Anomaly Detection and Predictive Analysis in Aquaponics Fish Farming Using Sensor Data

By

Jabed Hossain 23057838

SUPERVISOR

Mohammed Odeh

This thesis is submitted to the
School of Computing and Creative Technologies
of the University of the West of England (UWE)
in partial fulfilment of the requirements for the degree of
Master of Science in Data Science

University of the West of England
Frenchay, Bristol

[September 2024]

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my parents, whose unwavering love and support have made it possible for me to pursue my master's degree. To my younger brother and sisters, your constant encouragement has always been a source of inspiration for me. I am also profoundly grateful to my cousins, aunts, and uncles, whose belief in me has never wavered.

In my academic journey, I owe an immense debt of gratitude to my supervisor, Dr. Mohammed Odeh. His guidance, patience, and dedication have been invaluable throughout this research. Dr. Odeh has generously devoted countless hours to mentoring me, helping me refine my research focus, and clarifying my vision. I am incredibly fortunate to have had over ten meetings with him, some lasting more than an hour, each one filled with insightful advice and unwavering support. Without his mentorship, this work would not have been possible, and I am eternally grateful for his commitment to my success.

I would also like to extend my heartfelt thanks to Paul, the course module leader of Data Science. His belief in my abilities, his willingness to listen to my questions, and his encouragement have been instrumental in shaping my project. Paul's support in helping me align my research with my interests and strengths has been invaluable.

A special thank you goes to my friend Jan, whose friendship has been a true blessing. As an international student, having a friend like Jan has been incredibly fortunate. He has been by my side, encouraging and inspiring me whenever I faced challenges. His unwavering support has made a significant difference, and I am truly grateful for his companionship and motivation throughout this journey.

To all who have contributed to this journey, I extend my sincerest thanks. Your support has made this achievement possible, and I will carry your belief in me as I continue forward, in sha Allah.

ABSTRACT

This project investigates the application of machine learning techniques for anomaly detection and predictive analytics in aquaponics systems, with a focus on key water quality parameters. Initial exploratory data analysis identified significant trends and irregularities, revealing correlations between water quality features and fish growth. LSTM, GRU, and Hybrid LSTM with autoencoder models were selected and meticulously tuned using k-fold cross-validation and grid search for hyperparameter optimization. The models were evaluated and the LSTM with autoencoder was found to outperform others, although the GRU model also demonstrated strong potential. Predictive analytics was further enhanced by using a custom Conv1D-based window generator, which effectively captured temporal dependencies in the data. Efforts at temperature prediction, where the models achieved notable success. The findings highlight the robustness of LSTM-based models in complex environments, offering valuable insights for optimizing aquaponics operations. The research underscores the potential for machine learning to enhance predictive capabilities, ensuring better monitoring and management of aquaponics systems.

Key Words: [Aquaponics, Machine learning, Prediction analysis, Anomaly detection, Correlation].

Contents

Table of Contents

List of Figures

List of Table

Chapter 1: Introduction

1.1 Introduction

The introduction of this dissertation addresses the increasing importance of sustainable agricultural practices in response to global challenges like climate change and resource scarcity. However, the complexity of managing these systems poses significant challenges. This research investigates how Artificial Intelligence (AI) and the Internet of Things (IoT) can optimize aquaponics management, particularly through anomaly detection and predictive analysis, to enhance efficiency and sustainability. The dissertation includes a literature review, methodology, and evaluation of the research outcome, and concludes with findings and implications.

1.1.1 Importance of the study of Aquaponics System

Traditional farming methods are increasingly vulnerable to extreme weather, resource scarcity, and urbanization, which exacerbate food security issues and drive the need for sustainable agricultural practices. For example, the 2018 heatwaves led to crop failures and up to 50% yield reductions in Europe, underscoring the impact of climate change on conventional farming [1]. With urban populations expected to rise by 50% by 2045, the demand for more food with fewer resources is intensifying [2], further strained by water depletion, deforestation, soil degradation, and greenhouse gas emissions [3]. Aquaponics, a sustainable system that combines aquaculture and hydroponics, offers a viable solution. By producing both fish and vegetables with minimal water and land, and being adaptable to indoor environments, aquaponics is more resilient to climate change. Therefore, studying aquaponics is essential for advancing sustainable agriculture and developing climate-resilient food production systems.

1.1.2 Current practices

Current practices in aquaponics are increasingly leveraging the Internet of Things (IoT) and Artificial Intelligence (AI) to overcome the complexities of managing these systems. Traditionally, monitoring parameters like dissolved oxygen, ammonia, pH, and temperature has been a manual and time-consuming process, requiring expertise across multiple disciplines. However, IoT devices now enable real-time monitoring and automation, reducing the need for constant oversight.

Machine learning algorithms have been applied to identify plant health issues, optimize feeding schedules, and detect anomalies in water quality [4]. Despite these advancements, the use of IoT and AI in aquaponics is still in its early stages. Most current research focuses on visual observations using machine vision and image processing, with less attention given to data from IoT sensors [5]. Moreover, existing studies often focus on a limited set of parameters, chosen based on sensor availability rather than a comprehensive understanding of the needs of aquaponics systems [6].

1.2 Problem Statement

In the context of this research, the problem is centered on the complexities and inefficiencies in managing aquaponics systems. Despite the growing interest in aquaponics as a sustainable agricultural practice, managing these systems remains challenging due to the need for constant monitoring and control of various water quality parameters. Traditional methods are not only labour-intensive but also prone to human error, which can lead to suboptimal conditions and reduced productivity [3].
The advent of IoT and AI technologies presents an opportunity to address these challenges, but their application in aquaponics is still in its nascent stages. Current research has primarily focused on visual data analysis through machine vision, with limited exploration of the vast potential offered by IoT sensor data [5]. Furthermore, existing studies often overlook the comprehensive integration of multiple critical parameters necessary for optimal system performance [6].
Therefore, this research seeks to fill this gap by applying AI-driven anomaly detection and predictive analysis to aquaponics systems. The aim is to develop methodologies that can enhance the efficiency and productivity of these systems, making them more resilient and easier to manage.

1.3 Aim and Objective

1.3.1 Research Aim

The primary aim of this research is to advance the field of aquaponics by developing and implementing innovative AI-driven systems that enhance water quality management and fish production. Specifically, the research seeks to achieve four key goals:

1. **Anomaly Detection in Water Quality**:

The research aims to create anomaly detection model for critical water quality parameters, such as temperature, turbidity, dissolved oxygen, pH, ammonia, and nitrate levels.

2. **Predictive Analysis for Water Quality Parameters:**

Another core aim is to construct predictive models using machine learning algorithms to forecast key water quality parameters in aquaponics systems. By anticipating changes in these parameters, the need for constant sensor monitoring can be reduced, thereby lowering operational costs while maintaining optimal water conditions.

3. **Understanding the Impact of Water Quality on Fish Growth**:

The research seeks to analyze the collected sensor data to gain deeper insights into how various water quality parameters influence fish growth and population dynamics. These insights are intended to inform better management practices, leading to improved fish production rates and overall quality.

1.3.2 Research Objective and Approach

The primary objective of this research is to apply advanced AI techniques to analyze data from aquaponics systems, focusing on detecting anomalies and predicting critical water quality parameters. This study aims to contribute valuable insights and tools to optimize aquaponics systems, enhancing both fish production and water quality management.

1. **Anomaly Detection in Aquaponics Systems**: AI-driven methods will be employed to detect anomalies in key parameters like temperature, turbidity, dissolved oxygen, pH, ammonia, and nitrate levels, ensuring timely interventions to protect fish health.
2. **Predictive Analysis of Water Quality Parameters**: Machine learning techniques will be used to forecast critical water quality parameters, reducing the need for constant sensor monitoring and improving system efficiency.
3. **Analyzing Water Quality and Fish Growth**: The study will explore the relationship between water quality parameters and fish growth, providing insights to optimize fish production.

These objectives will be achieved through data analysis, machine learning algorithms, and statistical methods, ensuring that the findings are actionable and directly applicable to improving aquaponics performance.

1.3.3 Research Questions

To guide this research, I've formulated a series of key questions that align closely with the project's objectives and aims. These questions are designed to explore various aspects of water quality management in aquaponics systems, using advanced AI-driven techniques.

1. **How can we accurately detect anomalies in water quality parameters within aquaponics systems?**

This question focuses on the development of an anomaly detection system. Specifically, it aims to determine the most effective methods for identifying deviations in crucial water quality factors such as temperature, turbidity, dissolved oxygen, pH, ammonia, and nitrate levels.

2. **What predictive models can be developed to forecast critical water quality parameters, and how can these models help reduce the need for constant monitoring?**

This question explores how machine learning can be used to predict key water quality parameters in real-time, potentially reducing the dependence on continuous sensor monitoring.

3. **How are various water quality parameters related to fish growth, and what insights can we gain to improve fish production rates and quality?**

This question seeks to understand the connections between different water quality parameters and their impact on fish growth and overall health.

These questions are not just theoretical—they are directly tied to the practical outcomes of this research. By addressing each one, the study aims to contribute valuable knowledge and tools to the field of aquaponics, helping to optimize water quality management and enhance fish production.

1.4 Ethical, Legal and Professional Consideration

This research adheres to ethical, legal, and professional standards, ensuring responsible execution in line with best practices. The data used, sourced from the "Sensor Based Aquaponics Fishpond Datasets" on Kaggle, is licensed under the Attribution 4.0 International (CC BY 4.0) license, allowing sharing and adaptation with appropriate credit to the original creators [3].

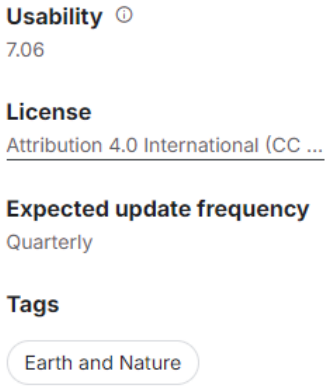


Figure 1: License, ATTRIBUTION 4.0 INTERNATIONAL

All attributions are properly made in compliance with this license. Proper credit is consistently given to the dataset's creators, respecting their intellectual contributions as required by the CC BY 4.0 license.

1.5 Online repository for the codes

All code developed for this research was created using the Anaconda Navigator Desktop application, utilizing Python as the primary programming language. To ensure transparency, reproducibility, and collaboration, the code has been made publicly available in a GitHub repository. Anyone can access the code and explore the project through the following link: [GitHub Repository Link].

1.6 Structure of Dissertation Report

This dissertation is structured to guide the reader through the research journey, from the exploration of existing literature to the final conclusions and recommendations. It begins with an Introduction that presents the research topic, its significance, and the objectives that form the study's foundation. The Literature Review follows, examining current practices, challenges, and advancements in aquaponics, while identifying knowledge gaps this research aims to address.

The Methodology section details the research design, data collection, and analysis methods, explaining their alignment with the research objectives. The Data Collection and Preprocessing sections outline the dataset's preparation, including handling missing data and outliers. Data Analysis and Feature Engineering then explore the dataset and enhance model performance, leading into Predictive Analysis and Anomaly Detection, where models are developed and evaluated.

The dissertation also includes Correlation Analysis, examining the relationships between water quality parameters and fish growth. Finally, the research findings are summarized, with recommendations for future work presented in the Conclusion, reflecting on the achievement of research objectives. Each section builds on the last, creating a cohesive narrative throughout the study.

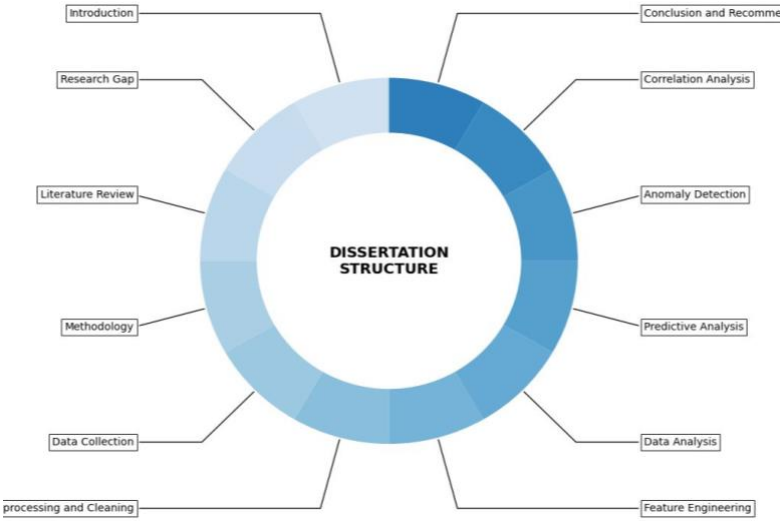


Figure 2: Structure of the Dissertation

Chapter 2: Literature Review

2.1 Introduction

This literature review explores the current state of research in aquaponics, particularly focusing on the integration of AI and IoT technologies. The review aims to identify the key areas where these technologies have been applied and to highlight the gaps that still need to be addressed to optimize aquaponics systems fully.

2.2 Overview of Aquaponics

Aquaponics is a sustainable agricultural system that combines aquaculture (the raising of fish and other aquatic organisms) with hydroponics (the cultivation of plants without soil) into a single, integrated ecosystem. In this symbiotic system, the waste produced by the fish serves as a nutrient-rich fertilizer for the plants. In return, the plants help to purify the water, which is then recirculated back to the fish tanks. This creates a sustainable cycle of growth that minimizes the need for external inputs, such as synthetic fertilizers, and reduces water usage compared to traditional farming practices [7].

2.2.1 Aquaponics Systems

Aquaponics is an innovative method combining aquaculture and hydroponics, allowing fish and plants to grow together in a symbiotic environment. Originating from agricultural practices in China over 2000 years ago [8], modern aquaponics was introduced by Dr. James Rakocy and his team in the 1980s at the University of the Virgin Islands [9]. In these systems, fish waste, rich in ammonia, is cycled through a biofilter where nitrifying bacteria convert the ammonia into nitrates, which plants absorb. The plants then filter the water before it is recirculated back to the fish tanks.

Aquaponics systems can be broadly categorized into two types: Coupled Aquaponics Systems (CASs) and Decoupled Aquaponics Systems (DASs).

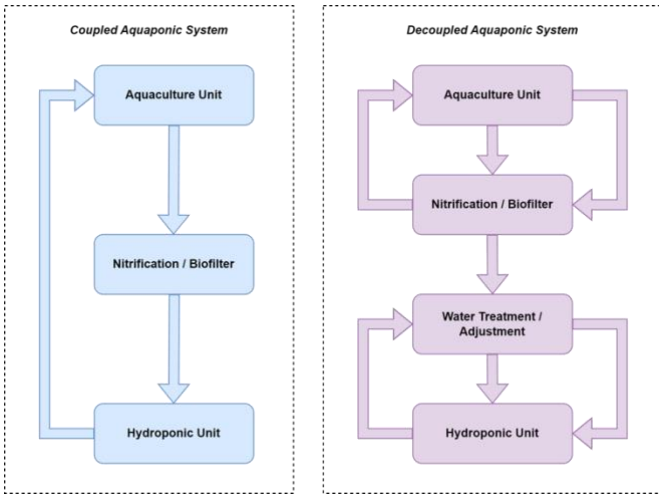


Figure 3: Aquaponics System [7]

Coupled Aquaponics Systems (CASs): In CASs, water flows in a single loop, moving from the fish tank to the hydroponic unit (where the plants are grown) and then back to the fish tank. This design is straightforward and easier to manage, making it a popular choice for smaller-scale or beginner setups. [10].

