

أحمد النهاري

Predicting fish growth using Internet of Things and Artificial Intelligence in Aquaculture systems_laast %282%291.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:339099397

Submission Date

Dec 28, 2025, 4:15 AM GMT+2

Download Date

Dec 28, 2025, 10:22 AM GMT+2

File Name

Predicting fish growth using Internet of Things and Artificial Intelligence in Aquaculture syste....docx

File Size

2.8 MB

67 Pages

12,369 Words

78,783 Characters

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **162 Not Cited or Quoted 18%**
Matches with neither in-text citation nor quotation marks
-  **3 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 14%  Internet sources
- 8%  Publications
- 12%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 162 Not Cited or Quoted 18%**
Matches with neither in-text citation nor quotation marks
- 3 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 14% Internet sources
- 8% Publications
- 12% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.geeksforgeeks.org	6%
2	Submitted works	Midocean University on 2025-09-05	<1%
3	Submitted works	Sunway Education Group on 2025-07-25	<1%
4	Internet	www.mdpi.com	<1%
5	Internet	medium.com	<1%
6	Submitted works	Université Internationale de Rabat on 2025-05-25	<1%
7	Publication	Soufiane Haddout, Tonni Agustiono Kurniawan. "Handbook of Artificial Intelligen...	<1%
8	Internet	pmc.ncbi.nlm.nih.gov	<1%
9	Submitted works	universititeknologimara on 2024-07-15	<1%
10	Publication	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computati...	<1%

11	Submitted works	University of Westminster on 2025-11-13	<1%
12	Internet	bora.uib.no	<1%
13	Internet	www.tandfonline.com	<1%
14	Submitted works	Islington College,Nepal on 2025-12-15	<1%
15	Internet	iieta.org	<1%
16	Internet	www.researchsquare.com	<1%
17	Submitted works	Liverpool John Moores University on 2025-04-25	<1%
18	Internet	khazna.ku.ac.ae	<1%
19	Submitted works	Liverpool John Moores University on 2024-09-05	<1%
20	Internet	mis.itmuniversity.ac.in	<1%
21	Internet	www.politesi.polimi.it	<1%
22	Publication	Sushil Kamboj, Pardeep Singh Tiwana. "Innovations in Computing", CRC Press, 2025	<1%
23	Submitted works	Swinburne University of Technology on 2025-11-04	<1%
24	Internet	www.diva-portal.org	<1%

25	Publication	"Advances in Computational Collective Intelligence", Springer Science and Busine...	<1%
26	Submitted works	The University of Texas at San Antonio on 2025-12-04	<1%
27	Publication	Fang, Ruijie. "Physiological Signal-Based Affective Computing System: Device, Too...	<1%
28	Publication	Ponson, Tori Armand. "A Comparative Study of Predictive Modeling in Human Im...	<1%
29	Publication	dos Santos, Ricardo Bruno Barbeiro. "Improving the Robustness of Multimodal AI...	<1%
30	Submitted works	Brunel University on 2025-12-17	<1%
31	Submitted works	University of Huddersfield on 2025-07-09	<1%
32	Internet	www.mecs-press.org	<1%
33	Internet	www.science.gov	<1%
34	Submitted works	Indian Institute of Technology Patna on 2025-12-18	<1%
35	Internet	www.ams.giti.waseda.ac.jp	<1%
36	Submitted works	Indian Institute of Technology, Kanpur on 2012-08-18	<1%
37	Submitted works	The Robert Gordon University on 2025-05-04	<1%
38	Submitted works	Aberystwyth University on 2025-09-11	<1%

39	Submitted works	Dublin City University on 2025-07-31	<1%
40	Submitted works	Eastern Michigan University on 2025-10-17	<1%
41	Publication	Lei Guo, Wei Luo, Qingqing Tian. "Water quality prediction in the Yellow River sou...	<1%
42	Submitted works	Universiti Brunei Darussalam on 2025-03-16	<1%
43	Publication	Zahra Jamshidzadeh, Mohammad Ehteram, Hanieh Shabanian. "A new hybrid mo...	<1%
44	Internet	academia.edu.pk	<1%
45	Internet	ir.nctu.edu.tw	<1%
46	Internet	run.unl.pt	<1%
47	Internet	www.impaq.ruc.dk	<1%
48	Submitted works	CSU, San Jose State University on 2023-04-22	<1%
49	Submitted works	The University of the West of Scotland on 2025-04-23	<1%
50	Internet	jonuns.com	<1%
51	Internet	repository.uigm.ac.id	<1%
52	Internet	uwspace.uwaterloo.ca	<1%

53	Submitted works	Ahsanullah University of Science and Technology on 2025-12-17	<1%
54	Submitted works	Islington College,Nepal on 2025-12-15	<1%
55	Publication	Suma Christal Mary Sundararajan, Yamini Bhavani Shankar, Sinthia Panneer Selv...	<1%
56	Internet	www.ijsr.net	<1%
57	Publication	Hamid Jahankhani, Gordon Bowen, Nitsa J. Herzog, David J. Herzog. "Symbiotic In...	<1%
58	Submitted works	Kotebe University on 2025-06-03	<1%
59	Publication	Narayanan, Usha. "Patient Behavior Prediction Using Big Data Analytics and Mac...	<1%
60	Publication	Rupali P. Shete, Amith Shekhar C., Yogeshwari V. Mahajan, Anupkumar M. Bongal...	<1%
61	Submitted works	University of Hertfordshire on 2023-05-15	<1%
62	Submitted works	University of Teesside on 2022-01-11	<1%
63	Publication	Yang Liu, Fanghao Wan. "Unveiling Temporal and Spatial Research Trends in Prec...	<1%
64	Internet	ijaer.in	<1%
65	Internet	pertanika2.upm.edu.my	<1%
66	Internet	www.pos.cps.sp.gov.br	<1%

67	Internet	www.researchgate.net	<1%
68	Publication	Akhilesh Kumar Yadav. "Computational Techniques in Environmental Engineerin...	<1%
69	Publication	Castillo, Robert. "Determining Best Sports Ranking Method With Machine Learnin...	<1%
70	Submitted works	Dublin City University on 2025-07-30	<1%
71	Submitted works	Glasgow Caledonian University on 2018-09-04	<1%
72	Internet	apps.dtic.mil	<1%
73	Internet	ijirt.org	<1%
74	Internet	sites.ualberta.ca	<1%
75	Internet	www.ijert.org	<1%
76	Publication	Abderahman Rejeb, Karim Rejeb, John G. Keogh. "The nexus of IoT and aquacultu...	<1%
77	Publication	Alphose Zingoni. "Engineering Materials, Structures, Systems and Methods for a ...	<1%
78	Publication	Dileep M. R., Shridhar Sanshi, Mahendra Pratap Singh, Maanak Gupta. "Integratin...	<1%
79	Submitted works	Dublin Business School on 2025-08-28	<1%
80	Submitted works	Midocean University on 2025-09-05	<1%

81	Publication	Mohamed Ould-Elhassen Aoueileyine, Mahdi Hamzaoui, Ridha Bouallegue. "chapt...	<1%
82	Publication	Ovidiu Vermesan, Alain Pagani, Paolo Meloni. "Charting the Intelligence Frontiers...	<1%
83	Internet	bmcmedinformdecismak.biomedcentral.com	<1%
84	Internet	cerv.aut.ac.nz	<1%
85	Internet	ebin.pub	<1%
86	Internet	repositorio.uchile.cl	<1%
87	Internet	repository.lib.ncsu.edu	<1%
88	Internet	vital.seals.ac.za:8080	<1%
89	Internet	www.biorxiv.org	<1%
90	Internet	www.nature.com	<1%
91	Publication	Adithya Challapalli, Guoqiang Li. "Artificial Intelligence Assisted Structural Optimi...	<1%
92	Submitted works	Universidad San Francisco de Quito on 2024-12-03	<1%
93	Submitted works	University of Carthage on 2025-06-19	<1%
94	Publication	Agostini, Augusto Cardoso. "A Univariate Transformer-Based Approach for Traffic...	<1%

95	Publication	Dinesh Goyal, Payal Bansal, Ritam Dutta, Madhav Sharma. "Intelligent Systems U...	<1%
96	Publication	Mithila Akter Mim, Mst. Rokeya Khatun, Muhammad Minoar Hossain, Wahidur Ra...	<1%
97	Submitted works	The University of the West of Scotland on 2025-08-13	<1%
98	Internet	booksc.en-proxy.com	<1%



7 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

70

I would like to say thank you to everyone who has been part of my graduation Thesis journey.

84

First and foremost, I owe my gratitude to my dear supervisor Dr Hager Saleh for their valuable guidance, endless patience and continued encouragement in all phases of this work. I appreciate your comments and constructive criticism that significantly improved the quality of this Thesis.

21

I would also like to express my heartfelt thanks to my family and friends for their support, understanding and motivation at the hardest days. Your faith in me has been that which keeps me going.

Finally, I would like to thank Midocean University / Science in Artificial Intelligence Department for outfitting me with every little thing that made this project a possibility.

No, this is not my accomplishment. It is the sum of other people's work and good deeds added on top of one another above.

Abstract

This thesis presents a data-driven framework which incorporates the IoT system and machine learning into aquaculture to enable predictive analysis and decision support. The framework is operational as an intelligent fishpond monitoring system which can observe the key water quality parameters such as temperature, pH level, dissolved oxygen, turbidity, ammonia level and nitrate value constantly. These parameters are acquired by a set of distributed sensors connected to an ESP-32 microcontroller, thereby ensuring accurate real-time data acquisition.

Structured post-sensing pre-processing routines including data validation and normalization are applied to the accumulated sensor data before storing it in a cloud-based storage system. The dataset generated is further taken for the predictive modeling. Different supervised regression methods are used to predict two basic variables related to fish growth: length and weight. The implemented models are Decision Tree, Random Forest, Support Vector Regression (SVR), Lasso Regression, Ridge Regression and XGBoost.

Model's performance is assessed based on three statistical measures: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2). Results show that ensemble learning methods perform better than single regression model in accuracy and stability. Specifically, Random Forest and XGBoost reach R^2 values above 0.99 for both growth-related targets, demonstrating the capability of these models in capturing nonlinear dependencies in IoT-based aquaculture data.

Furthermore, the paper presents a brief survey of fifteen recent papers published from 2018 to 2025 on IoT-based data analytics in aquaculture. The survey shows a tendency towards employing more hybrid learning approaches and ensemble methods, providing better predictive performance than traditional machine-learning approaches.

In conclusion, the results highlight the need to combine continuous IoT sensing with predictive machine learning for early risk detection, better pond management and higher fish survival. The proposed work serves as a steppingstone towards smart aquaculture by providing scalable and flexible approach that unites real-time sensing and data-driven prediction. Research focus Future works will involve enlarging the data set scale, multidimensional sensor fusion and automatic control mode etc., to realize fully intelligent and self-regulating traits of aquaculture system.

