UNIVERSITI TENAGA NASIONAL
The Energy University

Anomaly Detection for Smart Farming via Machine Learning

FINAL YEAR PROJECT S223/24
PANEL PRESENTATION

NAME : MUHAMMAD LUQMAN HAKIMI BIN MOHD ZIN
STUDENT ID : CS01081026
SUPERVISOR : ASSOC.PROF.TS.DR.ASMIDAR BTE ABU BAKAR
EXAMINER : TS.MARINA BTE MD. DIN

INTRODUCTION & BACKGROUND

Agriculture is one of important sector in economic growth, human rely on it to have their basic needs for sustainability. One of the branches of agriculture is aquaculture

Smart agriculture is the integration of agriculture with technology. To enhance human productivity, do complex tasks, and improve customer experiences.

The integration of technology (IoT) might possibly expose to cyber threats and disrupt daily operation of smart farming environment.

Many security mitigation applied to reduce the potential cyber security risk in IoT, one of it by using machine learning algorithms.

1 **Agriculture**
the science of cultivating the soil, and raising livestock in order to produce product.

2 **Aquaculture**
One of agriculture branches, focus on aquatic production either plants or animals

3 **Internet-of-Things (IoT)**
Use of technology to enhance productivity, and efficiency

4 **Security Issues**
IoT employs wireless connections, easy target for any attacks. Able to disrupt smart environment operation

PROBLEM STATEMENT

Context

Smart farming in aquaculture uses IoT for precise monitoring and management. Environmental factors like pH, temperature and turbidity is crucial for development of aquatic animals or plants.

Challenge

- IoT can possibly be expose to cyber threats or inadvertent threats.
- Some studies have demonstrated potential anomaly identifications like **(Moso et al, 2021), (Catalano et al., 2022), and (Petkovski & Shehu,.2023)**, but still lack of research about anomaly detection algorithm on smart aquaculture.

RESEARCH QUESTION

- Which machine learning algorithms are best for detecting anomalies in smart aquaculture systems?
- How do they compare in terms of characteristics and performance metrices?
- How effective is the selected machine learning model for anomaly detection?

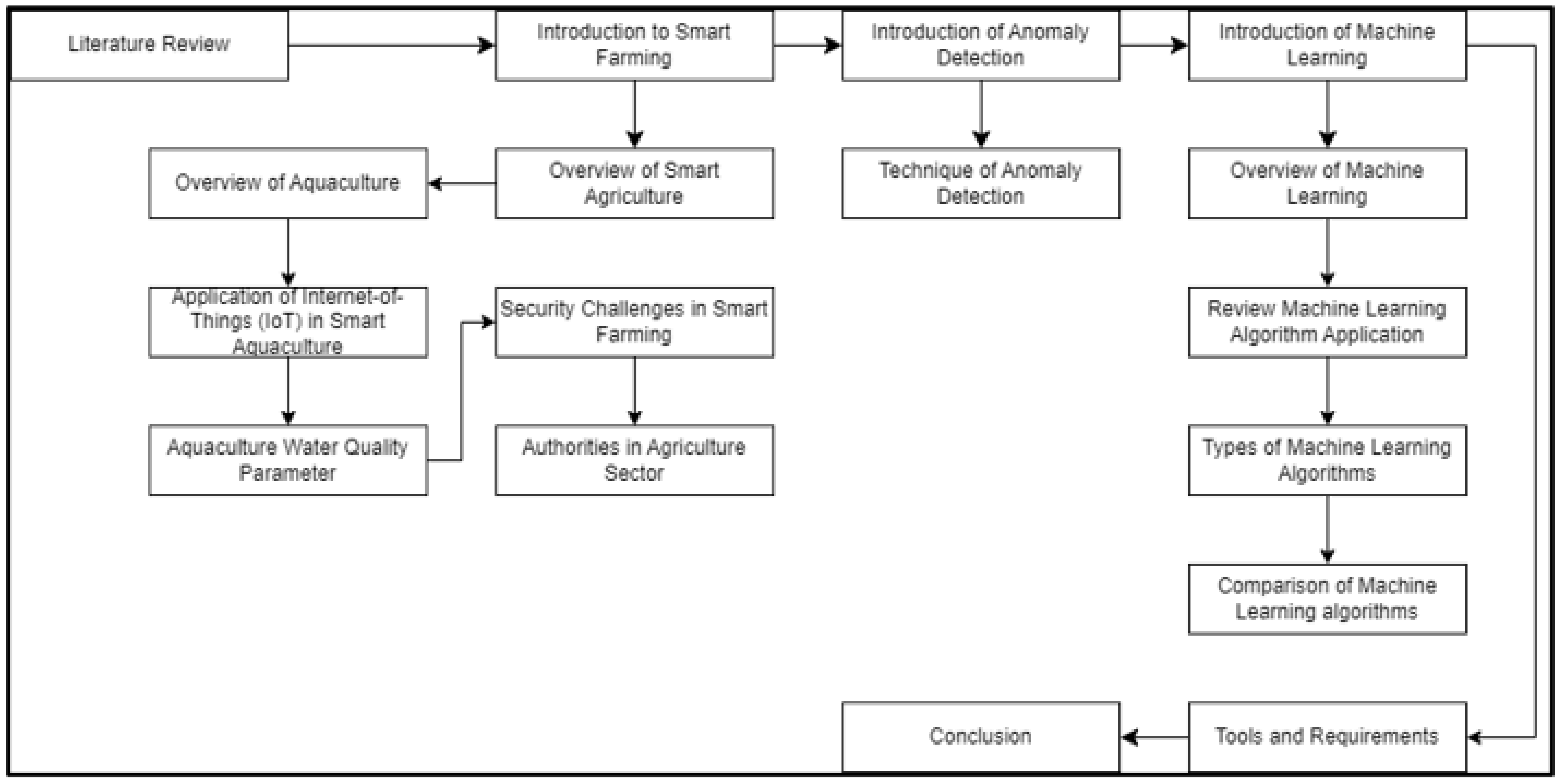
PROJECT OBJECTIVE

- To review and identify suitable machine learning algorithms for anomaly detection in smart farming focusing on smart Aquaculture.
- To conduct comparative analysis of machine learning algorithms.
- To analyze result from the selected machine learning model.

Research Scope

Review and identify machine learning algorithms used for anomaly detection and conduct proof-of-concept to analyze selected algorithm with the used of Smart Aquaculture dataset.

LITERATURE REVIEW



- Using technology to enhance farmer productivity
- “Zero-hunger” in SDGs
- Implementation of IoT and Machine Learning.

Overview of Smart Farming

Overview of Smart Aquaculture

- Raising aquatic creatures under regulated environments (pH, Temperature)
- Help farmer to do complex tasks classify fish species (Islam et al., 2021).
- Adaptation of climate change

SECURITY CHALLENGE
IN SMART FARMING

Social Engineering
Malicious activity can be accomplished through human interaction.

False Data Injections
When injection happen, it could lead to excessive damage to aquatic lives.

Malware Injection Attack
Malware might disrupt the process, and could possibly spread to other host.

Introduction

Anomaly Detection

Anomaly refers to an observation of the values that differs from most of the normal values.

Unsupervised Anomaly Detection

Detection without a label of "Normal" or "Anomalous" data beforehand. Learn the deviation by analyzing the trend and pattern.

Supervised Anomaly Detection

Contains labelled data called as a target or labelled data
A classifier will be trained using a dataset labelled as an 'anomaly' or 'normal'
(Boukerche et al, 2020)

Mixed Anomaly Detection

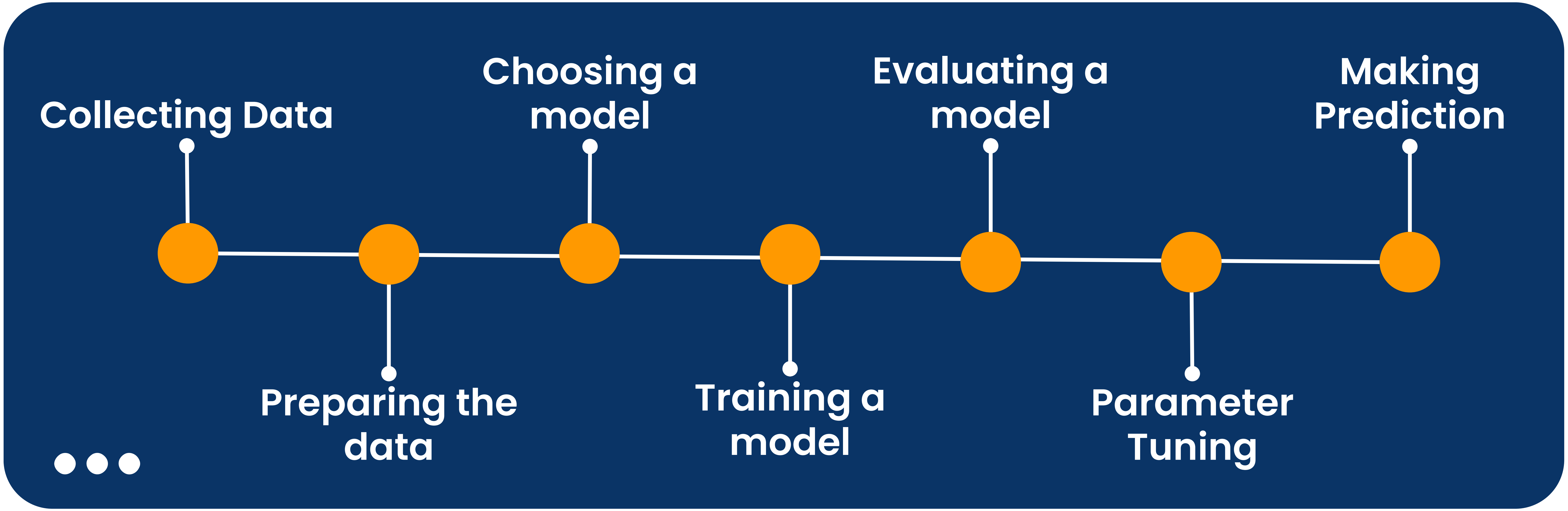
The combination of both techniques whereby usage of labelled and unlabeled data can overcome the disadvantages occurred from both techniques
PROBLEM: Data Heterogeneity

Introduction

Machine Learning

Computer intelligence to help human to do complicated tasks with the use of algorithms.

