regression, and Lasso Regression—IoT-enabled data can be transformed into accurate predictions that support early intervention and proactive pond management.

This thesis focuses on developing a data-driven framework for predicting water quality parameters in IoT-based fishponds using supervised machine learning techniques. The study leverages real IoT sensor data from fishpond environments to evaluate and compare the performance of several regression models. By identifying the most accurate models for water quality prediction, the research aims to support smarter aquaculture systems capable of improving fish health, reducing losses, and enhancing overall production efficiency.

1.1 Problem Definition

Water quality degradation remains one of the most critical issues in modern aquaculture. Sudden variations in pH levels, dissolved oxygen (DO) drops, temperature fluctuations, or increases in turbidity can drastically affect fish health and survival. In traditional aquaculture practices, water parameters are typically measured manually once or twice daily. Such low-frequency monitoring fails to capture rapid environmental changes, delaying corrective actions and leading to potential economic losses. Although IoT devices now provide continuous real-time measurements, fish farmers still struggle to interpret large streams of sensor data. Raw IoT data alone does not provide actionable insights unless processed and analyzed effectively. More importantly, farmers need predictive signals—not just historical values—to anticipate harmful trends before damage occurs.

Although Internet of Things (IoT) technologies now allow real-time monitoring through sensor networks, aquaculture practitioners continue to face difficulties in processing and interpreting large streams of raw data. Without proper analysis, IoT data provides descriptive rather than predictive insights. To achieve proactive pond management, predictive intelligence is essential — enabling farmers to anticipate water quality deterioration before it becomes critical.

Therefore, the core problem addressed in this study is:

How can IoT-based fish pond data be effectively processed and analyzed using machine learning algorithms to accurately predict key water quality parameters and enable early decision-making in aquaculture?

his research problem arises from several technical and practical limitations identified in current approaches:

- Raw IoT data is often noisy, incomplete, and requires preprocessing to be usable.
- Traditional statistical models fail to capture complex nonlinear interactions between water quality parameters.
- Existing monitoring frameworks primarily offer reactive responses rather than predictive insights.
- Limited research has compared multiple regression-based machine learning models using real IoT aquaculture datasets.
- These challenges highlight the need for a robust machine learning framework capable of evaluating, optimizing, and identifying high-performing predictive models suitable for real-time fishpond management systems .

1.2 Thesis Objectives

The primary objective of this thesis is to design, implement, and evaluate machine learning regression models capable of predicting fishpond water quality parameters using real-time IoT sensor data. The specific research objectives are as follows:

- To develop a predictive framework utilizing IoT-based fish pond datasets for accurate estimation of key water quality indicators.

- To perform comprehensive data preprocessing, including cleaning, handling missing values, and applying feature engineering techniques to improve model performance.
- To train and evaluate multiple regression-based machine learning algorithms — including Linear Regression, Random Forest, XGBoost, Support Vector Regression (SVR), Lasso, and other ensemble models.
- To identify the most significant environmental parameters influencing water quality dynamics and their impact on fish growth.
- To propose an optimized machine learning model suitable for integration into real-time IoT-based aquaculture management systems.
- To demonstrate how predictive analytics can provide early warnings and support proactive decision-making for sustainable fishpond operations.

1.3 Thesis Contribution

This thesis provides a set of scientific and practical contributions derived from the successful development, implementation, and evaluation of machine learning models for IoT-enabled fish pond management.

Developed and validated a machine learning framework for parameters (length and weight) using real IoT sensor data, including pH, dissolved oxygen, temperature, and turbidity.

Conducted a comprehensive comparison of six regression algorithms—Decision Tree (DT), Random Forest (RF), SVR, Ridge, Lasso, and XGBoost to predict fish growth using aquaponic data

Demonstrated the effectiveness of ensemble learning techniques, particularly Random Forest and XGBoost, in enhancing prediction stability and reducing errors, with RMSE values below 0.25 for length and 0.54 for weight.

Delivered a predictive analytics solution that enables fish farmers to monitor and forecast pond conditions with high accuracy, supporting early intervention and smarter management decisions.

Improved aquaculture productivity by transforming raw IoT sensor data into actionable insights that guide feeding, aeration, and water exchange strategies.

To compare the predictive accuracy of these models using standard evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE).

1.4 Thesis Organization

This thesis is organized into six chapters, each addressing a key component of the research process:

Chapter 1 – Introduction:

Provides background information on aquaculture and the significance of IoT in modern fish farming. It defines the research problem, outlines the objectives, highlights the study's contributions, and presents the overall structure of the thesis.

Chapter 2 – Related Work / Background:

Reviews fundamental concepts of the Internet of Things (IoT), cloud-based data acquisition, and machine learning techniques relevant to aquaculture. It explains the theoretical foundation upon which the proposed system is built.

Chapter 3 – Literature Review:

Discusses previous studies related to IoT applications in aquaculture, water quality monitoring systems, and existing machine learning approaches for predictive modeling in fish ponds. It also identifies current research gaps that form the basis for this study.

Chapter 4 – Methodology:

Describes the proposed framework in detail, including the IoT dataset used, sensor specifications, data preprocessing and normalization steps, feature selection, and implementation of various regression algorithms. It also presents the experimental setup and performance evaluation metrics.

Chapter 5 – Experimental Results and Discussion:

Presents the experimental findings and comparative performance of all regression models using error metrics and visual analysis. The discussion highlights which models achieved the best predictive accuracy and explains the underlying reasons based on data characteristics.

Chapter 6 – Conclusion and Future Work:

Summarizes the key findings of the research, outlines its major contributions to smart aquaculture, and provides recommendations for future studies and potential real-world implementations of IoT-driven predictive systems

Chapter 2 Overview

2.1 Introduction

Artificial Intelligence (AI) represents a broad and evolving scientific discipline focused on developing systems capable of performing tasks that typically require human intelligence. These tasks include learning from experience, recognizing patterns, understanding natural language, and making autonomous decisions. The recent growth of AI technologies has been fueled by several major subfields **Machine Learning (ML)**, **Deep Learning (DL)**, and **Explainable AI (XAI)** each of which contributes uniquely to advancing intelligent data analysis and automated decision-making.

The following subsections present an overview of these key AI concepts and their significance in contemporary data-driven applications.

2.2 Machine Learning (ML)

Machine Learning is a fundamental branch of AI centered on developing algorithms that learn patterns from historical data. Instead of being explicitly programmed for every task, ML models identify relationships within datasets and use them to make predictions, classifications, or recommendations.
ML is widely applied in fields that rely on structured or semi-structured data, offering capabilities such as:

- Predicting future outcomes,
- Detecting anomalies,
- Grouping similar observations,
- Supporting evidence-based decision-making.

ML techniques provide the foundation for most modern predictive systems, especially those that depend on real-time or continuous data streams.

2.3 Deep Learning (DL)

Deep Learning is a specialized subset of ML that uses artificial neural networks with multiple layers to extract abstract and complex features from data. Unlike traditional ML methods, DL models automatically learn high-level representations, making them highly effective for processing large-scale, nonlinear, or unstructured datasets such as images, audio signals, sensor readings, and natural language.

DL has achieved breakthroughs in fields like computer vision, speech recognition, natural language processing, and time-series prediction tasks that require capturing subtle patterns that conventional algorithms may miss.

2.4 Relevance of AI Concepts to Fishpond Water Quality Prediction

The concepts discussed above ML, DL, LLMs, and XAI form the technological foundation supporting modern environmental monitoring and smart aquaculture systems. In the context of **fishpond management**, AI contributes in several meaningful ways:

- **Machine Learning** enables models to predict key water quality parameters (such as pH, dissolved oxygen, temperature, and turbidity) based on historical sensor readings.
- **Deep Learning** can capture nonlinear relationships within complex aquatic environments, improving the accuracy of water quality forecasting.

By integrating these AI techniques, modern aquaculture systems can transition from reactive water management to **proactive, predictive, and data-driven practices**,

ultimately improving productivity, reducing fish mortality, and optimizing pond conditions.

2.5 Some of Models (Regression Models)

1. Decision Tree Regression (DT) A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. As you can see from the diagram below, a decision tree starts with a root node, which does not have any incoming branches. The outgoing branches from the root node then feed into the internal nodes, also known as decision nodes. Based on the available features, both node types conduct evaluations to form homogenous subsets, which are denoted by leaf nodes, or terminal nodes. The leaf nodes represent all the possible outcomes within the dataset. [17]

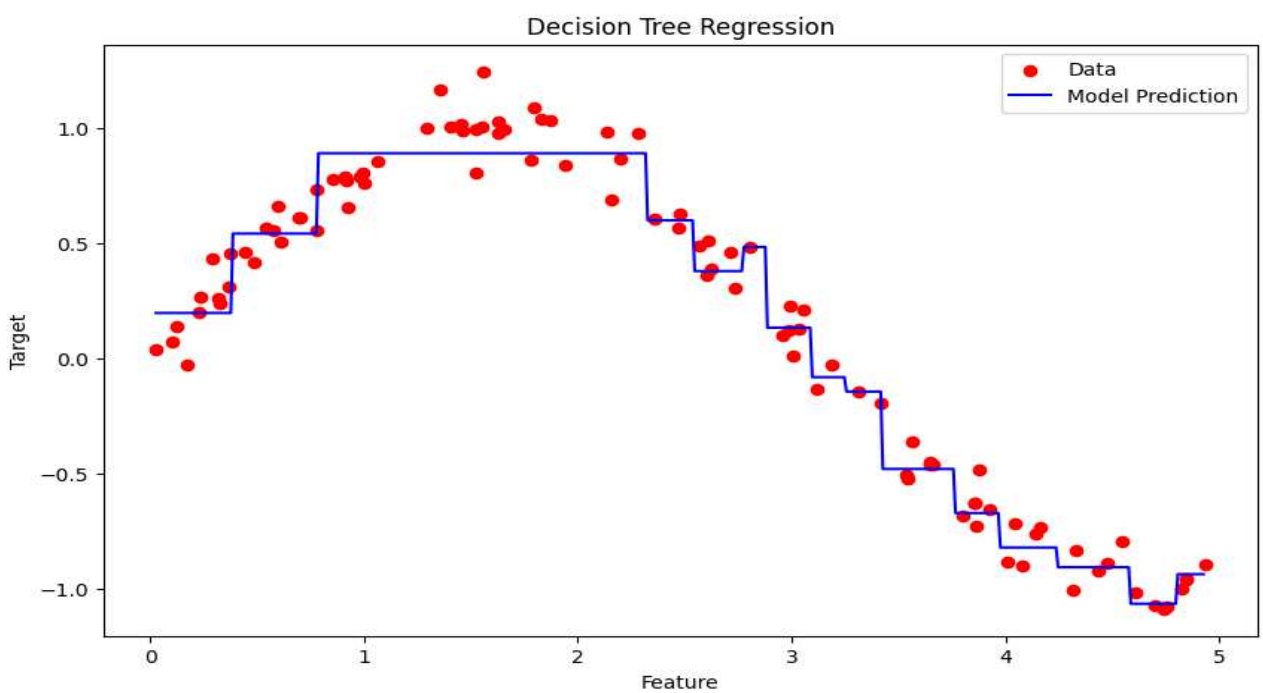

Figure 1 Decision Tree Regression (DT)

Application :

Robotics: RL is used to automate tasks in structured environments such as manufacturing, where robots learn to optimize movements and improve efficiency.

Games: Advanced RL algorithms have been used to develop strategies for complex games like chess, Go and video games, outperforming human players in many instances.

Industrial Control: RL helps in real-time adjustments and optimization of industrial operations, such as refining processes in the oil and gas industry.

Personalized Training Systems: RL enables the customization of instructional content based on an individual's learning patterns, improving engagement and effectiveness.

Advantages :

- Solves complex sequential decision problems where other approaches fail.
- Learns from real-time interaction, enabling adaptation to changing environments.
- Does not require labeled data, unlike supervised learning.
- Can innovate by discovering new strategies beyond human intuition.
  - Handles uncertainty and stochastic environments effectively.

Disadvantages :

- Computationally intensive, requiring large amounts of data and processing power.
- Reward function design is critical; poor design leads to unintended behaviors.
- Not suitable for simple problems where traditional methods are more efficient.
- Challenging to debug and interpret, making it hard to explain decisions.
- Exploration-exploitation trade-off requires careful balancing to optimize learning.