2 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	3
1.4 Thesis Organization.....	4

Chapter 2: Overview	6
2.1 Introduction	6
2.2 Machine Learning (ML).....	7
2.3 Deep Learning (DL)	8
2.4 Relevance of AI Concepts to Fishpond Water Quality Prediction	9
2.5 Some of Models (Regression Models)	9

Chapter 3: Related Work	24
3.1 Introduction	24
3.2 Literature Review.....	25
3.4 Comparison Table	36

Chapter 4: Chapter 4 Proposed Approach (Methodology)	38
4.1 Dataset Description	38
4.2 Data Pre-processing	40
4.3 Splitting the dataset	40
4.4 Models (Regression Models)	40
Chapter 5: Experimental Results and Discussion	44
5.1 Introduction	44
5.2 Experimental Setup	45
5.3 Training Parameters	46
5.4 Results and Discussion: Predicting Fish Length	47
5.5 Results and Discussion: Predicting Fish Weight	49
5.6 Performance analysis and overview	50
5.7 Comparative Analysis	53
Chapter 6: Conclusion and Future Work	54
6.1 Conclusion	54
6.2 Future Work	55
References	56

List of Figures

- Figure 1 Decision Tree Regression (DT) 11
- Figure 2 Random Forest Regressor (RF) 14
- Figure 3 Support Vector Regression (SVR)15
- Figure 4 Support Vector Machine (SVM)16
- Figure 5 Lasso Regression (LR)19
- Figure 6 Ridge Regression20
- Figure 7 XGBoost Regression21
- Figure 8 XGBoost Regression22
- Figure 9 D Tree Algorithm23
- Figure 10 Methodology overview38
- Figure 11 All algorithm Result Fish Length R^2 47
- Figure 12 All algorithm Result Fish weight49

List of Abbreviations

- Acc.: Accuracy
- IOT: Internet Of Thinks
- ANN: Artificial Neural Network
- 23 • ML: Machine Learning
- DL: Deep Learning
- XAI : Explainable Artificial Intelligence
- SVM: Support Vector Machine
- RF: Random Forest
- DT: Decision Tree
- 39 • LSTM: Long Short-Term Memory
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- 4 • LR: Logistic Regression
- KNN: K-Nearest Neighbors
- MAE: Mean Absolute Error
- MSE: Mean Squared Error
- R²: R - Squared
- LDA: Linear Discriminant Analysis
- 83 • LLM: Large Language Model
- NLP: Natural Language Processing
- EDM: Educational Data Mining

58

List of Tables

- Table 1 Comparison of previous studies 36
- Table 2 Performance Comparison of Regression Models for Fish
Length Prediction 46
- Table 3 Performance Comparison of Regression Models for Fish
Weight Prediction48
- Table 4 Comparative Performance of Regression Models for Fish
Length and Weight Prediction53

49

Chapter 1: Introduction

The demand for fish and fisheries throughout the world has stimulated growth in aquaculture, one of the most significant food production sectors. Freshwater aquaculture production is largely supported using fishponds which are dependent upon maintaining good water quality for sustained productivity and sustainability. Any slight changes in the parameters like temperature, pH, DO and turbidity can result in stress, diseases or mass extinction of fish populations. Conventional methods for water testing, employing manual sample collection and laboratory-based analyses, are slow, intermittent and unable to follow in real time the dynamic changes of an aquatic system.

86

The concept of Internet of Things (IoT) deployed in aquaculture brings a revolution to the way environmental monitoring is performed, as it allows continuous sensing, automatic data collection and remote control over fishponds. By means of distributed sensors and wireless network, the IoT systems are capable of generating big time series water quality data which can be used as ample resources to predictive analytics and intelligent decision making. Combined with machine learning (ML) models including RF, XGBoost, SVR, K-Nearest Neighbors, Linear Regression and Lasso Regression—IOT driven data can be converted to insightful predictions that are helpful for early prediction and proactive ponds management.

32

7

This thesis concentrates on constructing a data-driven framework to predict the values of water quality parameters in IoT-enabled fishponds based on supervised machine learning methods. The research capitalizes on actual IoT fishpond sensor data and investigates the performance of a range of regression models. Through selecting the best models for predicting water quality, the project is helping to develop more intelligent aquaculture systems that are better able to optimize fish health, minimize losses and increase production efficiency.

19

1.1 Problem Definition

Water pollution is still one of the most important problems in contemporary aquaculture. Low pH, sudden or significant drops in DO, fluctuating temperatures and murky water conditions can all have dramatic effects on the health and survival of fish. In common aquaculture, water conditions are periodically maintained by manual operation once or twice a day. However, due to the low frequency of monitoring irregular changes in the environment are not detected fast enough, resulting in economic losses. Fish farmers can access real-time measurements, albeit currently continuously, yet they have difficulties reading the streams of data provided by sensors. The IoT data hardly entails any actionable information without being processed and analyzed. Of even greater importance, however, farmers want predictive signs— not just historical numbers — so they can take action on the same curve or before any damage occurs.

Despite advances such as the ability to monitor in real time using sensor networks enabled by Internet of Things (IoT) technologies, aquaculture adopters currently struggle with making sense out of large raw data flows. Unanalyzed, IoT data offers composite rather than predictive information. To make pond management proactive, there is a need for predicting intelligence that enables farmers to forecast water quality degradation before it reaches critical point.

Hence, the problem of interest in this work is:

71

To what extent is it viable to analyze IoT-based fishpond data with the use of machine learning models to make precise prediction of the major water quality parameters that will assist in taking early decisions for improved aquaculture?

his research man exists due to several technical and practical limitations drawn from the current solutions; Raw IoT data is usually noisy, incomplete and not directly usable without preprocessing.

- Conventional statistical models do not account for complex nonlinear relationships among water quality factors.
- Current tracking systems typically provide a reactive instead of a predictive response.
- Few studies have compared multiple regression-based machine learning approaches through real IoT aquaculture datasets.

· Challenges These challenges point out the necessity of a strong machine learning framework to evaluate, optimize and identify performance-potent prediction models for realtime FSMSs.

24

1.2 Thesis Objectives

This thesis' main goal is to develop, implement and assess ML regression models to predict fishpond water quality parameters through real-time IoT sensor data.

Here are the specific research goals:

· To appear out a model based on IoT fishpond data in order to predict some essential water parameters.

· Update the data for all stations in the preprocessing step, including cleaning and missing value handling (including feature engineering to enhance model performance).

20

· To model and assess various models based on multiple regression- machine learning algorithms (e.g., Linear Regression, Random Forest, XGBoost Support Vector Regression (SVR), Lasso and others ensemble models)

· The determination of the major environmental factors that shape changes in water quality and their effects on growth of fish.

12

· To recommend an optimal machine learning model that can be used in real time IoT based aquaculture management system.

· To illustrate the abilities of predictive analytics to deliver early warnings and assist in proactive decision-making for sustainable fishpond farms.

1.3 Thesis Contribution

The thesis extends previous work in smart aquaculture by changing the emphasis of descriptions and real-time capturing of information to prediction models for fish evolution. In contrast to most of the existing work, which is based on instantaneous detection of water quality or manual inspection and decision-making, this research presents a purely data-driven prediction framework that can predict fish status indicators ahead.

The broad scientific achievements of this thesis can be synthesized as follow:

- A machine learning predictive model to predict the fish growth indicators i.e. the length of the fish and weight using real IoT sensor data obtained from fishponds.
- A multiple response framework is suggested that will account for water quality (temperature, pH, DO and TUR) in modelling approach to incorporate their joint effect on the growth of fish.
- Thorough comparison of several regression approaches such as Decision Tree, Random Forest, SVR, lasso R, ridge R, and XGBoost to find faction models that predict future growth over just real-time observations.
- Evidence that ensemble learning methods, especially Random Forest and XGBoost, have clear advantages over classic regression models in both prediction precision and stability can provide good estimation of fish length and weight.
- The conversion, from IoT sensor data as a surveillance resource to a predictive decision tool for proactive aquaculture management, early mild prevention and data planning oriented.
- Literature contribution by bridging the gap between water quality monitoring systems and predictive fish growth modeling, to shift aquaculture management from reactive and moment-by-moment consideration to intelligent and preventative control.

1.4 Thesis Organization

The thesis consists of six chapters; each one covering an important part of the study.

Chapter 1 – Introduction:

Gives an overview of aquaculture and the importance of IoT in contemporary fish farming. Problem formulation, goals and contribution of the thesis are introduced in chapter .

chapter 2 Overview / Background:

Introduces basic knowledge on the Internet of Things (IoT), cloud-based data collection and machine learning algorithms in relation to aquaculture. It describes the theoretical basis of the proposed system.

Chapter 3 – Literature Review:

76

Reviews related works on IoT applications in aquaculture, water quality monitoring system, and machine learning techniques used for predictive modeling in fish ponds. It also recognizes where there is still a void in knowledge, which is the starting point for this study.

Chapter 4 – Methodology:

Details the proposed framework such as the IoT dataset, sensor specifications, data preprocessing and normalization techniques, feature selection technique and different regression algorithms. It also introduces experimental configuration and performance measures.

74

Chapter 5 Experimental Results and Discussion:

Provides the results of experiments and compares performance for all regression models via different error metrics and visualization. The results are used to reveal which models performed best in terms of predictions, and the reasons for this based on characteristics of the data.

12

Chapter 6 – Summary and Further Work:

Summarizes the research findings and provides recommendations on future work and potential practical applications for this IoT-based predictive system in smart aquaculture.

Chapter 2 Overview

2.1 Introduction

Overall Background of Artificial Intelligence (AI) AI is a generic term for any kind of science. In other words, its objective is to create a machine system that would perform functions analogous to human beings. Such duties might be learning from experience, recognizing patterns, understanding the natural language and deciding without having "somebody" tell you what to do. AI's recent explosive development is based on a set of specialized subfields/concepts... -ML, -DL And Relationship between various AI concepts and Fishpond Water Quality Predictions and About Some of Regression Models.

In the following we will briefly review these basic concepts of AI and explain their importance in today's data-driven applications.

42

2.2 Machine Learning (ML)

29

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:

- Forecasting future trends, and
- Detecting anomalies,
- Aggregating like instances of observation,
- Enabling decision makers to base decisions on evidence.

Most present predictive systems, especially those relying on real-time or continuous data streams, are constructed around ML.

30

2.3 Deep Learning (DL)

Deep Learning is a subfield of ML, which uses artificial neural networks with many hidden layers to extract abstract and high-level features from data. Unlike traditional ML techniques that rely on manual feature extraction, the process of learning high-level representations is performed automatically by deep learning models, and hence DL

methods are effective for processing large scale nonlinear and unstructured datasets (e.g., images, audio signals, sensor reads, natural language).

65

DL has made breakthrough progress on tasks such as computer vision, speech recognition, NLP and time series prediction which are in need to capture subtle patterns that traditional algorithms cannot handle.

2.4 Relationship of AI Concepts to Fishpond Water Quality Predictions

The ideas mentioned on ML, DL, LLMs and XAI are the technology underlying modern environmental monitoring and smart aquaculture systems. In terms of fishpond functional operation, AI is helpful in the following aspects:

19

- Machine Learning allows models to estimate the main water quality indicators (pH, dissolved oxygen content, temperature and turbidity) in previous sensor measurements.

81

- The complex nonlinear relationships in complicated marine environments, which are further compounded by Deep Learning based modeling, can enhance the accuracy and performance of water quality prediction.

Combining AI component with conventional aquaculture system allows the modernization of aquaculture from reactive water management to proactive, predictive

and data-driven ecosystem monitoring and ultimately enhancing systems productivity, lowering fish mortality rates, and improving pond conditions.