OVERVIEW OF MACHINE LEARNING



Review Machine Learning Algorithm Application

TABLE 2. 1: Simplified literature review on machine learning application

Study	Approach	Proposed Method	Dataset	Performance Metrices	Significant Test
(Primartha & Tama, 2017)	Anomaly Detection	Random Forest	NSL-KDD, UNSW-NB15, and GPRS	Accuracy, False Alarm Rate	Yes
(N. Elmrabit et al, 2020)	Anomaly Detection	Logistic Regression (LR), Gaussian Naïve Bayes (GNB), K-nearest neighbour (KNN), Decision Tree (DT), Adaptive Boosting (AdaB), Random Forest (RF), Convolutional Neural Network (CNN), Long Short-term memory (LSTM), Gated Recurrent Units (GRU), Simple Recurrent Neural Network (RNN), Deep Neural Network (DNN)	CICIDS-2017, UNSW-NB15, and Industrial Control System (ICS) cyber-attack datasets	Accuracy, precision, Recall, and AUC	Yes
(Catalano et al., 2022)	Anomaly Detection	Multivariate Linear Regression (MLR) and Long Short-term Memory (LSTM)	Agriculture time-series dataset	Min Squared Error (MSE)	Yes
(Liu et al., 2008)	Anomaly Detection	Isolation Forest	http (KDDCUP99) and ForestCover dataset	AUC	Yes
(Islam et al., 2021)	Fish Survival Prediction	Random Forest	Aquaculture dataset (Real-time fish data)	Accuracy, True Positive (TP) rate and Kappa statistics	No
Hu et al. (2019)	Water Quality Prediction	Long Short-term Memory (LSTM)	Aquaculture dataset from designated sensor	Accuracy, Time cost	No
Petkovski & Shehu, (2023)	Anomaly Detection	K-means, Isolation Forest, Local Outlier Factor	Real-time time-series aquaculture dataset	Accuracy, Precision, Recall, and F-score	Yes
Ramteke & Monali Y., (2012)	Anomaly Detection	K-nearest neighbour (KNN)	CT scan images of normal and abnormal.	Sensitivity, Specificity, and Accuracy	Yes

Research Methodology

Literature Review

Explore existing research and studies related to smart farming, machine learning, and anomaly detection

Data Acquisition

Acquire Smart Aquaculture data from Kaggle to be used for Proof-of-Concept

Proof-of-Concept

Conduct experiment to observe how selected machine learning works and analyze the results

Comparative Analysis

Compare reviewed machine learning by conducting comparative analysis

Chapter: Analysis

DATA ACQUISITION

DATASET EXPLAINATIONS

01 Description

1. Acquire from Kaggle
2. Created by Researcher in Dhaka University.
3. Consist of 591 data

02 Feature

1. pH Value
2. Temperature
3. Turbidity
4. 11 Fish Species

03 Data Pre-processing

1. Filter fish species
2. Inject anomalies
3. Train Test Split (70:30)

pH	Temperature	Turbidity	Fish
6	27	4	katla
7.6	28	5.9	sing
7.8	27	5.5	sing
6.5	31	5.5	katla
8.2	27	8.5	prawn
...	...	...	...

Real-Time Pond Water Dataset for Fish Farming.

70% Training Set

30% Testing Set

CRISP DM

Cross Industry Process for Data Mining (CRISP-DM) one of standardize methodology in data mining, analytics, and data science applications.

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

Proof-of-Concept

This phase involves conducting an analysis on literature and experiment to observe the result

01 **Planning to develop a machine learning model using CRISP DM**

02 **Use Smart Aquaculture acquired to test the model**

03 **Analyze and evaluate the model after testing phase**

OVERVIEW OF STUDIES ON MACHINE LEARNING ALGORITHM IN SMART FARMING

TOTAL OF
8 paper study

6 paper
ANOMALY DETECTION

2 paper
OTHER APPLICATIONS

Anomaly Detection Application

- **(Primartha & Tama, 2017)**
  - using Random Forest for anomaly detection
- **(N. Elmrabit et al, 2020)**
  - Evaluate 12 algorithms to detect anomaly
- **(Catalano et al., 2022)**
  - Use MLR and LSTM for anomaly detection in smart agriculture
- **(Liu et al., 2008)**
  - Use iForest to detect network anomaly
- **Petkovski & Shehu, (2023)**
  - Use K-means, iForest, LOF to detect time series dataset anomaly.
- **Ramteke & Monali Y., (2012)**
  - Use KNN for detect anomaly image of CT scan

Others Applications

- **(Islam et al., 2021)**
  - Fish Survival Prediction
- **Hu et al. (2019)**
  - Water Quality Prediction

Performance Metrices

- **(Primartha & Tama, 2017)**
  - Accuracy, False Alarm Rate
- **(N. Elmrabit et al, 2020)**
  - Accuracy, precision, Recall, and AUC
- **(Liu et al., 2008)**
  - AUC
- **Petkovski & Shehu, (2023)**
  - Accuracy, Precision, Recall, and F1-score

Performance Matrices

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **AUC**

Analysis

01 Identify number of algorithms reviewed about application of machine learning

LSTM	proposed in (Catalano et al., 2022), Hu et al. (2019), (N. Elmrabit et al, 2020)
RF	proposed in (Primartha & Tama, 2017) , (Islam et al., 2021) , (N. Elmrabit et al, 2020)
IF	proposed in (Liu et al., 2008) , Petkovski & Shehu, (2023)
KNN	proposed in (Ramteke & Monali Y, 2012), (N. Elmrabit et al, 2020)
K-means & LOF	proposed in Petkovski & Shehu, (2023)
Others	proposed in (N. Elmrabit et al, 2020)

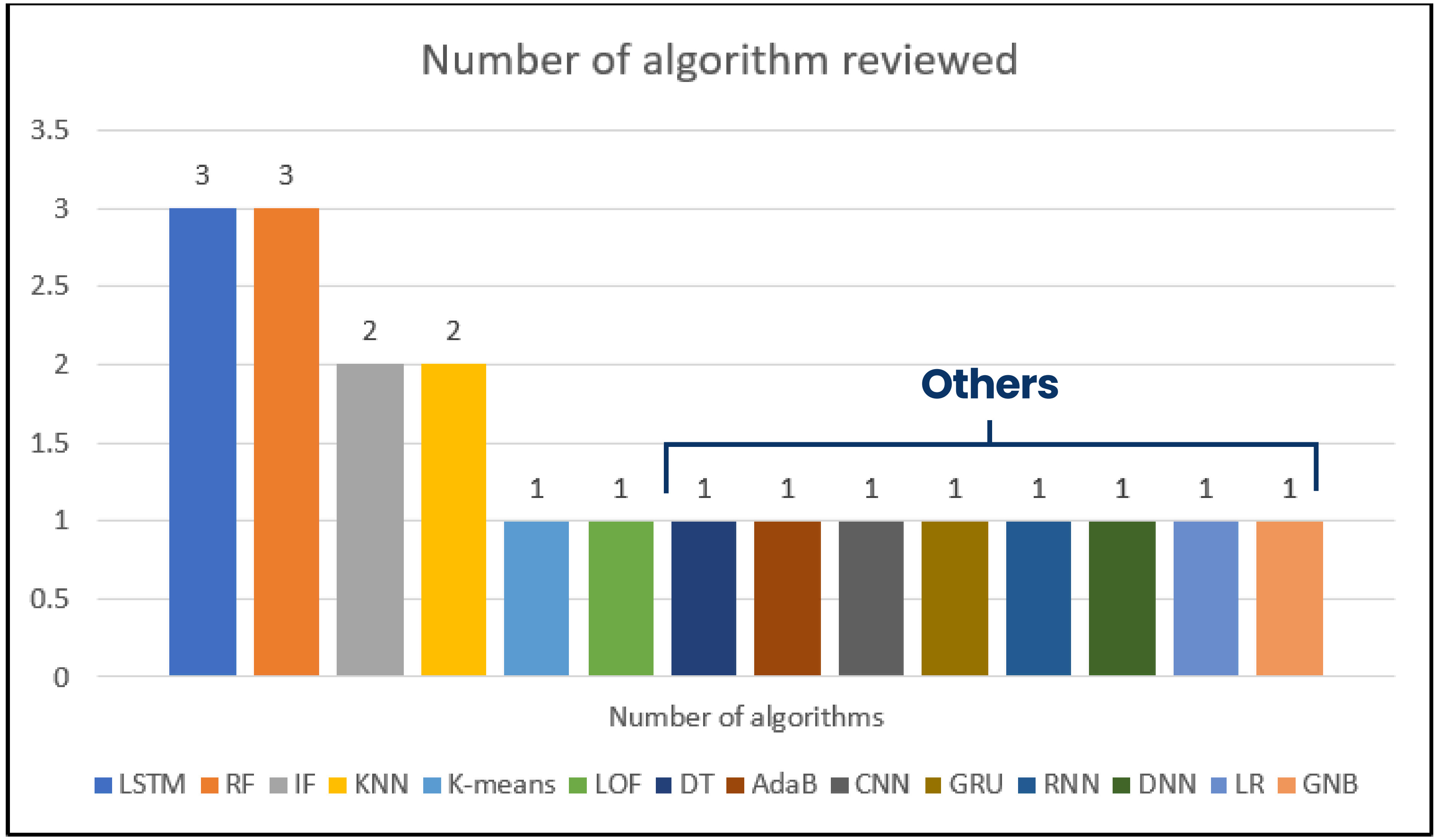


Figure 4. 2: Number of algorithms reviewed

Analysis

02 Identify number of algorithms reviewed that related with anomaly detection application

LSTM	proposed in (Catalano et al., 2022), (N. Elmrabit et al, 2020)
RF	proposed in (Primartha & Tama, 2017) , (N. Elmrabit et al, 2020)
IF	proposed in (Liu et al., 2008) , Petkovski & Shehu, (2023)
KNN	proposed in (Ramteke & Monali Y, 2012), (N. Elmrabit et al, 2020)
K-means & LOF	proposed in Petkovski & Shehu, (2023)
Others	proposed in (N. Elmrabit et al, 2020)