2. Random Forest Regressor (RF) A random forest is an ensemble learning method that combines the predictions from multiple decision trees to produce a more accurate and stable prediction. It can be used for both classification and regression tasks. In a regression task, we can use the Random Forest Regression technique for predicting numerical values. It predicts continuous values by averaging the results of multiple decision trees.[18]

Haw it Works Random Forest Regression works by creating multiple of decision trees each trained on a random subset of the data. The process begins with Bootstrap sampling where random rows of data are selected with replacement to form different training datasets for each tree. After this we do feature sampling where only a random subset of features is used to build each tree ensuring diversity in the models.

After the trees are trained each tree make a prediction and the final prediction for regression tasks is the average of all the individual tree predictions and this process is called as Aggregation**.**

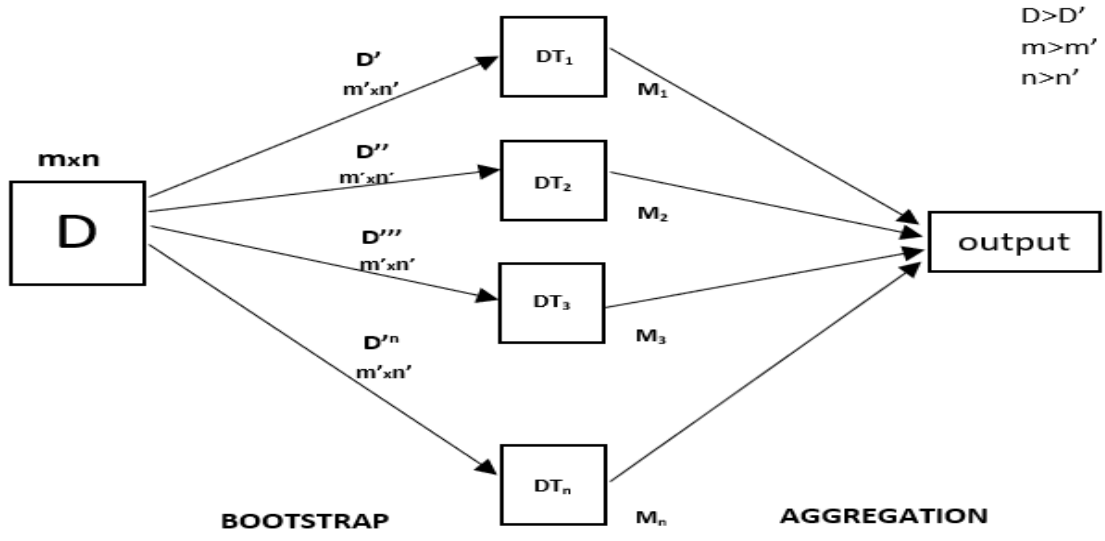

Figure 2 Random Forest Regressor (RF)

Applications of Random Forest Regression The Random forest regression has a wide range of real-world problems including:

- Predicting continuous numerical values: Predicting house prices, stock prices or customer lifetime value.
- Identifying risk factors: Detecting risk factors for diseases, financial crises or other negative events.
- Handling high-dimensional data: Analyzing datasets with a large number of input features.
- Capturing complex relationships: Modeling complex relationships between input features and the target variable.

Advantages of Random Forest Regression

- Handles Non-Linearity: It can capture complex, non-linear relationships in the data that other models might miss.
- Reduces Overfitting: By combining multiple decision trees and averaging predictions it reduces the risk of overfitting compared to a single decision tree.
- Robust to Outliers: Random Forest is less sensitive to outliers as it aggregates the predictions from multiple trees.
- Works Well with Large Datasets: It can efficiently handle large datasets and high-dimensional data without a significant loss in performance.
- Handles Missing Data: Random Forest can handle missing values by using surrogate splits and maintaining high accuracy even with incomplete data.
- No Need for Feature Scaling: Unlike many other algorithms Random Forest does not require normalization or scaling of the data.

Disadvantages of Random Forest Regression

- Complexity: It can be computationally expensive and slow to train especially with a large number of trees and high-dimensional data. Due to this it may not be suitable for real-time predictions especially with a large number of trees.
- Less Interpretability: Since it uses many trees it can be harder to interpret compared to simpler models like linear regression or decision trees.
- Memory Intensive: Storing multiple decision trees for large datasets require significant memory resources.
- Overfitting on Noisy Data: While Random Forest reduces overfitting, it can still overfit if the data is highly noisy especially with a large number of trees.
- Sensitive to Imbalanced Data: It may perform poorly if the dataset is highly imbalanced like one class is significantly more frequent than another.

3. Support Vector Regression (SVR) Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It tries to find the best boundary known as hyperplane that separates different classes in the data. It is useful when you want to do binary classification like spam vs. not spam or cat vs. dog. [19]

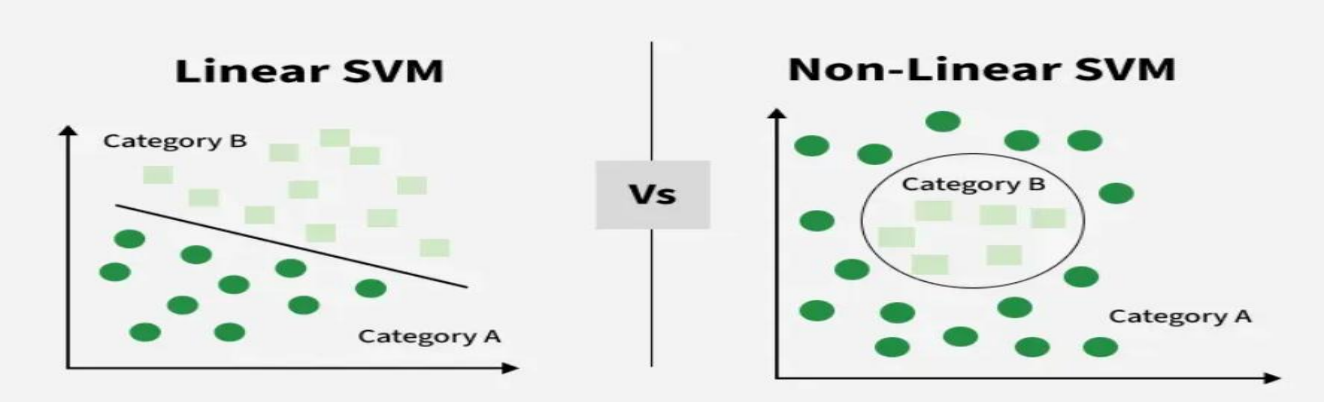

Figure 3 Support Vector Regression (SVR)

The main goal of SVM is to maximize the margin between the two classes. The larger the margin the better the model performs on new and unseen data.

How does Support Vector Machine Algorithm Work?

The key idea behind the SVM algorithm is to find the hyperplane that best separates two classes by maximizing the margin between them. This margin is the distance from the hyperplane to the nearest data points (support vectors) on each side.

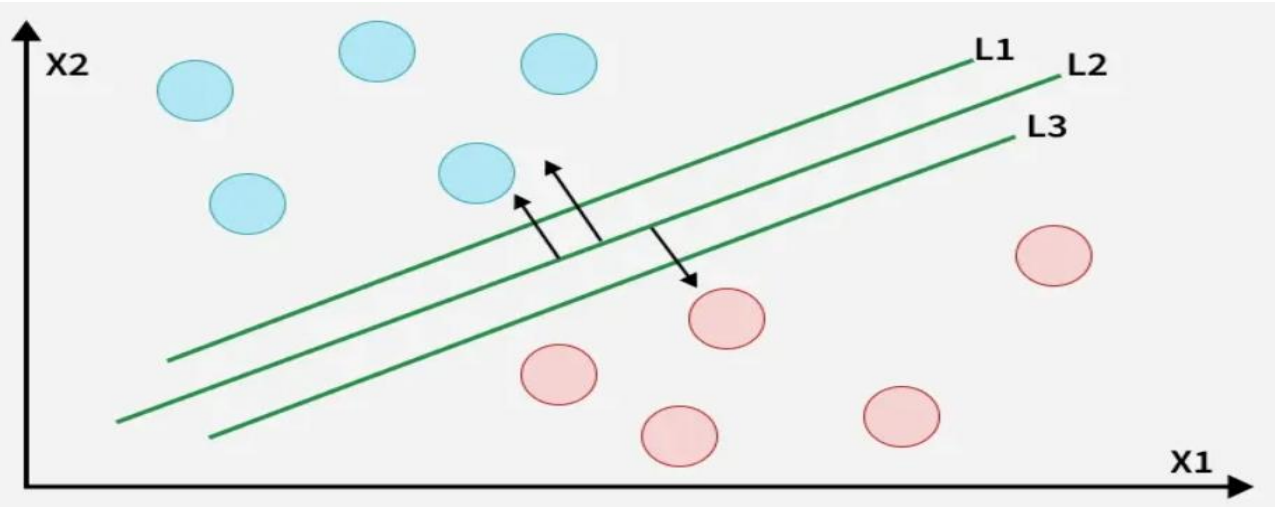

Figure 4 Support Vector Machine (SVM)

Advantages of Support Vector Machine (SVM)

- High-Dimensional Performance: SVM excels in high-dimensional spaces, making it suitable for image classification and gene expression analysis.
- Nonlinear Capability: Utilizing kernel functions like RBF and polynomial SVM effectively handles nonlinear relationships.
- Outlier Resilience: The soft margin feature allows SVM to ignore outliers, enhancing robustness in spam detection and anomaly detection.

- Binary and Multiclass Support: SVM is effective for both binary classification and multiclass classification suitable for applications in text classification.
- Memory Efficiency: It focuses on supporting vectors making it memory efficient compared to other algorithms.

Disadvantages of Support Vector Machine (SVM)

- Slow Training: SVM can be slow for large datasets, affecting performance in SVM in data mining tasks.
- Parameter Tuning Difficulty: Selecting the right kernel and adjusting parameters like C requires careful tuning, impacting SVM algorithms.
- Noise Sensitivity: SVM struggles with noisy datasets and overlapping classes, limiting effectiveness in real-world scenarios.
- Limited Interpretability: The complexity of the hyperplane in higher dimensions makes SVM less interpretable than other models.
- Feature Scaling Sensitivity: Proper feature scaling is essential, otherwise SVM models may perform poorly.

4. Lasso Regression (Least Absolute Shrinkage and Selection Operator)

Lasso Regression is a regression method based on Least Absolute Shrinkage and Selection Operator and is used in regression analysis for variable selection and regularization. It helps remove irrelevant data features and prevents overfitting. This allows features with weak influence to be clearly identified as the coefficients of less important variables are shrunk toward zero.[20]

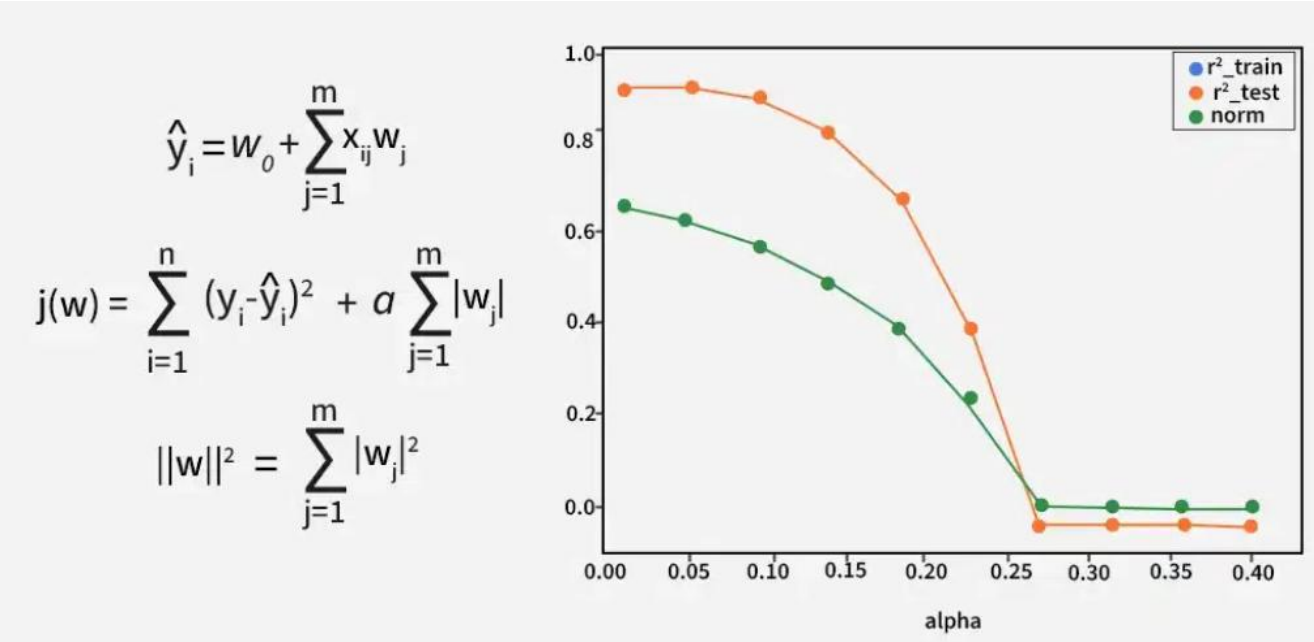

Figure 5 Lasso Regression (LR)

When to use Lasso Regression ?