2.5 Some of (Regression Models).

DT Regression (DT) Decision tree is a non-parametric supervised learning algorithm that can be employed in classification and regression problems. It is a hierarchical tree which includes a root node, branches, internal nodes and leaf nodes. As shown in the diagram below, a decision tree always begins with a root node with no incoming branches. Outgoing branches from the root node are followed by internal nodes (or decision nodes in this context). Both of node types perform their evaluation according to available features in such a way that homogenous subsets (referred as leaf nodes, or terminal nodes) are created. And the leaf nodes are the possible results in our dataset. [17]

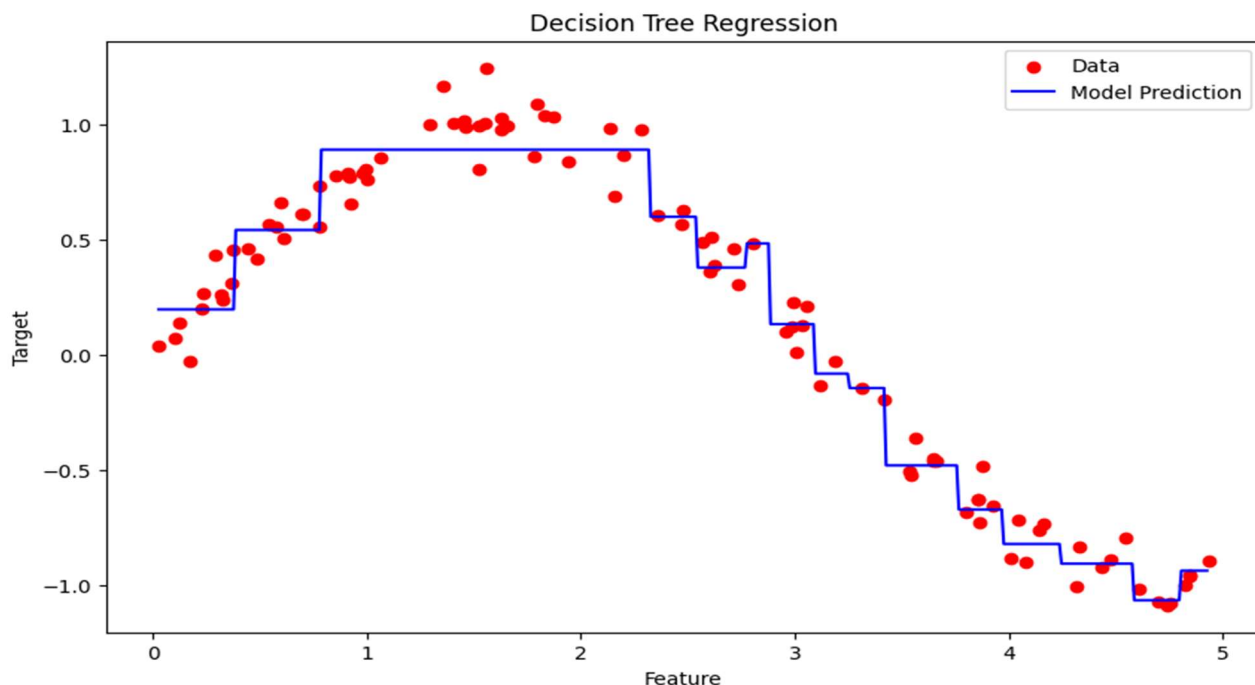


Figure 1 Decision Tree Regression (DT)

Application :

Robotics: RL is employed to automate repetitive tasks in organized environments, like manufacturing, with robots learning to refine movements and maximize effectiveness.

Games: Advanced RL algorithms have been employed to create policies for complex games such as chess, Go and video-games surpassing human players in many cases.

Industrial Control: RL facilitates in-the-moment fine tuning and optimization of industrial processes like refining process (oil & gas).

Personalized Training Regime: RL can be used to personalize the training curriculum of an individual based on her learning patterns, increasing engagement and efficacy.

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 technique which aggregates predictions derived from multiple decision trees to obtain a more accurate and stable prediction. It is applicable for classification as well as regression problem. If you are working on a regression problem, you can use Random Forest Regression method to predict the values. It makes predictions by averaging the predictions of several decision trees. [18]

How it Works Random Forest Regression works by making multiple of decision trees and training each of them on a random subset of the data. It starts with Bootstrap sampling, that is random sample of rows from data with replacement, to create a set of different training datasets for various decision trees. After this it's time for feature sampling, using only a random subset of the features to construct each tree to ensure that our models aren't too similar. Once the trees are trained every tree makes a prediction and your final prediction in regression task is nothing but the average of predictions made by all those individual trees and this process is known as Aggregation.

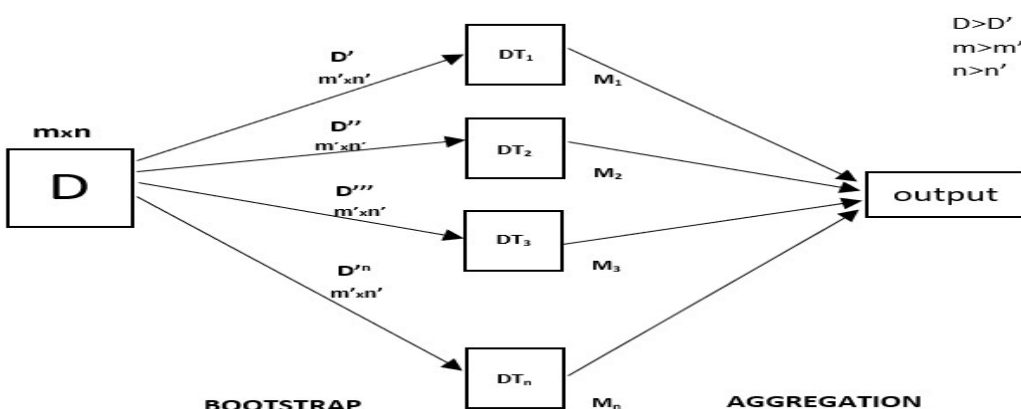


Figure 2 Random Forest Regressor (RF)

1

Applications of Random Forest Regression The Random forest regression can be effectively applied to real world problems such as:

1

§ Predict continuous numbers - e.g.: House Prices, Stock performance or customer lifetime value.

§ Identifying risk factors: Spotting the risk factors for diseases, financial crises or other bad things.

1

§ How to deal with high dimensional data: Datasets which contain a large number of input features.

§ Handling complex relationships : Modeling the complex relationship between input features and the target.

Advantages of Random Forest Regression

· Non-Linear Handling: It is able to capture complex,non-linear relationships in the data that other models may miss.

1

· Overfitting is reduced: since prediction is averaged over multiple decision trees, the risk of overfitting is lower clubbed to make a single decision tree.

· Outliers Tolerance: Random Forest is not very sensitive to outliers as it takes the mean of many trees.

· Works well with large data sizes: GMMs can handle large datasets and high-dimensional data without loss of efficiency.

1

· Dealing with Missing Data: Random Forest can deal with the missing values by using surrogate splits and still retaining high accuracy in presence of partial data.

1

No feature scaling Random Forest does not require normalization of data.

Disadvantages of Random Forest Regression

1

Complexity: It can be expensive and time consuming to train, especially for a large number of trees or high-dimensional data. This is a consequence of the fact that it may

not be an online suitable method for making real-time predictions particularly when we have a large number of trees.

· **Less Interpretable** : It becomes less interpretable because finally, it is an ensemble of trees which aren't easy to interpret compared to linear regression and decision trees.

· **Memory Usage**: Memory used are proportional to the number of decision trees for large data sets.

Overfitting on Noisy Data: Random Forest decreases the overfitting problem since it has large number of trees, which makes it's a decent model; but in this case solution often leads to overfitted model due to presence of noise data.

Sensitivity to Imbalanced Data: It may not work well if the dataset is extremely imbalanced, like one of the classes have more frequency than another.

3. Support Vector Regression (SVR) Support Vector Machine (SVM) is a supervised learning model for classification or regression challenges. It attempts to identify the best boundary called hyperplane that can segregate different classes on the basis of input

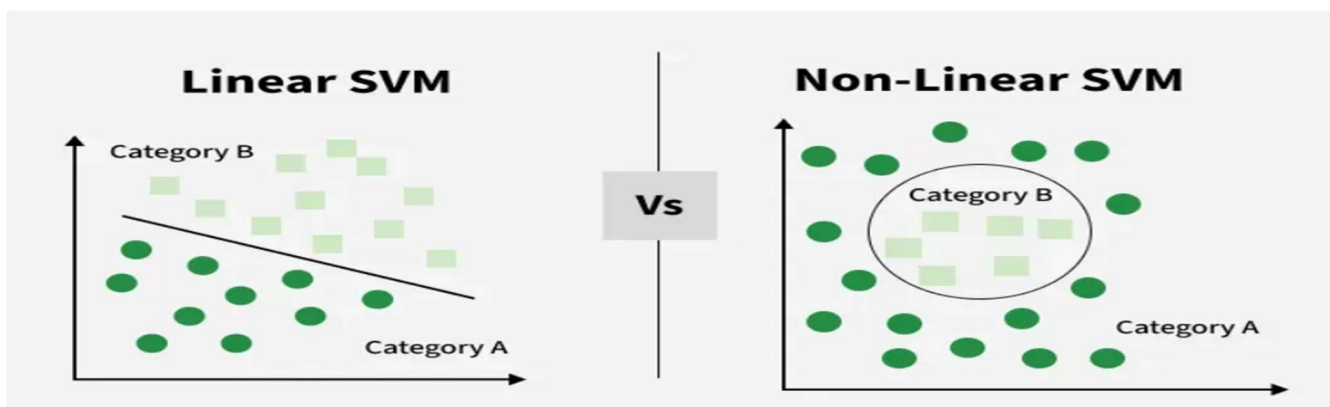


Figure 3 Support Vector Regression (SVR)

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.

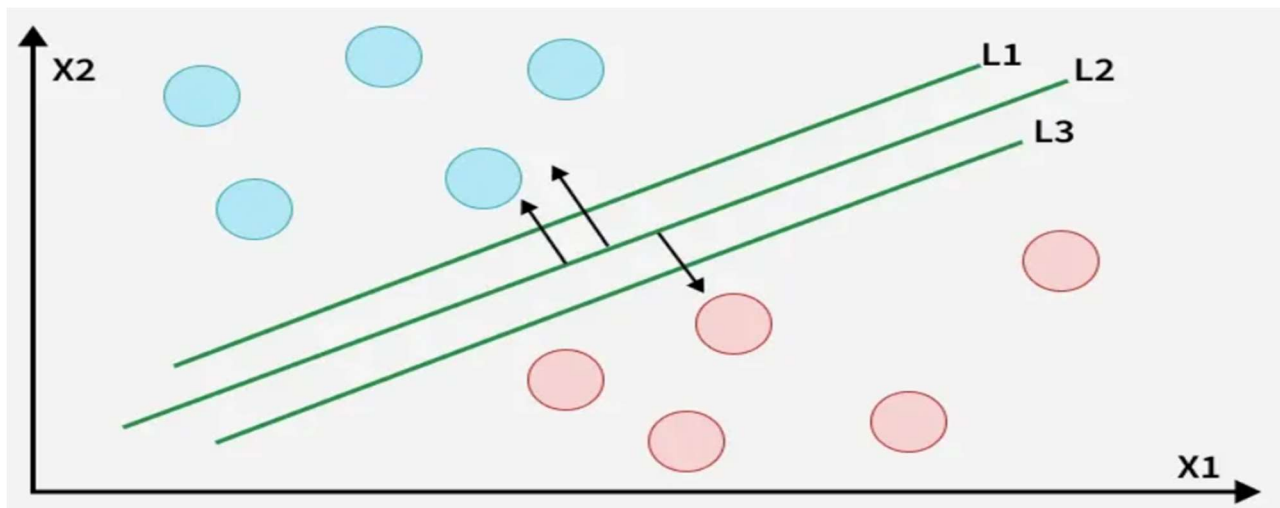


Figure 4 Support Vector Machine (SVM)

Advantages of Support Vector Machine (SVM)

- **High-Dimensional problems:** SVM is effective in high dimensional spaces especially for image classification and gene expression analysis.
- **Nonlinearity:** By implementing kernel functions (e.g., RBF, polynomials etc) SVM deals well with the non-linear relationship.
- **Outlier Tolerance:** As due to soft margin SVM can ignore outliers which makes it robust for spam detection and anomaly detection.
- **Recently, it was good for both binary classification and multiclass classification** Could be used as a compressed label to save memory usage in text classification.
- **Memory-Efficient:** Since it caters to the support vectors, this is memory efficient as compared to other algorithms.
- **Drawbacks of Support Vector Machine (SVM)**
- **Slow Training:** As SVM is slow for large datasets, performance will be affected in SVM in data mining.
- **Difficulty in Tuning of Parameters:** Choosing an appropriate kernel and the proper setting for parameters such as C - risk parameter, accentuates problem of SVM-based algorithms.