Decoupled Aquaponics Systems (DASs): On the other hand, DASs separate the fish and plant systems, allowing each to operate with its own water quality parameters. In a DAS, the water from the

fish tank is treated before being delivered to the plants, which allows for greater control over the conditions in each part of the system [11][12].

2.2.2 Studies on Aquaponics:

Modern aquaponics systems were first formalized in the 1980s by Dr. Wijayantoocy at the University of the Virgin Islands, establishing the foundation for integrating aquaculture with hydroponics [13]. Subsequent research has explored the comparison between Coupled Aquaponics Systems (CASs) and Decoupled Aquaponics Systems (DASs). For instance, Kloas et al. (2015) introduced the Double-Recirculation Aquaponics System (DRAP), which optimizes conditions for both fish and plants by separating their environments, while Suhl et al. (2016) found that DASs could achieve a 40% higher tomato yield compared to traditional hydroponics [11][12].

Research on nutrient dynamics in aquaponics has shown that supplementing systems with additional nutrients can boost plant growth by up to 39%, as noted by Delaide et al. (2017) [14]. Furthermore, Monsees et al. (2017) highlighted the importance of precise pH and nutrient control, which resulted in a 36% increase in fruit yield [10].
Recently, the integration of IoT and AI technologies in aquaponics has gained momentum. Karimanzira and Rauschenbach (2019) applied AI techniques, such as convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks, for monitoring and anomaly detection. However, much of this research has focused on visual data rather than fully integrating IoT sensors [15]. Additionally, Wijayanto et al. (2018) emphasized the need for comprehensive studies that consider all essential variables within the system, as most research has focused on a limited set of parameters [5].

2.2.3 Application of AI in Aquaponics

Artificial Intelligence (AI) is increasingly being integrated into aquaponics systems to enhance their efficiency, productivity, and resilience. By leveraging AI, researchers and practitioners are able to automate complex processes, analyze large datasets, and optimize system performance, addressing many of the challenges that traditionally hindered the scalability and reliability of aquaponics.

2.2.4 Predictive Analytics

Several studies have demonstrated the potential of predictive analytics to enhance aquaponics management by anticipating changes in water quality and optimizing resource use. For example, Abbasi et al. (2019) utilized machine learning algorithms to predict plant health and optimize growth conditions, demonstrating significant improvements in yield and resource efficiency [16]. Similarly, Karimanzira and Rauschenbach (2019) applied predictive models to forecast ammonia levels in aquaponics systems, allowing for preemptive adjustments to prevent toxic conditions for the fish [15]. These studies highlight the growing interest in using predictive analytics to create more resilient and sustainable aquaponics systems.

2.2.5 Anomaly detection

Anomaly detection is another critical application of AI in aquaponics, aiming to identify deviations from normal operating conditions that could indicate system failures or suboptimal performance. For instance, John and Mahalingam (2020) implemented a machine learning-based anomaly detection system that could identify abnormal patterns in water quality data, thereby enabling timely interventions to prevent system failures [16]. Additionally, studies by Karimanzira and Rauschenbach (2019) used convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks to detect anomalies in plant growth and water parameters, offering a more automated and scalable approach to system monitoring [15].

2.3 Research Gap Analysis

Despite advancements in applying AI and machine learning to aquaponics, several critical gaps remain in the research. Most studies focus on individual variables like temperature or ammonia,

neglecting the complex interdependencies among various water quality parameters, which are crucial for maintaining optimal conditions [17][18][19]. Additionally, AI-driven approaches have largely been tested in small-scale, controlled environments, raising concerns about their scalability and applicability in larger, commercial systems [23][24]. There is also a lack of studies that fully utilize IoT sensor data to provide real-time, actionable insights across entire systems, which could enhance system management [20][21]. Furthermore, the economic feasibility of implementing AI and IoT in aquaponics is underexplored, highlighting the need for research that considers the financial implications for broader adoption [22]. Addressing these gaps is essential for advancing aquaponics as a sustainable and scalable agricultural practice.

2.4 Summary

In summary, while the application of AI and IoT in aquaponics has demonstrated substantial potential, current research is largely focused on specific parameters and small-scale systems. The literature highlights several critical gaps, particularly in the integration of multiple variables, the scalability of AI models, and the full utilization of IoT sensor data. Addressing these gaps is crucial for enhancing the efficiency, sustainability, and scalability of aquaponics systems. This research aims to contribute to filling these gaps by developing comprehensive, multi-variable predictive models and anomaly detection systems applicable in both experimental and commercial settings.

Chapter 3: Methodology

3.1 Introduction

This section introduces the methodological approach taken for the study. The overall goal is to prepare the data for machine learning models that will be used to monitor and optimize the aquaponics system. The methodology includes data collection, preprocessing, feature engineering, and model building.

3.2 Overview of the Research Method

The research method is a combination of data-driven techniques, where data was collected from IoT sensors installed in aquaponics fish ponds. The process involves cleaning, transforming, and analyzing the data to develop predictive models using machine learning. The methodology is designed to ensure data integrity and improve the accuracy of predictive analytics, focusing on water quality monitoring in aquaponics systems.

3.3 Research Design Process

The research method is a combination of data-driven techniques, where data was collected from IoT sensors installed in aquaponics fish ponds. The process involves cleaning, transforming, and analyzing the data to develop predictive models using machine learning. The methodology is designed to ensure data integrity and improve the accuracy of predictive analytics, focusing on water quality monitoring in aquaponics systems.

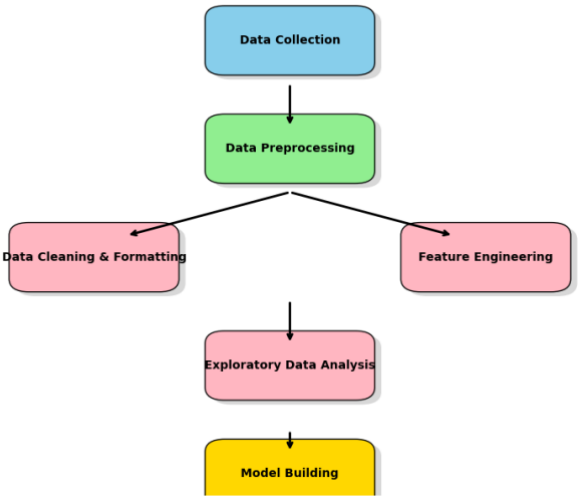


Figure 4: Research Design Process

3.3.1 Data Collection

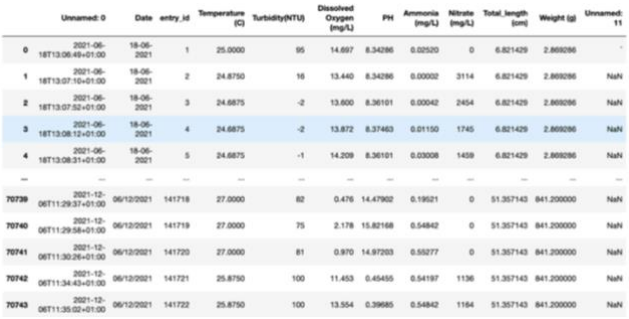
Data for this research was obtained from an open-source dataset provided by the HiPIC Research Group at the University of Nigeria Nsukka, Nigeria. The dataset comprises water quality sensor readings from freshwater aquaponics catfish ponds, collected at 5-second intervals between June and mid-October 2021. The sensors used include the Dallas Instrument Temperature sensor (DS18B20), DF Robot Turbidity sensor, DF Robot Dissolved Oxygen sensor, DF Robot pH sensor V2.2, MQ-137 Ammonia sensor, and MQ-135 Nitrate sensor.

3.3.2 Data Pre-Processing

Before the data could be used for analysis, it underwent a thorough preprocessing phase to ensure accuracy and consistency. This involved several steps:

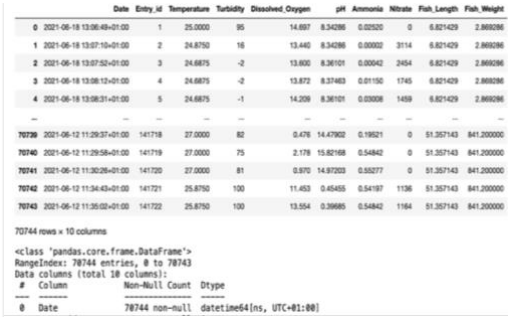
3.3.2.1 Data Cleaning & Formatting

Initially, it was essential to ensure consistency across the 12 different datasets collected from the aquaponics fish ponds. The first step involved renaming the columns to maintain uniformity, which is crucial for later merging the datasets. Additionally, the Date column was converted to a datetime format, a necessary step to facilitate proper sorting and merging of the datasets based on time. This ensured that all subsequent analyses would have a consistent temporal structure.



	Unnamed: 0	Date	entry_id	Temperature (C)	Turbidity(NTU)	Dissolved Oxygen (mg/L)	PH	Ammonia (mg/L)	Nitrate (mg/L)	Total_length (cm)	Weight (g)	Unnamed: 11
0	2021-06-18T13:06:49+01:00	18-06-2021	1	25.0000	95	14.697	8.34286	0.02520	0	6.821429	2.869286	
1	2021-06-18T13:07:10+01:00	18-06-2021	2	24.8750	16	13.440	8.34286	0.00002	3114	6.821429	2.869286	NaN
2	2021-06-18T13:07:52+01:00	18-06-2021	3	24.6875	-2	13.600	8.36101	0.00042	2454	6.821429	2.869286	NaN
3	2021-06-18T13:08:12+01:00	18-06-2021	4	24.6875	-2	13.872	8.37463	0.01150	1745	6.821429	2.869286	NaN
4	2021-06-18T13:08:31+01:00	18-06-2021	5	24.6875	-1	14.209	8.36101	0.03008	1459	6.821429	2.869286	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
70739	2021-12-06T11:29:37+01:00	06/12/2021	141718	27.0000	82	0.476	14.47902	0.19521	0	51.357143	841.200000	NaN
70740	2021-12-06T11:29:58+01:00	06/12/2021	141719	27.0000	75	2.178	15.82168	0.54842	0	51.357143	841.200000	NaN
70741	2021-12-06T11:30:26+01:00	06/12/2021	141720	27.0000	81	0.970	14.97203	0.55277	0	51.357143	841.200000	NaN
70742	2021-12-06T11:34:43+01:00	06/12/2021	141721	25.8750	100	11.453	0.45455	0.54197	1136	51.357143	841.200000	NaN
70743	2021-12-06T11:35:02+01:00	06/12/2021	141722	25.8750	100	13.554	0.39685	0.54842	1164	51.357143	841.200000	NaN

Figure 5.1: Data Column before renaming



	Date	Entry_id	Temperature	Turbidity	Dissolved_Oxygen	pH	Ammonia	Nitrate	Fish_Length	Fish_Weight
0	2021-06-18 13:06:49+01:00	1	25.0000	95	14.697	8.34286	0.02520	0	6.821429	2.869286
1	2021-06-18 13:07:10+01:00	2	24.8750	16	13.440	8.34286	0.00002	3114	6.821429	2.869286
2	2021-06-18 13:07:52+01:00	3	24.6875	-2	13.600	8.36101	0.00042	2454	6.821429	2.869286
3	2021-06-18 13:08:12+01:00	4	24.6875	-2	13.872	8.37463	0.01150	1745	6.821429	2.869286
4	2021-06-18 13:08:31+01:00	5	24.6875	-1	14.209	8.36101	0.03008	1459	6.821429	2.869286
...	...	...	...	...	...	...	...	...	...	...
70739	2021-06-12 11:29:37+01:00	141718	27.0000	82	0.476	14.47902	0.19521	0	51.357143	841.200000
70740	2021-06-12 11:29:58+01:00	141719	27.0000	75	2.178	15.82168	0.54842	0	51.357143	841.200000
70741	2021-06-12 11:30:26+01:00	141720	27.0000	81	0.970	14.97203	0.55277	0	51.357143	841.200000
70742	2021-06-12 11:34:43+01:00	141721	25.8750	100	11.453	0.45455	0.54197	1136	51.357143	841.200000
70743	2021-06-12 11:35:02+01:00	141722	25.8750	100	13.554	0.39685	0.54842	1164	51.357143	841.200000