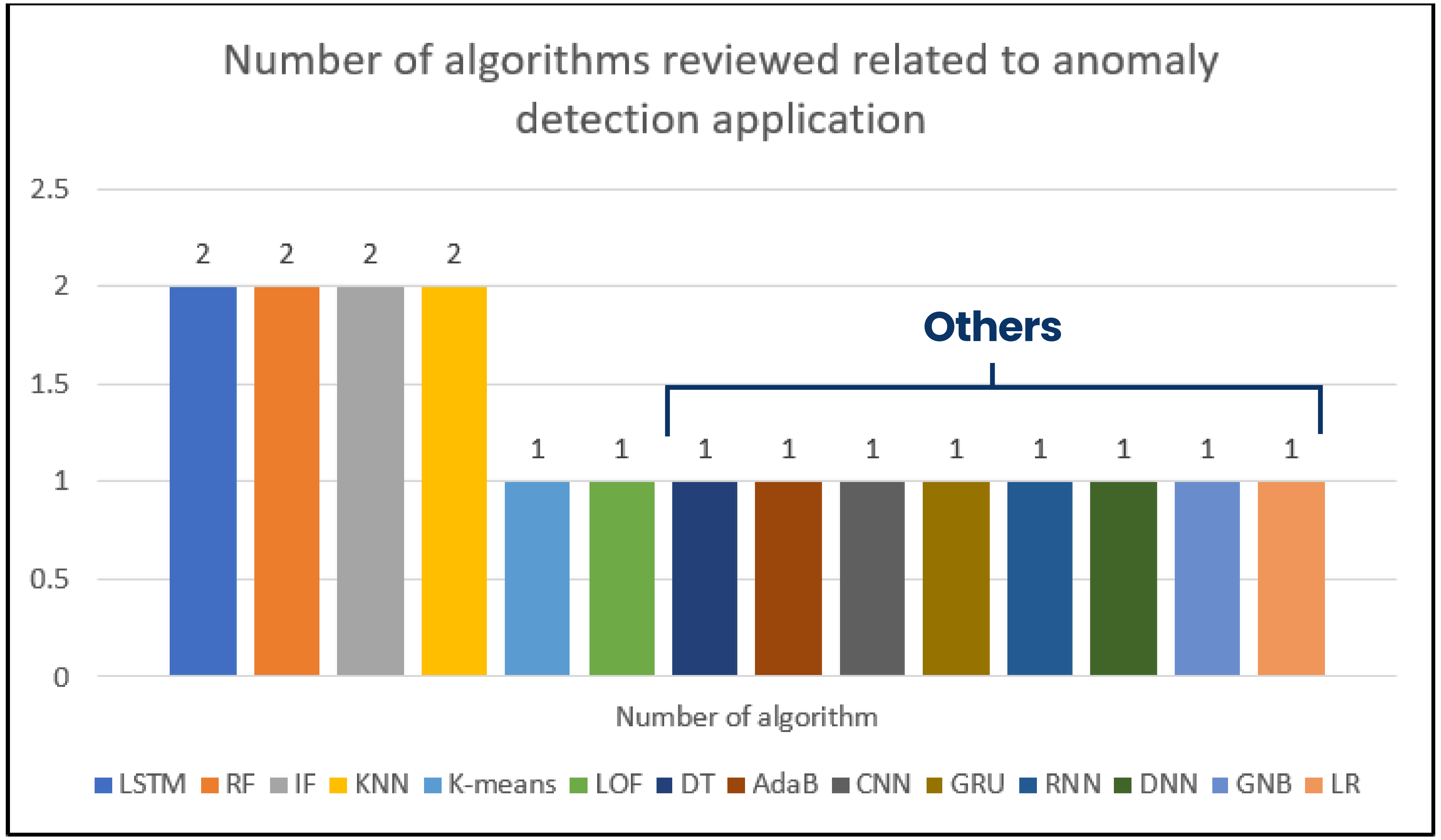


Figure 4. 3: Number of papers reviewed (in context of Anomaly Detection)

Comparison of Machine Learning

Table 2. 3: Comparison of Selected Machine Learning Algorithms

Attributes	LSTM	RF	IF	KNN
Accuracy	High for sequential data	High for structured data	High for anomaly detection	Varies, can be high with proper feature selection
Computational Cost	High due to complex neural network architecture	Moderate; depends on the number of trees	Low due to linear time complexity	Low during prediction, high during training
Speed	Slow due to intensive computation	Moderate; faster than LSTM but slower than Isolation Forest	Fast, especially efficient for large datasets	Slow for large dataset due to distance calculations
Advantages	Excellent for time-series and sequential data	Handles large datasets well; robust to overfitting	Efficient with low memory requirement; scalable	Simple to implement
Disadvantages	Requires large datasets and significant computational resources	Can be memory-intensive, less effective with noisy data	May not perform well on very small datasets; not suited for all types of data	Computationally intensive for large datasets
Robustness	Robust to varying time lags in data	Robust to overfitting and noisy data	Robust to irrelevant attributes and high-dimensional data	Moderate, sensitive to noise and irrelevant features.

Analysis

03 Compare the selected algorithm characteristics from Literature Review

LSTM	High Computational Cost Low on Speed
RF	High Accuracy, Moderate on Computational Cost & Speed
IF	Low Computational Cost High Accuracy & Speed
KNN	Low Computational Cost, Speed, and Robustness

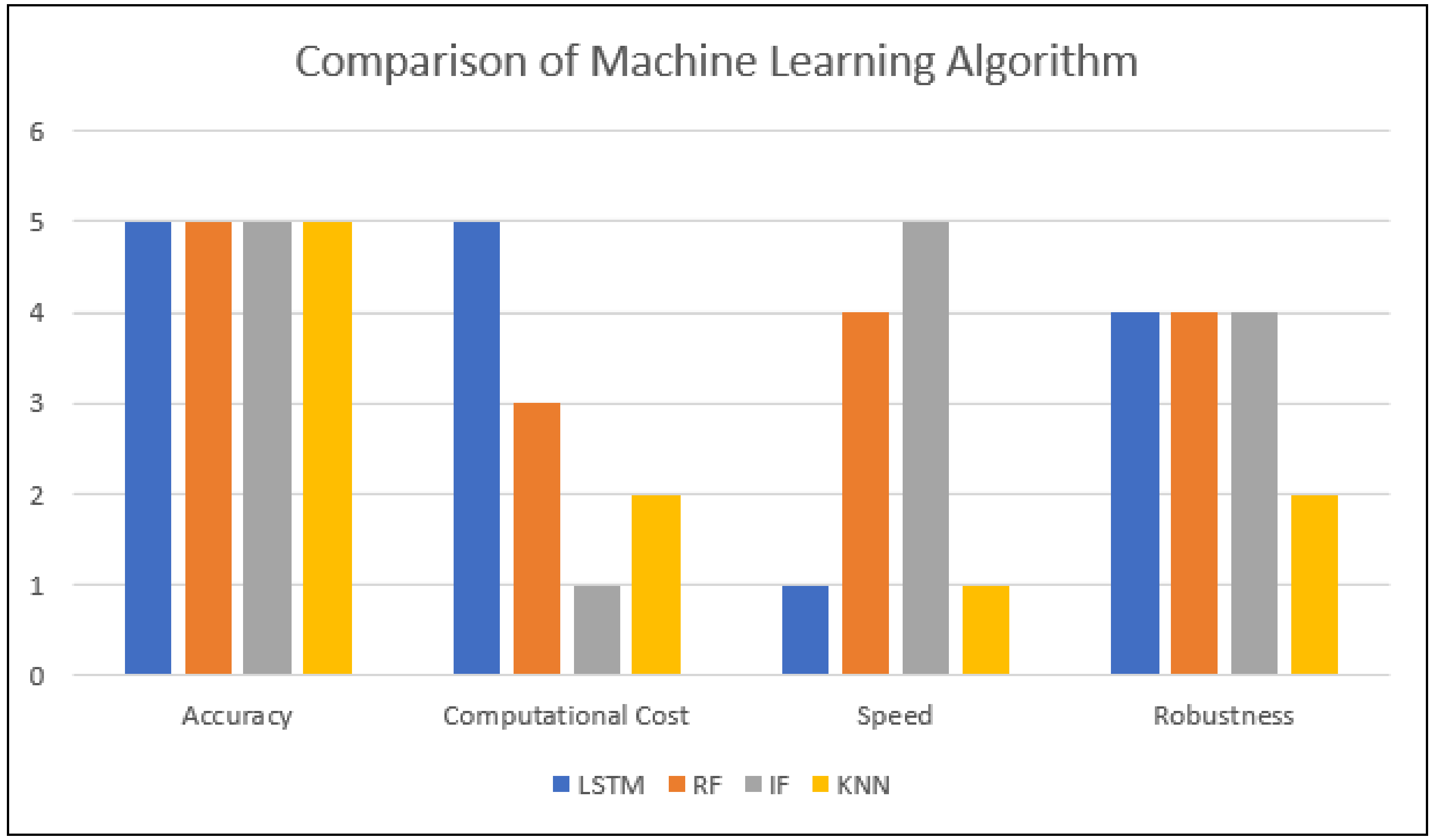


Figure 4. 4: Bar graph of algorithms analysis

Analysis Conclusion

- LSTM is high on computational cost, low speed and not suitable for this smart aquaculture dataset.
- RF, IF, KNN suitable for classifications. Continue the comparative analysis with the dataset and proof-of-concept.