Lasso Regression is useful in the following situations:

- Feature Selection: It automatically selects most important features by reducing the coefficients of less significant features to zero.
- Collinearity: When there is multicollinearity, it can help us by reducing the coefficients of correlated variables and selecting only one of them.
- Regularization: It helps prevent overfitting by penalizing large coefficients which is useful when the number of predictors is large.
- Interpretability: Compared to traditional linear regression models that have all features lasso regression generates a model with fewer non-zero coefficients making model simpler to understand.

Advantages of Lasso Regression

- Feature Selection: It removes the need to manually select most important features hence the developed regression model becomes simpler and more explainable.
- Regularization: It constrains large coefficients so a less biased model is generated which is robust and general in its predictions.
- Interpretability: This creates another models helps in making them simpler to understand and explain which is important in fields like healthcare and finance.
- Handles Large Feature Spaces: It is effective in handling high-dimensional data such as images and videos.

Disadvantages:

- Selection Bias: Lasso may randomly select one variable from a group of highly correlated variables which leads to a biased model.
- Sensitive to Scale: It is sensitive to features with different scales as they can impact the regularization and affect model's accuracy.
- Impact of Outliers: It can be easily affected by the outliers in the given data which results to overfitting of the coefficients.
- Model Instability: It can be unstable when there are many correlated variables which causes it to select different features with small changes in the data.
- Tuning Parameter Selection: Analyzing different λ (alpha) values may be problematic but can be solved by cross-validation.

5. Ridge Regression

Ridge regression, also known as L2 regularization, is a technique used in linear regression to address the problem of multicollinearity among predictor variables. Multicollinearity occurs when independent variables in a regression model are highly correlated, which can lead to unreliable and unstable estimates of regression coefficients. [21]

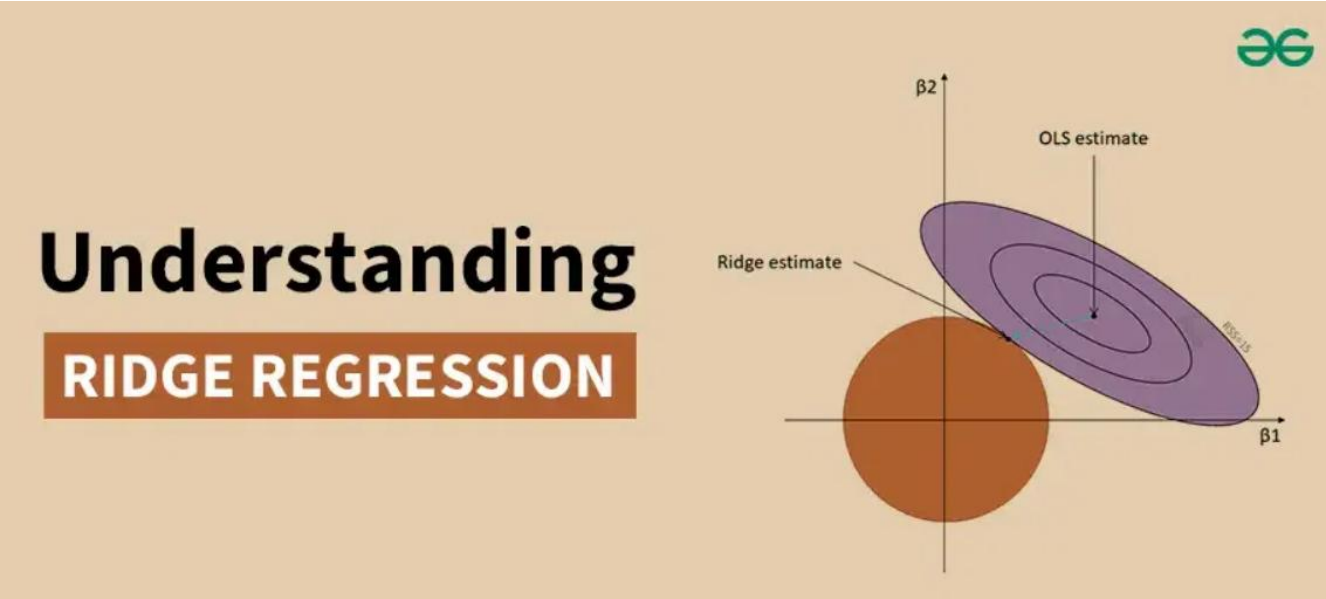

Figure 6 Ridge Regression

Applications of Ridge Regression :

- Forecasting Economic Indicators: Ridge regression helps predict economic factors like GDP, inflation, and unemployment by managing multicollinearity between predictors like interest rates and consumer spending, leading to more accurate forecasts.

- Medical Diagnosis: In healthcare, it aids in building diagnostic models by controlling multicollinearity among biomarkers, improving disease diagnosis and prognosis.
- Sales Prediction: In marketing, ridge regression forecasts sales based on factors like advertisement costs and promotions, handling correlations between these variables for better sales planning.
- Climate Modeling: Ridge regression improves climate models by eliminating interference between variables like temperature and precipitation, ensuring more accurate predictions.
- Risk Management: In credit scoring and financial risk analysis, ridge regression evaluates creditworthiness by addressing multicollinearity among financial ratios, enhancing accuracy in risk management.

Advantages of Ridge Regression

- Stability: Ridge regression provides more stable estimates in the presence of multicollinearity.
- Bias-Variance Tradeoff: By introducing bias, ridge regression reduces the variance of the estimates, leading to lower MSE.
- Interpretability: Unlike principal component regression, ridge regression retains the original predictors, making the results easier to interpret.

Disadvantages of Ridge Regression

- Bias Introduction: The introduction of bias can lead to underestimation of the true effects of the predictors.

- Parameter Selection: Choosing the optimal ridge parameter k can be challenging and computationally intensive.
- Not Suitable for Variable Selection: Ridge regression does not perform variable selection, meaning all predictors remain in the model, even those with negligible effects.

6. XGBoost Regression (Extreme Gradient Boosting)

Traditional machine learning models like decision trees and random forests are easy to interpret but often struggle with accuracy on complex datasets. XGBoost short form for extreme Gradient Boosting is an advanced machine learning algorithm designed for efficiency, speed and high performance. [21]

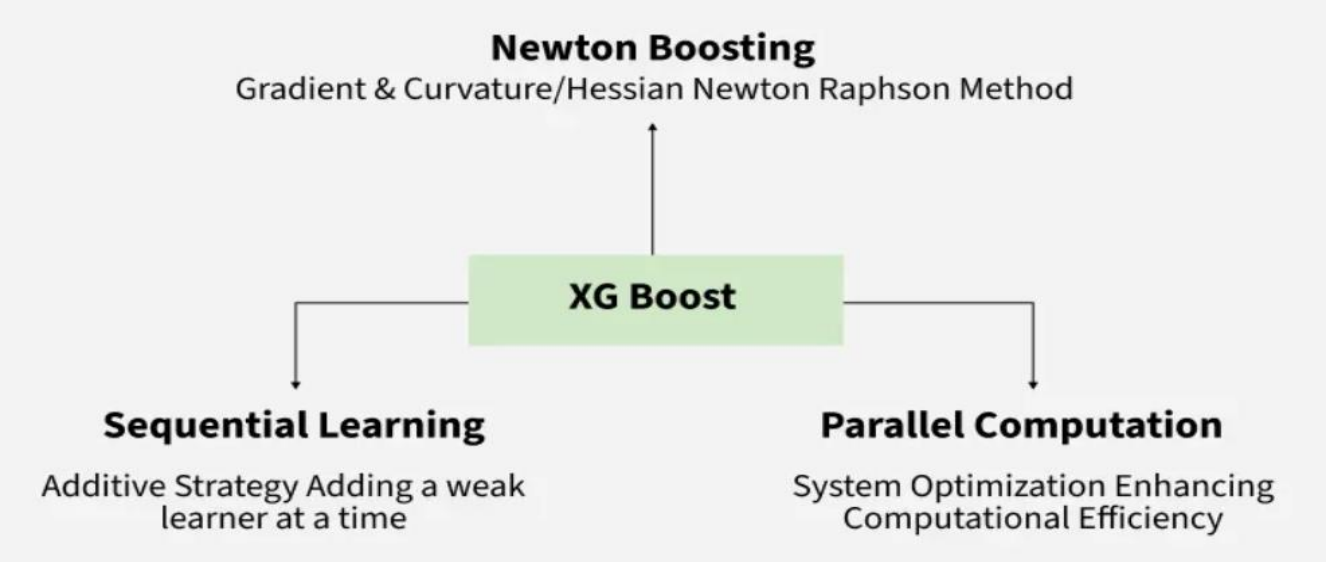

Figure 7 XGBoost Regression

To efficiently handle large datasets, XGBoost uses an approximate method to find optimal splits.

- **Weighted quantiles:** Quickly estimate the best split without checking every possibility.
- **Efficiency:** Reduces computational overhead while maintaining accuracy.
- **Benefit:** Ideal for large datasets where full evaluation is costly.

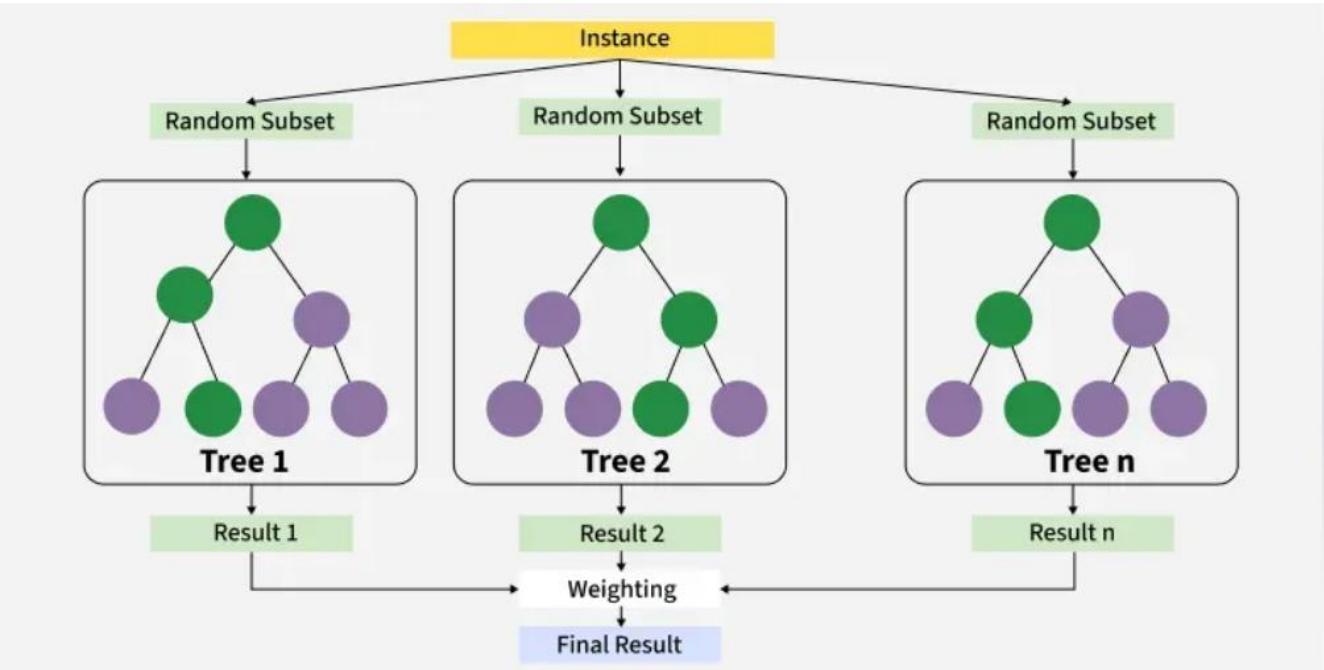

Figure 8 XGBoost Regression

How XGBoost Works?