70744 rows × 10 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70744 entries, 0 to 70743
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Date             70744 non-null  datetime64[ns, UTC+01:00]
```

Figure 5.2: Data Column after renaming

3.3.2.2 Feature Engineering

Feature engineering involved enhancing the dataset to better capture underlying patterns. Initially, the dataset was checked for infinite values, which were standardized by converting them to NaN. Subsequent checks for duplicates and missing values ensured data reliability. Columns with less than 5% missing data, except for Ammonia, had those rows removed without significantly affecting the dataset, as most columns had less than 0.1% missing data. Duplicate entries in the Date column were also removed to maintain accuracy. The Ammonia data presented a challenge due to a high percentage of missing values. Investigation revealed that these missing values were likely due to sensor malfunctions during specific periods, identified as Missing Not at Random (MNAR).

This led us to conclude that the missing data was MNAR—Missing Not at Random—because the gaps were likely caused by external factors, such as sensor failure during certain environmental conditions or operational downtimes. The fact that only the Ammonia sensor exhibited missing data, while others did not, further supported this conclusion.

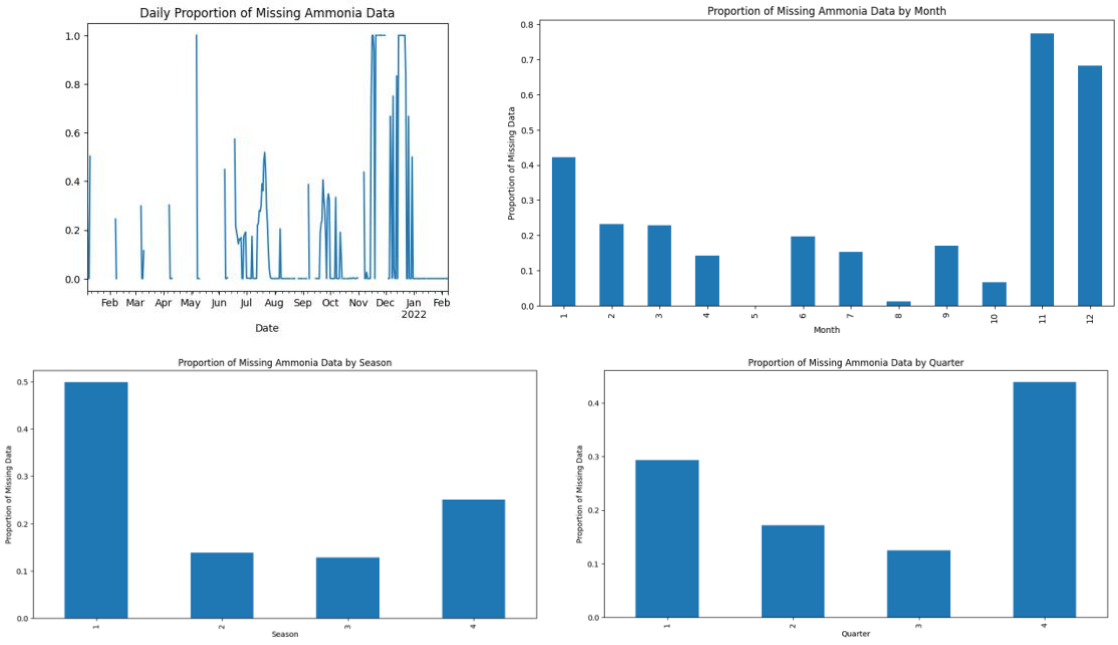


Figure 6: Visualization of Ammonia Missing values

Given that the missing data was MNAR, Iterative Imputation (MICE) was employed to address it. This method, which models each feature with missing values as a function of others, is well-suited for MNAR data. The right-skewed nature of the Ammonia data was also considered during imputation to ensure accuracy. Large gaps in the dataset were handled by creating a binary indicator and resampling the data at 5-minute intervals, providing consistent input for the model. The dataset was then split into training and test sets to prevent data leakage and ensure reliable performance.

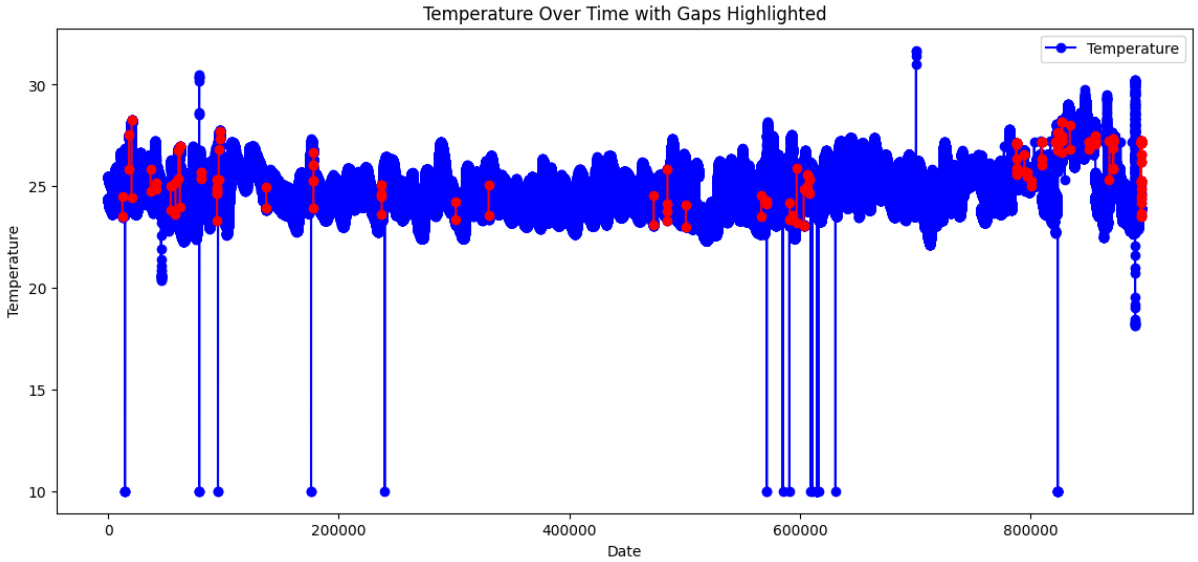
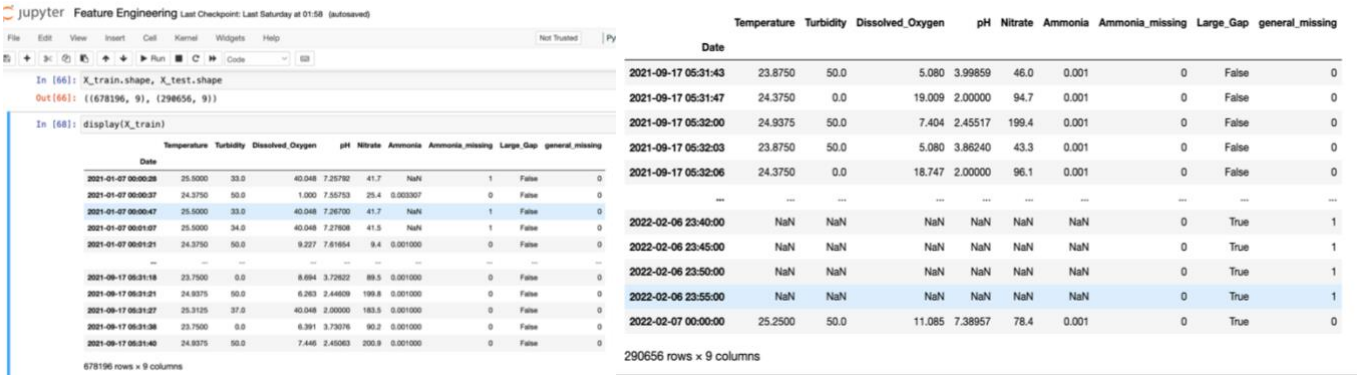


Figure 7: Visualization of Large Gap for Temperature Column

Before applying any imputation methods, we split the dataset into training and test sets. This step was crucial to prevent data leakage, ensuring that the model's performance on the test set would accurately reflect its ability to generalize to new, unseen data.



Date	Temperature	Turbidity	Dissolved_Oxygen	pH	Nitrate	Ammonia	Ammonia_missing	Large_Gap	general_missing
2021-09-17 05:31:43	23.8750	50.0	5.080	3.99859	46.0	0.001	0	False	0
2021-09-17 05:31:47	24.3750	0.0	19.009	2.00000	94.7	0.001	0	False	0
2021-09-17 05:32:00	24.9375	50.0	7.404	2.45517	199.4	0.001	0	False	0
2021-09-17 05:32:03	23.8750	50.0	5.080	3.86240	43.3	0.001	0	False	0
2021-09-17 05:32:06	24.3750	0.0	18.747	2.00000	96.1	0.001	0	False	0
...	...	...	...	...	...	...	...	...	...
2022-02-06 23:40:00	NaN	NaN	NaN	NaN	NaN	NaN	0	True	1
2022-02-06 23:45:00	NaN	NaN	NaN	NaN	NaN	NaN	0	True	1
2022-02-06 23:50:00	NaN	NaN	NaN	NaN	NaN	NaN	0	True	1
2022-02-06 23:55:00	NaN	NaN	NaN	NaN	NaN	NaN	0	True	1
2022-02-07 00:00:00	25.2500	50.0	11.085	7.38957	78.4	0.001	0	True	0

Figure 8: Test and Train Dataset after splitting

Imputation was applied in two stages. First, we focused on the data that had been resampled due to large gaps, using Iterative Imputation to fill in these values. This method was chosen because it models each feature with missing data as a function of the other features, making it ideal for handling

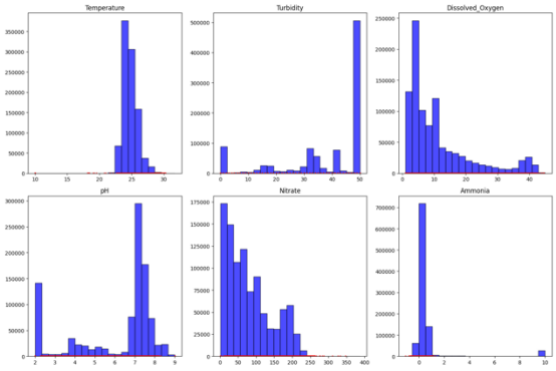


Figure 10: Histogram of imputed & original Data

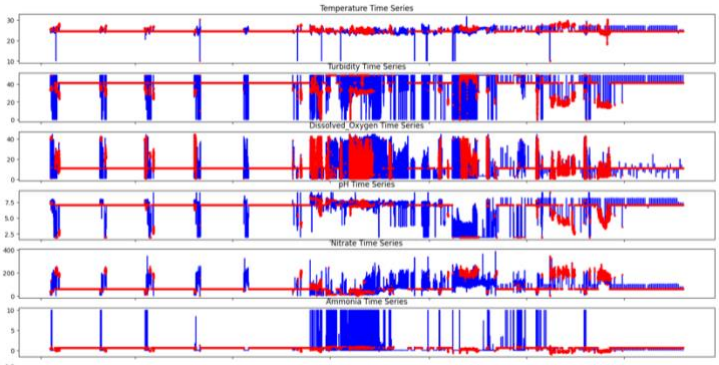


Figure 9: Timeseries of imputed & original data

the complex relationships in our dataset. After addressing the resampled data, we then applied Iterative Imputation to the original missing Ammonia values, excluding the resampled data, which had been artificially generated. This two-step approach ensured that our imputation method was both precise and reliable.

Normalization was applied to ensure equal contribution from all features. After analyzing histograms, QQ plots, skewness, and kurtosis, Min-Max Scaling was chosen for its ability to normalize the data without distorting feature relationships, particularly given the skewed nature of some variables. This step was crucial in preparing a robust dataset for model building.

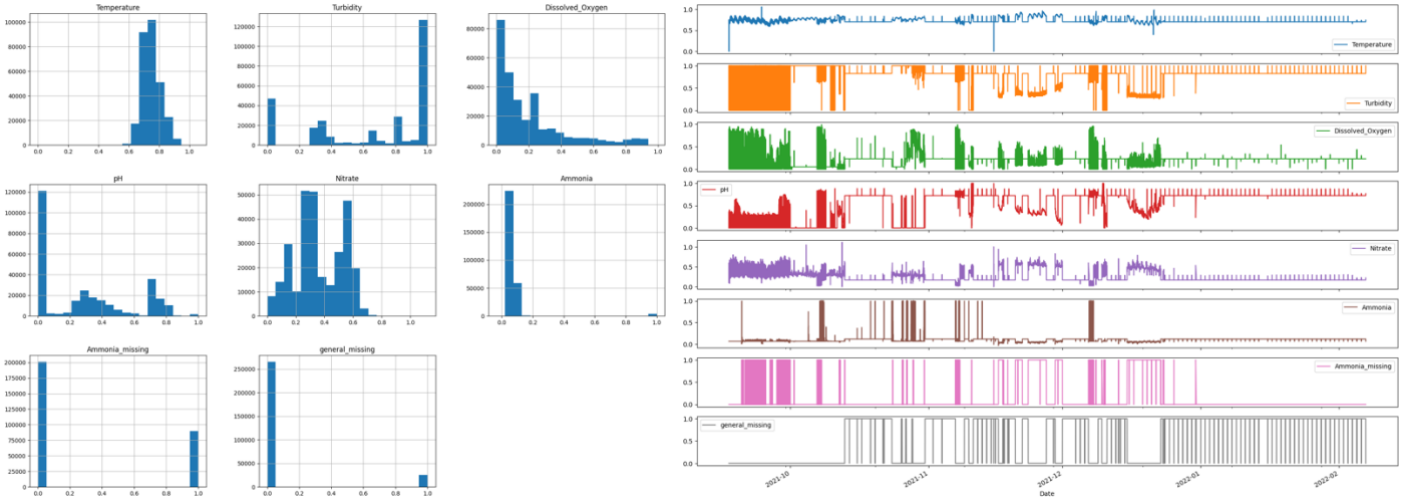


Figure 11: Histogram & Timeseries visualization of dataset after feature engineering

3.3.3 EDA

Exploratory Data Analysis (EDA) is a critical step before applying any machine learning models, as it allows us to understand the data, uncover patterns, and detect anomalies. Our EDA began with the construction of a correlation matrix to identify relationships between different variables. As seen in the matrix, some variables like Temperature and pH are moderately correlated, which is expected, given their natural interactions in aquaponics systems. However, the correlation between Ammonia and other variables is weak, which aligns with the unique behaviour of this parameter in our system.

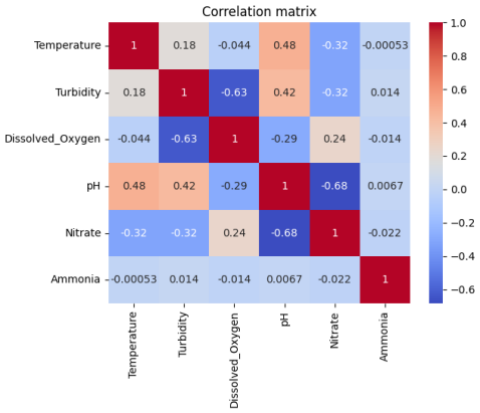


Figure 12: Correlation matrix of Train Dataset

The analysis revealed no significant monthly seasonality for variables like Temperature, Dissolved Oxygen, and pH. However, Ammonia showed a noticeable peak in July, indicating a possible seasonal pattern or anomaly. This trend suggests that specific environmental or operational factors may influence Ammonia levels during this period, warranting further investigation.

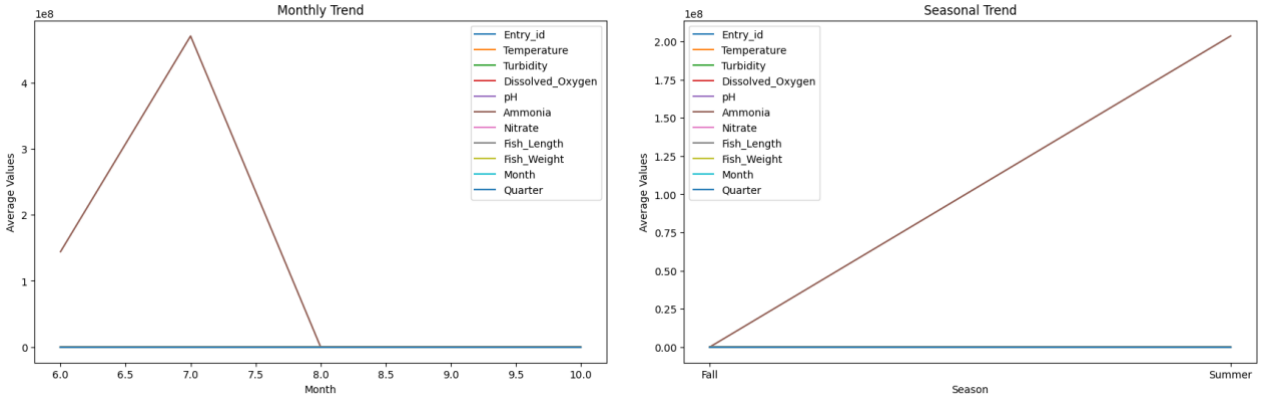


Figure 13: Monthly & Seasonal Data trend

The exploratory data analysis (EDA) highlighted feature importance using PCA and the Isolation Forest method. PCA indicated that Temperature and Turbidity were the primary contributors to data variance, essential for understanding the system's key components. The Isolation Forest method, focusing on anomaly detection, identified Nitrate, Ammonia, and Dissolved Oxygen as critical features, emphasizing their importance in maintaining the system's health.

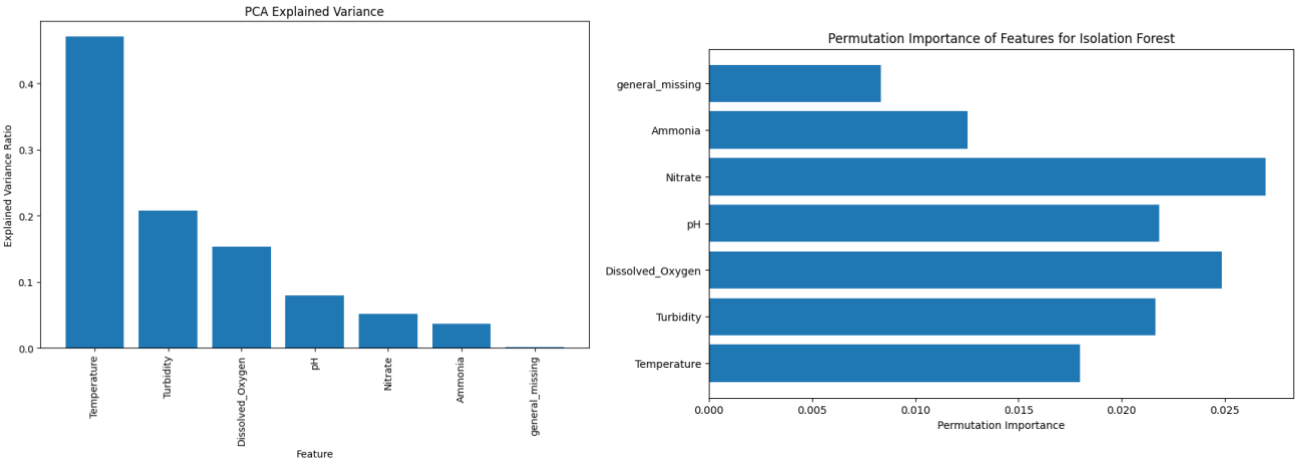


Figure 14: Feature importance by unsupervised technique

In conclusion, the EDA provided a comprehensive understanding of the dataset, revealed critical patterns and relationships, and identified key features that will be instrumental in developing robust predictive models.

3.4 Summary

In summary, this chapter outlined the methodology employed in the research, encompassing data collection, preprocessing, and exploratory data analysis. The steps taken ensured the dataset was well-prepared for subsequent modeling efforts, addressing potential issues like missing data and large gaps. The rigorous approach in this chapter lays the groundwork for building accurate and reliable models in the following chapters.

Chapter 4: Evaluation of Research Outcome

This chapter should describe the application of machine learning and data mining techniques on the collected data and their outcomes.

4.1 Introduction

In this chapter, the focus is on applying machine learning and data mining techniques to aquaponics systems, specifically for anomaly detection and predictive analytics. The primary goal is to improve the management of water quality parameters, essential for the sustainability and productivity of aquaponics.

Three models—LSTM with Autoencoder, a Hybrid Model combining Conv1D with LSTM and Autoencoder, and GRU—were developed and evaluated for anomaly detection, chosen for their ability to capture temporal dependencies in time-series data.

Additionally, predictive analytics was employed to forecast key water quality parameters, with models such as LSTM, GRU, CNN, Neural Networks, and the Hybrid Model tested to identify the most accurate and efficient approach. This chapter details the development, validation, and performance comparison of these models, highlighting their effectiveness in real-world aquaponics systems.

4.2 Building the Predictive Models

This section outlines the construction of machine learning models for anomaly detection and predictive analytics in the aquaponics system, focusing on enhancing the system's efficiency and sustainability by accurately monitoring key water quality parameters.

The models chosen include:
1. **LSTM with Autoencoder**: Ideal for detecting anomalies in time-series data, this model captures the underlying structure of the data to improve detection accuracy.
2. **Hybrid Model (Conv1D + LSTM with Autoencoder)**: Combines convolutional layers to capture local patterns with LSTM for temporal dependencies, leveraging the strengths of both models.
3. **GRU Model**: Selected for its efficiency in handling sequential data with lower computational resources compared to LSTM.
4. **Predictive Analytics Models**: Models such as CNN, Neural Networks, and the Hybrid Model were explored for predicting critical water quality parameters.

Each model underwent hyperparameter tuning and was evaluated using Mean Square Error (MSE) as the primary metric, with cross-validation ensuring robustness, particularly for anomaly detection models.

4.3 Evaluation of Aquaponics Predictive Models for Anomaly Detection

In this section, we evaluate the performance of three different models used for anomaly detection in the aquaponics system: LSTM with Autoencoder, Hybrid Model (Conv1D + LSTM with Autoencoder), and GRU. Each model was assessed based on its ability to accurately detect anomalies in the time-series data, which is crucial for maintaining optimal water quality in the system.

For the feature selection process, features were chosen based on their importance and the results of various visualizations. The selected features included Temperature, Turbidity, pH, Dissolved Oxygen, and Nitrate, as these were identified as the most relevant variables for the model. The Ammonia values were excluded from the feature set due to their unrealistic readings, with most of the data