01 Machine Learning Design

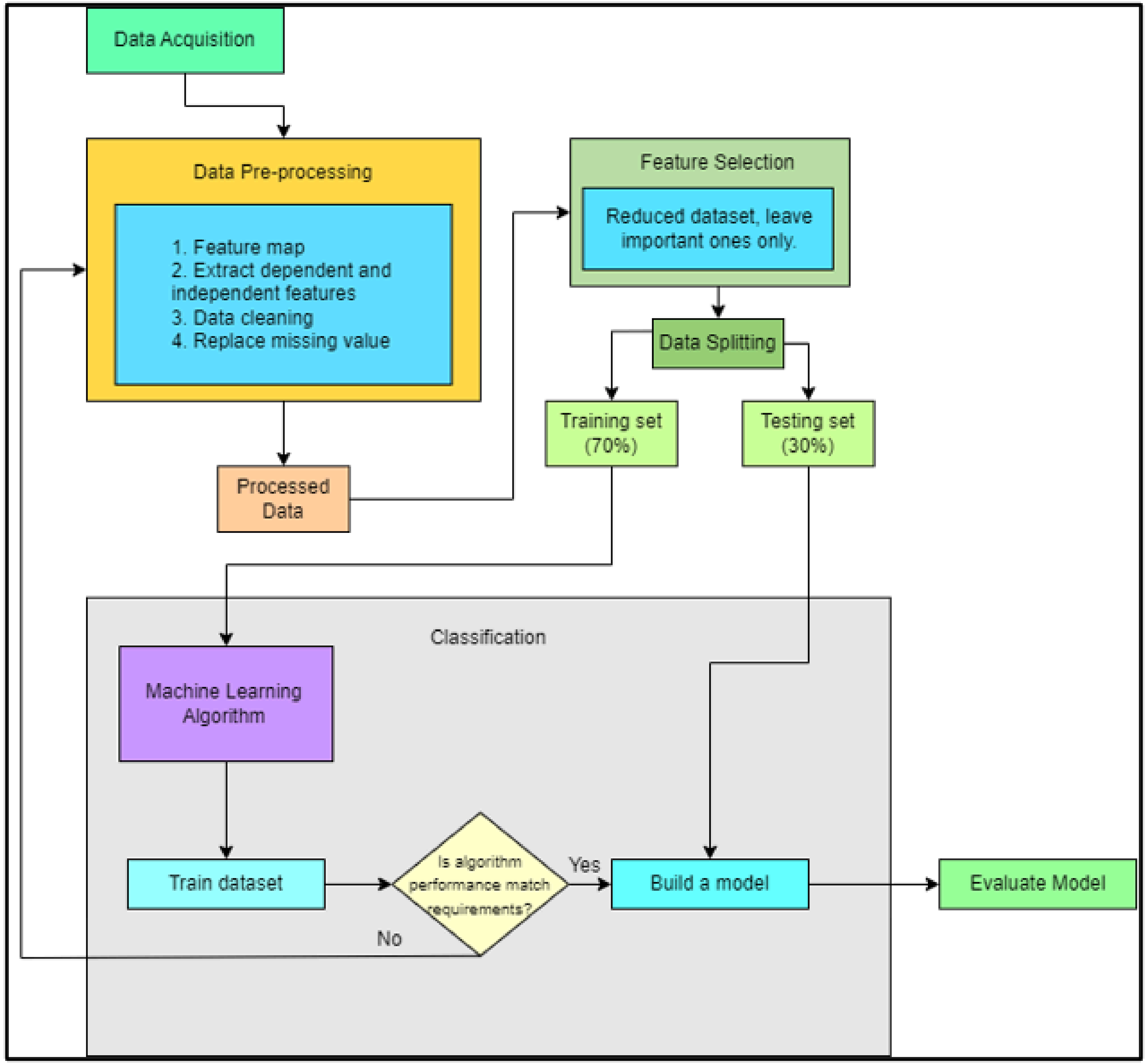


Figure 4. 5: Proposed machine learning flowchart

Design for Proof-of-Concept

- **Machine Learning Design**
- Anomaly Detection Design
- Data Visualization Design
- Dataset Design

01 Anomaly Detection Design

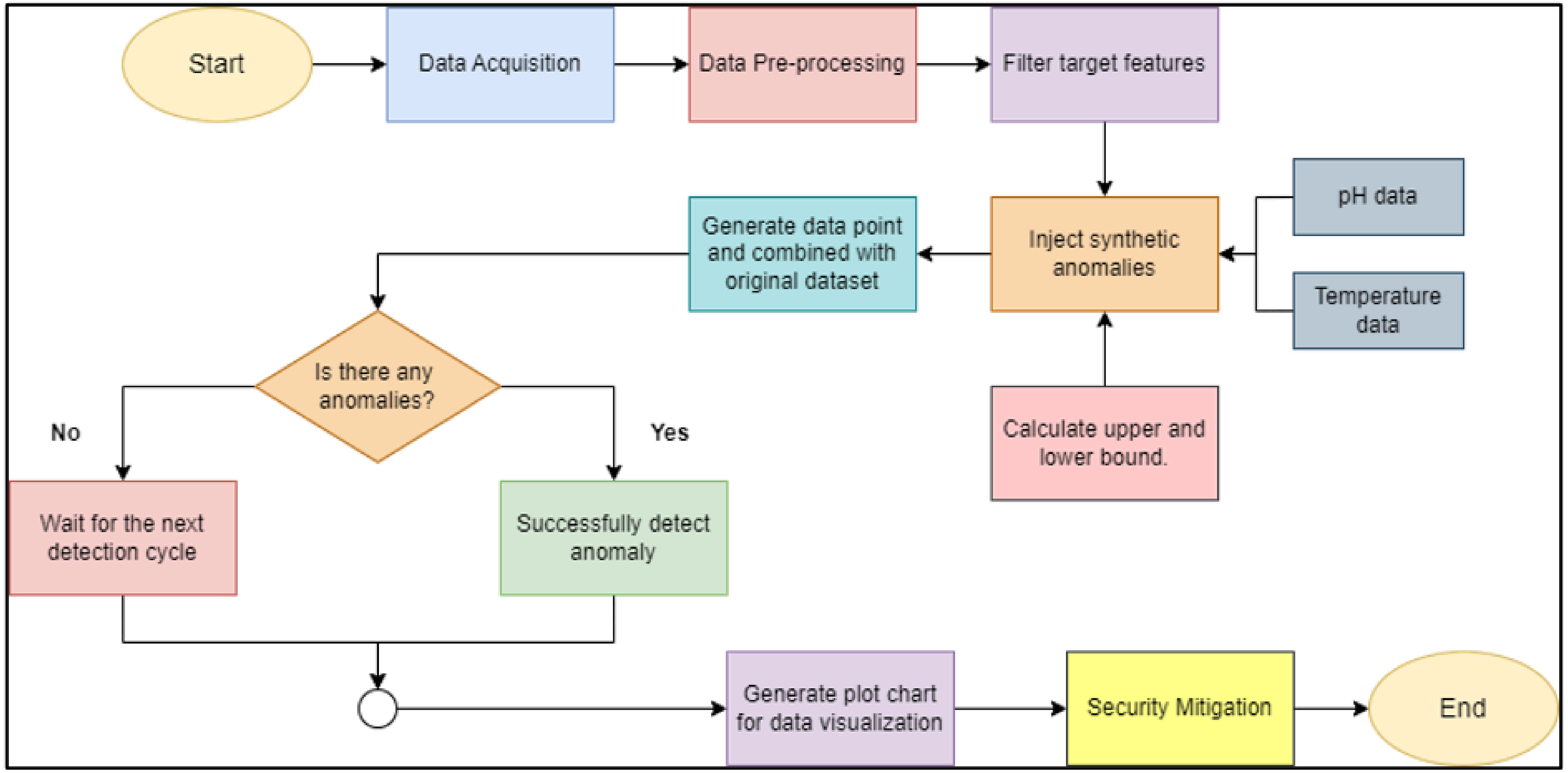


Figure 4. 6: Proposed anomaly detection flowchart

Design for Proof-of-Concept

- Machine Learning Design
- **Anomaly Detection Design**
- Data Visualization Design
- Dataset Design