It builds decision trees sequentially with each tree attempting to correct the mistakes made by the previous one. The process can be broken down as follows:

1. **Start with a base learner**: The first model decision tree is trained on the data. In regression tasks this base model simply predicts the average of the target variable.

2. **Calculate the errors**: After training the first tree the errors between the predicted and actual values are calculated.
3. **Train the next tree**: The next tree is trained on the errors of the previous tree. This step attempts to correct the errors made by the first tree.
4. **Repeat the process**: This process continues with each new tree trying to correct the errors of the previous trees until a stopping criterion is met.
5. **Combine the predictions**: The final prediction is the sum of the predictions from all the trees.

Advantages of XGBoost :

- Scalable for large datasets with millions of records.
- Supports parallel processing and GPU acceleration.
- Offers customizable parameters and regularization for fine-tuning.
- Includes feature importance analysis for better insights.
- Available across multiple programming languages and widely used by data scientists.

Disadvantages of XGBoost

- Computationally intensive; may not be suitable for resource-limited systems.
- Sensitive to noise and outliers; careful preprocessing required.
- Can overfit, especially on small datasets or with too many trees.
- Limited interpretability compared to simpler models, which can be a concern in fields like healthcare or finance.t^{th}

Figure 9 Tree Algorithm

Chapter 3 Related Work

3.1 Introduction:

Aquaculture faces significant challenges in maintaining optimal conditions for fish growth, particularly with respect to water resource availability and quality. Water quality directly impacts fish health and productivity, yet traditional manual monitoring methods are labor-intensive and often inadequate for early detection of problems. A persistent issue in conventional fish pond management has been the lack of comprehensive data to guide decisions on breeding, feeding, and pond maintenance. In this context, the advent of Internet of Things (IOT) technology offers new opportunities for continuous, automated monitoring in aquaculture environments. By deploying networks of sensors and connected devices, IOT systems can provide high-frequency, real-time data on water conditions, enabling timely interventions and data-informed management. One persistent gap has been the absence of comprehensive datasets and continuous data streams to guide fish farmers in data-driven decision-making for breeding and pond maintenance. In this context, the advent of IOT technology offers unprecedented opportunities for continuous, automated monitoring in aquaculture environments, enabling timely detection of water quality issues and collection of rich real-time data .

Aquaponics, a sustainable farming system, exemplifies the integration of aquaculture with hydroponics (soilless plant cultivation). In an aquaponic fish pond, fish waste nutrients are taken up by plants, which in turn help filter and reuse the water for fish. This symbiotic setup significantly reduces the need for frequent water replacement, addressing one challenge of conventional fish ponds. To further optimize such systems, IOT-based monitoring can be implemented for real-time water quality management. [22],[23]

Anumber of studies in recent years have explored IoT deployments in fish farming alongside machine learning algorithms to predict water quality variables or other aquaculture outcomes.

3.2 Literature Review:

Study 1

MS-CAGRU for Water Quality Prediction

Sundararajan S. C. M. et al. (2025) developed an IoT-based water quality prediction model for an aquaponic fish pond by using a multiscale feature fusion approach with deep learning. In their study, an IoT sensor network collected water parameters which were then used to train a hybrid deep learning model combining a convolutional autoencoder (CAE) for feature extraction and a gated recurrent unit (GRU) for sequence learning. The proposed model (termed MS-CAGRU) effectively captured multi-scale patterns in the water quality data. The authors applied different ML models: random forest (RF), support vector machine (SVM), decision tree (DT), K-nearest neighbor (KNN), and logistic regression (LR) to classify different types of fish using water quality parameters: pH, temperature, and turbidity. The dataset was collected from AquaBot, It attained approximately 94% classification accuracy with an F1-score of 0.94, outperforming baseline recurrent neural networks (standard RNN and plain GRU) by 2–5%. The study noted that regularization techniques further improved the model's generalization. This work demonstrates that fusing convolutional feature extraction with recurrent prediction can yield high predictive accuracy for aquaculture water quality, providing a foundation for intelligent pond management. [1]

Study 2

IOT & Advanced ML Algorithms (Tilapia Ponds, 2025)

Hridoy, M. A. A. M. et al. (2025) applied IoT monitoring and advanced machine learning algorithms to improve water quality prediction in tilapia aquaculture ponds. The researchers deployed a custom IoT system to continuously record six months of sensor data (including temperature, dissolved oxygen, pH, and turbidity) from several tilapia ponds. A variety of predictive models were evaluated on these time-series data, ranging from decision tree ensembles to deep neural networks. Notably, the authors leveraged ensemble techniques such as **bagging and boosting** to enhance predictive performance. By analyzing temporal patterns and combining diverse learners, the models could forecast key water quality parameters with high precision. The best results were obtained with a **custom ensemble model** that integrated decision tree methods with boosting algorithms, achieving nearly **99% accuracy** in predicting trends like dissolved oxygen fluctuations. This far exceeded the accuracy of individual models (which were in the 88–93% range), underscoring the benefit of ensemble learning. The study concluded that integrating IoT-sensed data with advanced ML algorithms enables near real-time water quality forecasting with minimal error, which is crucial for proactive pond management.[2]

Study 3

IOT + Random Forest with Quantum Optimization (Tropical Pond, 2025)

Baena-Navarro, R. et al. (2025) described an intelligent water quality monitoring framework that integrates IoT sensing with machine learning for sustainable aquaculture management. They implemented an IoT system in a **tropical fish pond** to continuously collect environmental data (temperature, DO, pH, turbidity, etc.), and used these data to train predictive models. A *random forest regression* approach was employed to forecast water quality variables, and the novelty of this study lay in applying a **quantum-inspired optimization algorithm (QAOA)** to improve the model's training efficiency and accuracy. The Random Forest model optimized with QAOA achieved a *near-perfect fit* for dissolved oxygen predictions, with a coefficient of determination $\mathbf{R^2 \approx 0.999}$ and root mean square error (RMSE) on the order of 0.1 mg/L. Such accuracy indicates that the model's predictions almost exactly matched the observed DO levels. The integrated system allowed rapid, on-site adjustments (e.g. activating aerators) to maintain fish survival rates above 90%. The best-performing approach was the QAOA-enhanced Random Forest, demonstrating that coupling IoT data streams with cutting-edge optimization techniques can yield extremely accurate predictive models for aquaculture.[3]

Study 4

Hybrid EEMD–MLR–LSTM Model (Salmon Farm, 2023)

Eze, E. et al. (2023) explored a hybrid deep learning strategy for multivariate water quality forecasting in an offshore salmon farm. Their study introduced a composite model that combined Ensemble Empirical Mode Decomposition (EEMD) for signal preprocessing, multivariate linear regression (MLR) for initial approximation, and a deep Long Short-Term Memory (LSTM) network for sequence learning. Using time-series sensor data from a Scottish salmon aquaculture site, the hybrid model was trained to predict indicators of harmful algal blooms (e.g., chlorophyll-a levels) and other water quality metrics. The EEMD component decomposed complex time-series into simpler sub-series, which improved the LSTM's ability to learn patterns at multiple scales. The proposed EEMD–MLR–LSTM model outperformed several baseline models (including a standalone LSTM and a stacked autoencoder-LSTM) in predictive accuracy. Notably, it achieved a very low RMSE of around 0.05 for chlorophyll-a concentration forecasts, indicating high precision. By effectively capturing both short-term fluctuations and long-term trends in the data, this hybrid approach minimized prediction error and was identified as the best model. The results highlight the value of multi-scale feature extraction combined with deep learning to forecast complex water quality dynamics in aquaculture.[4]

Study 5

Yu, X. et al. (2025) focused on dissolved oxygen prediction in freshwater aquaculture using feature selection and attention-based recurrent neural networks. The authors first applied Pearson **correlation analysis** to identify which water quality variables were most influential for predicting DO levels. By filtering the input features to only the most relevant (e.g., selecting temperature, pH, etc. that correlate strongly with DO), they reduced noise and improved model focus. They then designed an LSTM network augmented with an **attention mechanism** (denoted LSTM-AM) to forecast dissolved oxygen several days into the future. The attention layer enabled the model to weight the most important time steps and features when making predictions. Trained on an IoT-sensed dataset from a freshwater fish farm, the enhanced model (called *Pearson-LSTM-AM*) achieved substantially better accuracy than baseline models (including standard LSTM, CNN-LSTM, and bidirectional LSTM). It attained an R^2 of **0.859** and an RMSE of **0.379** for 5-day DO predictions, representing about an 8–10% reduction in error compared to the next best model. The results demonstrate that combining statistical feature selection with an attention-enhanced LSTM can greatly improve mid-term water quality forecasting in aquaculture. The Pearson-LSTM-AM was singled out as the top-performing model, underscoring the importance of focusing on key input variables and temporal patterns.[5]

Study 6

Xu, L. et al. (2025) proposed a hybrid deep learning framework for real-time dissolved oxygen prediction in intensive aquaculture systems. Their model pipeline incorporated multiple deep learning components arranged to capture different aspects of the time-series data. First, a one-dimensional Convolutional Neural Network (CNN) was used to extract local temporal features from the sensor input sequences. Next, a self-attention mechanism was applied to these features to dynamically weight the most informative parts of the sequence. Finally, the sequence modeling was performed by a recurrent layer using a Bidirectional Simple Recurrent Unit (Bi-SRU) architecture. The Bi-SRU, a variant of standard RNNs, processed the sequence in both forward and backward directions, allowing the model to utilize past and future context. The hybrid CNN–SA–BiSRU model was evaluated on an IoT dataset from an intensive fish pond in Guangzhou, China. It achieved very high accuracy in short-term DO prediction, with a mean absolute error on the order of 0.03–0.04 mg/L and RMSE ≈ 0.047 mg/L. The coefficient of determination was $R^2 \approx 0.9765$, indicating that the model explained over 97% of the variance in DO levels. This performance was substantially better than traditional machine learning models like random forest or basic neural nets. The authors identified their CNN–SA–BiSRU hybrid as the best model tested, demonstrating the efficacy of combining convolutional feature extraction, attention, and bidirectional recurrent units for accurate real-time predictions in aquaculture.[6]

Study 7

Support Vector Machine vs Neural Networks (Industrial Fish Farm, 2022):

Li, T. et al. (2022) conducted a comparative study of machine learning approaches to predict multiple water quality parameters in an **industrial fish farming system**. They compiled a dataset comprising both literature-derived and experimentally measured water quality data (including DO, pH, ammonia, nitrate, nitrite levels) for intensive aquaculture tanks. Four different modeling techniques were implemented and compared: a **backpropagation neural network (BPNN)**, a **radial basis function neural network (RBFNN)**, a standard Support Vector Machine, and a Least-Squares SVM (an optimized variant of SVM). The performance was evaluated across all five water quality indicators. The study found that the **Support Vector Machine (SVM)** consistently delivered the most accurate and stable predictions for each parameter. Specifically, the SVM model achieved about **99% accuracy** in simulating water quality metrics, with near-perfect correlation coefficients (~0.99) between predicted and observed values. In contrast, the neural network models (BPNN and RBFNN) showed more variable and lower accuracy, likely due to their susceptibility to local minima and parameter tuning challenges in the given data regime. The authors concluded that SVM, with appropriate kernel selection and tuning, was the superior model for this task, and they recommended its use for reliable water quality prediction in similar industrial aquaculture scenarios.[7]

Study 8

AquaSense IoT Framework + Dilated CNN–BiLSTM (Smart Aquaculture, 2024):

Iniyan Arasu, M. et al. (2024) introduced an IoT-driven framework called "AQUASENSE" for smart aquaculture pond management, combined with a novel deep learning architecture for water quality classification. The AQUASENSE system deployed a network of autonomous **IoT sensors** in aquaculture ponds to continuously measure parameters such as temperature, pH, dissolved oxygen, and salinity. Data from these sensors were transmitted to a cloud platform for processing. The authors developed a custom deep learning model using a **dilated CNN–BiLSTM** architecture to analyze the streaming data. The dilated CNN component captured patterns at multiple temporal scales by using dilated convolutional filters, while the BiLSTM (bidirectional LSTM) component analyzed the sequence forward and backward to improve temporal context understanding. The model's task was to classify the water quality status (e.g., optimal, suboptimal) and raise alerts for anomalies. In evaluations, the AQUASENSE framework achieved an overall classification accuracy of **96.98%**, substantially outperforming earlier IoT-based aquaculture monitoring systems (which achieved ~88–93% accuracy). In particular, the dilated CNN–BiLSTM model exceeded the accuracy of prior frameworks like ISAS and AquaStat by 3–8 percentage points. This study's best model was the integrated AQUASENSE deep learning classifier, demonstrating how tailored neural network designs can leverage IoT data for highly accurate water quality assessment and proactive management in aquaculture[8] .