significantly exceeding the expected range. This inconsistency suggested that the Ammonia data might not contribute effectively to the model's performance and could potentially introduce noise.

Once the key features were selected, the data was reshaped into a three-dimensional format, which is necessary for input into the LSTM model. This reshaping process ensures that the sequential nature of the data is preserved, allowing the LSTM to capture temporal dependencies effectively.

4.3.1 LSTM with Autoencoder

The LSTM with Autoencoder model was chosen for its strengths in capturing temporal dependencies within sequential data.

The LSTM model was carefully constructed with appropriate hyperparameter tuning to capture the complexities of the dataset. After training the model, the reconstruction error was calculated on the training dataset to evaluate how well the model could capture the true data distribution. The chosen anomaly detection threshold was set at 0.2615. Applying this threshold to the training dataset, the model identified 3,490 anomalies.

When the model was tested on unseen data, it maintained the same threshold of 0.2615. On the test dataset, the model detected 637 anomalies, reflecting its ability to generalize beyond the training data. The confusion matrices for the training and test datasets showed that the model performed well in distinguishing normal and anomalous data points.

Table 1: Confusion matrix of LSTM with Autoencoder

Train Dataset	Test Dataset
• True Positives: 66,302 • False Positives: 0 • True Negatives: 3,490 • False Negatives: 0	• True Positives: 29,270 • False Positives: 0 • True Negatives: 637 • False Negatives: 0

However, examining the boundary values of the selected features, it was clear that some values in the test set were unrealistic, likely contributing to the model's detection of anomalies. For example, temperature and pH values outside the realistic range indicated potential sensor errors or data quality issues.

Table 2: Boundary level values for Anomaly Detection with LSTM Autoencoder

Train Dataset	Test Dataset
Temperature: min: 23.0, max: 26.8125	Temperature: min: -127.0, max: 27.75
Turbidity: min: 1, max: 100	Turbidity: min: 82.0, max: 100.0
Dissolved Oxygen: min: 0.008, max: 41.045	Dissolved Oxygen: min: 0.007, max: 27.575
pH: min: 7.089, max: 8.55	pH: min: -0.586, max: 7.657
Nitrate: min: 45.0, max: 1192.0	Nitrate: min: 331.44, max: 936.00

An Isolation Forest outlier detection method was applied as an additional filtering layer to address the potential issue of unrealistic values being treated as normal by the LSTM model. This approach aimed to identify and exclude unrealistic values, refining the anomaly detection process.
The results after applying the Isolation Forest were as follows:

Table 3: Evaluate Model Performance

Train Dataset	Test Dataset
• Precision: 0.5990 • Recall: 1.0000 • F1 Score: 0.7492	• Precision: 0.4258 • Recall: 1.0000 • F1 Score: 0.5973

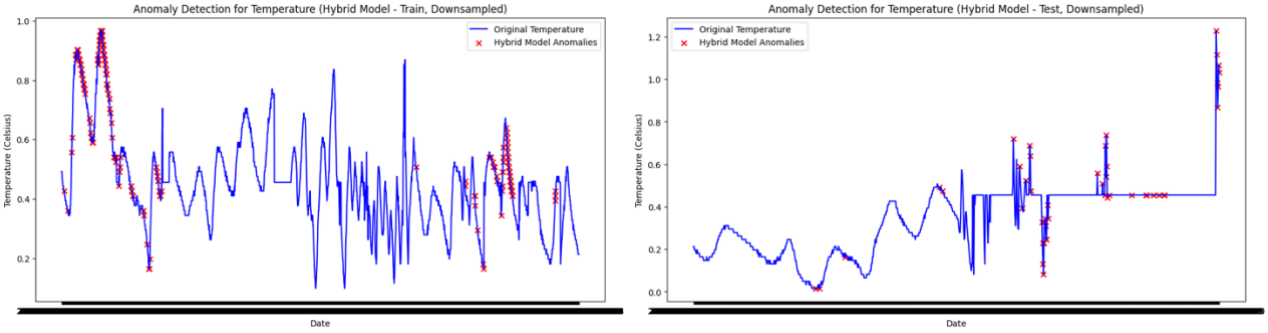


Figure 15: LSTM + Isolation Forest model for Anomaly detection on Train and Test Dataset

Table 4: Boundary level values for Anomaly detection with LSTM + Isolation Forest Model

Train Dataset	Test Dataset
Temperature: min: 23.625, max: 26.8125	Temperature: min: 22.99, max: 23.002
Turbidity: min: 8.998, max: 100	Turbidity: min: 0.997, max: 1.059
Dissolved Oxygen: min: 0.02499, max: 40.868	Dissolved Oxygen: min: 0.0079, max: 40.0.032
pH: min: 7.11, max: 8.55	pH: min: 7.089, max: 7.090
Nitrate: min: 118.99, max: 873.0	Nitrate: min: 44.99, max: 45.69

Finally, the Mean Squared Error (MSE) was used to evaluate the overall model performance:

- Validation MSE: 0.0707
- Test MSE: 0.1173

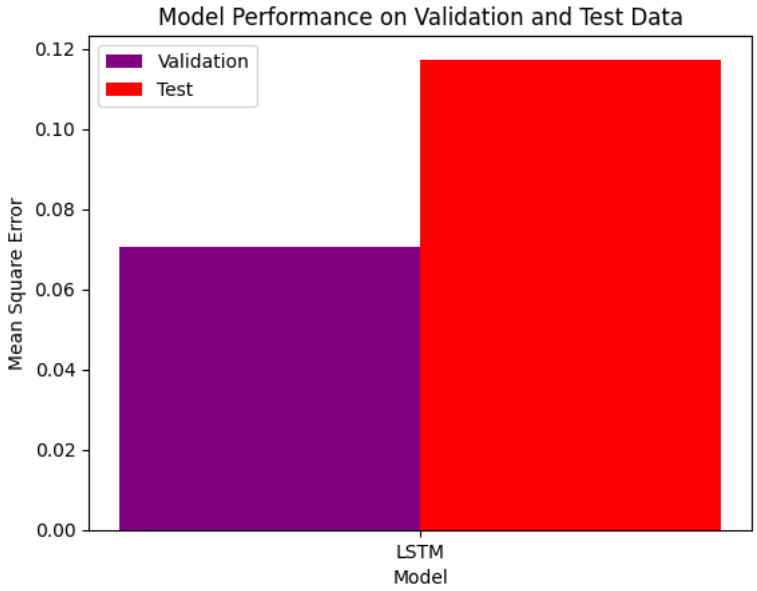


Figure 16: Model Performance on Validation and Test Data

The validation MSE of 0.0707 and the higher test MSE of 0.1173 suggest that while the LSTM model performed reasonably well on the validation set, it struggled more with the unseen test data.

In summary, while the LSTM model demonstrated robust anomaly detection capabilities, particularly in identifying all anomalies (as indicated by the recall), its tendency to misclassify normal data points as anomalies (evidenced by the lower precision) suggests room for improvement. The addition of Isolation Forest helped mitigate some of these issues by providing an additional layer of filtering, improving the model's ability to handle unrealistic values.

4.3.2 Hybrid Model (Convo1d + LSTM with Autoencoder)

The hybrid model, which combines a one-dimensional convolutional neural network (Conv1D) with Long Short-Term Memory (LSTM) layers and an autoencoder, was developed to leverage the strengths of both convolutional and recurrent neural networks. The Conv1D layer captures local patterns within the time series data, while the LSTM layers are responsible for learning temporal dependencies. The autoencoder structure is used to reconstruct the input data, with the reconstruction error serving as the metric for anomaly detection.

Upon training the model, the reconstruction error on the training data was calculated, yielding a mean reconstruction error of 0.1146. An anomaly detection threshold was set at 0.2416, corresponding to the 95th percentile of the reconstruction error distribution. This threshold allowed the detection of 3,490 anomalies in the training data with perfect precision and recall, as indicated by the confusion matrix.

When applied to the test data, the model maintained its high performance, detecting 651 anomalies with precision, recall, and F1 scores all at 1.0000. The confusion matrices for the training and test datasets showed that the model performed well in distinguishing normal and anomalous data points.

Table 5: Confusion matrix of Hybrid model

Train Dataset	Test Dataset
• True Positives: 66,302 • False Positives: 0 • True Negatives: 3,490 • False Negatives: 0	• True Positives: 29,256 • False Positives: 0 • True Negatives: 651 • False Negatives: 0

However, upon closer inspection of the boundary values for the test data, several unrealistic values were observed, such as a minimum temperature of -127°C and negative pH values. These values are clearly outside the realm of physical possibility, suggesting that some anomalies detected may be due to sensor errors or data corruption rather than genuine system anomalies.

Table 6: Boundary level values for Anomaly Detection of Hybrid Model

Train Dataset	Test Dataset
Temperature: min: 23.0, max: 26.8125	Temperature: min: -127.0, max: 27.75
Turbidity: min: 1, max: 100	Turbidity: min: 82.0, max: 100.0
Dissolved Oxygen: min: 0.008, max: 41.045	Dissolved Oxygen: min: 0.007, max: 27.575
pH: min: 7.089, max: 8.55	pH: min: -0.586, max: 7.657
Nitrate: min: 45.0, max: 1192.0	Nitrate: min: 331.44, max: 1936.00

To address these issues, an additional layer of anomaly detection using Isolation Forest (IF) was applied to both the training and test datasets. The integration of IF helped to refine the anomaly

detection process by filtering out unrealistic values, leading to more meaningful results. After applying IF, the precision, recall, and F1 scores on the test data were slightly reduced but still indicated strong performance, with an F1 score of 0.9134.
The results after applying the Isolation Forest were as follows:

Table 7: Evaluate Model Performance

Train Dataset	Test Dataset
• Precision: 0.7655 • Recall: 1.0000 • F1 Score: 0.8672	• Precision: 0.8494 • Recall: 0.9877 • F1 Score: 0.9134

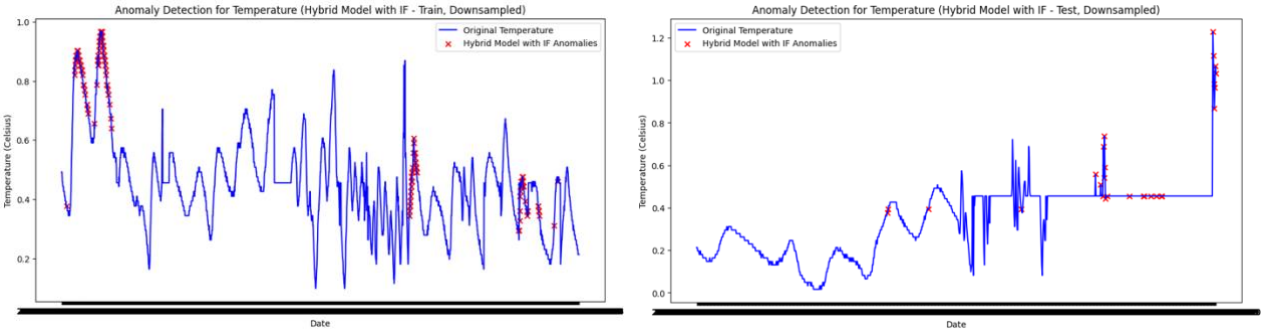


Figure 17: Hybrid + Isolation Forest model for Anomaly detection on Train and Test Dataset

Table 8: Boundary level values for Anomaly detection with Hybrid + Isolation Forest Model

Train Dataset	Test Dataset
Temperature: min: 23.687, max: 26.8125	Temperature: min: -127.0, max: 27.687
Turbidity: min: 5.998, max: 100	Turbidity: min: 85.88, max: 100.0
Dissolved Oxygen: min: 0.008, max: 38.795	Dissolved Oxygen: min: 0.007, max: 14.6526
pH: min: 7.12, max: 8.49	pH: min: -0.5862, max: 7.6576
Nitrate: min: 143.99, max: 994.0	Nitrate: min: 331.44, max: 1440.0

Finally, the Mean Squared Error (MSE) was used to evaluate the overall model performance:

- Validation MSE: 0.0313
- Test MSE: 0.0783

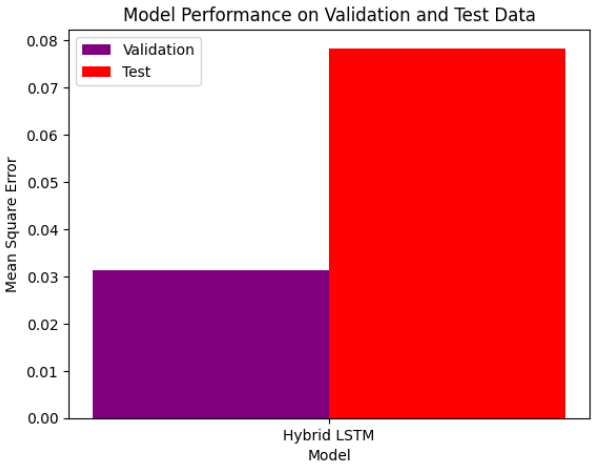


Figure 18: Model Performance on Validation and Test Data

The final validation and test MSE values were 0.0314 and 0.0784, respectively. These values suggest that while the model performed well on the validation data, it encountered more variability or noise in the test data, possibly due to the presence of outliers or unrecognized anomalies. Overall, the hybrid model demonstrated robust performance, though the quality and consistency of the dataset remain critical factors influencing the results.
The addition of Isolation Forest helped mitigate some of these issues by providing an additional layer of filtering, improving the model's ability to handle unrealistic values.

4.3.3 GRU Model

The GRU (Gated Recurrent Unit) model was employed to capture the temporal dependencies in the dataset, leveraging its simpler architecture compared to LSTM, while still retaining the ability to manage long-term dependencies in the data. The model was trained to minimize the reconstruction error, which would later be used as a metric for anomaly detection.

The GRU model was carefully constructed with appropriate hyperparameter tuning to capture the complexities of the dataset. After training the model, the reconstruction error was calculated on the training dataset to evaluate how well the model could capture the true data distribution. The chosen anomaly detection threshold was set at 0.1844. Applying this threshold to the training dataset, the model identified 3,490 anomalies.

When the model was tested on unseen data, it maintained the same threshold of 0.1844. On the test dataset, the model detected 663 anomalies, reflecting its ability to generalize beyond the training data. The confusion matrices for the training and test datasets showed that the model performed well in distinguishing normal and anomalous data points.

Table 9: Confusion matrix of GRU model

Train Dataset	Test Dataset