01 Data Visualization Design

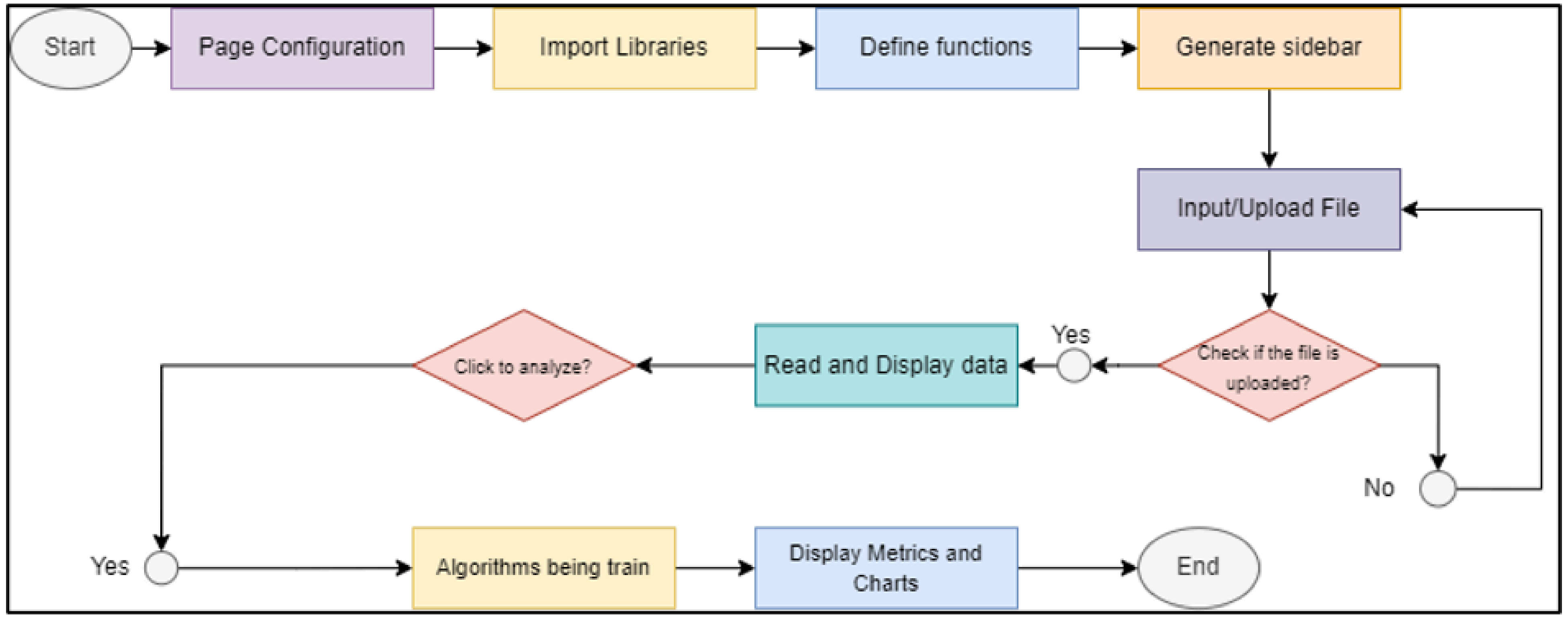


Figure 4. 9: Diagram of Data Visualization Process flowchart

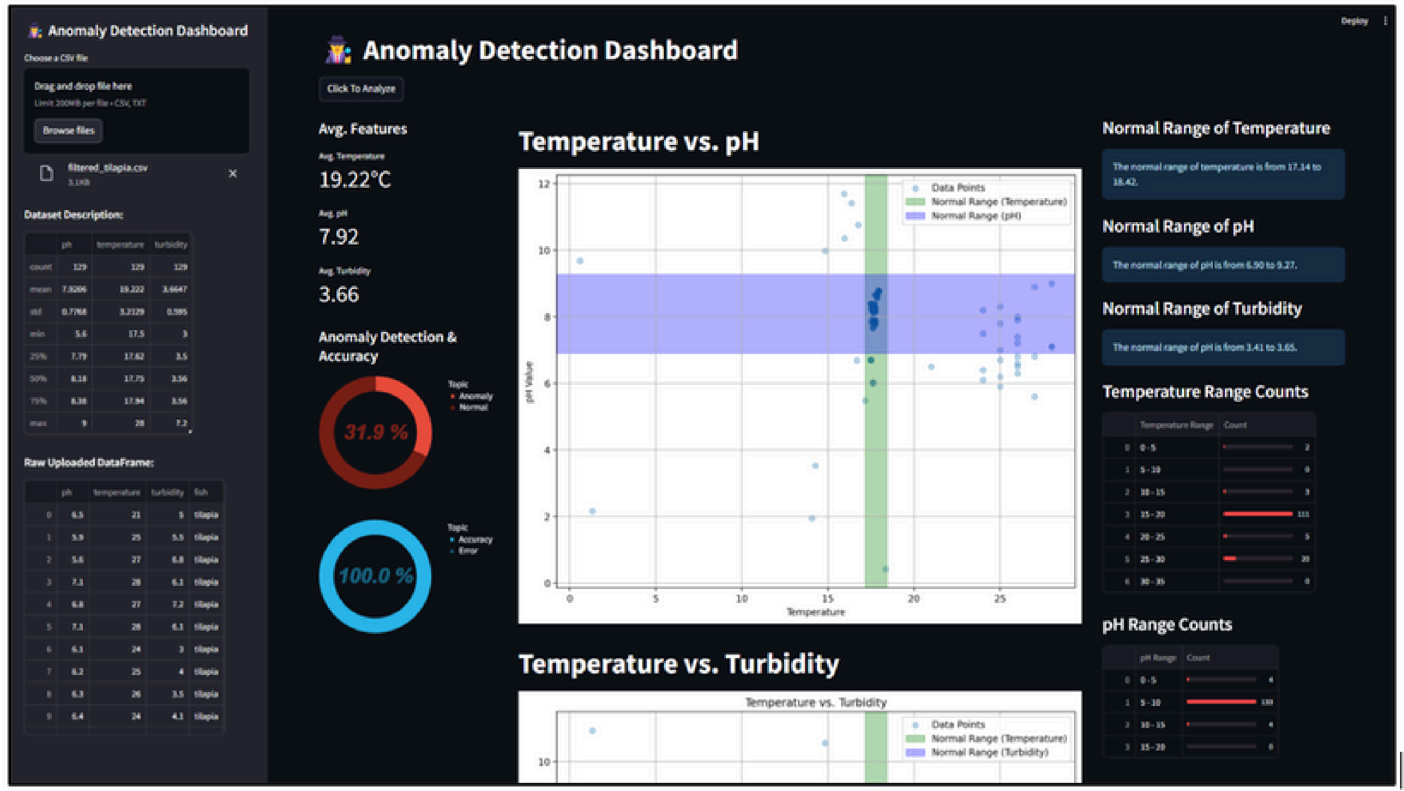


Figure 4. 10: Anomaly Detection Dashboard

Design for Proof-of-Concept

- Machine Learning Design
- Anomaly Detection Design
- **Data Visualization Design**
- Dataset Design

01 Dataset Design

Table 4. 5: realfishdataset.csv (Real-Time Pond Water Dataset for Fish Farming, 2021)

01

pH	Temperature	Turbidity	Fish
6	27	4	katla
7.6	28	5.9	sing
7.8	27	5.5	sing
6.5	31	5.5	katla
8.2	27	8.5	prawn
…	…	…	…

TABLE 4. 6: Result after filter and injecting synthetic anomalies for Tilapia species

02

pH	temperature	turbidity	status
7.4	26	3	anomaly
8.9	27	3	anomaly
9	28	3.3	anomaly
7.67	17.62	3.56	normal
…	…	…	…

Table 4. 7: Final design of dataset

03

ph	temperature	turbidity	status
7.4	26	3	0
8.9	27	3	0
9	28	3.3	0
7.67	17.62	3.56	1
…	…	…	…

Design for Proof-of-Concept

- Machine Learning Design
- Anomaly Detection Design
- Data Visualization Design
- Dataset Design

Analysis

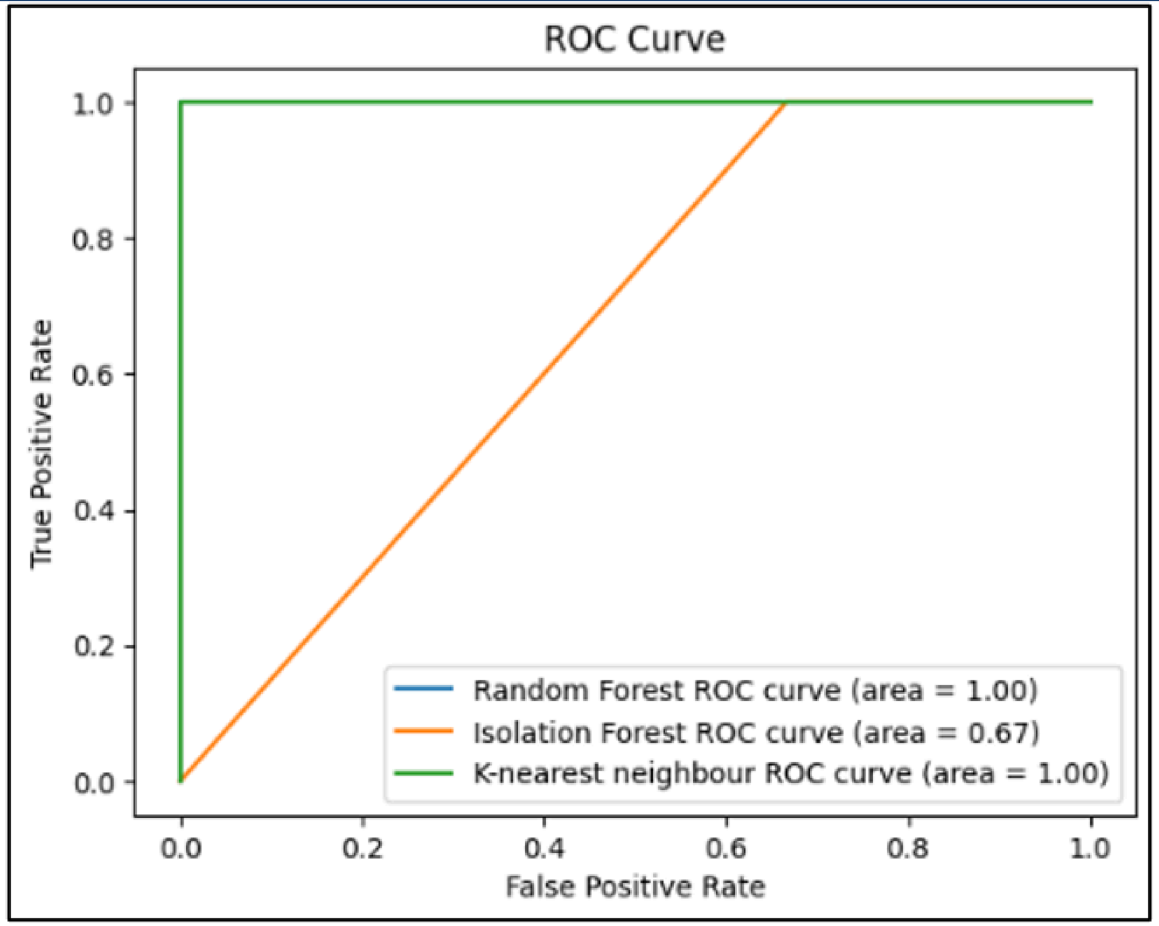


Figure 5. 2: ROC Curve of tested algorithms

01 ROC Curve visualization

02 Classification Report

- RF and KNN give the same result, which indicates the algorithms classify the anomaly perfectly.
- KNN got more disadvantages than RF, therefore RF will be selected.

Table 5. 3: Evaluation of tested algorithms.

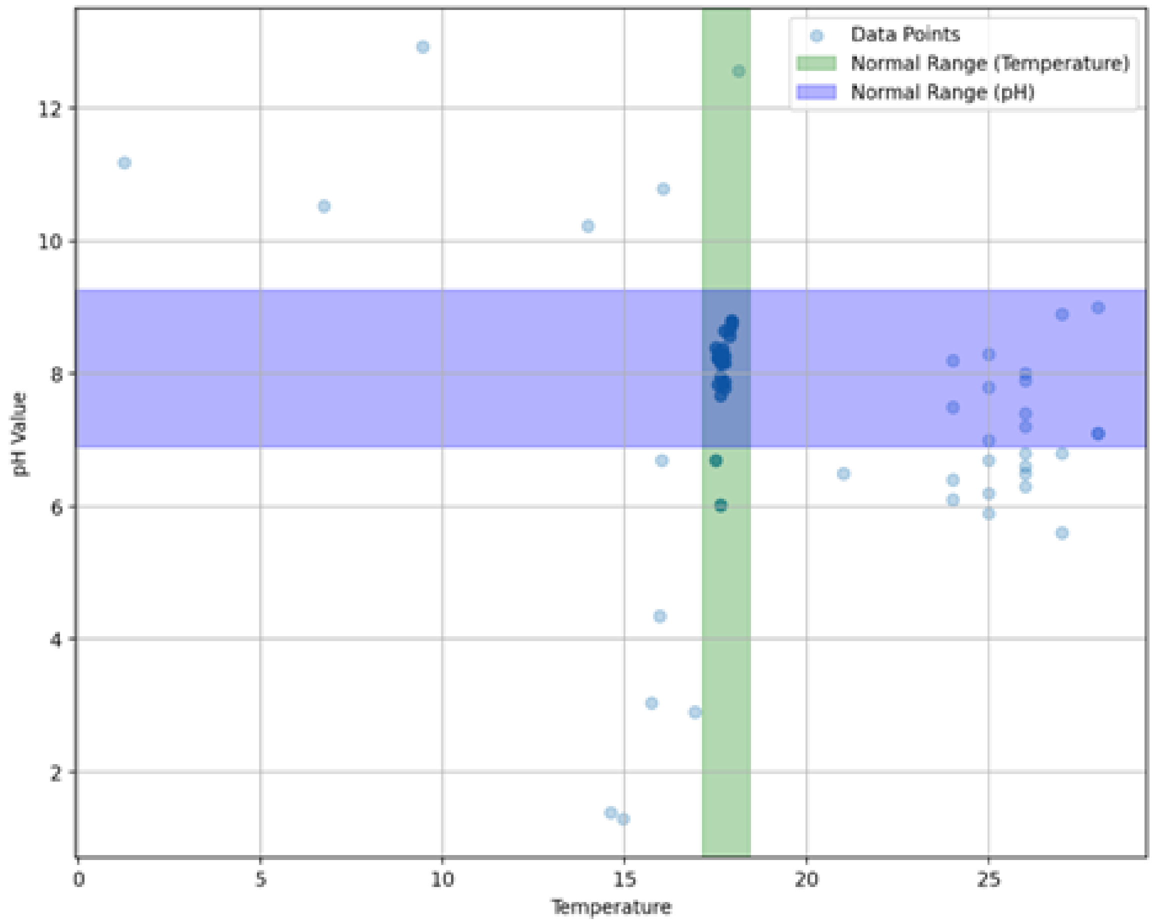
	Accuracy	Precision		Recall		F1-Score		ROC-AUC
		0	1	0	1	0	1	
RF	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
IF	0.77	1.00	0.74	0.33	1.00	0.50	0.85	0.67
KNN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Anomaly Detection Dashboard Demo