Study 9

AquaBot IOT System + Ensemble ML (Smart Fish Pond, 2024):

Hemal, M. M. et al. (2024) developed an integrated smart pond management system, named "AquaBot," which combines IoT-based monitoring with ensemble machine learning to support fish farming decisions. AquaBot features an **IoT-enabled robotic platform** that traverses fish ponds to measure water parameters at various locations (including pH, temperature, and turbidity) in real time. The collected data are sent to a central server that runs multiple machine learning algorithms to analyze water quality and recommend suitable fish species or interventions for the current water conditions. The authors trained several classifiers on an annotated dataset of water parameters and recommended fish species, including Random Forest, SVM, decision tree, k-Nearest Neighbors, logistic regression, and a **custom stacked ensemble** that combines these models. Among these, the ensemble model (which aggregated the predictions of the individual classifiers) delivered the highest performance, with roughly **94% accuracy** as well as high precision and recall in multi-class classification of fish species suitability. This was notably better than any single algorithm (individual models ranged from ~88% to 93% accuracy). The results highlight that the ensemble approach in AquaBot effectively captured complementary strengths of different algorithms. The study's best-performing model – the stacked ensemble – provided reliable recommendations, demonstrating the potential of IoT data fused with ML to automate decision-making in aquaculture (such as suggesting species that can thrive given the real-time water quality).[9]

Study 10

IoT Sensors + Decision Tree Classifier (Fish Behavior Analysis, 2023):

Patro, K. S. K. et al. (2023) investigated the use of IoT sensors and machine learning for monitoring fish behavior, specifically aiming to detect stress or health issues in ornamental fish via environmental changes. In their experimental setup, goldfish in a controlled tank were continuously observed under varying conditions of water temperature and dissolved oxygen, while an IoT system recorded these environmental parameters. The fish behaviors (e.g., normal swimming, erratic movement, surface gasping) were classified and time-synchronized with the sensor readings. The researchers then trained four different classifiers – a decision tree, Naïve Bayes, k-nearest neighbor (k-NN), and Linear Discriminant Analysis (LDA) – to predict the fish's behavioral state (stress vs. normal) from the water quality indicators. After evaluating the models via cross-validation, the decision tree classifier emerged as the most accurate, achieving about 86.2% accuracy (approximately 13.8% prediction error). It outperformed k-NN and Naïve Bayes, which had error rates around 28–30%. The decision tree's interpretable rules also identified specific environmental thresholds associated with stress behaviors: for instance, the fish exhibited erratic swimming or gasping when water temperature exceeded ~40.5 °C or DO dropped below ~6.6 mg/L. This study's best model, the decision tree, thus enabled automated detection of fish stress with high accuracy, illustrating how IoT sensor data can be used not only for water quality management but also for real-time animal welfare monitoring in aquaculture.[10]

Study 11

Dilated CNN + BiGRU with Attention (Indoor Mariculture, 2024)

Arepalli, P. G. & Naik, K. J. (2024) reported on a water quality prediction method for indoor recirculating aquaculture that employs a dual attention-based deep learning model. They formulated a multi-step dissolved oxygen forecasting model by integrating Variational Mode Decomposition (VMD), a CNN-BiGRU network, and attention mechanisms. First, VMD was applied to the raw sensor time-series to decompose it into components capturing different frequency patterns (e.g., daily cycles vs. rapid fluctuations). Each decomposed sequence was then fed into a CNN–BiGRU model: a one-dimensional CNN extracted features, and a bidirectional GRU processed the temporal sequence both forward and backward. Moreover, the model included *dual attention mechanisms*: one to focus on important VMD components and another temporal attention within the BiGRU to highlight critical time steps. Testing this approach on data from a lab-scale indoor mariculture tank, the authors found that the hybrid VMD-CNN-BiGRU-Attention model met practical accuracy requirements for DO prediction. It consistently produced lower error than conventional single-network models like LSTM or plain GRU, especially for multi-hour forecasts. Qualitatively, the predicted DO trends from the model closely matched the actual measurements, indicating that the method can reliably serve real-time monitoring needs. By leveraging signal decomposition and attention, this model effectively captured both global trends and local fluctuations, making it the top performer among those evaluated.[11]

Study 12

Dual-Attention ResNet + LSTM (Offshore Aquaculture, 2025): Kunapinun, A. et al. (2025) sought to improve predictions of biomass growth and water quality in integrated multi-trophic aquaculture systems by combining data-driven models with physical knowledge. They developed a hybrid modeling approach that integrates a bidirectional LSTM (Bi-LSTM) neural network with physics-based constraints derived from aquaculture system dynamics. The context was a multi-trophic system (where fish, plants, and possibly other species coexist), and the model was tasked with forecasting variables like biomass yield and key water parameters. By incorporating physical equations or known system invariants into the model's training (often termed a *physics-informed neural network* approach), the Bi-LSTM was guided to produce realistic predictions that honor fundamental constraints (e.g., mass balance or known growth kinetics). Training and testing on data from an experimental aquaculture setup, the physics-informed Bi-LSTM demonstrated superior long-term prediction stability compared to purely data-driven models. In one case, it provided accurate biomass forecasts over extended periods where ordinary Bi-LSTM models drifted. Overall, the hybrid Bi-LSTM+physics model yielded improved prediction accuracy for both biomass and water quality metrics, especially for longer-range forecasts where purely statistical models tend to degrade. This approach underscores the benefit of marrying domain knowledge with machine learning – the best model leveraged both sources, achieving reliable performance and offering insights that are consistent with aquaculture science.[12]

Study 13

Hybrid Multi-Scale Feature Model (DO Prediction, 2018):**Li, C. et al. (2018)** presented a hybrid modeling approach for predicting dissolved oxygen in aquaculture ponds by exploiting **multi-scale features** in the time-series data. They observed that DO levels in ponds exhibit both short-term fluctuations (due to factors like aerator cycles or feeding) and long-term trends (due to seasonal changes or biomass growth). To capture this, the authors decomposed the original DO time-series into multiple frequency components using techniques such as wavelet transform or empirical mode decomposition. Each resulting component highlighted patterns at a particular timescale. These features were then used to train a deep neural network (in their implementation, an LSTM or a CNN-LSTM hybrid) to forecast future DO levels. The multi-scale hybrid model was tested on historical pond data and compared against conventional single-scale models that used the raw time-series. The results showed a clear advantage for the multi-scale approach: the hybrid model achieved significantly higher precision, reducing RMSE by approximately **20–30%** relative to a standard LSTM that did not utilize multi-scale inputs. This improvement indicates that the model more effectively captured the complex dynamics of DO by integrating information across fast and slow oscillations. Thus, the best-performing model was the one that combined **multi-scale feature extraction** with deep learning, demonstrating superior accuracy and stability for DO prediction in aquaculture.[13]

Study 14

Deep Belief Network (Recirculating Aquaculture, 2020): Ren, Q. et al. (2020) investigated the use of deep learning for dissolved oxygen prediction in recirculating aquaculture systems (RAS), focusing specifically on Deep Belief Networks (DBNs). RAS environments are highly controlled and dynamic, making DO forecasting vital for automated aeration control. The authors developed a DBN-based regression model to map various inputs (such as water temperature, feed rate, aeration settings, etc.) to future DO concentrations. The DBN, being a stacked neural network of Restricted Boltzmann Machine layers, learns a hierarchical representation of the data through unsupervised pre-training and fine-tuning. In comparisons with more traditional models (including simple neural networks and mechanistic water quality models), the DBN achieved faster convergence during training and higher prediction accuracy on the test data. For instance, in trials on data from marine fish RAS facilities, the DBN predictor yielded a high coefficient of determination (R^2) and low prediction error, indicating it could closely track observed DO fluctuations. The authors noted that the DBN's deep layered structure allowed it to capture complex nonlinear relationships between inputs and DO that simpler models missed. This led to more stable and precise forecasts, especially for short-term DO variations that are critical for timely aeration adjustments. Therefore, the DBN-based model was selected as the best-performing approach in this study, highlighting the promise of deep learning methods for enhancing water quality management in RAS aquaculture.[14]

Study 15

Jamshidzadeh, Z. et al. (2024) proposed a hybrid deep learning model to predict water quality parameters in a coastal aquifer relevant to aquaculture, introducing a novel metaheuristic optimization in the training process. In this study, two hybrid model architectures were formulated: one combining a unidirectional LSTM network with Support Vector Machine regression (LSTM-SVM), and another combining a Bidirectional LSTM with SVM (BiLSTM-SVM). To fine-tune the hyperparameters of these models and select optimal input features, the authors employed the "Ali Baba and the Forty Thieves" (ABFT) algorithm – a recent nature-inspired optimization technique. The case study focused on predicting Total Dissolved Solids (TDS) and Electrical Conductivity (EC) in groundwater adjacent to a coastal aquaculture site in northern Iran, using daily water quality records from 2017–2020. The performance of various models was evaluated for one-day, two-day, and three-day ahead predictions. The results revealed that the BiLSTM-SVM hybrid model, after ABFT-driven optimization, consistently outperformed all other models across prediction horizons. For one-day ahead forecasts, it achieved very low mean absolute percentage error (MAPE) — around 2% for TDS — and a high R^2 of 0.94 for EC, indicating excellent predictive accuracy Although accuracy naturally declined for longer-term predictions, the BiLSTM-SVM maintained the lowest error rates compared to alternatives like standalone LSTM, standalone SVM, or the LSTM-SVM without bidirectionality. The study concluded that the optimized BiLSTM-SVM was the top model, demonstrating how hybridizing different learning algorithms (deep learning + SVM) and using advanced optimization (ABFT) can yield highly precise water quality prediction tools for aquaculture resource management.[15]

3.4 Comparison Table :

Table 1 Comparison of previous studies

Papers	Year	Methods	Datasets	The best model	Results
Study [1]	2025	Applied Conv. Autoencoder + GRU (hybrid deep learning)	Aquaponic fish pond IOT sensors [24]	MS-CAGRU (CAE+GRU hybrid)	~94% accuracy; F1 ≈ 0.94; ~5% higher accuracy vs. baseline RNNs.
Study [2]	2025	Ensemble ML (bagging & boosting); time-series analysis	Tilapia pond IOT data (6 months)	Custom ensemble (tree-based)	~99% accuracy in DO prediction; ensemble outperformed individual models by ~5–10%.
Study [3]	2025	Random Forest regression; Quantum optimization (QAOA)	Tropical pond IoT sensors (continuous) [25]	Random Forest + QAOA optimized	$R^2 \approx 0.999$ for DO; RMSE ~0.1 mg/L (near-perfect fit); >90% survival rate maintained.
Study [4]	2023	Applied EEMD (signal decomposition) + LSTM (hybrid model) + MLR	Salmon farm sensor time-series	EEMD–MLR–LSTM hybrid	Lowest error (RMSE ~0.05); outperformed LSTM
Study [5]	2025	Applied Pearson correlation + LSTM + Attention	Freshwater pond (DO focus)	Pearson-LSTM-AM (Attn.)	$R^2 \approx 0.859$; RMSE 0.379 (best among tested)
Study [6]	2025	CNN + Self-Attention + Bi-SRU (RNN variant)	Intensive pond (Guangdong, China)	CNN–SA–BiSRU (hybrid DL)	RMSE 0.047; $R^2 \approx 0.9765$ (excellent DO prediction)
Study [7]	2022	Applied different BPNN, RBFNN, SVM, LSSVM (model comparison)	Industrial fish farm data	SVM model (regression)	~99% accuracy; correlation ~0.99 on key params
Study [8]	2024	IOT sensor network; Dilated CNN +	Smart aquaculture pond	Dilated CNN–BiLSTM	96.98% accuracy; +3–8% vs prior methods

