Probabilistic Modeling of Fish Growth in Smart Aquaculture Systems

Jongwon Kim[1], Eunbi Park[1], Sungyoon Cho[2], Kiwon Kwon[2], and Young Myoung Ko[1,3*]

[1] Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang, 37673, Gyeongbuk, Republic of Korea
[2] Korea Electronics Technology Institute, Seoul, South Korea
[3] Open innovation Big Data Center, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang, 37673, Gyeongbuk, Republic of Korea
[e-mail: pioneer0517@postech.ac.kr, ebpark@postech.ac.kr, sycho@keti.re.kr, kwonkw@keti.re.kr, youngko@postech.ac.kr]
*Corresponding author: Young Myoung Ko

Received February 27, 2023; revised July 5, 2023; accepted July 27, 2023; published August 31, 2023

Abstract

We propose a probabilistic fish growth model for smart aquaculture systems equipped with IoT sensors that monitor the ecological environment. As IoT sensors permeate into smart aquaculture systems, environmental data such as oxygen level and temperature are collected frequently and automatically. However, there still exists data on fish weight, tank allocation, and other factors that are collected less frequently and manually by human workers due to technological limitations. Unlike sensor data, human-collected data are hard to obtain and are prone to poor quality due to missing data and reading errors. In a situation where different types of data are mixed, it becomes challenging to develop an effective fish growth model. This study explores the unique characteristics of such a combined environmental and weight dataset. To address these characteristics, we develop a preprocessing method and a probabilistic fish growth model using mixed data sampling (MIDAS) and overlapping mixtures of Gaussian processes (OMGP). We modify the OMGP to be applicable to prediction by setting a proper prior distribution that utilizes the characteristic that the ratio of fish groups does not significantly change as they grow. We conduct a numerical study using the eel dataset collected from a real smart aquaculture system, which reveals the promising performance of our model.

Keywords: Smart aquaculture, fish weight estimation, probabilistic modeling, multi-modal distribution, wireless sensor network

A preliminary version of this paper was presented at ICONI 2022, and was selected as an outstanding paper.
This research was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00225, Development of digital aqua twin core platform for optimal aquafarm design and operation) , in part by Korea Institute for Advancement of Technology (KIAT) grant by the Korea Government (MOTIE) (P0008691), and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C1094699 and NRF-2021R1A4A1031019)

http://doi.org/10.3837/tiis.2023.08.017 **ISSN : 1976-7277**

1. Introduction

As reported by the 2022 edition of the State of World Fisheries and Aquaculture (SOFIA) [1], due to the growth in aquaculture, particularly in Asia, total fishery and aquaculture production reached an all-time high of 214 million tonnes in 2021. Moreover, aquaculture has experienced more remarkable growth over the past two years compared to capture fisheries, and this gap is expected to widen over the next decade. With this growth of the aquaculture industry, smart aquaculture also has undergone significant innovations in recent years with the upcoming IoT services [2, 3, 4, 5]. Smart aquaculture is essential for increasing fish farming efficiency because it maximizes fish growth within limited resources and reduces wasted resources unnecessarily by managing the fishery environment with automatically collected data. In smart aquaculture, wireless sensors and cameras collect data. Then the feeder makes decisions based on that data and operates remote machines for ecological environment management and feeding schedule planning. Progress in remote devices has replaced laborious tasks such as feeding [6], water pumps [7], etc. Advances in sensor technology make gathering data faster and creating an explosion of data [8]. This vast dataset enables the smart aquaculture system to make autonomous decisions. These days, many researchers try to implement machine learning methods for the automatic decision-making process and develop end-to-end automation models that eliminate the need for manual human interaction with the aquaculture system [9, 3].

Smart aquaculture data is categorized into two types: images and tabular datasets. Image datasets help monitor and automatically quantify the fish's visual status (length, behavior, etc.). Image-based models analyze feeding status [10], classify fish behavior [11], detect abnormal behavior [12, 13], or track fish's trajectory [14]. Image datasets are also used to estimate the weight of the fish [15, 16, 17]. However, a model based on an image can only return the weight of the fish in the picture, not the weight of all fish in a tank. Furthermore, the model has limitations in that they are designed for analysis, not for prediction. Tabular datasets mainly consist of ecological environment and feeding data, which can be controlled. To develop an effective fish growth model, it's crucial to integrate tabular data with fish weight data. However, the process of collecting weight data is conducted manually by human workers, making it not only costly but also prone to errors. Consequently, it remains challenging to collect weight data, whereas environmental data from IoT sensors are readily available. As a result, prior studies on smart aquaculture with tabular datasets have primarily focused on tasks like predicting specific environmental conditions, such as water quality [18] and dissolved oxygen levels [19], using only environmental data. For this study, we meticulously collect a weight dataset alongside an environmental dataset from a collaborating smart aquaculture system, which allows us to design a growth model based on this combined dataset.