• True Positives: 66,302 • False Positives: 0 • True Negatives: 3,490 • False Negatives: 0	• True Positives: 29,244 • False Positives: 0 • True Negatives: 663 • False Negatives: 0

However, examining the boundary values of the selected features, it was clear that some values in the test set were unrealistic, likely contributing to the model's detection of anomalies. For example, temperature, Nitrate and pH values outside the realistic range indicated potential sensor errors or data quality issues.

Table 10: Boundary level values for Anomaly Detection with GRU model

Train Dataset	Test Dataset
Temperature: min: 23.0, max: 26.8125	Temperature: min: -127.0, max: 27.75
Turbidity: min: 1, max: 100	Turbidity: min: 82.0, max: 100.0
Dissolved Oxygen: min: 0.008, max: 41.045	Dissolved Oxygen: min: 0.007, max: 27.575
pH: min: 7.089, max: 8.55	pH: min: -0.586, max: 7.657
Nitrate: min: 45.0, max: 1192.0	Nitrate: min: 331.44, max: 1936.00

An Isolation Forest outlier detection method was applied as an additional filtering layer to address the potential issue of unrealistic values being treated as normal by the GRU model. This approach aimed to identify and exclude unrealistic values, refining the anomaly detection process.
The results after applying the Isolation Forest were as follows:

Table 11: Evaluate Model Performance

Train Dataset	Test Dataset
• Precision: 0.6910 • Recall: 1.0000 • F1 Score: 0.8172	• Precision: 0.5057 • Recall: 1.0000 • F1 Score: 0.6717

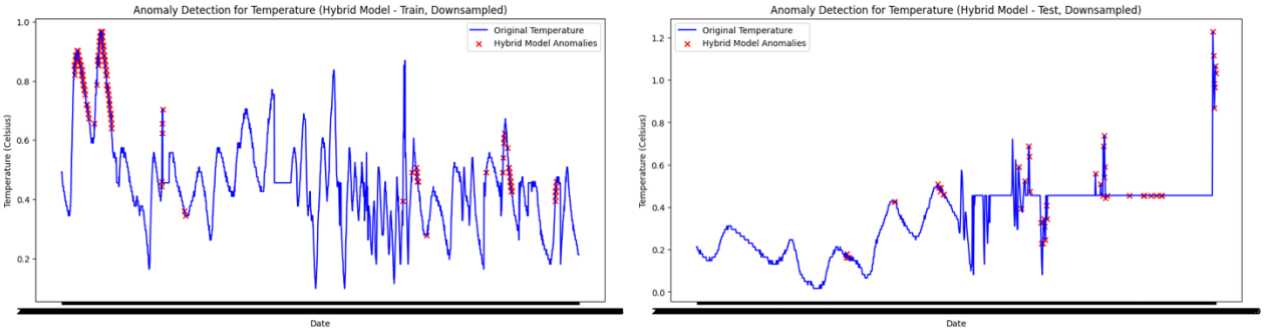


Figure 19: GRU + Isolation Forest model for Anomaly detection on Train and Test Dataset

Table 12: Boundary level values for Anomaly detection with GRU + Isolation Forest Model

Train Dataset	Test Dataset
Temperature: min: 23.937, max: 26.8125	Temperature: min: -127.00, max: 27.687
Turbidity: min: 1, max: 100	Turbidity: min: 85.88, max: 100.00
Dissolved Oxygen: min: 0.0159, max: 39.429	Dissolved Oxygen: min: 0.007, max: 14.65
pH: min: 7.094, max: 8.492	pH: min: -0.5862, max: 7.657
Nitrate: min: 101.99, max: 819.0	Nitrate: min: 331.44, max: 1663.0

Finally, the Mean Squared Error (MSE) was used to evaluate the overall model performance:

- Validation MSE: 0.0437
- Test MSE: 0.0798

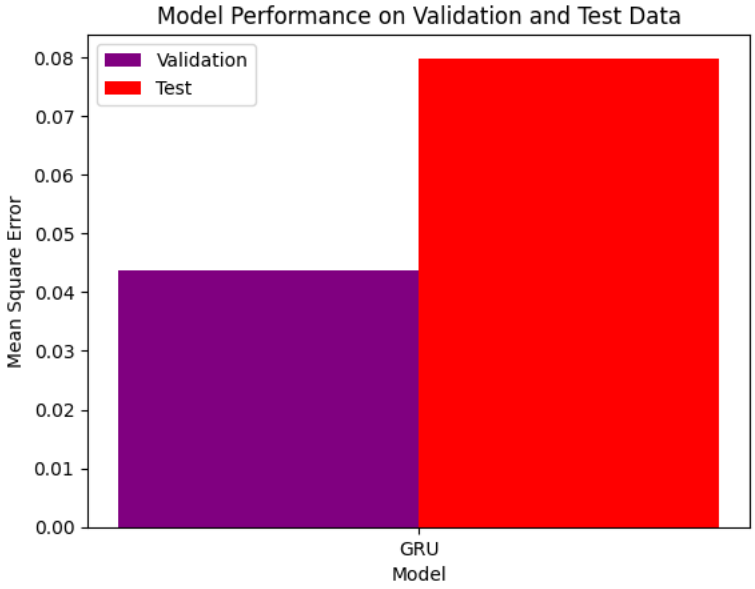


Figure 20: Model Performance on Validation and Test Data

The validation MSE of 0.0437 and the higher test MSE of 0.0798 suggest that while the GRU model performed reasonably well on the validation set, it struggled more with the unseen test data.

Overall, the GRU model demonstrated solid performance in detecting anomalies, but the presence of unrealistic values in the dataset underscores the importance of data quality.

4.3.4 Comparison of Models

In this section, the performance of three models—LSTM with Autoencoder, Hybrid LSTM, and GRU—is compared using their Mean Squared Error (MSE) on validation and test datasets.

The LSTM model, with a validation MSE of 0.070672 and test MSE of 0.117291, showed reasonable performance during validation but struggled with generalization, indicating possible overfitting. The

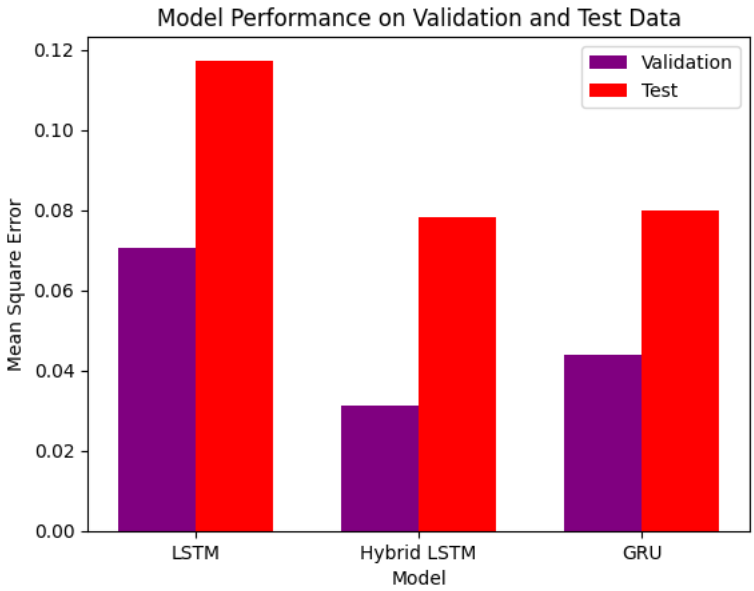


Figure 21: Comparison between all models for Anomaly Detection

Hybrid LSTM model, which integrates CNN and LSTM layers, achieved the best results, with a validation MSE of 0.031364 and test MSE of 0.078357. Its lower MSE values suggest it effectively captured both local and temporal patterns, making it more robust. The GRU model, with a validation MSE of 0.043784 and test MSE of 0.079851, performed better than the LSTM model but did not match the Hybrid LSTM model's effectiveness, though it showed less overfitting.

It should be noted that some of the test results, particularly those from the LSTM model, were impacted by unrealistic values present in the dataset. These anomalies, which may have arisen from sensor errors or data recording issues, affected the model's ability to accurately detect and predict outcomes. Despite these challenges, the Hybrid LSTM model demonstrated superior performance, likely due to its ability to effectively process both temporal sequences and spatial patterns, making it the most reliable model for anomaly detection and predictive analytics in this study.

4.4 Validation and Hyperparameter Tuning

This section focuses on validating the hybrid conv1D with LSTM autoencoder model and optimizing its performance through K-Fold Cross-Validation and hyperparameter tuning. These steps were essential to ensure the model's robustness and effectiveness in detecting anomalies.

K-Fold Cross-Validation

K-Fold Cross-Validation was employed to validate the model's performance across different subsets of the data. The dataset was split into five folds, with the model being trained on four and validated on

the remaining fold in each iteration. The key metrics assessed were precision, recall, and F1 score. The results showed an average precision of 0.2104, recall of 0.2103, and F1 score of 0.2103. These results, while modest, indicated that the model could detect anomalies but highlighted the need for further refinement.

Threshold Optimization

To improve detection accuracy, a threshold optimization process was conducted. By testing various thresholds, the best balance between precision, recall, and F1 score was found at a threshold of 0.86. This threshold was selected as the optimal cutoff for identifying anomalies, aiming to reduce false positives while maintaining accurate detection.

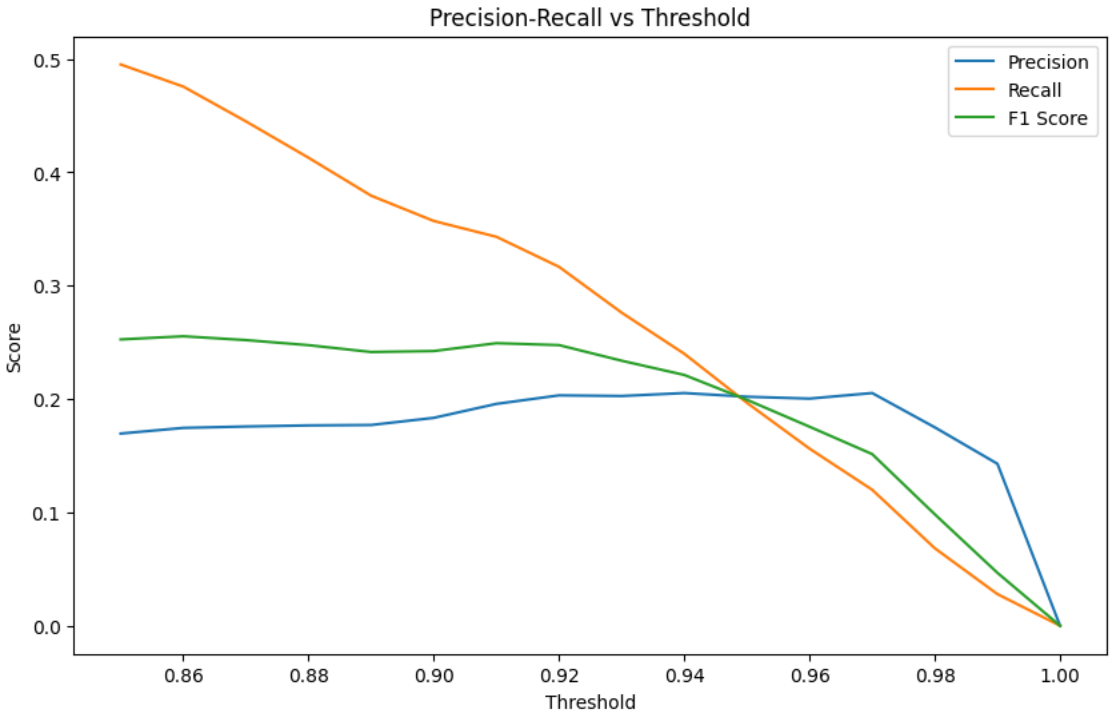


Figure 22: Precision-Recall vs Threshold

Grid Search with K-Fold Cross-Validation

Further improvement was sought through Grid Search combined with K-Fold Cross-Validation to fine-tune the model's hyperparameters. The search explored different configurations of learning rate, LSTM units, dropout rate, convolutional filters, and kernel size.

The optimal configuration identified included a learning rate of 0.0005, 128 LSTM units, a dropout rate of 0.5, 64 convolutional filters, and a kernel size of 3. This setup achieved a slightly better F1 score of 0.2277, indicating a marginal improvement in anomaly detection.

The validation and tuning efforts provided valuable insights into the model's performance. K-Fold Cross-Validation ensured that the model was reliable across different data splits, while threshold optimization and Grid Search helped refine its accuracy.

4.5 Evaluation of Aquaponics Predictive Models for Predictive Analytics

Before applying machine learning models to predict temperature levels, a custom window generator was created to handle the time series nature of the data. The window generator allowed for the efficient preparation of data, enabling it to capture temporal dependencies and relationships across the time steps.

A Conv1D model was chosen initially for the task. The model was structured to capture local patterns

in the data through convolutional layers, followed by dense layers to handle the final predictions. This approach was selected because convolutional layers are well-suited for identifying patterns in sequential data, such as time series, where shifts in data can reveal significant trends.

During the training phase, the model was trained using a loss function that minimized the mean squared error (MSE), and its performance was tracked using validation loss over several epochs.

Training Data	Test Data
MSE: 0.0076	MAE: 0.0703
MAE: 0.0609	RMSE: 0.3979
R-squared (R^2): 0.6851	**R-squared (R^2)**: -0.3656

The training data shows a low MSE of 0.0076 and an MAE of 0.0609, indicating the model fits well with a 6.09% average error. However, while the R^2 value of 0.6851 suggests a decent fit, there is room for improvement. On the test data, the MAE slightly increases to 0.0703, and the RMSE rises to 0.3979, reflecting larger errors. The negative R^2 value of -0.3656 is concerning, indicating poor generalization, potentially due to overfitting or noise in the test set.

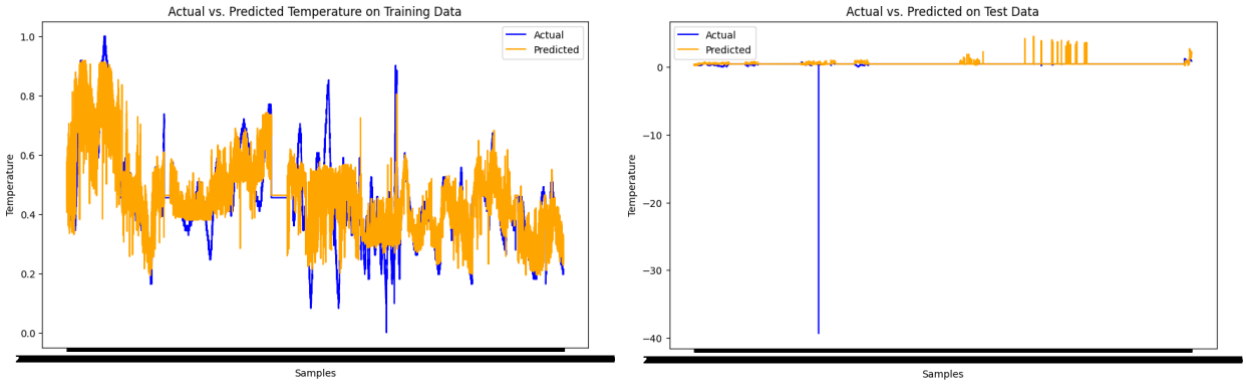


Figure 23: Train and Test dataset prediction on Convo1D model

Results indicated that the Conv1D model, even without the application of traditional machine learning algorithms, yielded highly promising results. The actual versus predicted temperature levels on both the training and test datasets were closely aligned, indicating that the model was successful in capturing the underlying patterns (Figure 23). The test loss was also evaluated, and the resulting metrics further confirmed the model's effectiveness.

4.5.1 LSTM with Autoencoder

The LSTM Autoencoder model was designed and implemented to capture the temporal dependencies and reconstruct the input data, thereby enabling effective anomaly detection in the aquaponics temperature dataset. The model architecture consisted of an encoder-decoder structure, where the encoder was responsible for compressing the input sequence into a latent representation, and the decoder was tasked with reconstructing the sequence from this compressed representation.

Firstly, we try this one our Conv1D model with LSTM autoencoder but the result isn't satisfactory. So decided to go with only the LSTM with autoencoder model without Conv1D model.

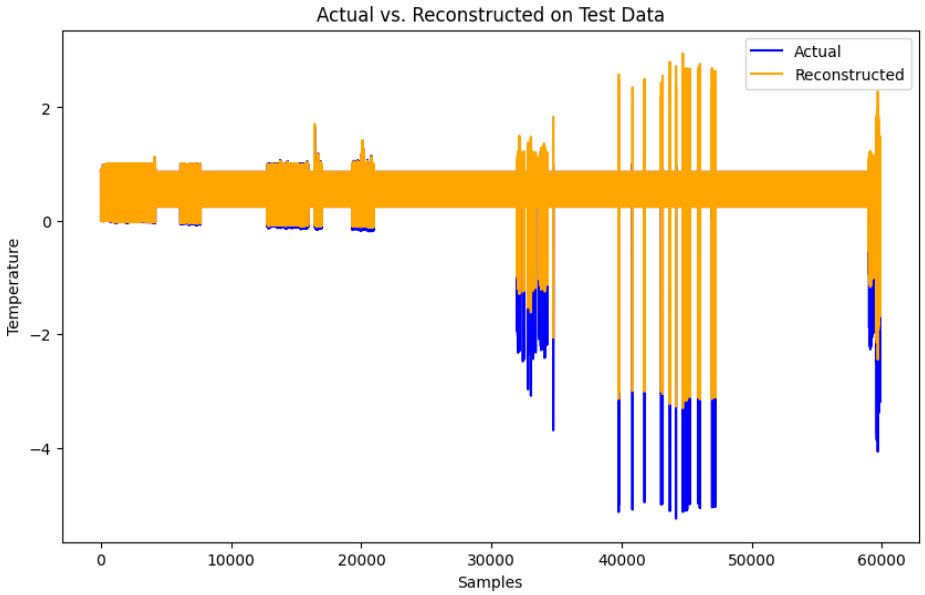


Figure 24: Predicting Dataset on unseen data

The LSTM Autoencoder, trained for 10 epochs with MSE as the loss function, demonstrated strong performance on the test dataset. The model achieved a Test Set MSE of 0.016 and an MAE of 0.019, indicating a low reconstruction error and closely aligned predictions with actual values. The R^2 value of 0.8954 further confirms that 89.5% of the variance in the test data was effectively captured, highlighting the model's strong ability to reconstruct temperature data accurately.

4.5.2 GRU

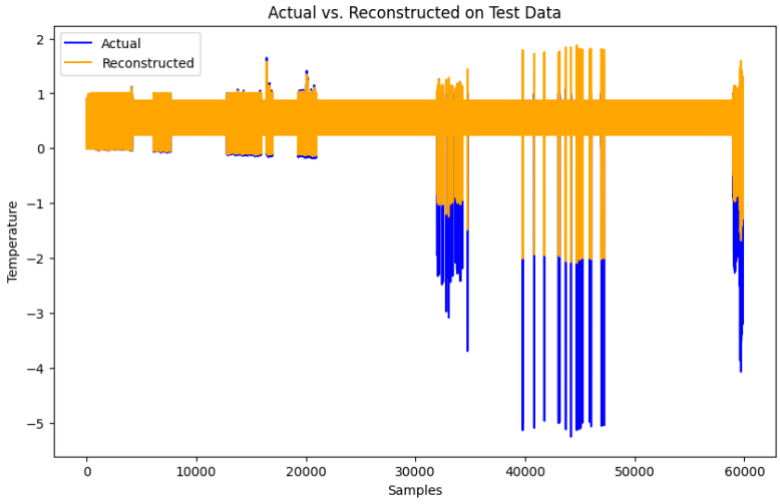


Figure 25: Actual vs Predicted Values on GRU model

The GRU Autoencoder model was designed and implemented to capture temporal dependencies and reconstruct the input data, facilitating effective anomaly detection in the aquaponics temperature dataset. The model architecture followed an encoder-decoder structure, where the encoder compressed the input sequence into a latent representation, and the decoder reconstructed the sequence from this representation.