		BiLSTM DL	(multi-sensor)		
Study [9]	2024	Applied IOT robot + ML (RF, SVM, k-NN, ensemble)	Smart pond (Bangladesh) IOT data	Custom Ensemble (stacked)	~94% accuracy /precision/recall; best recommender
Study [10]	2023	IoT env. sensors + ML classifiers (DT, KNN, etc.)	Ornamental fish tank (behavior)	Decision Tree classifier	~86% accuracy; identified stress behavior triggers
Study [11]	2024	VMD + CNN + BiGRU + Attention (hybrid deep model)	Indoor RAS dataset (lab-scale)	VMD–CNN–BiGRU–Attention	Met real-time DO forecasting needs; lower error than LSTM/GRU; attention improved multi-step accuracy.
Study [12]	2025	Dual-channel ResNet + LSTM with dual attention	Offshore farm IoT (multi-param)	Dual-Attention ResNet-LSTM	~93–94% Nash efficiency for DO/pH; +10% vs single-channel
Study [13]	2018	Multi-scale feature extraction + deep network	Aquaculture pond (DO time-series)	Hybrid multi-scale DL model	~20–30% error reduction vs single-scale LSTM
Study [14]	2020	Deep Belief Network (DBN) regression	Recirculating aquaculture (marine)	DBN predictor	High R^2 in DO forecasts; outperformed legacy models
Study [15]	2024	BiLSTM + SVM hybrid models; ABFT optimization algorithm	Coastal aquifer data (2017–2020 daily series)	Optimized BiLSTM-SVM	MAPE ~2% for TDS; R^2 ~0.94 for EC (1-day ahead); best performance across 1–3 day forecasts; outperformed LSTM or SVM alone.

Chapter 4 Proposed Approach (Methodology)

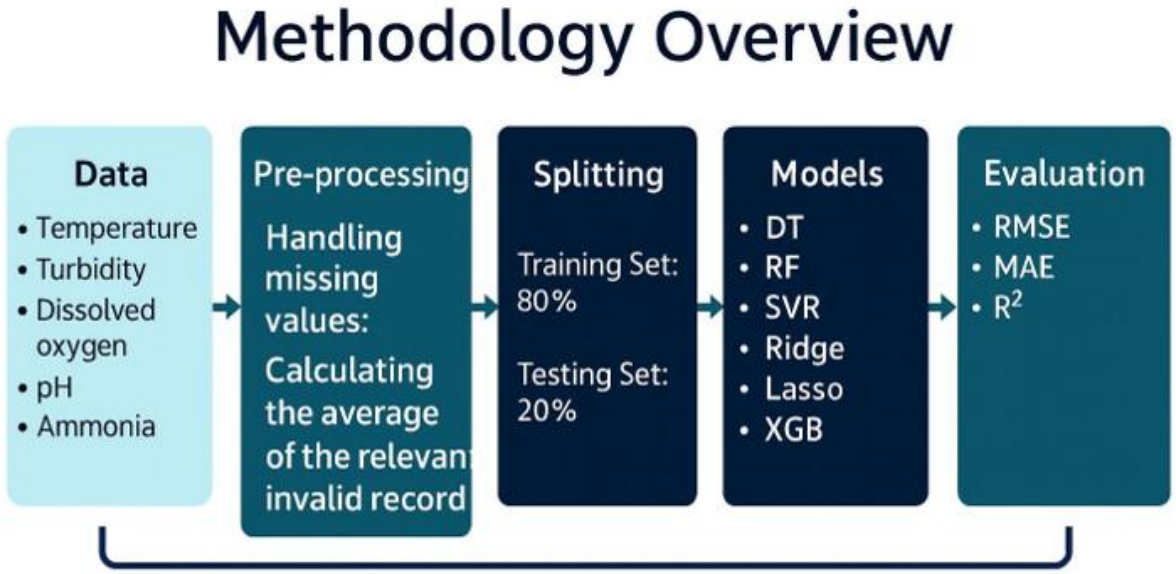

Figure 10 Methodology overview

4.1 Dataset Description

The dataset used in this research is the Fish Pond IoT Dataset [27] , which contains various water-quality measurements collected through Internet of Things (IoT) sensors installed in a smart aquaculture environment.

These sensors continuously monitor the physical and chemical properties of the pond water to ensure optimal fish growth conditions.

The main features (independent variables) included in the dataset are as follows :

- **Temperature (°C):** Represents the water temperature, a key factor influencing the metabolic rate and growth of fish.
- **Turbidity (NTU):** Indicates the clarity of water; higher turbidity levels may reduce environmental quality and light penetration.
- **Dissolved Oxygen (g/ml):** One of the most critical parameters affecting fish health, respiration, and feeding behavior.
- **pH:** Determines the acidity or alkalinity of water, directly influencing fish metabolism and enzyme activity.
- **Ammonia (g/ml):** Acts as an indicator of organic pollution in the water; excessive levels are toxic to fish.
- **Nitrate (g/ml):** Elevated nitrate concentrations can negatively impact water quality and fish development.

The **target variables (dependent variables)** in this study are:

- **Fish Length (cm)**
- **Fish Weight (g)**

Both targets were predicted using the above environmental variables through machine learning regression models to analyze and forecast fish growth patterns in IoT-enabled aquaculture systems.

4.2 Data Pre-processing

Before training the model, several data processing steps were applied to ensure data quality, reliability, and model accuracy. These steps included handling missing values, where the dataset was carefully examined for any missing or null entries. Missing values were then addressed either by calculating the average of the relevant property or by removing invalid records, as appropriate.

4.3 Splitting the dataset

The dataset was divided into two main subsets for model development and evaluation:

- **Training Set:** 80% of the data was used to train the regression models.
- **Testing Set:** 20% of the data was reserved for evaluating model performance on unseen samples.

This split was conducted using **random sampling** with a fixed random state (42) to ensure reproducibility.

4.4 Models used (Regression Models)

1. Decision Tree Regression (DT)

Decision Trees divide the data into smaller segments based on simple decision rules.
Strengths:

- Highly interpretable and easy to visualize.
- Can model nonlinear relationships effectively.

Limitations:

- Prone to overfitting, especially with noisy data.
- Performance may fluctuate due to its reliance on a single tree structure.

2. Random Forest Regression (RF)

Random Forest combines many Decision Trees into an ensemble, improving stability and predictive accuracy.
Strengths:

- Reduces overfitting by averaging multiple trees.
- Generally provides strong and consistent performance.

Limitations:

- Less interpretable than a single tree.
- Computational cost increases with a higher number of trees.

3. Support Vector Regression (SVR)

SVR uses support vector principles and kernel functions to fit a regression line within a defined margin of error.

Strengths:

- Well-suited for nonlinear relationships.
- Performs effectively in high-dimensional feature spaces.

Limitations:

- Computationally intensive, especially with large datasets.
- Requires careful tuning of kernel type and model parameters.

4. Lasso Regression

Lasso applies L1 regularization, shrinking some coefficients to zero and simplifying the model.

Strengths:

- Useful for feature selection by eliminating less important features.
- Helps prevent overfitting in high-dimensional data.

Limitations:

- Assumes linear relationships and cannot capture complex interactions.
- May remove relevant variables if the regularization is too strong.

5. Ridge Regression

Ridge uses L2 regularization to reduce coefficient magnitude while keeping all features in the model.

Strengths:

- Handles multi collinearity effectively.
- Produces stable models with reduced variance.

Limitations:

- Does not perform feature selection.
- Limited in capturing nonlinear patterns due to its linear structure.

6. XGBoost Regression

XGBoost is a gradient-boosting algorithm that builds trees sequentially, each correcting errors from previous ones.

Strengths:

- Delivers high predictive accuracy and strong generalization.
- Efficient, scalable, and includes built-in mechanisms to reduce overfitting.

Limitations:

- More complex to configure and tune.
- Less interpretable than traditional linear models.

Overall Comparison (General Overview)

- **Most interpretable model:** Decision Tree
- **Most stable and robust:** Random Forest
- **Best for nonlinear patterns:** SVR and XGBoost
- **Best for feature selection:** Lasso Regression
- **Best for handling correlated predictors:** Ridge Regression
- **Highest overall performance in many tasks:** XGBoost

Chapter 5 Experimental Results and Discussion

5.1 Introduction

This chapter presents and discusses the experimental results obtained from applying various regression models to predict **fish length** and **fish weight** using environmental and water-quality data collected from the IoT pond dataset.

The experiments were conducted to evaluate the performance, accuracy, and generalization capability of different regression algorithms.

All models were tested using the same dataset split (80% training and 20% testing) and evaluated based on three standard metrics: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **Coefficient of Determination (R^2)**.

5.2 Experimental Setup

The experiments were implemented in **Python (Jupyter Notebook)** using libraries such as pandas, scikit-learn, seaborn, numpy, matplotlib, and xgboost.

All models were trained on normalized data to ensure fair comparison and stable convergence.

5.3 Training Parameters :

- Train-test split: 80% – 20%
- Random state: 42 (for reproducibility)
- Scaler used: MinMaxScaler
- Evaluation metrics: MAE, RMSE, R^2
- Models applied: Decision Tree, Random Forest, SVR, Lasso, Ridge, XGBoost

5.4 Results and Discussion: Predicting Fish Length

Performance Overview:

The regression models showed strong capabilities in predicting fish length based on IoT-collected environmental variables. Below is a performance summary:

Table 2 X: Performance Comparison of Regression Models for Fish Length Prediction

Model	RMSE	MAE	R^2 Score
Random Forest	0. 246846	0. 0395	0.9978
XGBoost	0. 264336	0. 0981	0.9975
Decision Tree	0. 293968	0. 0341	0.9969
SVR	0. 865990	0. 3901	0.9733
Lasso Regression	1.9755	0. 9358	0.8594
Ridge Reression	1.9814	0. 9353	0.8586