- Noise Sensitivity: The SVM performs poorly on noisy data sets and with overlapping class, it is less effective in practical application.
- Limited Interpretability: The hyperplane in higher dimensions is so complex that it is very difficult for users to have a thorough insight to it.
- Feature Scaling Sensitivity: Proper feature scaling is essential, otherwise SVM models may perform poorly.

1

72

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

Lasso Regression is a regression analysis method that performs both variable selection and regularization. It is useful in eliminating irrelevant data features and to avoid overfitting. This enables insignificant features of weak influence to be judged in contrast, as the coefficient values for unimportant variables are penalized towards zero. [20]

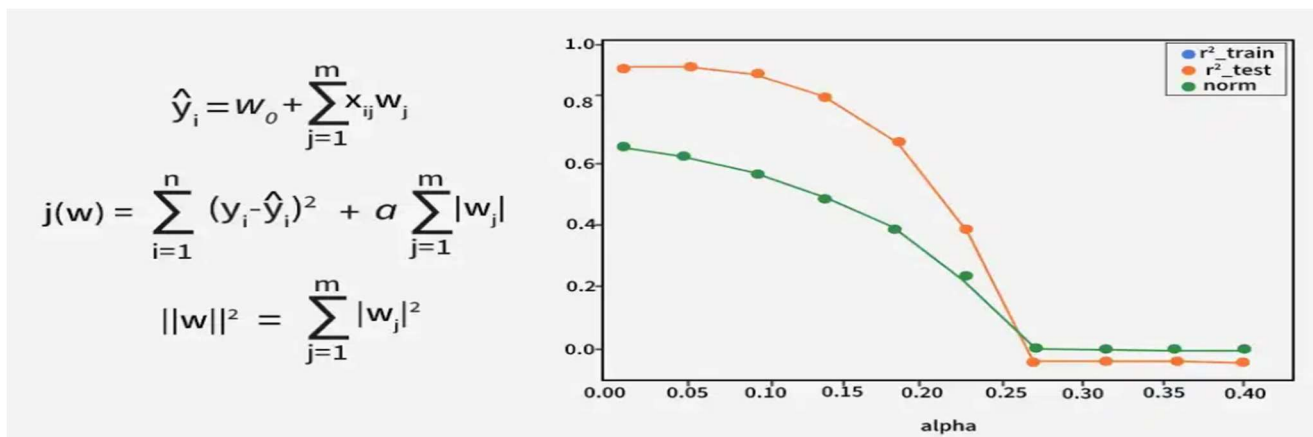


Figure 5 Lasso Regression (LR)

When to use Lasso Regression ?

Lasso Regression is helpful where:

Feature Selection: It selects most important features while setting the coefficients of less significant ones to zero.

- Collinearity: In case of multicollinearity, it assists us in getting rid of the coefficients to control and keeps one of the variables.
- Regularization: It works by penalising large coefficients, especially useful when you have a large number of predictors.
- Interpretability: As compared to classic linear models which has all the features Lasso regression produce a model with less number of non-zero coefficients, so it is easier to interpret.

Advantages of Lasso Regression

- No feature selection required: Most important features do not need to be identified manually, resulting in a simpler and more interpretable regression model.
- Regularization: It limits the large coefficients so that we obtain a less-biased model which gives stable and generalized predictions.
- Interpretability - This also builds an alternate models to make them more interpretable and explainable, which is very critical in domains like healthcare or finance.

Handles Large Feature Spaces: It works with high dimensional data such as images and videos.

Disadvantages:

- Selection Bias: Lasso might randomly choose one variable amongst many correlated variables, and as a result provide a biased model.
- Sensitive to Scale: It is sensitive to features that have different scales because values of such features can influence the regularization and accuracy of the model.
- Effect of Outliers: It can be greatly impacted by the outliers in the data which proves to overfit the coefficients.
- Model instability: It may unstable when there are a large number of correlated variables since the model is obtained by selecting features, which can vary with just small difference in the data.
- The Value of λ Selection: Tuning a number of λ values is not suitable but solvable unless for cross-validation.

5. Ridge Regression

Ridge regression (L2 regularization) is a linear regression method used to reduce the multicollinearity among predictors. Multicollinearity refers to the case of several independent variables (IVs) in a regression model being strongly correlated, which might incur coefficients suffering from non-robustness and instability. [21]

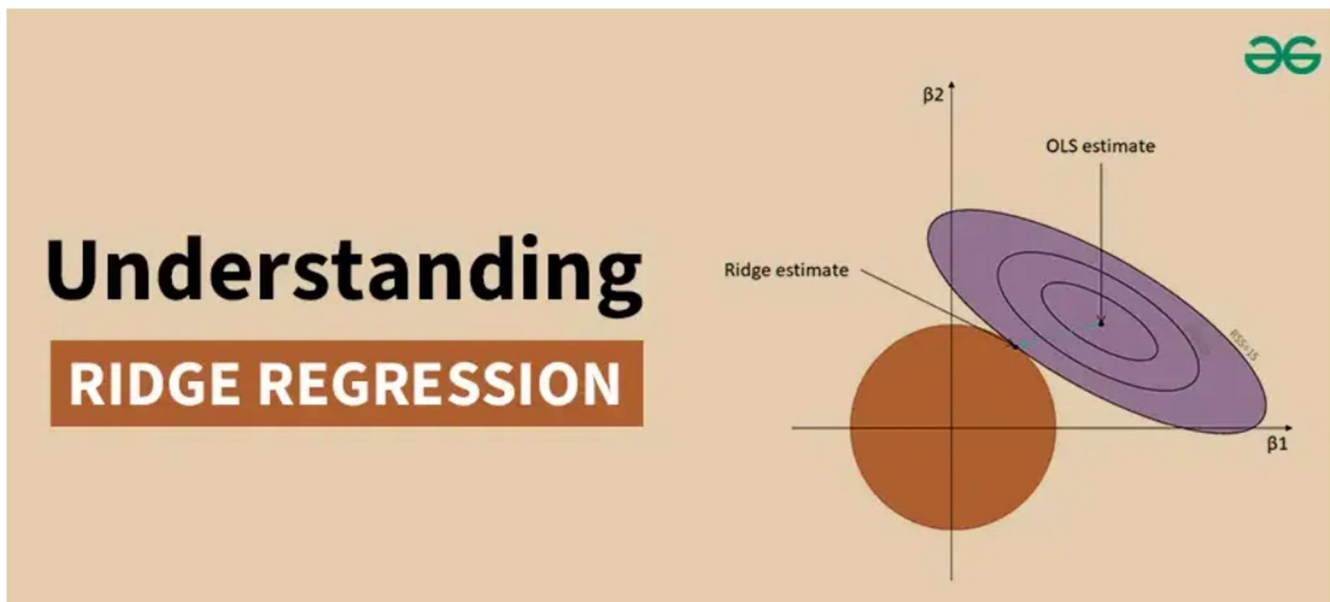


Figure 6 Ridge Regression

1

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

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]

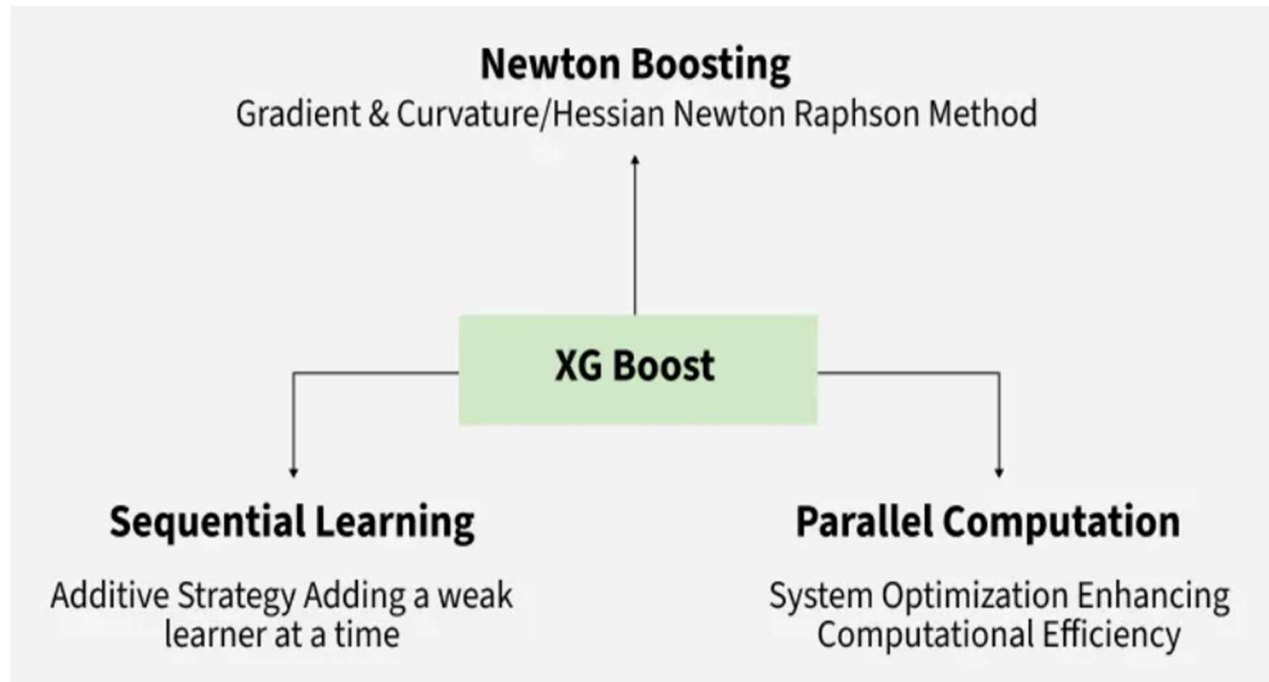


Figure 7 XGBoost Regression

in order to deal with big data sets, XGBoost leverages an approximate way in finding the best splits.

- Quantiles with weights: Make a quick guess of the ideal split without having to check all possibilities.
- Efficiency: Minimizes overhead of computations while preserving accuracy.
- Advantages : Suitable for large dataset where full evaluation is very expensive.

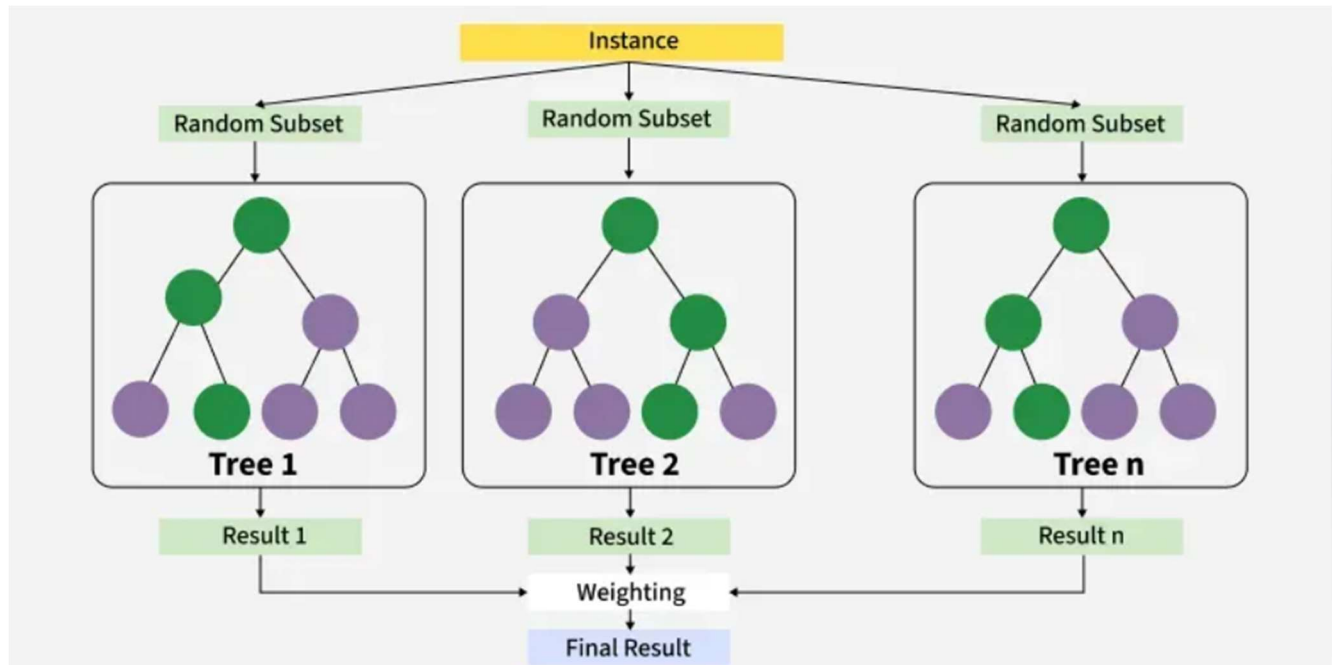


Figure 8 XGBoost Regression

How XGBoost Works?