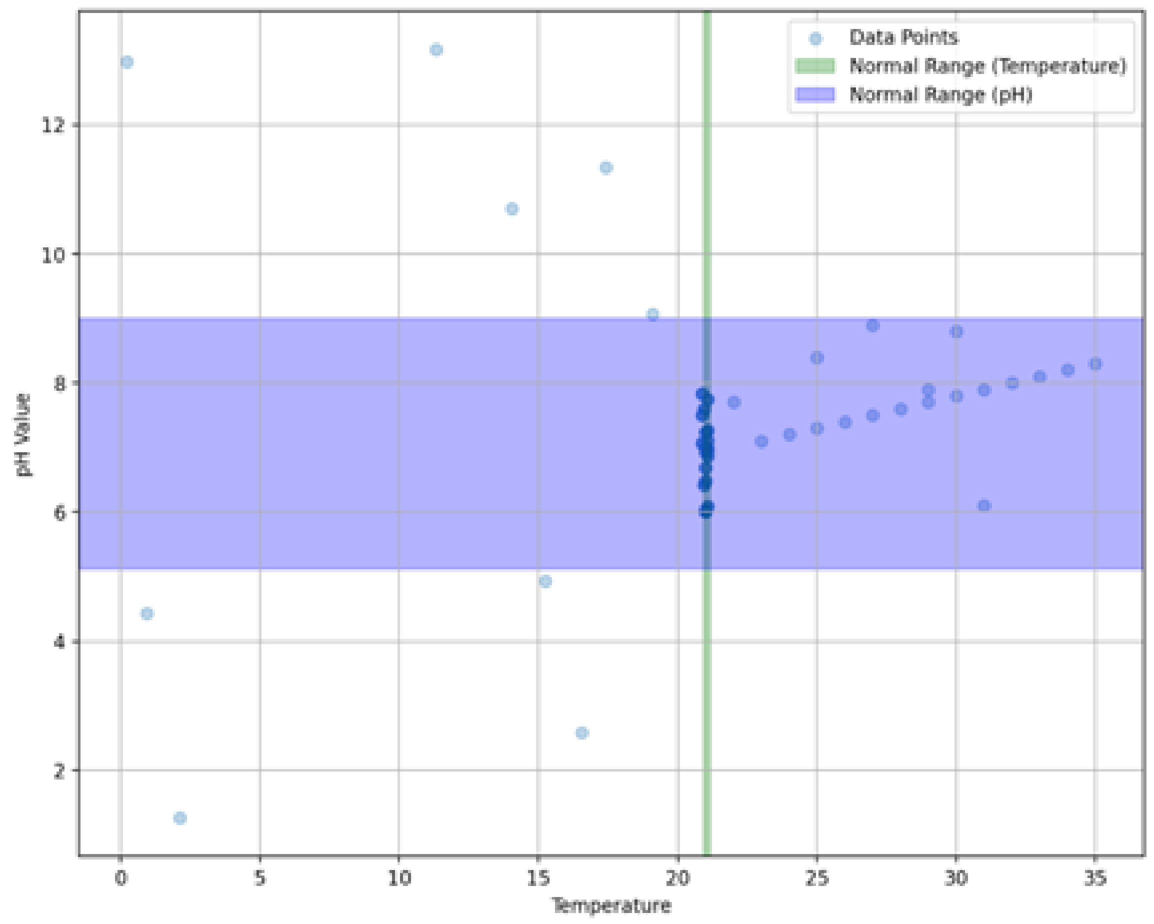
Data **analysis** and visualization

For analysis, 3 fish species dataset has been extracted from the collected dataset and injected with synthetic anomaly. The species is Tilapia, Rui, and Pangas. Feature to be observe is pH, Temperature and Turbidity