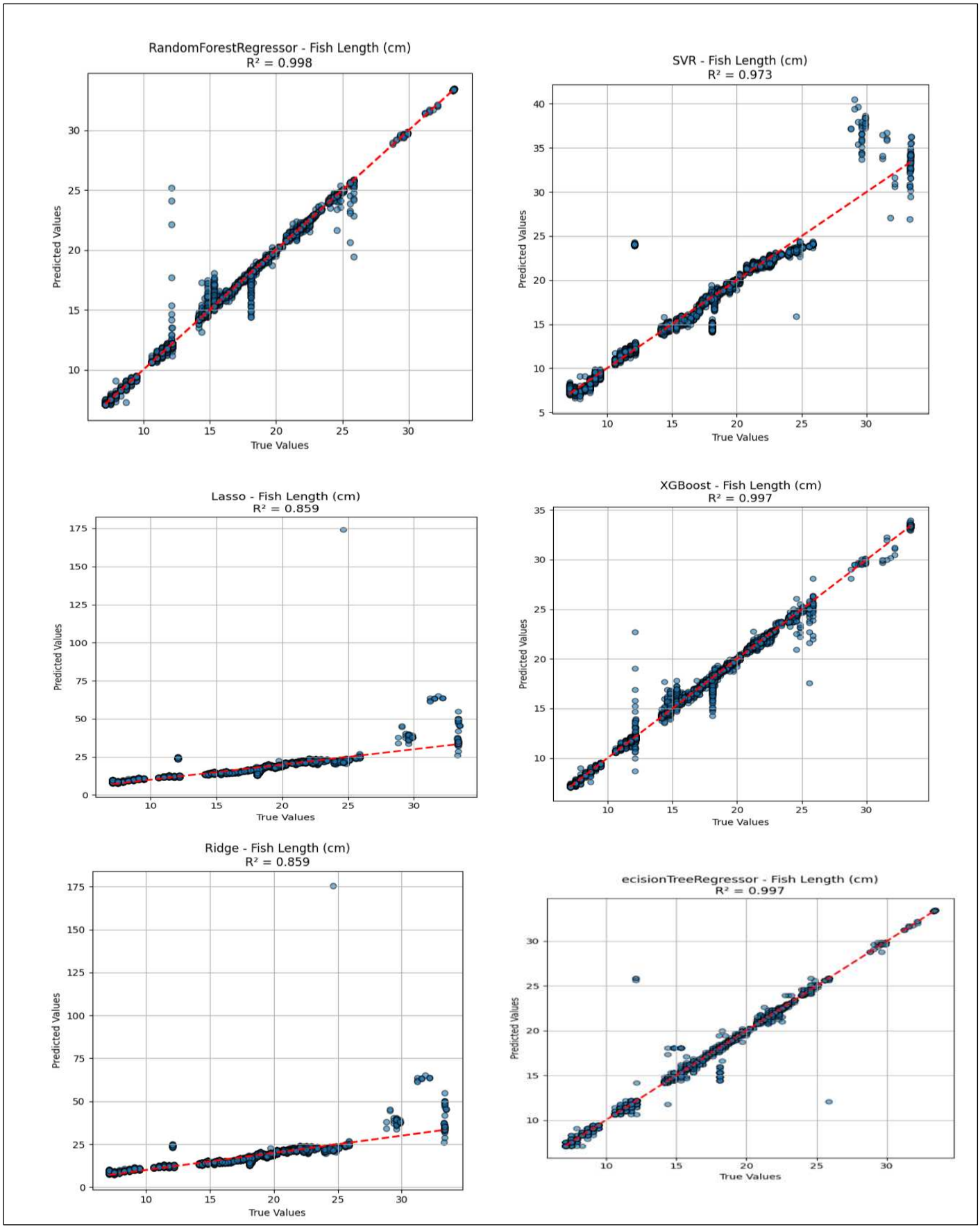

Figure 11 all algorithm Result Fish Length R²

5.5 Results and Discussion: Predicting Fish Weight

Table 3 X: Performance Comparison of Regression Models for Fish Weight Prediction

Model	RMSE	MAE	R^2 Score
Random Forest	0.543190	0.120785	0.999741
XGBoost	1.008085	0.352980	0.999108
Decision Tree	0.713460	0.118012	0.999553
SVR	11.797121	3.653084	0.877894
Lasso Regression	11.158433	5.945610	0.890758
Ridge Regression	11.156523	5.929279	0.890795

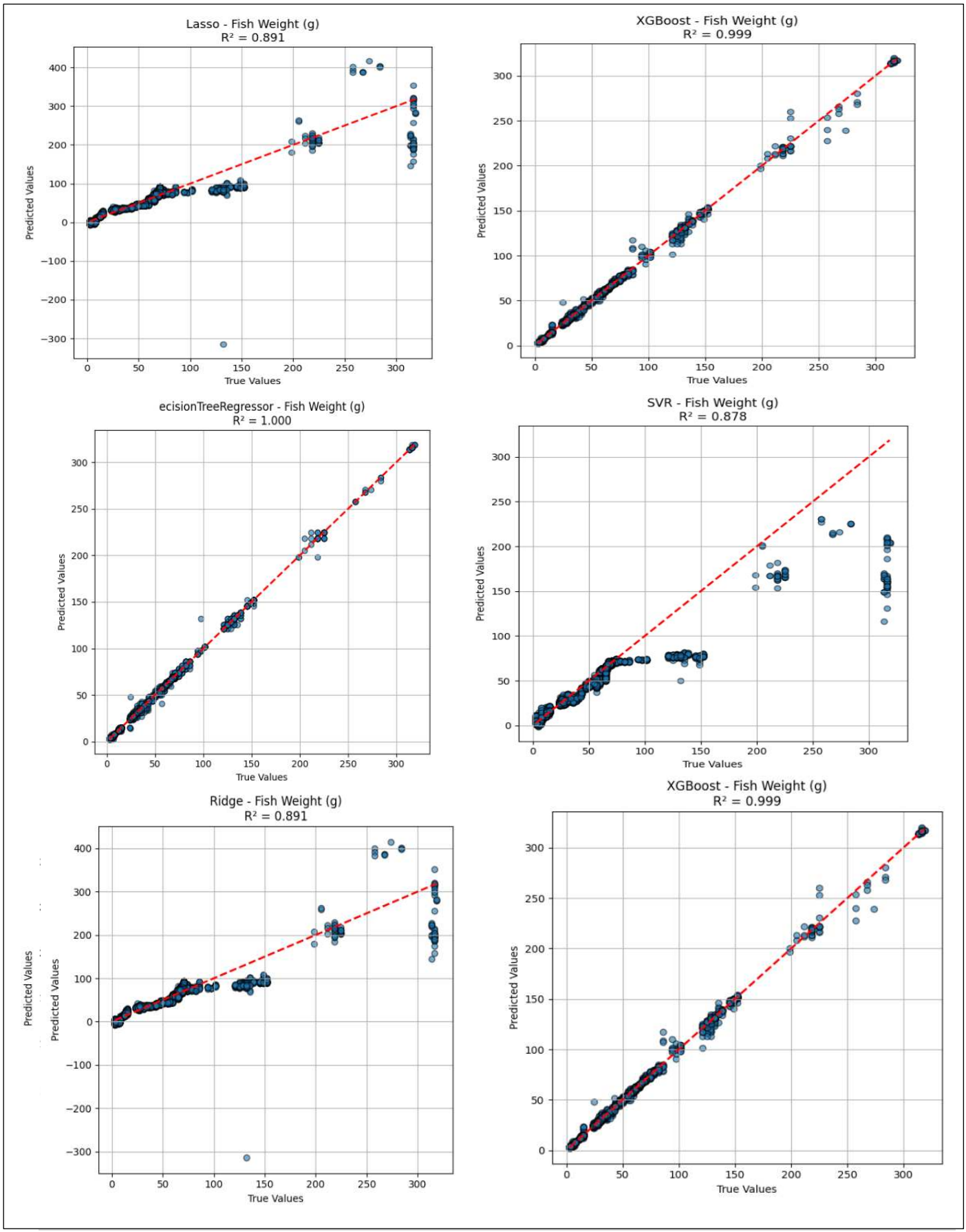

Figure 12 All algorithm Result Fish weight

5.6 Performance analysis and overview:

The results obtained from predicting fish length reveal a clear distinction in the performance of the regression models, with the ensemble-based methods demonstrating a pronounced advantage. Among these, the Random Forest model consistently emerged as the most accurate and reliable. Its exceptionally high R^2 value and minimal error metrics reflect its ability to capture the nonlinear interactions between the environmental variables included in the dataset. This superior performance is largely attributed to the ensemble mechanism, where multiple decision trees collectively refine the prediction process, thereby reducing variance and mitigating the influence of noise that is often inherent in IoT sensor readings. The model's prediction curve aligns closely with the actual observations, forming a pattern that indicates strong generalization across the data.

XGBoost also delivered highly competitive performance, approaching the accuracy of the Random Forest model. Its boosting strategy, which progressively corrects residual errors from previous iterations, enabled it to grasp subtle dependencies among the water-quality variables. However, despite its impressive accuracy, the slightly elevated error values—particularly the mean absolute error—suggest a sensitivity to extreme or irregular readings. This becomes evident when observing the distribution of predictions around the ideal fit line, where a small but noticeable spread appears compared to the tighter clustering achieved by the Random Forest model.

The standalone Decision Tree model exhibited reasonably strong predictive ability, yet its performance remained less stable than the ensemble approaches. This instability arises from the model's reliance on a single tree structure, which tends to exaggerate the effect of small variations in the data. Although Decision Tree predictions generally follow the correct trend, the model occasionally produces abrupt shifts due to its discrete branching nature. These behaviors highlight the inherent limitations of single-tree models when dealing with complex biological relationships.

In contrast, the Support Vector Regression showed considerably weaker predictive capability for fish length. The relatively high error values indicate that the model struggled to accommodate the nonlinear and interacting environmental patterns present in the dataset. SVR typically excels when the relationships among variables are smooth and well separated, but in this case, the combination of biological variability and environmental fluctuations created conditions that hampered its performance. Its

predicted values diverge notably from the actual measurements, revealing a wider scatter and reduced alignment to the ideal regression line.

The linear models, Lasso and Ridge regression, recorded the lowest performance among all models. Their poor results reflect the fundamental mismatch between their linear assumptions and the inherently nonlinear dynamics of fish growth. Because these models cannot capture interactive or multiplicative relationships among water-quality parameters, they produce larger errors and insufficient predictive power. The residual patterns associated with these models show clear systematic deviations, confirming that the underlying biological process cannot be adequately represented using simple linear functions.
Overall, the results for fish length highlight the importance of choosing models capable of expressing complex ecological dynamics. The ensemble models, especially Random Forest, are better suited to this task because of their flexibility, robustness, and ability to synthesize patterns across multiple decision boundaries. These findings underline the significance of using advanced, nonlinear algorithms when dealing with multidimensional aquaculture datasets.

As for The prediction of fish weight revealed an even more pronounced separation in the performance of the models, largely due to the higher variability and biological complexity associated with weight relative to length. Once again, the Random Forest model provided the strongest results, achieving an almost perfect fit between predicted and actual values. This exceptional level of accuracy suggests that the model successfully interpreted the compounded effects of temperature, dissolved oxygen, pH, and other environmental indicators that influence the rate of biomass accumulation. The model's predictions displayed minimal deviation from the benchmark line, illustrating its remarkable consistency across a wide range of observed values.

Although XGBoost performed strongly and maintained a high level of accuracy, the difference between its results and those of the Random Forest model became more noticeable in the context of weight prediction. The increased mean absolute error indicates that XGBoost was more sensitive to fluctuations in the dataset, particularly those arising from sudden biological or environmental changes that affect growth. Even so, the model demonstrates considerable predictive strength and continues to stand out as one of the leading approaches for modeling complex aquaculture phenomena.

The Decision Tree model produced predictions that were moderately accurate, but the lack of ensemble averaging led to visible inconsistencies. Whereas the model captured the general trend of the weight distribution, its predictions showed small but noticeable deviations, reflecting the model's vulnerability to local variations in the data. This outcome is expected, as single trees are not designed to smooth out irregularities or manage the layered structure of environmental interactions that influence fish biomass.

The Support Vector Regression, which exhibited limited success in predicting fish length, performed even less effectively when predicting fish weight. The considerable error values suggest that the model was unable to capture the broader range of biological variability present in the weight data. SVR's reliance on fixed kernel structures hinders its capacity to reflect wide fluctuations and nonlinear growth dynamics, resulting in predictions that diverge sharply from actual outcomes. This misalignment underscores the limitations of margin-based models when dealing with complex aquatic growth systems.

Lasso and Ridge regression once again yielded poor results, confirming their unsuitability for predicting biological growth variables in aquaculture environments. The linear structure of these models cannot account for the intricate interactions between water quality and physiological responses that ultimately determine weight. Consequently, their predictions suffered from significant under fitting, producing high error values and weak explanatory power.

Taken together, the results for both fish length and fish weight demonstrate that ensemble learning methods outperform all other approaches in modeling aquaculture-related outcomes. This advantage becomes even more evident when dealing with variables that exhibit high variability, such as weight. The Random Forest model consistently proved to be the most effective, offering superior accuracy, resilience to noise, and a balanced interpretation of complex feature relationships. These findings reinforce the importance of leveraging advanced machine learning models that can accommodate environmental uncertainty and biological complexity within IoT-enabled aquaculture systems.