It creates decision trees on working one after the other, where each new tree helps to let down the error of its predecessor. It can be decomposed as:

Begin with a base learner: It learns the tree model, which is first learning decision tree based on the data. In regression, this base model predicts just the average value of the target variable.

Measure the errors: After training of one tree calculate difference between actual values and predicted values.

Train the next tree: Train the next tree on the errors of previous tree ($y_2 - y_1$). This tries to correct for the mistakes that have been made by the first tree.

Continue reiterating this process: This sequence of operation will be continued with additional tree, each attempting to correct the errors made by its predecessor trees until a stopping rule is satisfied.

Aggregate the predictions: The final prediction is obtained by adding up the predictions of all trees.

Advantages of XGBoost :

- Scalable for large datasets of millions of record.
- Supports parallel processing and GPU, so everything runs lightening fast!
- Provides parameters and regularization that can be fine-tuned.
- Comes with feature importance analysis for more insights.
- Compatible with many languages and has a broad appeal to data scientists.

Disadvantages of XGBoost

- Computationally intensive; not appropriate for resource-constrained devices.
- Sensitive to the noise and outliers need careful pre-processing.
- Can overfit, especially with small datasets or too many trees.
- Interpretability may be limited relative to simpler models, which can be a concern in e.g. healthcare or finance.

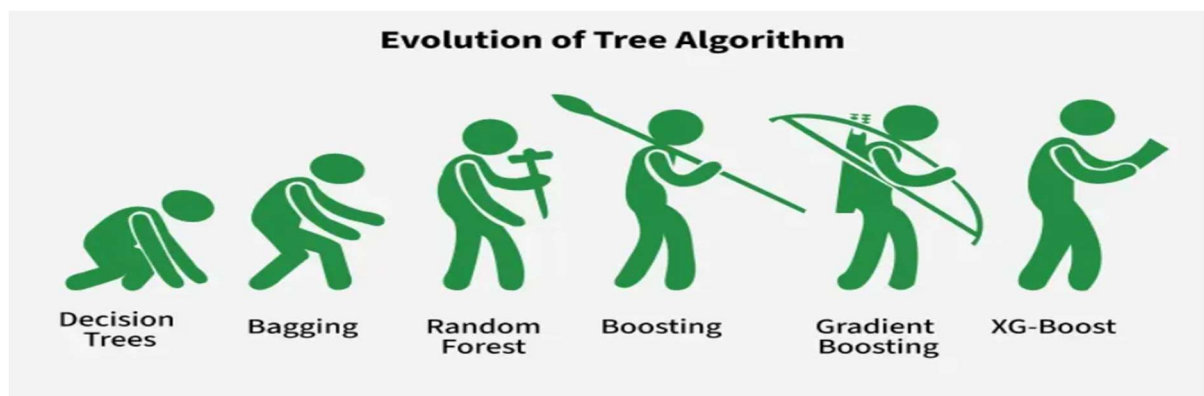


Figure 9 Tree Algorithm

Chapter 3 Related Work

3.1 Introduction:

There are persistent difficulties in aquaculture systems concerning the control of environmental conditions such as water availability and water quality for healthy growth of fish. Fish survival and growth are closely linked to water quality, but traditional monitoring methods involve a lot of manual work on a time-delay basis that are not adequate for early detection of water quality condition. Consequently, management consideration such as feeding regime, stocking density and pond managements are usually taken and will often be scant of specificity to dynamisms. One of the most significant constraints facing fishpond managers is the absence of continuous and high frequency data necessary for informed decision making. Farmers often rely on infrequent sampling, which that doesn't account for the ephemeral nature of water and limits the ability to take preventative action in regard to product safety. This lack of data has historically been the nemesis of predictive and data driven management in aquaculture.

The adoption of Internet of Things (IoT) technology offers new opportunities to enhance environmental monitoring within aquaculture environments. Using networked sensors and wireless communication technologies, IoT solutions allow for ongoing, automated monitoring of water quality data with high temporal resolution. The systems enable the monitoring of key environmental factors in real time, making possible prompt responses and better management decisions. However, even with real-time data being streamed, many of the current IoT use cases are preventative in nature as opposed to predictive.

Aquaponics is a form of sustainable agriculture that combines aquaculture with soilless plant farming, such as hydroponic cultivation, in one integrated system. In aquaponic fishpond nutrients made from fish waste are fed to plants which in turn purify water before it is returned to the fish tanks.

This synergistic interaction reduces the need for frequent water replacement and enhances resource efficiency compared to conventional fishpond systems. As a result, aquaponics has attracted increasing research attention as a promising model for environmentally sustainable aquaculture, particularly when combined with smart monitoring and data analytics technologies. To further optimize such systems, IOT-based monitoring can be implemented for real-time water quality management. [23]

7 Number 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

55 MS-CAGRU for Water Quality Prediction

11 **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]

73

41

8

78

Study 2

IOT & Advanced ML Algorithms (Tilapia Ponds, 2025)

Hridoy, M. A. A. M. et al. (2025) combined IoT monitoring and advanced machine learning to enhance water quality forecast in tilapia aquaculture on ponds. The authors used an in-house built IoT system to monitor the sensor data (temperature, dissolved oxygen, pH and turbidity) for six months from several tilapia ponds. Several predictive models were then compared on these time-series data, including tree ensembles and deep neural networks. It is noteworthy that the authors employed ensemble methods like bagging and boosting for better predictive performance. The models demonstrated a strong predictive power for the major water quality constituents through learning temporal patterns and employing multiple learners. The ensemble of the structure with decision-tree based method and boosting algorithm performed well, which could reach about 99% in predicting trends in dissolved oxygen fluctuations. This was significantly better accuracy than individual models could achieve (each individual model had 88%-93% accuracy), demonstrating the advantage of ensemble learning. It was observed that in the integrated viewpoint between IoT-sensed data and advanced ML algorithms, water quality can be predicted almost in real time with minimum error; it is necessary 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 installed an IoT system in a tropical fish pond to acquire environmental data (such as temperature, DO, pH and turbidity) continuously from the water, thus providing inputs to train prediction-based models. Water quality parameters were predicted using a random forest regression, yet the innovation of this study focused on using quantum-inspired optimization algorithm (QAOA) to optimize and bypass regular training complexity. The performance of the Random Forest model, optimized with QAOA's parameters showed an excellent fit for dissolved oxygen predictions ($R^2 \approx 0.999$) and a RMSE on the order of 0.1mg /L which demonstrated that the model was almost capable to perfectly predict based

upon observations; this level of precision is too high success as you can imagine. The implemented system permitted fast adjustments on-site, for example to switch on aerators to sustain survival of the fish at levels >90%. The most successful method found is QAOA-enhanced Random Forest, showing that combining IoT features and state-of-the-art optimization techniques you can obtain highly accurate predictive models for aquaculture monitoring. [3]

Study 4

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

Eze, E. et al. (2023) introduced a hybrid deep learning approach for multivariate prediction of water quality in an offshore salmon farm. They proposed a concatenated model, integrating the EEMD signal processing with an initial approximation using multivariate linear regression (MLR), and sequence learning with deep Long Short-Term Memory (LSTM) network. The model training approach was then applied to the prediction of harmful algal blooms (chlorophyll-a) and other water quality parameters from time-series sensor data collected within a Scottish salmon aquaculture site. The complex time-series were decomposed into simpler sub-series by the EEMD component, which enhanced the LSTM to learn patterns at multiple scales. The introduced EEMD–MLR–LSTM approach achieved superior prediction performance over several baseline models (including a single LSTM and stacked autoencoder-LSTM). Of particular interest, an extremely low RMSE of about 0.05 was realized in the chlorophyll-a concentration predictions, representing a very high precision. By efficiently recognizing the short-term dynamics and longtime trends inherent in the data, this mixed model reduced the prediction error as much as possible and was selected as the optimal one. The findings demonstrate the importance of multi-scale feature extraction and deep learning for predicting complex water quality dynamics in aquaculture. [4]

Study 5

37 Yu, X. et al. (2025) presented the prediction of dissolved oxygen in freshwater aquaculture through feature selection and attention-based RNNs. 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 oxygen dissolved several days into the future. 37 The attention layer enabled the model to weigh the most important 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 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

79 Xu, L. et al. (2025) developed a hybrid deep-learning approach for real-time prediction of dissolved oxygen in intensive aquaculture systems. Their model pipeline was a combination of several deep learning models stacked to learn different things about the time-series data. First, a one-dimensional CNN was adopted to get local temporal features of the sensor input sequences. Then, a self-attention was utilized for these features to assign different weight on different parts of the sequence adaptively. Finally, the sequence modeling is handled by a recurrent layer with Bi-SRU structure. The Bi-SRU, which is a modified form of the standard RNNs processed the sequence forward and backward to enable the model access both past and future context. The combined CNN-SA-BiSRU model was tested on an IoT dataset of an intensive fishpond in Guangzhou, China. The highest

accuracy was achieved in the short-term prediction of DO, with a MAE on the order of 0.03–0.04 mg/L and RMSE $\approx$ 0.047 mg/L; $R^2 = 98\%$ of the variance are explained by the model ($R^2 \approx 0.9765$). This was well beyond state-of-the-art performance of comparable traditional machine learning models such as random forest or simple neural nets. The authors found their CNN–SA–BiSRU hybrid to be the best model out of those trialled, confirming that combining convolutional feature extraction, selective attention and bidirectional recurrent units can provide accurate real time predictions in aquaculture. [6]

Study 7

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

Li, T. et al. (2022), did a comparative analysis over machine learning techniques for the estimation of multiple water quality parameters in an industrial fish farm system. train test literature and experimentally measured (DO, pH, ammonia, nitrate and nitrite) data among the intensive aquaculture tanks. Four modeling methods, i.e., BPNN (backpropagation neural network), RBFNN (radial basis function neural network), the standard SVM and S-LSSVM which was an improved version of least-squares SVM were employed to compare with each other. Performance was tested with all five types of water quality. The results show that Support Vector Machine (SVM) based models always yield the most accurate and stable predictions for all parameters. In particular, the prediction accuracy for water quality metrics was 99% based on SVM with almost perfect correlation coefficients (~ 0.99) between theoretically predicted and observed values. On the other hand, BPNN/RBFNN are more prone to local minima, and it is difficult for tuning the parameters in such data regime which further leads them to yield variable and lower levels of accuracy. The authors suggested that SVM was the better model for this task, given proper kernel choice and tuning, and they advocated using it for a dependable water quality prediction in such industrial aquaculture environments. [7]

Study 8

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

Iniyan Arasu, M. et al. (2024) proposed an IoT-based architecture “AQUASENSE” for intelligent pond managing in aquaculture and designed a deep-learning model for water quality classification. In AQUASENSE, a grid of independent water IoT sensors was installed in the ponds to constantly monitor variables (temperature, pH, oxygen dissolved and salinity). Information collected by these sensors was sent to the cloud for analysis. Authors created a custom deep learning model with dilated CNN-BiLSTM architecture to process the streaming data. The dilated CNN part extracted features at different temporal spans by taking advantage of the use of dilation convolutional filters and the BiLSTM (bidirectional LSTM) component captured information in both forward and backward directions to have a better understanding on sequence-wise context. It was the model's job to classify for water quality status (e.g., optimal, suboptimal), and alert if anything was awry. In our experiments, the accuracy of overall classification reached 96.98% in the AQUASENSE framework, far higher than those on the previous IoT-based aquaculture monitoring systems (~88–93%). Specifically, the extended CNN–BiLSTM model outperformed previous frameworks, such as ISAS and AquaStat by 3–8 percent point. The treatment performance by the sensitivity analysis (for this study's best model was provided by the combined AQUASENSE deep learning classifier, showing that custom neural network porting strategies can utilize IoT data for effective waterbody quality diagnostic and preventative management in aquaculture [8].