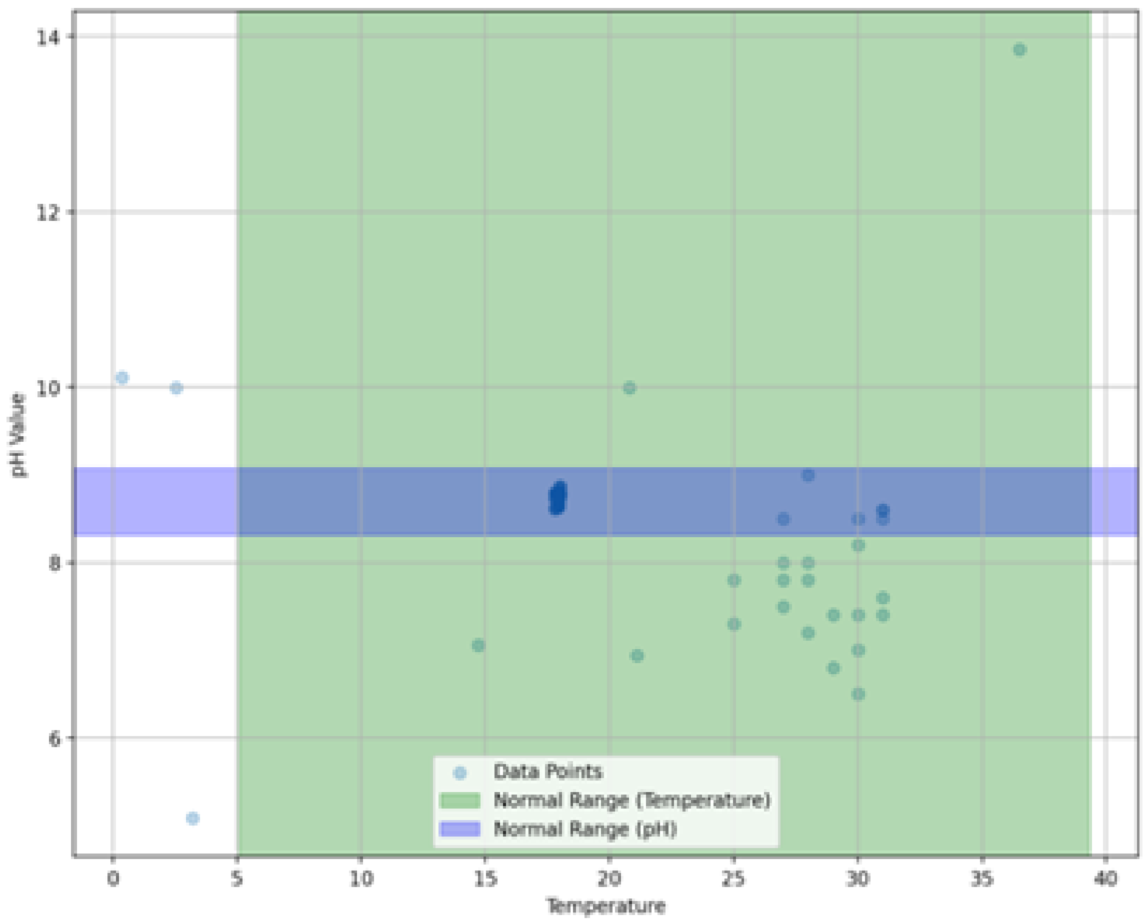
Temperature vs pH Values



Tilapia



Rui

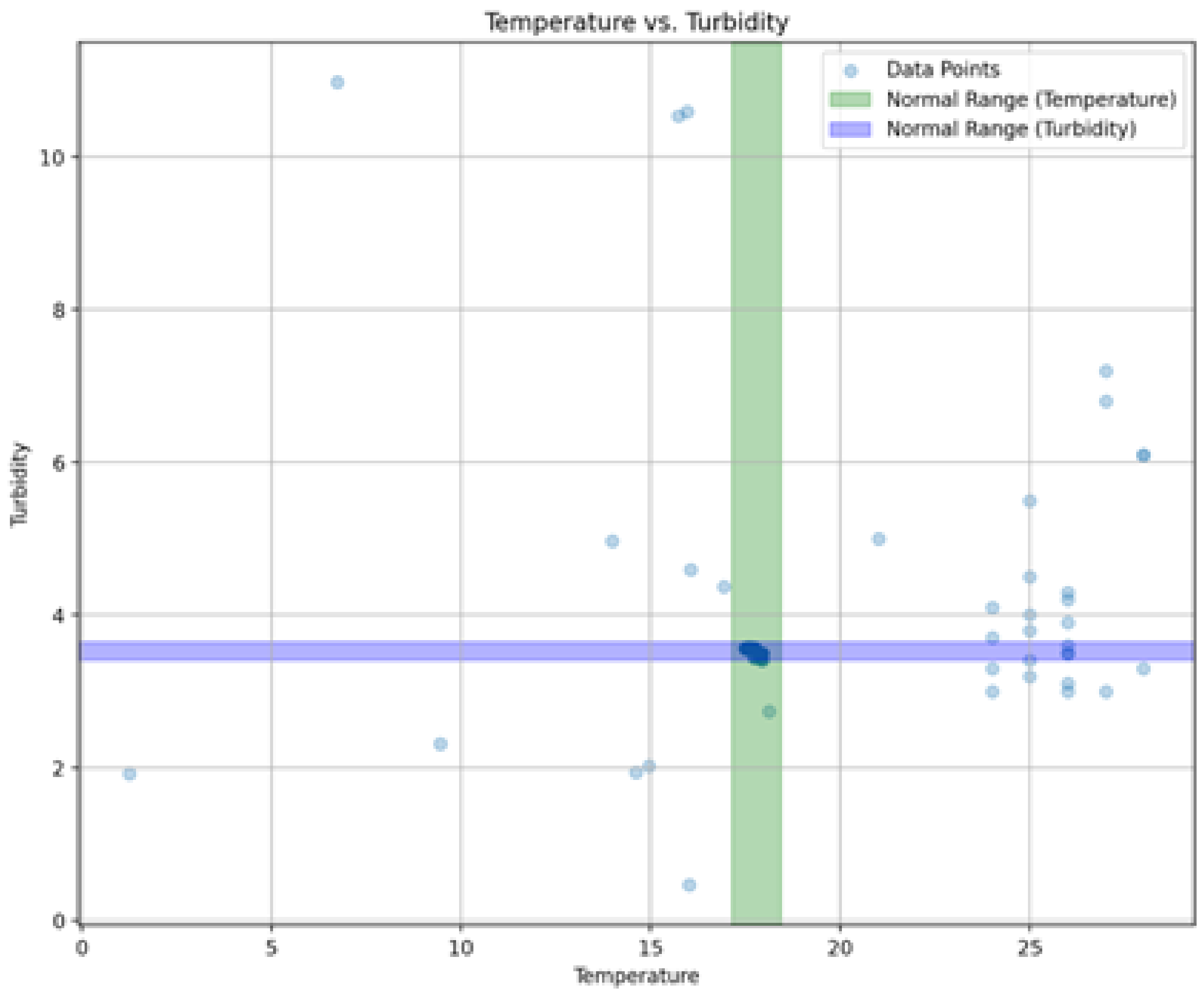


Pangas

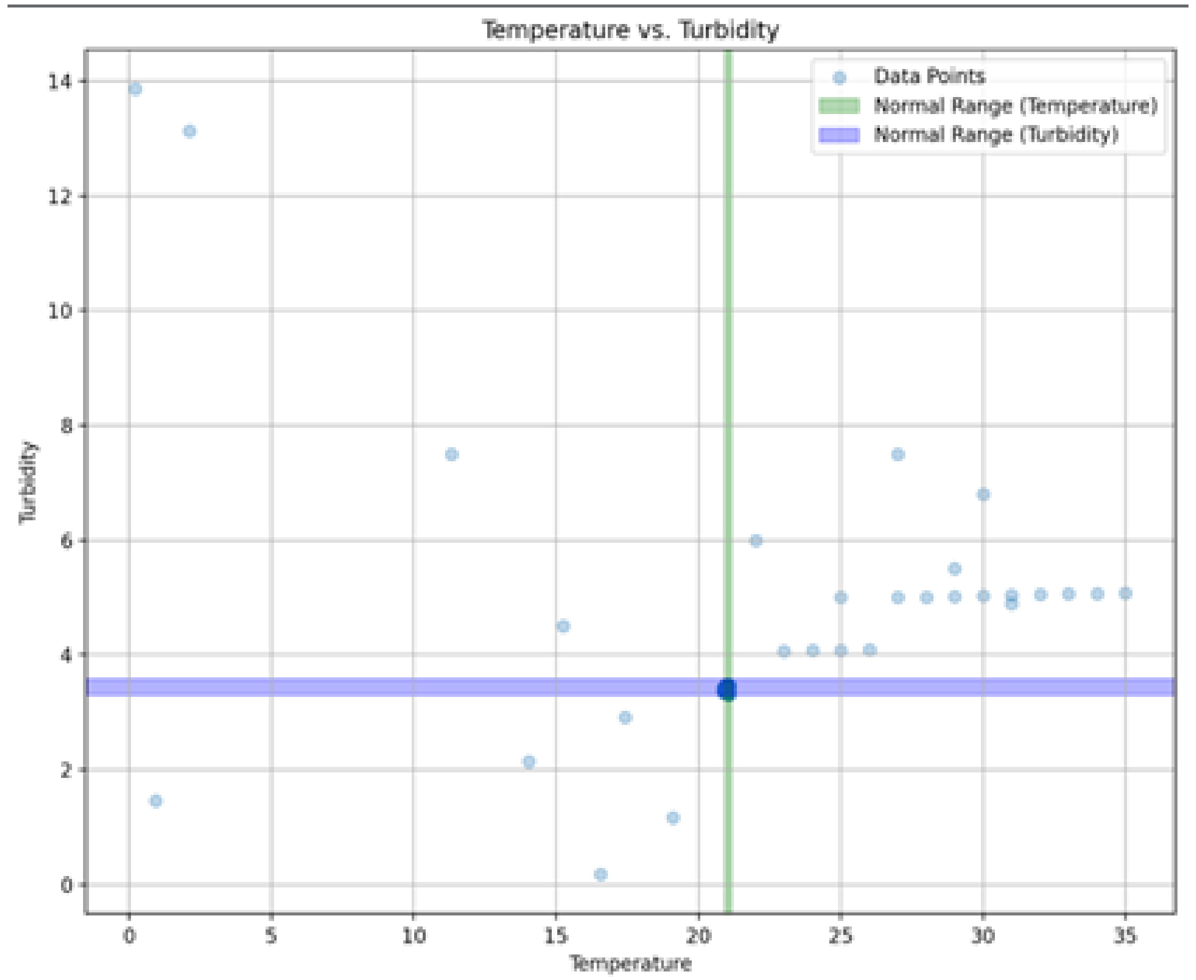
Data **analysis** and visualization

For analysis, 3 fish species dataset has been extracted from the collected dataset and injected with synthetic anomaly. The species is Tilapia, Rui, and Pangas. Feature to be observe is pH, Temperature and Turbidity