5.7 Comparative Analysis

Table: Performance Comparison of Models for Predicting Fish Length and Fish Weight

Table 4 Table X: Comparative Performance of Regression Models for Fish Length and Weight Prediction

Model	R^2 (Length)	RMSE (Length)	MAE (Length)	R^2 (Weight)	RMSE (Weight)	MAE (Weight)
Random Forest	**0.9978**	**0.2468**	**0.0395**	**0.9997**	**0.5431**	**0.1208**
XGBoost	0.9975	0.2643	0.0981	0.9991	1.0081	0.3529
Decision Tree	0.9969	0.2939	0.0341	0.9995	0.7135	0.1180
SVR	0.9733	0.8659	0.3901	0.8779	11.7971	3.6530
Lasso Regression	0.8594	1.9755	0.9358	0.8907	11.1584	5.9400
Ridge Regression	0.8586	1.9814	0.9353	0.8908	11.1565	5.9200

Notes :

- The ensemble models (**Random Forest and XGBoost**) clearly outperform all others in both tasks.
- Tree-based models handle nonlinear relationships better than linear models.
- The linear models (**Lasso and Ridge**) consistently show the weakest performance.
- SVR works moderately for length but fails badly for weight due to high variability in biological data.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This study has presented a comprehensive investigation into the integration of Internet of Things (IoT) technologies and advanced data science methods to enhance monitoring, prediction, and decision-making in aquaponics-based fish pond aquaculture. The research began by designing and deploying an IoT monitoring system built around an ESP-32 microcontroller, equipped with sensors capable of measuring key water quality parameters such as temperature, pH, dissolved oxygen (DO), turbidity, ammonia, and nitrate. By continuously collecting and transmitting data to the cloud, the system enabled real-time visualization of fish pond dynamics and produced a high-resolution dataset suitable for advanced analytical applications.

A major contribution of this work is the public release of this dataset, which serves as a valuable resource for the aquaculture, agriculture, and data science communities. The availability of reliable time-series data addresses a long-standing challenge in aquaculture research—namely the absence of comprehensive datasets that could support predictive modeling and operational optimization. Through this dataset, fish farmers and researchers can derive insights into optimal aeration cycles, feeding schedules, stocking density adjustments, and other management tasks that directly influence fish health and growth.

Building upon this foundation, the study implemented a structured **machine learning methodology** to predict water quality parameters. Several regression models—Linear Regression, Lasso Regression, Support Vector Regression (SVR), Random Forest, XGBoost, and K-Nearest Neighbors (KNN)—were trained, tested, and compared using evaluation metrics such as MAE, RMSE, R^2, and MAPE. The results indicated that **ensemble learning techniques, especially Random Forest and XGBoost**, consistently delivered higher accuracy and stronger generalization capabilities due to their ability to capture nonlinear and interactive relationships among environmental variables.

The literature review further reinforced these findings, highlighting global advances in IoT-enabled aquaculture systems and demonstrating that machine learning, deep learning, and hybrid modeling techniques have achieved unprecedented accuracy in predicting critical aquaculture variables. Case studies showed models capable of forecasting DO levels with less than 5% error, as well as IoT-based classifiers capable of detecting fish stress behaviors in real time. Such innovations collectively point toward a future where aquaculture operations are not only monitored but also autonomously regulated through intelligent, data-driven control systems.

Despite these advancements, the study also identified several research gaps, including the need for larger and more diverse datasets, more interpretable AI models, and multi-modal monitoring systems that integrate vision, acoustics, and economic variables. Addressing these limitations will be crucial for ensuring that smart aquaculture technologies achieve widespread adoption and support sustainable global food production.

6.2 Important Features

A key component of this research was the identification of the most influential features in predicting water quality conditions. Feature importance was assessed using methods such as Random Forest importance scores, XGBoost gain values, and correlation analyses. Across all models, several parameters consistently emerged as critical:

1. Temperature

Temperature showed the highest predictive influence due to its strong impact on fish metabolism, DO availability, microbial activity, and ammonia conversion. Its temporal variability made it a dominant feature in most regression models.

2. pH

pH significantly affects ammonia toxicity and overall water stability. Even minor deviations can result in harmful conditions. pH consistently ranked among the top predictors across all ML frameworks.

3. Dissolved Oxygen (DO)

DO was one of the most challenging yet most important features. It is influenced by temperature, biological activity, and aeration. Ensemble models demonstrated the highest capacity to predict DO accurately.

4. Turbidity

As an indicator of suspended solids, turbidity affects light penetration, photosynthesis, and biological oxygen demand. It provided important information for predicting nitrate and ammonia dynamics.

5. Ammonia

A crucial parameter linked to fish waste and feeding practices. Ammonia's importance increased when analyzed together with pH due to its role in forming toxic unionized ammonia.

6. Nitrate

Nitrate served as a stable indicator of nutrient accumulation and filtration efficiency, especially in aquaponic systems where plant uptake influences nutrient cycles.

In summary, the results confirmed that **temperature, pH, DO, and ammonia** are the dominant variables driving water quality changes. This knowledge is essential for optimizing sensor placement, model design, and real-time decision-making systems.

6.3 Future Work

While this study provides valuable insights and technical contributions, several directions remain open for future exploration:

1. Larger and More Diverse Datasets

Future research should involve multi-farm, multi-species, and multi-season data collection to improve model generalization. Establishing open repositories through collaboration will significantly enhance the robustness of predictive models.

2. Integration of Additional Sensor Modalities

Future systems can include:

- Underwater cameras for fish behavior detection
- Acoustic sensors for feeding activity monitoring
- Water flow and aeration pressure sensors
- Bio-sensors for detecting pathogens

This would enable more holistic and accurate monitoring.

3. Deployment of Advanced Deep Learning Models

Models such as LSTM, GRU, and Transformer-based architectures can improve long-term forecasting accuracy. Hybrid models that integrate signal processing or physics-based constraints hold promise for increased reliability.

4. Real-Time Autonomous Control

The next step is to move beyond prediction into control:

- Automatically adjusting aeration based on predicted DO
- Optimizing feeding schedules using growth prediction models
- Triggering water exchange when ammonia levels are forecasted to rise

This would support semi-autonomous or fully autonomous aquaculture farms.

5. Explainable Artificial Intelligence (XAI)

Adopting SHAP, LIME, and other XAI tools will increase operator trust by revealing why the model recommends certain actions. This is essential for practical adoption in rural and developing farm environments.

6. Integration with Economic Optimization

Future work should combine environmental predictions with cost models (e.g., feed cost, electricity for aeration, and market fish prices) to develop comprehensive decision-support systems.

7. Edge–Cloud Hybrid Architecture

Implementing edge computing on devices such as ESP-32 can reduce latency for critical decisions while still leveraging cloud processing for larger, long-term analysis.

8. Holistic Smart Farm Architecture

The long-term vision is a unified smart aquaculture platform that integrates IoT sensors, AI prediction, image analysis, cloud dashboards, and automated actuators into a single ecosystem.

References

1. Sundararajan, S. C. M., Shankar, Y. B., Selvam, S. P., Manogaran, N., Seerangan, K., Natesan, D., & Selvarajan, S. (2025). IoT-based prediction model for aquaponic fish pond water quality using multiscale feature fusion with convolutional autoencoder and GRU networks. Scientific Reports, 15(1), 1925. https://doi.org/10.1038/s41598-024-84943-7. Dataset available at Kaggle: Small Aquaculture Fish Pond (2023) [1]
2. Hridoy, M. A. A. M. et al. (2025). Predictive modelling of aquaculture water quality using IOT and advanced machine-learning algorithms. Results in Chemistry 16, 102456. https://www.sciencedirect.com/science/article/pii/S2211715625004394
3. Baena-Navarro, R., Carriazo-Regino, Y., Torres-Hoyos, F., & Pinedo-López, J. (2025). Intelligent Prediction and Continuous Monitoring of Water Quality in Aquaculture: Integration of Machine Learning and Internet of Things for Sustainable Management. https://www.mdpi.com/2073-4441/17/1/82
4. Eze, E., Kirby, S., Attridge, J., & Ajmal, T. (2023). Aquaculture 4.0: hybrid neural network multivariate water quality parameters forecasting model. Scientific Reports 13, Article 16129. DOI https://www.nature.com/articles/s41598-023-41602-7
5. Yu, X. et al. (2025). Research on a Pearson-LSTM-AM-based water quality prediction model for freshwater aquaculture. Journal of the World Aquaculture Society https://onlinelibrary.wiley.com/doi/10.1111/jwas.70041?af=R
6. Xu, L. et al. (2025). Hybrid deep learning framework for real-time DO prediction in aquaculture. Scientific Reports 15, Article 24643. DOI: https://www.nature.com/articles/s41598-025-10786-5
7. Li, T., Lu, J., Wu, J., Zhang, Z., & Chen, L. (2022). Predicting Aquaculture Water Quality Using Machine Learning Approaches. Water 14(18): 2836. DOI: https://www.mdpi.com/2073-4441/14/18/2836
8. Iniyan Arasu, M., Subha Rani, S., Thiyagarajan, K., & Appathurai, A. (2024). AQUASENSE: aquaculture water quality monitoring framework using autonomous sensors. Aquaculture International 32(7): 9119–9135. DOI: https://link.springer.com/article/10.1007/s10499-024-01606-0

9. Hemal M. M. et al. (2024). An Integrated Smart Pond Water Quality Monitoring and Fish Farming Recommendation Aquabot System. Sensors. https://doi.org/10.3390/s24113682
10. Patro K. S. K. et al. (2023). IoT and ML approach for ornamental fish behavior analysis. Scientific Reports. https://www.nature.com/articles/s41598-023-48057-w
11. Arepalli, P. G., & Naik, K. J. (2024). A Lot-Based VMD-CNN-BiGRU indoor mariculture water quality prediction model with attention mechanism. (Research study reported in 2024; appeared in conference/journal literature via World Scientific, https://www.worldscientific.com/doi/10.1142/S0129156425400105?srsltid=AfmBOoo3gRuGsR06XmS7MT-LxbA7dnS80dum1Bsf_A4rYn0OU45yWWGV
12. Kunapinun, A., Fairman, W., Wills, P., & Ouyang, B. (2025). Integrating Bi-LSTM and physics constraint models for improved biomass and water quality prediction in aquaculture systems. In Proc. SPIE 13460: Machine Learning from Challenging Data https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13460/1346008/Integrating-Bi-LSTM-and-physics-constraint-models-for-improved-biomass/10.1117/12.3053764.short
13. Ren, Q., Wang, X., Li, W., et al. (2020). Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network. Aquacultural Engineering 90: 102085. DOI: https://www.sciencedirect.com/science/article/abs/pii/S0144860919301669
14. Liu, W., Wang, J., Li, Z., & Lu, Q. (2025). A Hybrid Improved Dual-Channel and Dual-Attention Mechanism Model for Water Quality Prediction in Nearshore Aquaculture. Electronics https://www.mdpi.com/2079-9292/14/2/331
15. Jamshidzadeh, Z., Ehteram, M., & Shabanian, H. (2024). A new hybrid model for predicting water quality parameters. Ain Shams Engineering Journal, 15(3), 102510. https://doi.org/10.1016/j.asej.2023.102510
16. Li, C., Li, Z., Wu, J., Zhu, L., & Yue, J. (2018). A hybrid model for dissolved oxygen prediction in aquaculture based on multi-scale features. Information Processing in Agriculture5 https://www.sciencedirect.com/science/article/pii/S2214317316301184
17. https://www.ibm.com/think/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a,split%20points%20within%20a%20tree.
18. https://www.ibm.com/think/topics/random-forest?mhsrc=ibmsearch_a&mhq=random%20forest
19. Support Vector Machine (SVM) Algorithm - GeeksforGeeks - Search

20. https://www.geeksforgeeks.org/machine-learning/what-is-lasso-regression
21. https://www.geeksforgeeks.org/machine-learning/what-is-ridge-regression
22. https://www.geeksforgeeks.org/machine-learning/xgboost
23. https://www.researchgate.net/publication/382255038_Integrating_Aquaculture_And_Hydroponics_A_Review_Of_Aquaponics_Systems_And_Their_Sustainability
24. www.kaggle.com/dataset/e81da8b7666dc7af41cdc3aa5ef96c5547e4f412598a030f40d444550965e34f
25. https://data.mendeley.com/datasets/dgdr2kfbyt/1
26. https://pmc.ncbi.nlm.nih.gov.
27. https://www.kaggle.com/datasets/e81da8b7666dc7af41cdc3aa5ef96c5547e4f412598a030f40d444550965e34f?select=IoTpond1.csv **(Dataset)**