Study 9

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

Hemal, M. M. et al. (2024) proposed an automated pond management solution, AquaBot that uses both IoT based monitoring and machine learning ensembles to assist farmers with fish breeding choices. 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

56 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 top-performing model of the study (stacked ensemble) yielded consistent recommendations evidencing that IoT data combined with ML offers promise to automate decision-making in aquaculture, e.g., suggesting species that can survive based on the real-time water quality. [9]

Study 10

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

50 **Patro, K. S. K. et al. (2023)** proposed the use of IoT sensors and machine learning for fish behavior monitoring, in particular, to detect stress or health complications in ornamental fish through environmental variations. In an experimental condition fish (goldfish) maintained in a controlled tank were constantly monitored at different water temperature × dissolved oxygen scenarios, and corresponding environmental parameters were stored by means of an IoT platform. The fish behavior (e.g., normal swimming, tumbling, surfacing) was categorized and time-stamped using the sensor outputs. The researchers then used it to train four classifiers: a decision tree, Naïve Bayes, k-nearest neighbor (k-NN), and Linear Discriminant Analysis (LDA) to predict from the water indicators if the fish was stressed or not. Then we cross-validated the models and found that decision tree performed best with a total accuracy of around 86.2% (or Bayesian error to around 13.8%). It did better than k-NN, and Naïve Bayes, which had error rates that ranged from 28 to 30%. The decision tree's interpretable rules also revealed distinct environmental thresholds at which stress behaviors were exhibited (e.g., in the case of water temperature rising above 40.5°C or DO dropping below 6.6mg/L that fish displayed

erratic swimming or gasping). The best model produced by this study, a decision tree, therefore allowed for automatic identification of fish stress with high accuracy and demonstrated how IoT sensor data can support not only water quality management but also real-time animal welfare monitoring in aquaculture. [10]

Study 11

Dilated CNN + BiGRU w/ Attention (Indoor Mariculture, 2024)

Arepalli, P. G. & Naik, K. J. (2024) published a water quality prediction technique for indoor recirculating aquaculture using a dual attention based deep learning model. They built a multi-step DO prediction model using VMD (Variational Mode Decomposition), CNN-BiGRU network, and attention mechanism. First, the VMD was used on the raw sensor time-series to separate it into 'a set of components each capturing different frequency content' (e.g., diurnal patterns vs. quick changes). Each decomposed sequence was input to a CNN-BiGRU model, in which the 1D CNN layers extracted features and bidirectional GRU layers were used to model the temporal sequence bi-directionally. Also, a BiGRU optimized to emphasize significant timesteps was embedded an additional time attention and VMD quality attention for the detection of relevant portions of VMD, as well as training on highly selective components. In a validation experiment with laboratory indoor mariculture tank data, the proposed VMD-CNN-BiGRU-Attention model satisfied practical precision requirements for DO prediction. It resulted in consistently lower error than single-network models such as LSTM or a plain GRU, particularly for multi-hour forecasts. The predicted DO behavior simulated by the model was in good agreement with both measured DO data qualitatively, demonstrating its potentially suitability for online use. This model achieved the best performance among all models tested as it effectively exploited signal decomposition and attention to the trends and fluctuations of global and local levels. [11]

Study 12

Dual-Attention ResNet + LSTM (Offshore Aquaculture, 2025): Kunapinun., A. et al. (2025) who used data-driven models combined with their knowledge of the physics to improve predictions about the growth of biomass and water quality in IMTA. They proposed a hybrid modeling method which combined a Bi-LSTM network with physics-based constraints of aquaculture system dynamics. It was a multi-trophic system (fish/plant and any possible species) where the model should have predicted farming resources selection (biomass yield) and water key parameters. Through the inclusion of physical equations or some system invariants into the training of this model (usually referred as physics-informed neural network approach) Bi-LSTM was driven to provide physically sound predictions obeying basic constraints (mass balance, known growth kinetics). Using data from an experimental aquaculture system for training and testing, the physics-informed Bi-LSTM achieved better stability in long-term prediction than purely data-driven models. On one case, it resulted in correct biomass predictions during long time where classical Bi-LSTM models diverged. In general, the hybrid model Bi-LSTM+physics performed better in terms of prediction accuracy for both biomass [-] and overall water quality metrics, particularly improving long-term forecasts which are timescales that plain statistical models tend to fall apart. This indicates the advantage of combining machine learning with domain knowledge, because the best model on both established ensemble sources performed reliably, and its interpretations are consistent with aquaculture science. [12]

Study 13

Hybrid Multi-Scale Feature Model (DO Prediction, 2018): Li, C. et al. (2018) proposed a hybrid modelling framework to forecast dissolved oxygen in aquaculture pond, using multiple scale time series characteristics. They reported that the DO level in ponds has both short-term variations (in relation to, for example, aeration patterns or feeding) and long-term trends (relating to seasonal changes or biomass development). To convey this, the authors decomposed the initial DO time-series signal into several frequency components using methods such as wavelet transform or empirical mode decomposition. Each of the resulting components featured patterns at a unique temporal scale. These characteristics were used to train a deep neural network (they used an LSTM or CNN-LSTM hybrid) for predicting future DO concentrations. The multi-scale hybrid

model was evaluated using historical pond data and compared with classical single scale ones that were built directly from the raw time-series. The results indicated the superiority of using multi-scale data for short-term traffic flow forecasting: the hybrid model provided significantly higher prediction accuracy with a RMSE reduction by around 20–30% compared to a vanilla LSTM that did not use multi-scale input. This gain implies that the model was better at characterizing the complex dynamics of DO by combining information from fast and slow accelerations. Therefore, the best model was regarded as the combination between multiscale extraction of features and deep learning, which had better accuracy and stability in DO prediction for aquaculture. [13]

Study 14

Deep Belief Network (Recirculating Aquaculture, 2020) (Haqiq et al., 2013): Ren, Q. and Ullah, M.H Deep Belief Network based image Recognition of fish in Recirculating Aquaculture System using an aquatic vision sensor, vol. (2020) provided an overview of the application of deep learning for dissolved oxygen prediction in RAS and, more specifically, on DBNs. RAS systems are also heavily regulated and dynamic, thus DO predictions are necessary for automated aeration control. The authors' previous study proposed a DBN which was used as regression model to predict the DO concentrations for various inputs including water temperature, feed rate and aeration settings. As an example, a layered model like the DBN is trained (in unsupervised manner) to learn a hierarchical representation of data using unsupervised pre-training and fine-tuning, because it is composed of stacked restricted Boltzmann machine layers. The DBN compared well with a more standard alternative (simple neural networks and a mechanistic water quality model), showing faster convergence rates during training and superior predictive accuracy on test data. For example, in testing using data from marine fish RAS systems the DBN predictor achieved good coefficient of determination (R^2) and prediction error that was low implying that it had potential for closely following observed DO fluctuations. The authors added that the deep architecture of the DBN facilitated learning complex, non-linear input-DO relationships that are invisible to shallower model. This resulted in more accurate and reliable predictions, particularly of the short-term DO changes that were important for timely decision-making for modifications in aeration application. Hence, the DBN-based model was identified as the most effective model in this study, which also demonstrated a potential for deep learning models to support water quality control of RAS aquaculture. [14]

Study 15

Jamshidzadeh, Z. et al. (2024), a hybrid deep learning model was proposed for the prediction of water quality parameters in a coastal aquifer for aquaculture, incorporating the use of a novel metaheuristic optimization during training. Two hybrid models were proposed in this paper, one fusing a unidirectional LSTM network with Support Vector Machine regression (LSTM-SVM), and the other combining a Bidirectional LSTM with Support Vector Machine (Bi LSTM-SVM). To adjust the hyperparameters of such models and identify the most relevant input features, they used a recent nature-inspired algorithm called “Ali Baba and the Forty Thieves” (ABFT). The objective of the case study was to predict TDS and EC in groundwater close to a coastal aquaculture location on the northern coast of Iran, based on daily water quality readings from 2017-2020. The models were tested for one day, two days and three days ahead of the forecast. The experiments demonstrated that the Bi LSTM-SVM hybrid model performed favorably compared to all competing models in terms of prediction horizons. For 1-day-ahead forecast, it performed with low level of mean absolute percentage error (MAPE) – about 2% for TDS – and decent R^2 – above 0.94 in EC – which describe good predictive accuracy. As expected, drop was observed as prediction horizon increased over time but the BiLSTM-SVM outperformed alternative models (standalone LSTM, SVM and LSTM-SVM excluding bidirectionality), resulting in lower errors throughout the higher time spans. The authors find that the optimized Bi LSTM-SVM is their best model for the task, thereby showing that hybridization of learning algorithms as well as advanced optimization (ABFT) are capable to provide us with highly accurate water quality prediction tool in the form 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 $\approx$ 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 + BiLSTM DL

Smart aquaculture pond (multi-sensor)

Dilated CNN–BiLSTM

96.98% accuracy; +3–8% vs prior methods

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)

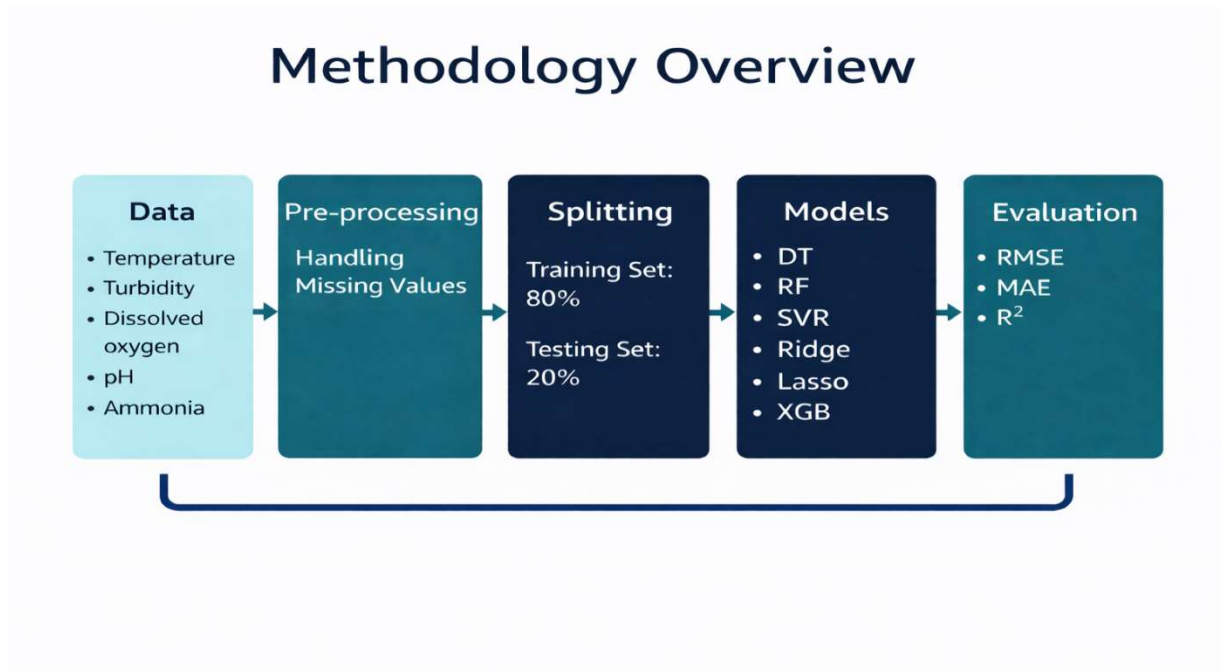


Figure 10 Methodology overview

4.1 Dataset Description

The dataset employed in this study is derived from an IoT-enabled fishpond monitoring system, referred to as the *Fishpond IoT Dataset*. The dataset consists of continuous measurements collected by Internet of Things sensors deployed within a smart aquaculture environment. These sensors are designed to monitor both physical and chemical characteristics of pond water, providing essential information for maintaining suitable conditions for fish growth. Also, this dataset consists of approximately 8300 rows and 6 columns.