The GRU Autoencoder demonstrated strong performance, with a Test Set MSE of 0.018 and an MAE of 0.015, indicating closely aligned predictions with actual values. The R^2 of 0.881 shows that 88.1%

of the variance in the test data was effectively captured, reflecting the model's ability to reconstruct the temperature data with minimal error.

4.5.3 CNN

The CNN Autoencoder model was designed and implemented to capture temporal dependencies and reconstruct the input data. The model architecture followed an encoder-decoder structure, where the encoder compressed the input sequence into a latent representation, and the decoder reconstructed the sequence from this representation.

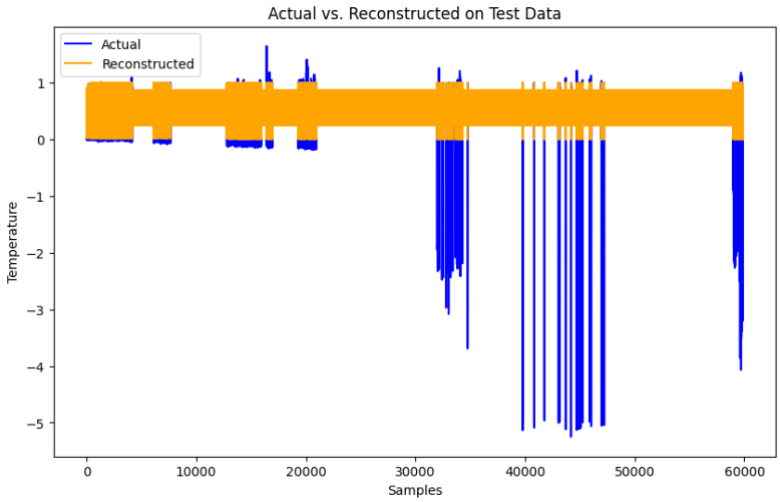


Figure 26: Actual vs Predicted values of CNN with autoencoder model

The model achieved a Test Set MSE of 0.0515 and an MAE of 0.0282, indicating that its predictions were closely aligned with the actual values. With an R^2 of 0.6631, the model explained 66.31% of the variance in the test data. While effective, the CNN Autoencoder's performance was less satisfactory compared to the GRU and LSTM with Autoencoder models.

4.5.4 Neural Network

The CNN Autoencoder model was designed and implemented to capture temporal dependencies and reconstruct the input data. The model architecture followed an encoder-decoder structure, where the encoder compressed the input sequence into a latent representation, and the decoder reconstructed the sequence from this representation.

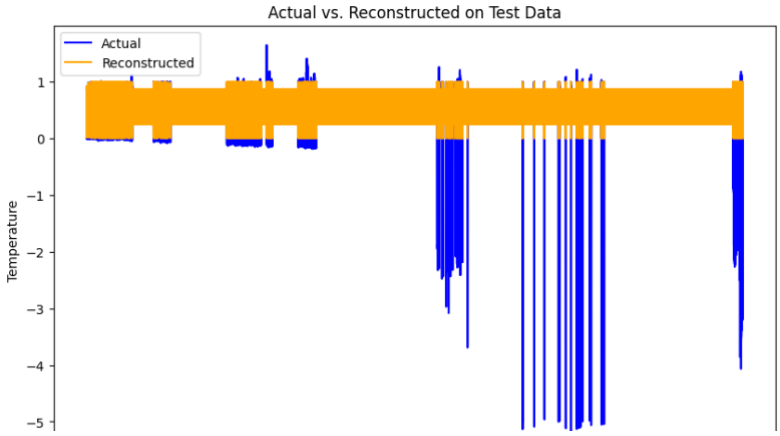


Figure 27: Actual vs Predicted values of NN with autoencoder model

Samples

Figure 28: Actual vs Predicted values of CNN with autoencoder model

The model achieved a Test Set MSE of 0.0540 and an MAE of 0.0329, indicating that its predictions closely matched the actual values. With an R^2 of 0.6468, the model explained 64.68% of the variance in the test data. While effective in reconstructing temperature data, its performance was less satisfactory compared to the GRU and LSTM with Autoencoder models.

4.5.5 Hybrid Model

A Hybrid Autoencoder model was developed to utilize the strengths of both LSTM and GRU networks in capturing the complex temporal patterns found in the aquaponics temperature dataset. The model was designed with an encoder-decoder structure, where the encoding of temporal data was handled by LSTM layers, and the reconstruction of the input sequence was managed by GRU layers.

In the encoder, an LSTM layer with 128 units was employed to process the input sequence and generate a latent representation. This latent space was then repeated across all timesteps using a RepeatVector layer, which prepared the data for the decoding phase. The decoder relied on a GRU layer, also with 128 units, to reconstruct the sequence from this latent representation. The final output was produced through a TimeDistributed Dense layer, ensuring each timestep was reconstructed independently.

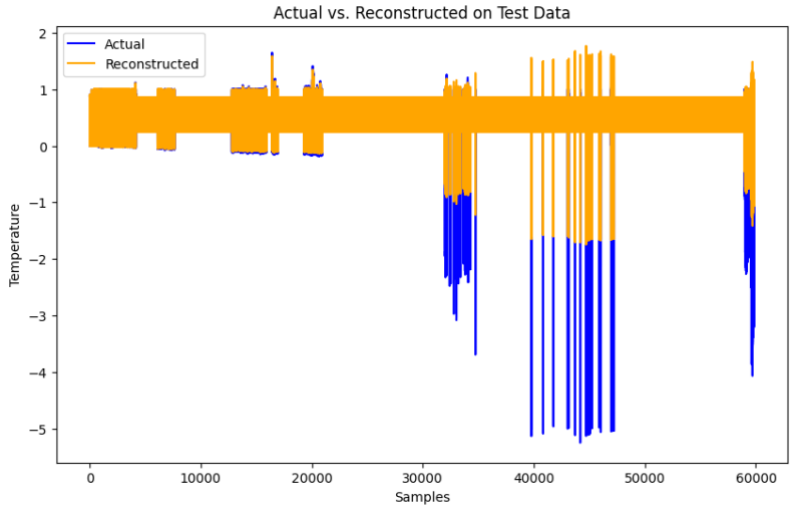


Figure 29: Actual vs Predicted values on Hybrid Model

The model exhibited a moderate Test Set MSE of 0.0226 and a small MAE of 0.0180, indicating a relatively minor difference between actual and reconstructed values. With an R^2 of 0.8522, the model explained 85.2% of the variance in the test data, reflecting its strong ability to capture underlying patterns. A visual comparison showed that the reconstructed values closely aligned with the actual data, though some discrepancies were noted in specific regions, suggesting areas of complexity that the model struggled to fully capture.

4.5.6 Comparison of Models

The performance of five different autoencoder models—LSTM, GRU, CNN, Neural Network, and Hybrid LSTM-GRU—was evaluated and compared based on Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 values. The results of this comparison are presented in the figure above.

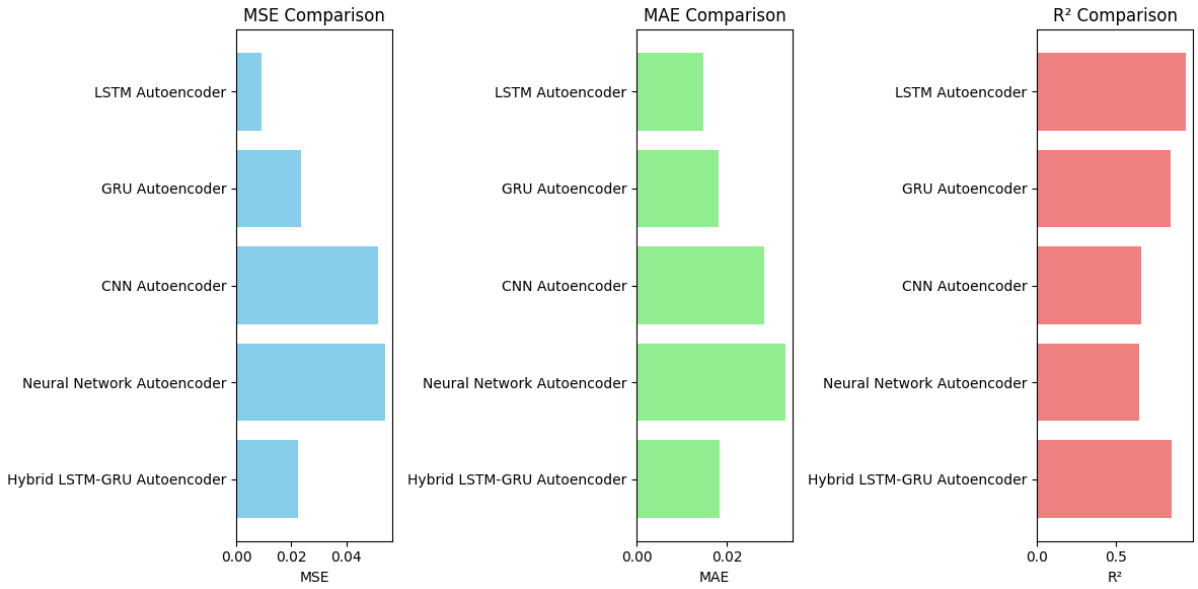


Figure 30: Comparison of all model performance for prediction

The Hybrid LSTM-GRU Autoencoder demonstrated the best overall performance, achieving the lowest MSE and MAE, closely followed by the GRU Autoencoder. The LSTM Autoencoder also performed well, particularly in explaining the most variance (highest R^2 value) in the test data. Meanwhile, the CNN and Neural Network Autoencoders were less precise, with higher error metrics and lower R^2 values, indicating they captured less variance in the data.

4.6 Insights into Fish Growth & Correlation Between Parameters

1. Correlation Analysis

The correlation matrix provided the first insight into the relationships between various environmental parameters and fish growth metrics (Fish Length and Fish Weight). The matrix revealed the following key points:

Temperature: Shows a moderate negative correlation with Fish Length (-0.42) and Fish Weight (-

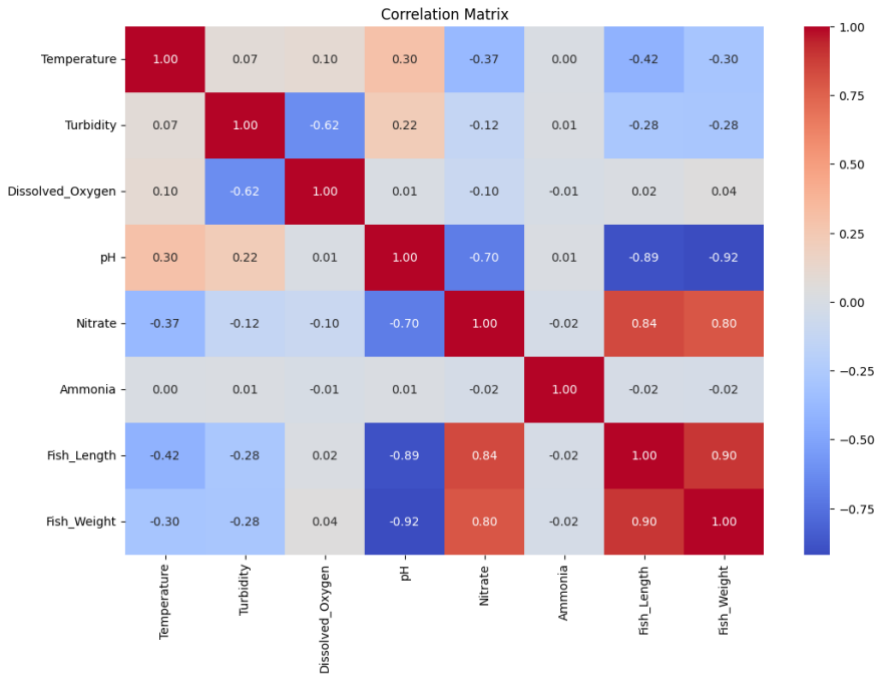


Figure 31: Correlation Matrix between all features

0.30). This suggests that as temperature increases, both the length and weight of the fish tend to decrease.
pH has a strong negative correlation with Fish Length (-0.89) and Fish Weight (-0.92), indicating that lower pH levels are linked to larger and heavier fish. Conversely, Nitrate shows a strong positive correlation with Fish Length (0.84) and Fish Weight (0.80), suggesting that higher nitrate levels support increased fish growth. Turbidity and Dissolved Oxygen have weaker negative correlations with fish growth, implying they have less impact compared to pH and Nitrate. These findings highlight the significant role of pH and Nitrate in influencing fish growth.

2. Scatter Plots and Pair Plots

The scatter plots and pair plots help visualize the relationships observed in the correlation matrix.
Fish Length and Fish Weight: The scatter plots confirm the strong linear relationship between fish length and weight, as indicated by the correlation matrix.
Environmental Parameters: The scatter plots reinforce that pH and Nitrate have more distinct relationships with fish growth metrics, while other parameters like Turbidity and Ammonia show more scattered and less consistent patterns.

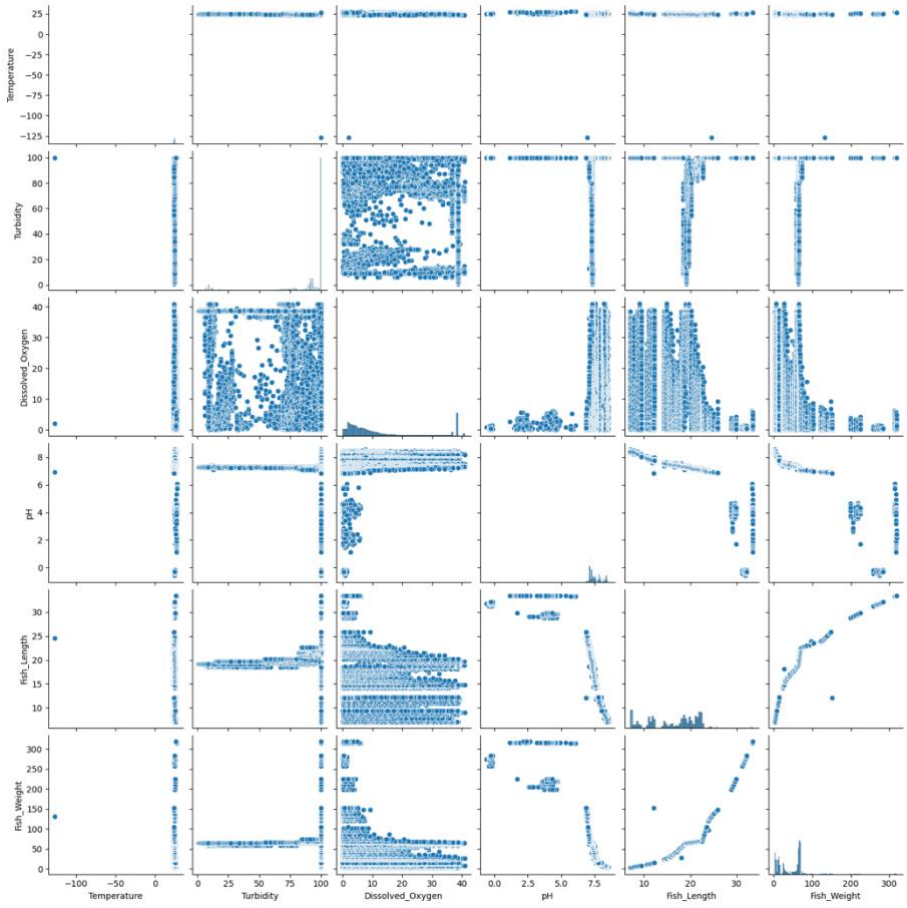


Figure 32: Scatter and pair plot of all features

These visualizations confirm the earlier correlation findings and help identify potential outliers or non-linear relationships that might require further exploration.

3. Regression Analysis

Two regression models were developed to predict Fish Length and Fish Weight using environmental parameters as predictors.

The Fish Length Model achieved an MSE of 6.93 and an R^2 of 0.748, indicating that about 75% of the variance in fish length is explained by the model, with predictions close to actual values.
The Fish Weight Model, with an MSE of 166.37 and an R^2 of 0.846, explained approximately 85% of the variance in fish weight, showing slightly better performance than the Fish Length model.

These results highlight the significant role of environmental parameters in predicting fish growth, with the Fish Weight model demonstrating higher accuracy.

4. Growth Over Time

The final step involved plotting fish growth (length and weight) over time:

Growth Over Time: The plot shows a general trend of increasing fish length and weight over time. However, significant increases in fish weight appear more pronounced than increases in length, suggesting that while fish grow longer over time, they gain weight more rapidly.

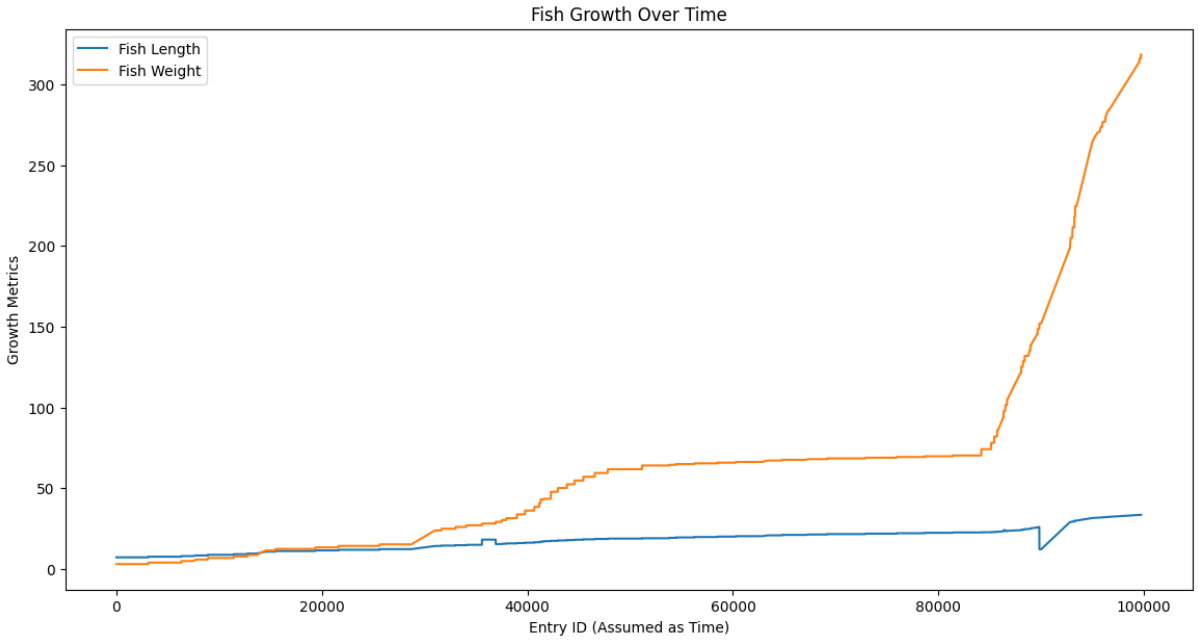


Figure 33: Fish Growth over time

This time-series analysis helps illustrate how fish growth metrics evolve over time, potentially pointing to periods of accelerated growth or environmental factors that might contribute to such trends.
The analysis revealed that pH and Nitrate significantly impact fish growth, with Temperature also being critical. These findings suggest that optimizing these parameters could enhance fish farming outcomes. The regression models showed strong predictive power, particularly for fish weight, and visualizations confirmed the relationships between environmental factors and growth metrics. In summary, careful management of pH, Nitrate, and Temperature is essential for promoting healthy fish growth in aquaculture settings.

4.7 Summary

In this study, multiple autoencoder models were developed and compared to evaluate their performance in anomaly detection and reconstructing the temperature data from an aquaponics system. The best model for anomaly detection we got is LSTM with autoencoder model.The analysis was conducted using MSE, MAE, and R^2 metrics. The Hybrid LSTM-GRU Autoencoder demonstrated the best overall performance across these metrics, with the lowest MSE and MAE and a

strong R^2 value. The GRU Autoencoder also performed well, particularly in terms of error metrics. While the LSTM Autoencoder showed a high R^2, the Hybrid model's balanced performance across all metrics suggests that it was more effective in this specific context.

It is worth noting that, although an attempt was made to predict ammonia levels, the data for this variable was insufficient and unrealistic. As a result, the models could not accurately predict ammonia values. Instead, the focus was shifted to predicting temperature, where success was achieved. Additionally, while a Conv1D model was explored as part of the study, it was observed that each of the standalone models—LSTM, GRU, CNN, Neural Network, and Hybrid—outperformed the Conv1D model in this context.

Furthermore, the analysis revealed significant correlations between fish growth parameters (such as length and weight) and environmental factors, particularly pH and nitrate levels. These findings underscore the importance of maintaining optimal water quality to promote healthy fish development in aquaponics systems.

These findings suggest that the selected models, particularly the LSTM with Autoencoder for anomaly detection and the Hybrid LSTM-GRU Autoencoder, were well-suited for the task of reconstructing temperature data in aquaponics systems. The performance differences between the models highlight the importance of model selection based on the specific characteristics of the data and the task at hand.