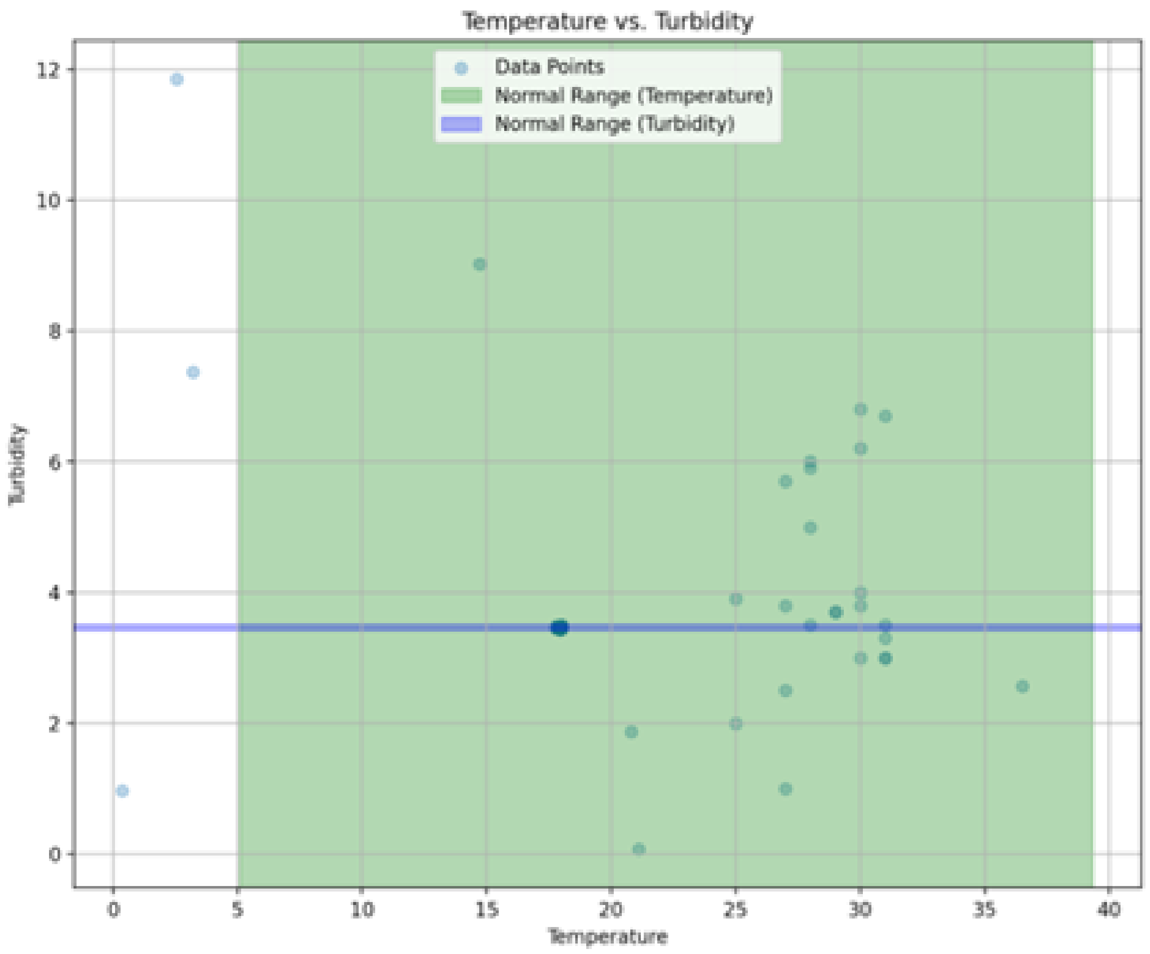
Temperature vs Turbidity



Tilapia



Rui

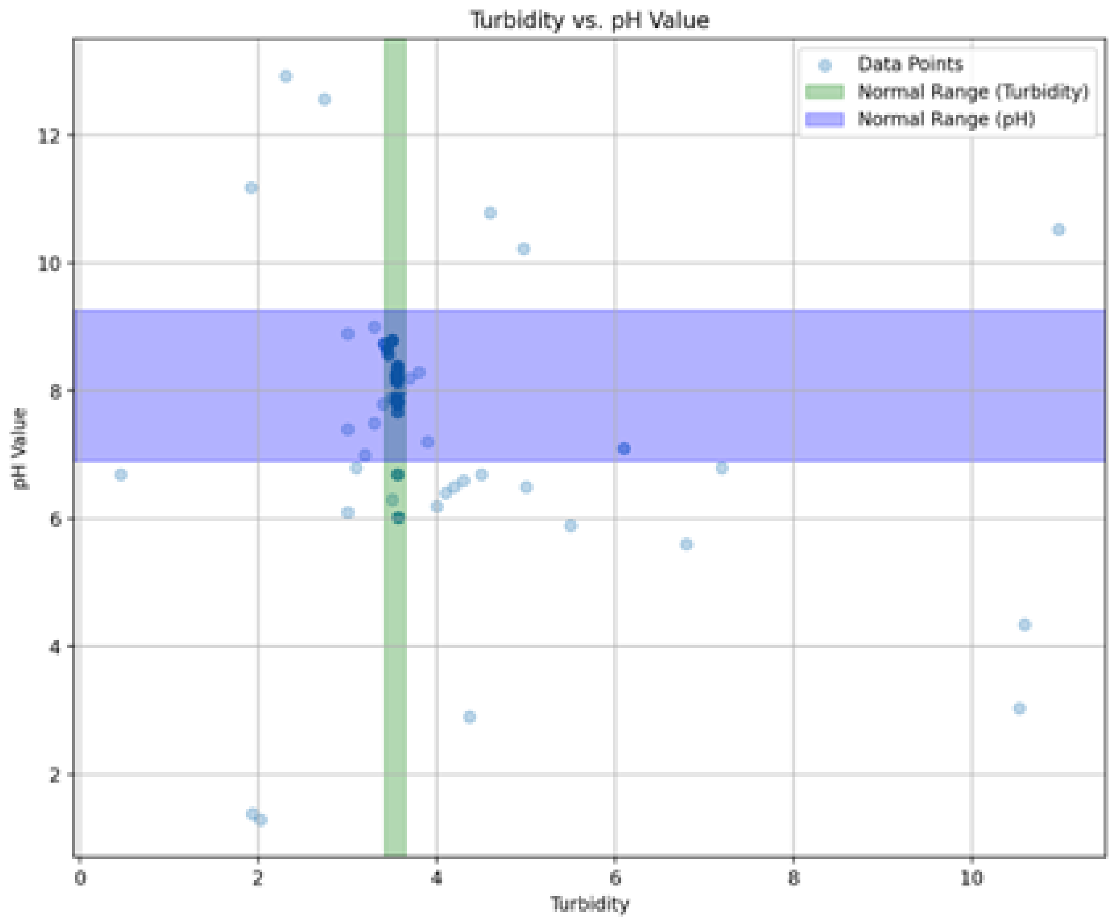


Pangas

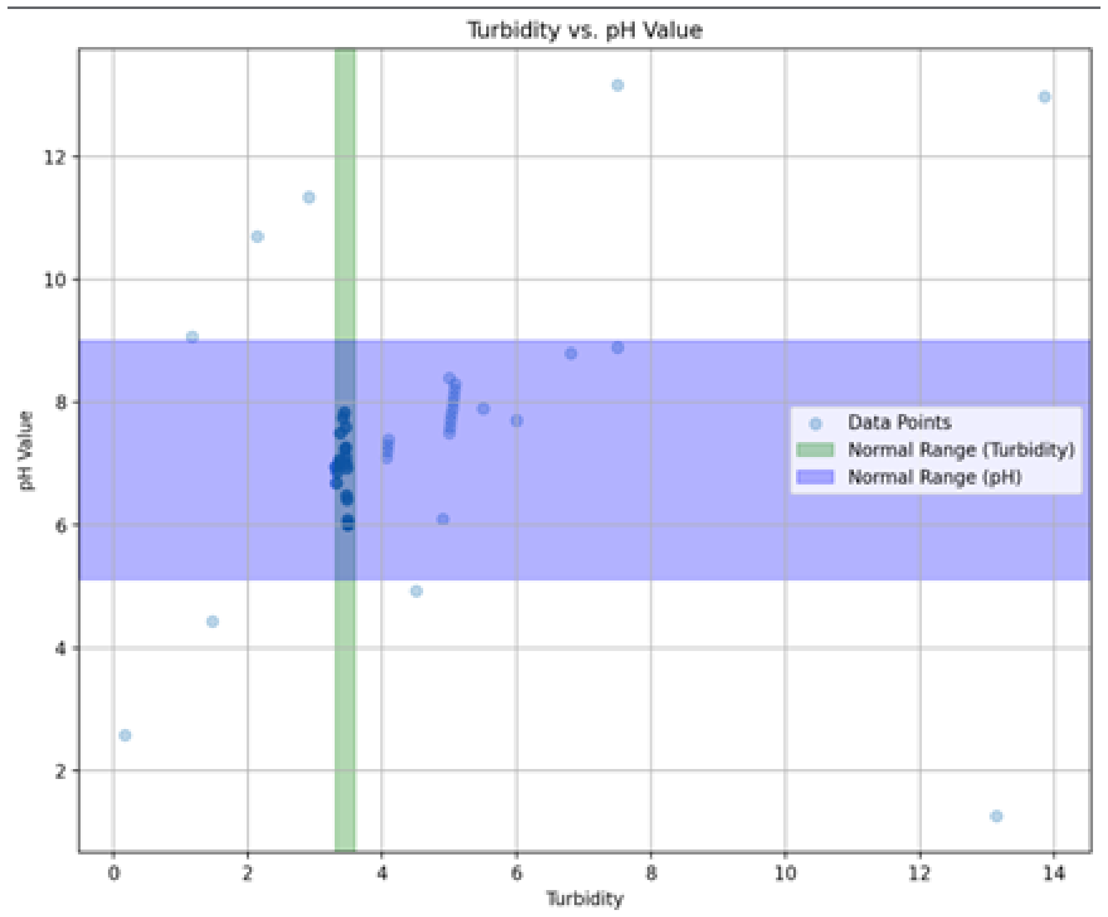
Data analysis and visualization

For analysis, 3 fish species dataset has been extracted from the collected dataset and injected with synthetic anomaly. The species is Tilapia, Rui, and Pangas. Feature to be observe is pH, Temperature and Turbidity

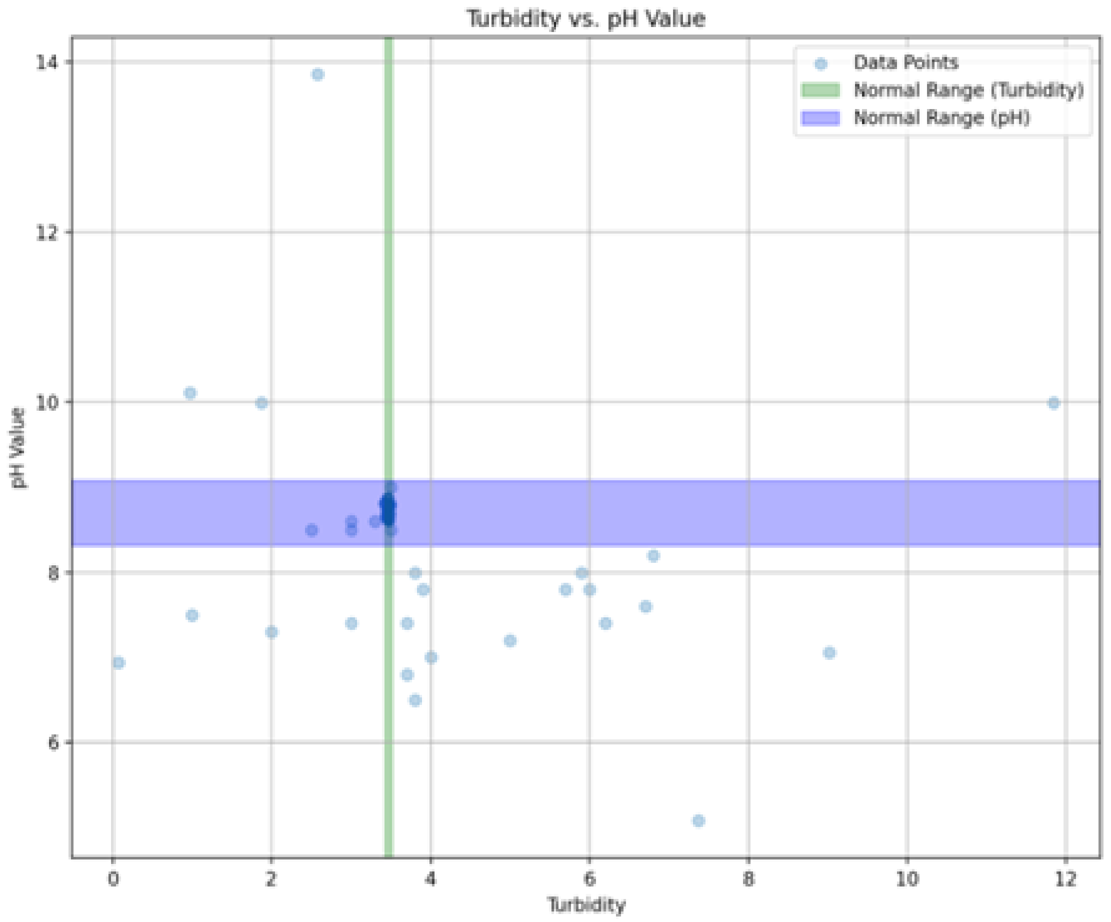
Temperature vs pH Values



Tilapia



Rui



Pangas

Risk Mitigation via Anomaly Detection in Smart Farming environment

A next plan after anomalies has been detection is by conducting a risk mitigation plan. There a several risk mitigation plan can be performed.

Incident Response and Recovery

Endpoint Security

Access Control and Authentication

CONCLUSION

1 Suitable machine learning algorithm for anomaly detection

- Random Forest give the best result compared to other algorithms.
- Achieve high accuracy, precision, recall, f1-score, and ROC-AUC for smart aquaculture dataset.

2 Enhancement for IoT security risk mitigation

- Goals to make smart aquaculture become more secure from possible threats

CHALLENGE & FUTURE WORK

01 Understanding the fundamental of machine learning

02 Learn new framework such as Streamlit, but still can successfully develop a complete program

---

01 Real-time Anomaly Detection
- Using real sensors and monitor the data
- Set up a small environment, and perform real mitigation plan

02 Enhanced data collection and integration
- Additional environmental factors such as light intensity, CO_2 levels
- Integrate with weather forecast to improve anomaly detection

CONCLUSION

Objective	Description	Slides
Objective 1	Review suitable machine learning algorithms for anomaly detection in smart farming	Slide 10: Review Machine Learning Algorithm Application Slide 14: Overview of studies on machine learning algorithm in smart farming
Objective 2	Conduct comparative analysis of machine learning algorithms	Slide 17: Comparison of Machine Learning Slide 15, 16, 18: Analysis of studies and characteristic comparison
Objective 3	Analyze result from the selected machine learning model	Slide 25, 26, 27: Data analysis and visualization

ANOMALY DETECTION FOR SMART FARMING VIA MACHINE LEARNING

Q&A SESSION

FINAL YEAR PROJECT 23/24

ANOMALY DETECTION FOR SMART FARMING VIA MACHINE LEARNING

THANK YOU!

FINAL YEAR PROJECT 23/24