4

The input features used in the predictive models represent key environmental parameters that influence fish health and development. These parameters include:

- Water Temperature (°C): A critical factor affecting fish metabolism, growth rate, and overall physiological activity.
- Turbidity (NTU): Reflects water clarity and suspended particles, which can influence light penetration and aquatic conditions.
- Dissolved Oxygen (DO): One of the most important indicators of water quality, directly affecting fish respiration and feeding behavior.
- pH Level: Represents water acidity or alkalinity, influencing biochemical processes and fish metabolism.
- Ammonia Concentration: Serves as an indicator of organic waste accumulation; elevated levels can be harmful or toxic to fish.
- Nitrate Concentration: High nitrate levels may degrade water quality and negatively impact fish development.

The output variables targeted in this research correspond to two fundamental indicators of fish growth:

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

These growth-related variables are predicted using the above environmental parameters through supervised machine learning regression models to analyze and forecast fish growth trends in IoT-based aquaculture systems.

91

4.2 Data Preprocessing

Before training models, a data preprocessing process was carried out to enhance the reliability of and accurate prediction in data. First, the dataset was scanned for missing or erroneous data points. Records with missing information were cautiously processed by reasonable methods (e.g., replaced statistically according to the corresponding characteristic distribution, or discarded) if needed.

These preprocessing stages were necessary to decrease noise and increase intra-habitat data consistency and reduce the effect of poor-quality data on model generalization.

4.3 Dataset Splitting

The dataset is separated into two subsets to facilitate objective evaluation and model development:

- Training Set: 80% of the available observations used for training the models.
- Test Set: 20% of the data to test model against new data.

The splitting was done by random sampling, and achievable with fixed random seed to make the experiment results reproducible and maintain model performance comparison fairness.

4.4 Regression Models

Various regression algorithms were developed for estimating the length and weight of fish as a function of water quality parameters.

1. Decision Tree Regression

The regression model of Decision Tree splits the data into subsets through hierarchical decision rules.

Advantages:

- Easy to interpret and visualize.
- Capable of modeling nonlinear relationships.

Limitations:

- Sensitive to noise and prone to overfitting.
- Performance depends heavily on tree structure.

1. Random Forest Regression

Random Forest is an ensemble learning method that aggregates multiple decision trees to improve prediction stability.

Advantages:

- Reduces overfitting through averaging.
- Provides strong and reliable predictive performance.

Limitations:

- Less interpretable than single-tree models.
- Increased computational complexity with more trees.

2. Support Vector Regression (SVR)

SVR applies kernel-based techniques to model nonlinear relationships within a defined error margin.

Advantages:

- Effective for nonlinear and high-dimensional data.
- Robust to outliers when properly tuned.

48

Limitations:

- Computationally expensive for large datasets.
- Requires careful selection of kernel functions and parameters.

3. Lasso Regression

Lasso regression employs L1 regularization to penalize model complexity by shrinking coefficients.

Advantages:

- Performs automatic feature selection.
- Reduces overfitting in high-dimensional datasets.

Limitations:

- Assumes linear relationships.
- May discard relevant features under strong regularization.

4. Ridge Regression

Ridge regression applies L2 regularization to stabilize coefficient estimates.

Advantages:

- Handles multicollinearity effectively.
- Produces stable and consistent models.

Limitations:

- Does not eliminate irrelevant features.
- Limited in capturing nonlinear interactions.

44

5. XGBoost Regression

XGBoost is a gradient boosting algorithm that builds decision trees sequentially to minimize prediction errors.

Advantages:

- High predictive accuracy and strong generalization.
- Efficient and includes built-in regularization mechanisms.

Limitations:

- Requires careful hyperparameter tuning.
- Less transparent compared to simpler models.

Overall Comparison:

- 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

46

Chapter 5 Experimental Results and Discussion

5.1 Introduction

This chapter provides detailed analysis of experimental outcome by deploying multiple regression-based machine learning models to predict fish growth indicators which are fish length and fish weight. It is implemented based on the environmental and water quality measurement of an IoT fishpond monitoring system.

The principal goal of these experiments is to investigate the prediction performance, stability and generalization capability of the chosen regression algorithms across three datasets using the same experimental setting. To give a fair and unbiased comparison, all models were trained and tested on one partition of the dataset with 80% samples for training and 20% testing.

The model performance was quantitatively assessed by means of three most common statistical measures, i.e., the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). These measures are complimentary and provide 'dual axis' views regarding the size of prediction error, sensitivity to large changes and the fraction of variance explained by each model, respectively. The results presented in this chapter are preliminary and serve as an initial step to compare model performance and determine the best approaches for predictive management of fishponds.

5.2 Experimental Setup

The experimental implementation was carried out using the Python programming language within the Jupyter Notebook environment. Several well-established data science and machine learning libraries were employed to support data processing, model development, and result visualization, including pandas and NumPy for data

manipulation, scikit-learn for machine learning algorithms and evaluation, XGBoost for gradient boosting models, and Matplotlib for graphical analysis.

40

Prior to model training, all input features were normalized to ensure consistent scaling across variables and to improve numerical stability during the learning process. This preprocessing step was essential to guarantee a fair comparison among models, particularly those sensitive to feature magnitude. Each regression model was trained under identical conditions, allowing the observed performance differences to be attributed to algorithmic characteristics rather than experimental bias.

5.3 Training Parameters:

- Train-test split: 80% – 20%
- Random state: 42 (for reproducibility)
- Scaler used: Min, Max Scaler
- 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 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