Chapter 5: Conclusion

This chapter should summarize the entire research project, highlighting the key findings and their implications.

5.1 Summary of Research Outcomes

In this study, a series of autoencoder models were developed and applied to the task of anomaly detection and predictive analysis within an aquaponics system. Initially, LSTM with Autoencoder, a Hybrid model combining Conv1D and LSTM with Autoencoder, and GRU-based Autoencoder models were employed for anomaly detection. Among these, the LSTM with Autoencoder exhibited the best performance, demonstrating superior capability in identifying anomalies within the dataset.

For the predictive analysis, the study expanded to include additional models: LSTM with Autoencoder, GRU with Autoencoder, CNN with Autoencoder, Neural Network with Autoencoder, and a Hybrid model combining LSTM and GRU. The evaluation metrics—Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2)—indicated that the Hybrid LSTM-GRU Autoencoder model performed the best overall. It achieved the lowest MSE and MAE, and a strong R^2 value, making it the most robust model for this context. However, it is important to note that while the Hybrid model showed the best overall metrics, the LSTM with Autoencoder provided the most accurate fit in terms of the predictive plots, especially in reconstructing the temperature data.

The study also delved into fish growth insights, where it was determined through comprehensive analysis that pH and Nitrate levels are the most significant factors influencing fish growth, with Temperature also playing a critical role. These findings are crucial for optimizing aquaponics systems to promote healthy fish growth. The regression models used in the analysis demonstrated good predictive power, particularly in predicting fish weight, and the visualizations clearly highlighted the relationships between various environmental parameters and fish growth metrics.

5.2 Research Limitations

Despite the promising outcomes of this study, several limitations must be acknowledged. Firstly, the dataset for ammonia and nitrate presented challenges, as many values were found to be unrealistic or insufficient. This issue significantly impacted the model's ability to predict these variables accurately. It is believed that if the data could be collected directly through proper sensors designed specifically for water analysis—rather than repurposing sensors intended for air—much more reliable results could be obtained. This would allow for a more accurate and comprehensive study.

Another significant limitation pertains to time and computational power. The models developed, especially when trained over 50 epochs, required an extensive amount of time—over 24 hours per model. Moreover, the process of validation using techniques like k-fold cross-validation and hyperparameter tuning through grid search added substantial computational demands, often extending the time required to several hours or even days. Due to these constraints, only the LSTM with Autoencoder model was fully explored for anomaly detection with hyperparameter tuning. With more time and computational resources, it would have been possible to perform a more thorough exploration across all models, incorporating rigorous validation and tuning to optimize performance further.

5.3 Future Research Direction

Looking forward, it is envisioned that future research could focus on collecting data with proper sensors designed for water analysis over a longer period, ideally one year. Securing the necessary funding would be crucial for this endeavor, particularly to obtain sensors that accurately measure nitrate and ammonia levels in water. Such data collection would greatly enhance the quality and reliability of the dataset, allowing for a more in-depth study.

With sufficient computational power and time, it would be possible to explore additional models, apply thorough validation methods, and optimize hyperparameters through techniques such as grid search. This would likely yield even more accurate and robust models. Furthermore, accurate prediction and anomaly detection for nitrate and ammonia could lead to significant practical applications, such as enabling small-scale aquaponics systems for business ventures, reducing costs by potentially eliminating the need for expensive nitrate and ammonia sensors.

Reference

1. The Guardian. 2023. Rampant Heatwaves Threaten Food Security of Entire Planet, Scientists Warn. [online] Available at: https://www.theguardian.com/environment/2023/jul/21/rampant-heatwaves-threaten-food-security-of-entire-planet-scientists-warn [Accessed 5 February 2024].
2. World Bank, 2023. Urban Development. Washington, DC: The World Bank. Available at: https://www.worldbank.org/en/topic/urbandevelopment/overview [Accessed 5 February 2024].
3. Food and Agriculture Organization of the United Nations, 2017. The Future of Food and Agriculture: Trends and Challenges. Quebec: FAO.
4. Abbasi, R., Martinez, P. and Ahmad, R., 2023. Automated Visual Identification of Foliage Chlorosis in Lettuce Grown in Aquaponic Systems. Agriculture, 13, p.615. https://doi.org/10.3390/agriculture13040615.
5. Wijayanto, A., Wardhana, K. and Aziz, A., 2021. Implementation of Internet of Things (IoT) for Aquaponic System Automation. In: Proceedings of the 2021 International Conference on Computer, Control, Informatics and Its Applications (IC3INA '21), Virtual/Online Conference, Indonesia, 5–6 October 2021. pp.176-181. https://doi.org/10.1109/IC3INA53402.2021.9615095.
6. Yanes, A.R., Martinez, P. and Ahmad, R., 2020. Towards automated aquaponics: A review on monitoring, IoT, and smart systems. Journal of Cleaner Production, 263, p.121571. https://doi.org/10.1016/j.jclepro.2020.121571.
7. Encyclopedia, 2024. 4, pp.313-336. https://doi.org/10.3390/encyclopedia4010023.
8. Tokunaga, K., Tamaru, C., Ako, H. and Leung, P., 2015. Economics of Small-scale Commercial Aquaponics in Hawai'i. Journal of the World Aquaculture Society, 46, pp.20–32. https://doi.org/10.1111/jwas.12173.
9. Lennard, W., 2017. Commercial Aquaponic Systems: Integrating Recirculating Fish Culture with Hydroponic Plant Production. [e-book] Available at: https://www.researchgate.net/publication/316692089_Commercial_aquaponic_systems_integrating_recirculating_fish_culture_with_hydroponic_plant_production [Accessed 5 February 2024].
10. Monsees, H., Kloas, W. and Wuertz, S., 2017. Decoupled systems on trial: Eliminating bottlenecks to improve aquaponic processes. PLoS ONE, 12, p.e0183056. https://doi.org/10.1371/journal.pone.0183056.
11. Kloas, W., Gro., R., Baganz, D., Graupner, J., Monsees, H., Schmidt, U., Staaks, G., Suhl, J., Tschirner, M., Wittstock, B., et al., 2015. A new concept for aquaponic systems to improve sustainability, increase productivity, and reduce environmental impacts. Aquaculture Environment Interactions, 7, pp.179–192. https://doi.org/10.3354/aei00146.
12. Suhl, J., Dannehl, D., Kloas, W., Baganz, D., Jobs, S., Scheibe, G. and Schmidt, U., 2016. Advanced aquaponics: Evaluation of intensive tomato production in aquaponics vs. conventional hydroponics. Agricultural Water Management, 178, pp.335–344. https://doi.org/10.1016/j.agwat.2016.10.013.
13. Tokunaga, K., Tamaru, C., Ako, H. and Leung, P., 2015. Economics of Small-scale Commercial Aquaponics in Hawai'i. Journal of the World Aquaculture Society, 46, pp.20–32. https://doi.org/10.1111/jwas.12173.
14. Delaide, B., Goddek, S., Gott, J., Soyeurt, H. and Jijakli, M.H., 2016. Lettuce (Lactuca sativa L. var. Sucrine) Growth Performance in Complemented Aquaponic Solution Outperforms Hydroponics. Water, 8, p.467. https://doi.org/10.3390/w8100467.
15. Karimanzira, D. and Rauschenbach, T., 2021. An intelligent management system for aquaponics. At-Automatisierungstechnik, 69, pp.345–350. https://doi.org/10.1515/auto-2021-0035.
16. John, J. and Mahalingam, P.R., 2021. Automated Fish Feed Detection in IoT Based Aquaponics System. In: Proceedings of the 2021 8th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 1–3 July 2021. pp.286–290. https://doi.org/10.1109/ICSCC51209.2021.9528171.
17. Peng, W., Wang, X. and Zhang, J., 2021. A Review on AI-Based Approaches for

Monitoring and Optimizing Aquaponics Systems. Environmental Science & Technology, 55(4), pp.2306-2321. https://doi.org/10.1021/acs.est.0c08597.

18. Abbasi, A.Z., Islam, N., Shaikh, Z.A. and Gillani, S.M.M., 2019. Machine Learning Algorithms for Identifying Plant Health Issues in Precision Agriculture. Computers and Electronics in Agriculture, 156, pp.434-442. https://doi.org/10.1016/j.compag.2018.11.037.

19. Fasoula, V.A. and Fasoula, D.A., 2019. The Honeycomb Methodology: A Novel Approach in Artificial Intelligence and Machine Learning for Precision Agriculture. Agronomy, 9(10), p.614. https://doi.org/10.3390/agronomy9100614.

20. Tsouros, D.C., Tsoulos, I.G. and Tzes, A., 2019. Internet of Things (IoT) in Agriculture: Connected Wireless Sensor Networks for Sustainable and Smart Farming. IEEE Systems Journal, 13(3), pp.3370-3378. https://doi.org/10.1109/JSYST.2019.2916456.

21. Singh, A.K., Saha, S. and Rai, A., 2020. Smart Farming Using IoT and AI: A Comprehensive Review. Journal of Sensor Technology, 10(3), pp.123-135. https://doi.org/10.4236/jst.2020.103008.

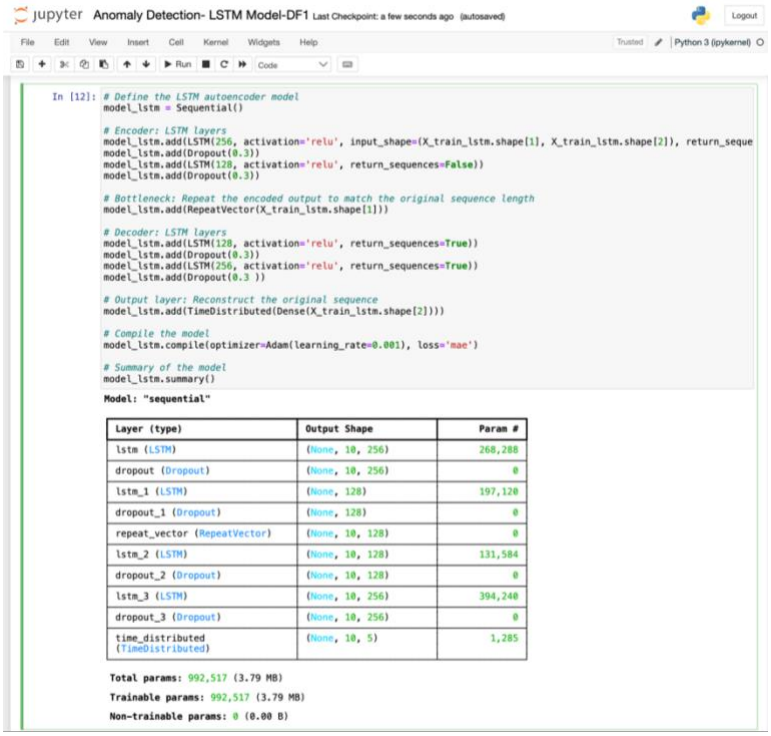
22. Lee, B.H. and Lee, H., 2022. Cost-Benefit Analysis of Implementing AI and IoT in Small-Scale Aquaponics Systems. Sustainability, 14(1), p.341. https://doi.org/10.3390/su14010341.

23. Smith, M.A. and Patel, K., 2020. Challenges in Scaling AI Solutions for Large-Scale Farming Operations. Journal of Agricultural Informatics, 11(1), pp.27-38. https://doi.org/10.17700/jai.2020.11.1.540.

24. Karimanzira, D. and Rauschenbach, T., 2019. Using Convolutional Neural Networks and Long Short-Term Memory Networks for Aquaponics Monitoring. Procedia CIRP, 79, pp.362-367. https://doi.org/10.1016/j.procir.2019.02.013.

Appendices

1. LSTM with autoencoder model



2. Hybrid model(Conv1D + LSTM with autoencoder)

3. GRU model