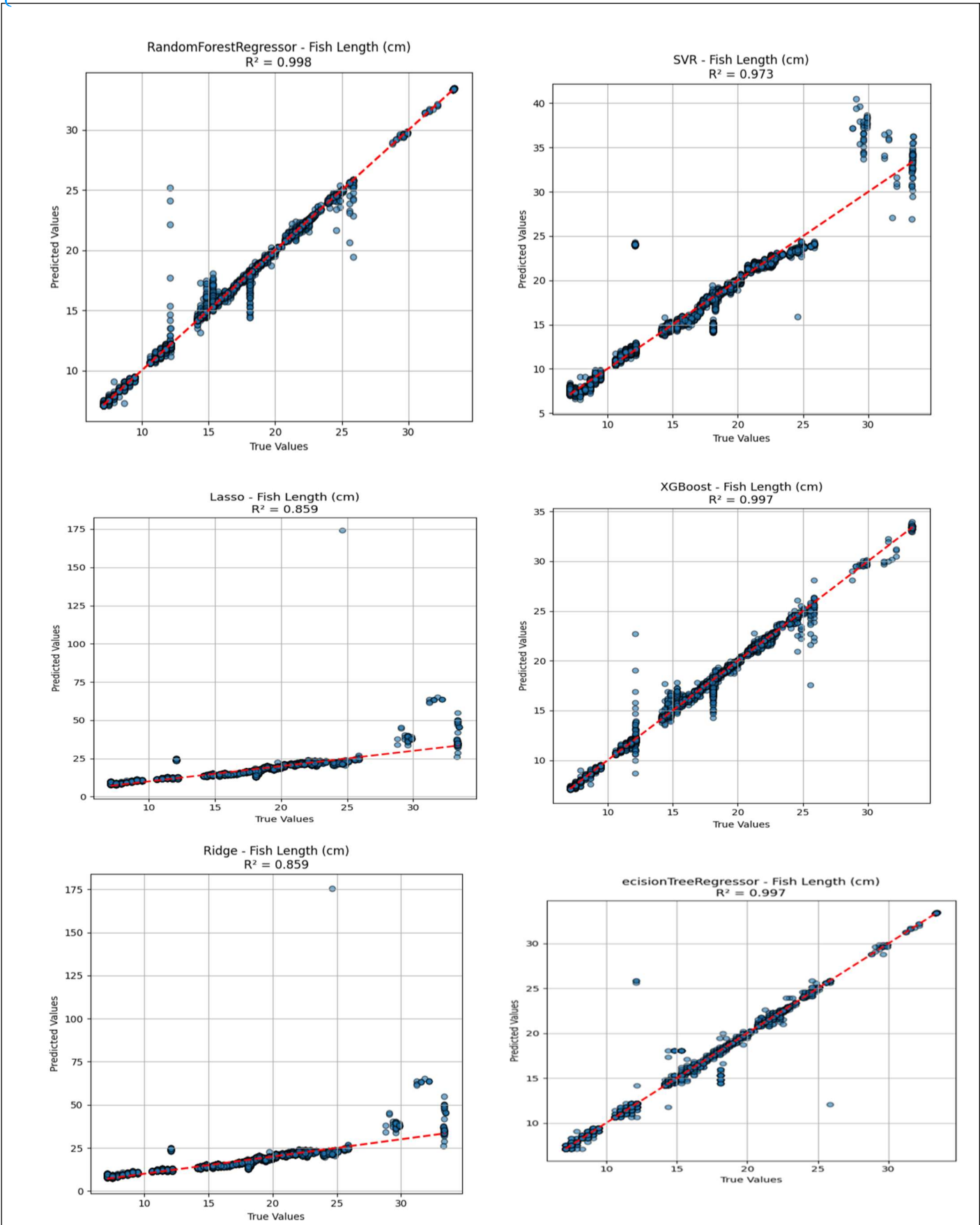


Figure 11 all algorithm Result Fish Length R^2

5.5 Results and Discussion: Predicting Fish Weight

Table 3 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

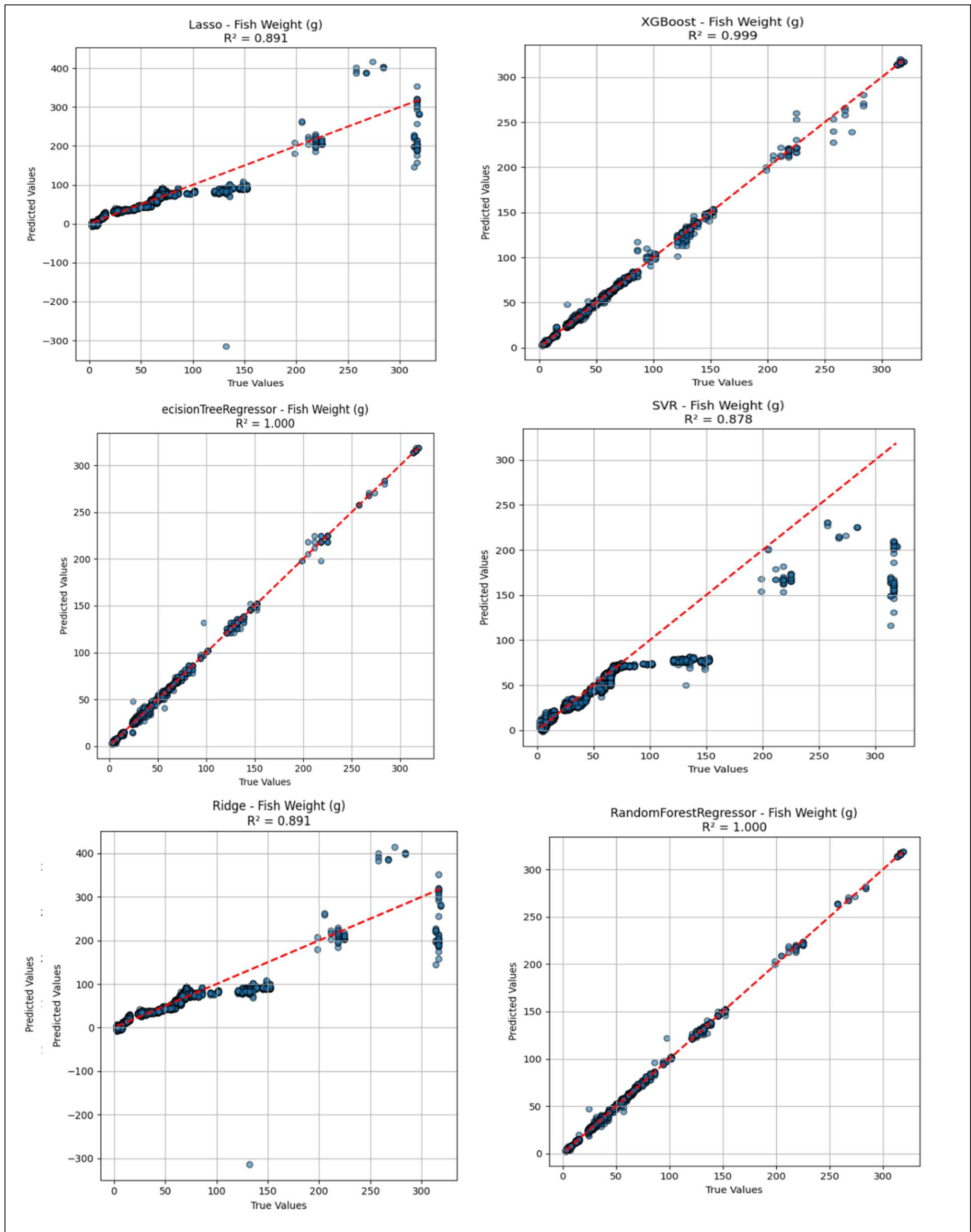


Figure 12 All algorithm Result Fish weight

5.6 Performance analysis and overview:

Furthermore, in the prediction of fish length, superior homogeneity and heterogeneous model-based approaches are discriminated to show up clear differences concerning predictive performance. Of these, the Random Forest model was consistently the most predictive and trustworthy. It is observed that very high R^2 values with lastly error metrics are estimates the interactions among environmental factors involved in the dataset nonlinearity. Much smaller variance and deviant prediction have been observed, which is mostly owing to the ensemble mechanism that a number of decision trees cooperatively facilitate the prediction under such circumstances, thus coving up the influence of noise existed in IoT sensor readings. The model's predicted curve can be seen closely following the actual observations, demonstrating low sensitivity to noise in the data, and high generalization ability.

XGBoost performed very well also, it was behind the Random Forest model in terms of accuracy. Its boosting strategy, iteratively adjusting the residual errors of previous iterations, made it possible to capture the subtle dependencies among water-quality factors. Nevertheless, some of the relatively higher error values (especially for mean absolute errors) indicate that it may still be sensitive to outliers or abnormal readings. This is clear from looking at the spread of predictions around the optimum fit line which is less tight as opposed to those clustered by the Random Forest model.

The base model from Decision Tree had modestly good prediction strength and less stability compared to ensemble models. This instability stems from using only a single tree shape for the model that can create exaggerated impact of local tiny diversity in data. Decision Tree predictions are usually correct trend but come with sharp turns in between because of their discrete nature of splitting. These behaviors illustrate the poor capability of single-tree models, in presence of complicated biological relationships.

By contrast, the predictive ability of the Support Vector Regression model for fish length was very week. The high values of error indicate that it was difficult for the model to include the nonlinear and interaction patterns on environment present in data base. In general, SVR shows higher classification power in the case of smooth and well-separated spectra; however, biological variability along with environmental fluctuation caused a degradation of its performance in this study. Its forecast values obviously differ from direct measurements, indicating more spread and less correlation to the ideal regression line.

The low performance of linear models, namely Lasso and Ridge regression was captured by all models. Their dismal performance is an explicit consequence of the fact that such

simple representations are fundamentally wrong since they impose linear-thinking tools on a process which is intrinsically non-linear, i.e. fish growth. As these model types cannot represent relations between water-quality parameters by means of interaction or multiplication, the models thus result in excessive errors and poor predictive power. The residual patterns observed in these models exhibit evident systematics, thus indicating that the underlying biological process cannot be fully characterized by linear functions of transferrin receptors.

In general, the length-at-age fits for fish illustrate that selecting among a suite of models capable of representing complex ecological dynamics is imperative. The ensemble models, in particular Random Forest one can better cope with this since they are more flexible, strong adapted and good at capturing patterns and variations across decision boundaries. Such results emphasize the importance to apply sophisticated nonlinear algorithms when facing multi-dimensional aquaculture datasets.

Regarding He.D's fish weight prediction, the models exhibited even more clear-cut differences between them, mainly because of the greater biological variability and complexity associated with this attribute as opposed to length. As before, the Random Forest model reported the best performance, with just a slightly off fit of prediction versus actual. This very high level of precision indicates that the model was properly translating temperature, DO, pH and other environmental parameters that affect biomass accumulation rate. Indeed, the model's predictions were quite close to the benchmark line, indicating it had done a fantastic job in terms of an overall range of observed values. While XGBoost showed high performance and good accuracy, a comparison of its and Random Forest results became clearly distinct when we made predictions regarding weight. The higher mean absolute error suggests that XGBoost was more sensitive to perturbations in the data, which are likely due to rapid internal and external conditions occurring within growth. Nevertheless, the model presented has good predictive capabilities and remains an excellent candidate for taking complex aquaculture causes into account.

The predictions of the Decision Tree model were fairly accurate, but there were some graphical disparities since the ensemble averaging was not employed. Although the rough trend of weight distribution was captured by the model, predictions made were still relatively close to actual values and some deviation could be observed which pointed out that our model was sensitive to local changes in data as well. This result is not so surprising because single trees are not designed to eliminate fluctuations or handle the discontinuous nature of interactions at different spatial scales in the fish biomass.

The Support Vector Regression approach, which yielded little success for predicting fish length, was less successful in estimating fish weight. Results showed large error values, indicating that the model could not capture the broader biological variability observed in weight data. Thus, it becomes difficult for SVR to represent the broad changes and non-linear growth characteristics, leading to severe fluctuation between its prediction and observed values. This discrepancy highlights that margin-based approaches have difficulties in handling complex aquatic growth systems.

Lasso and Ridge regression again performed poorly, further supporting their limited capacity for predicting biological growth variables in an aquaculture setting. Figure 1B and D The protein to water ratios Each of these models are linear, so they do not represent complex processes by which water quality interacts with physiological responses that ultimately regulate weight. Therefore, they made a lot of under fitted predictions leading to high errors and low explanative power.

Overall, results for fish length and weight, both suggest that ensemble learning techniques should be preferred with respect to other attempts in modeling aquaculture-related responses. This is even more advantageous for outcome variables that have a high amount of variability, like weight. Model-wise, Random Forest was generally the best performing model by providing higher accuracy, more resistant to noise and balanced interpretation of complex feature relationships. These results underline the necessity of using sophisticated machine learning models that may handle environmental uncertainty and biological complexity in IoT-based aquaculture systems.

5.7 Comparative Analysis

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

Table 4 Table 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.

45

Chapter 6 Conclusion and Future Work

60

9

68

32

6.1 Conclusion

This thesis has published a story that covers the implementation of Internet of Things (IoT) in combination with machine learning to improve monitoring, forecasting and decision support in aquaponics-based fishpond system. The research was undertaken with the development and implementation of an IoT monitoring system, based on an ESP-32 microcontroller, and sensors that measured essential water quality conditions such as temperature, pH; dissolved oxygen (DO); turbidity; Ammonium (NH_4^+); Nitrate (NO_3^-) at a minimum. Real-time access to pond conditions were made possible through continuous data collection and cloud storage, with a high-fidelity dataset for advanced analytics.

Significance One of the main contributions of this work is offering structured and dependable IoT-based data, addressing an enduring problem in aquaculture research regarding the lack of continuous, high-quality time-series information. The dataset offers a solid basis for predictive modeling and operational analysis, providing researchers as well as practitioners with new possibilities to understand pond dynamics and management.

10

22

10

This data foundation was leveraged by the study to conduct a systematic machine learning workflow for predicting fish growth indicators, namely fish length and weight according to environmental parameters. Standard performance measures were used to train, test and compare multiple regression models. The experimental results showed the ensemble learning methods, such as Random Forest and XGBoost often performed better in terms of prediction accuracy and generalization capability. This advantage is owed to their capability of simulating nonlinear relationships and interdependencies among the water quality variables typically found in aquaculture ecosystems.

The results of this study provide further evidence in favor of the larger body's literature showcasing IoT-based predictive analytics for aquaculture. This work demonstrates potential for proactive aquaculture management in terms of both fish health, operational risks and production efficiency by moving from the current state of real-time monitoring towards data-driven forecasting. The results, however, suggest that there is still room for growth in the development of intelligent aquaculture systems especially in scalability, interpretability and system integration.

6.2 Future Work

Although this work makes significant contributions from technical and scientific aspects, some open issues are left to be studied in future:

Scale and Diversity of the Dataset:

There is a need to use data from more farms, species and seasons for construction of more robust and generalizable models.

Integration of Additional Sensing Modalities:

Optical sensors (e.g., vision-based systems), sonar or acoustic sensors, flow and aeration sensors, and biosensors for detection of pathogens should also be added to achieve full monitoring capability in aquaculture.

Advanced Deep Learning Models Used:

Time-based models (e.g., LSTMs, GRUs, Transformer architectures) may be good options to increase long-term prediction accuracy when used in conjunction with hybrid or physics-informed methods.

Transition Toward Autonomous Control:

Future generations of systems could progress from prediction to automatic control actions like adaptive aeration, optimized feeding routines, or predictive water exchange policies.

59

Explainable Artificial Intelligence (XAI):

Transparency via explainability tools such as SHAP and LIME can improve the user trust which is critical for deployment.

Economic and Operational Optimization:

Combining environmental forecasting with economic models (for example, feed price, energy use and market price) would facilitate complete decision support systems.

Edge–Cloud Hybrid Architectures:

Edge computing in devices like ESP-32 can reduce latencies for mission-critical decisions, and cloud resources utilized for largescale analysis.

Unified Smart Aquaculture Platforms:

The goal is to build a society consisting of IoT sensing, AI-driven prediction, image analysis, cloud dashboards and automatic actuators in single intelligent aquaculture.

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?srsId=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 Agriculture* 5 <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. [Integrating Aquaculture And Hydroponics A Review Of Aquaponics Systems And Their Sustainability](https://www.researchgate.net/publication/382255038)
24. www.kaggle.com/dataset/e81da8b7666dc7af41cdc3aa5ef96c5547e4f412598a030f40d444550965e34f
25. <https://data.mendeley.com/datasets/dgdr2kfby/1>
26. <https://pmc.ncbi.nlm.nih.gov>.
27. <https://www.kaggle.com/datasets/e81da8b7666dc7af41cdc3aa5ef96c5547e4f412598a030f40d444550965e34f?select=IoTpond1.csv> **(Dataset)**