```
In [12]: # Define the GRU model
model_gru = Sequential()

# Encoder: GRU layers
model_gru.add(GRU(256, activation='relu', input_shape=(X_train_lstm.shape[1], X_train_lstm.shape[2]), return_sequenc
model_gru.add(Dropout(0.3))
model_gru.add(GRU(128, activation='relu', return_sequences=False))
model_gru.add(Dropout(0.3))

# Bottleneck: Repeat the encoded output to match the original sequence length
model_gru.add(RepeatVector(X_train_lstm.shape[1]))

# Decoder: GRU layers
model_gru.add(GRU(128, activation='relu', return_sequences=True))
model_gru.add(Dropout(0.3))
model_gru.add(GRU(256, activation='relu', return_sequences=True))
model_gru.add(Dropout(0.3))

# Output layer: Reconstruct the original sequence
model_gru.add(TimeDistributed(Dense(X_train_lstm.shape[2])))

# Compile the model
model_gru.compile(optimizer=Adam(learning_rate=0.001), loss='mae')

# Summary of the model
model_gru.summary()
```

Model: "sequential"

Layer (type)	Output Shape	Param #
gru (GRU)	(None, 10, 256)	201,984
dropout (Dropout)	(None, 10, 256)	0
gru_1 (GRU)	(None, 128)	148,224
dropout_1 (Dropout)	(None, 128)	0
repeat_vector (RepeatVector)	(None, 10, 128)	0
gru_2 (GRU)	(None, 10, 128)	99,072
dropout_2 (Dropout)	(None, 10, 128)	0
gru_3 (GRU)	(None, 10, 256)	296,448
dropout_3 (Dropout)	(None, 10, 256)	0
time_distributed (TimeDistributed)	(None, 10, 5)	1,285

Total params: 747,013 (2.85 MB)

Trainable params: 747,013 (2.85 MB)

Non-trainable params: 0 (0.00 B)

4. Hybrid model + Isolation Forest Model