A growth model estimates the length or weight of fish over time. (These two variables, length and weight, can easily be converted into each other using a simple formula [20].) The majority of existing growth models are built upon the von Bertalanffy model [21, 22], which utilizes the age of the fish to predict its growth. However, being a non-probabilistic model, it can't be applied to a population of fish without assuming uniformity in growth parameters among all fish. In an attempt to yield probabilistic results, there have been studies estimating the distribution of this model's parameters [23, 24, 25, 26, 27] in a Bayesian sense, or those that used a weighted sum of certain statistical models [28, 29] with Bayesian weight parameters [30]. With the recent development of smart aquaculture systems, there has been a study on growth models that incorporate environmental data as input [31]. However, they relied on data from a test system and failed to capture the realistic attributes found in real-

world smart aquaculture systems, such as sorting and grading, and the periodic absence of data. Additionally, they did not consider fish growth in a probabilistic way. **Table 1** provides a summary of the existing fish growth models.

Table 1. Summary of Existing Fish Growth Models

Name	Dataset	Probabilistic	References
von Bertalanffy based model	length-at-age dataset	X	[21, 22]
Other statistical models	length-at-age dataset	X	[28, 29]
Bayesian von Bertalanffy model	length-at-age dataset	O	[23, 24, 25, 26, 27, 30]
ANN-based growth model	environment features in test smart aquaculture system	X	[31]
Ours	environment features in actual smart aquaculture system	O	

This paper focuses mainly on exploring the unique characteristics of tabular datasets in smart aquaculture and developing a preprocessing step and fish growth model while considering the following data characteristics. 1) Mixed frequency: we predict fish weight with a temporal tabular dataset, where the fish weight is manually collected less frequently than automatically collected sensor data. 2) Multi-modality: fish grow at different rates even under the same environment [32]. So, a probabilistic approach is more appropriate than a point prediction method. Furthermore, we observe that our empirical distribution of the weight data exhibits multi-modality. 3) Fragmented timeseries: In smart aquaculture, sorting and grading refer to the process of separating fish by weight and redistributing them into their corresponding tank. This process is typically done to create more uniform groups of fish, which can help with feeding, disease management, and other aspects of fish farming. However, this process creates fragmented time series that are difficult to manage and analyze. 4) Missing data: because of sensor replacements and sorting operations, long periodic missing data frequently arises. Additionally, the data is fragmented along the time axis, and the number of data is insufficient. We mainly discuss the above characteristics in detail in Section 2.3 and demonstrate the importance of considering the characteristics from numerical analysis.

The rest of this paper is organized as follows. Section 2 explains the data characteristics in smart aquaculture. Section 3 describes the model we propose. Section 4 conducts numerical experiments to show the performance of the proposed method. Section 5 makes a concluding remark with a discussion for future work.

2. Dataset collected from smart aquaculture

2.1 Notation

Throughout this paper, we will use the following notation: boldfaced lowercase letters (**x, y, z**) denote vectors, boldfaced capital letters (**X, Y, Z**) denote matrices, and ordinary letters (*b*, *n*, *L*) denote scalars. Vectors are column vectors unless we mention explicitly that they are row vectors. The subscript below $\mathbf{x}_i$ means the i-th element of **x**.

2.2 Data description

We collected eel data from a smart aquaculture system during the period of 2021/07 to 2021/12. The data comprises sensor data, growth data, and feeding data observed from five tanks. Each tank contains tens of thousands of fish, which are not mixed before the grading and sorting period, and it is impossible to track which fish are moved from which tank during the grading and sorting process. Throughout the period, there are two sorting times, 08/08, and 10/31, in which all fish are taken out of the tank and redistributed according to the similar weight. The sensor data of each tank was collected in real time, and they include dissolved oxygen (DO), water temperature, pH, oxidation reduction potential (ORP), CO_2, oxygen, and light. The feeding data includes variables such as the amount of food and water in the morning and afternoon, recorded daily. The growth data is composed of the weight of 20 randomly selected fish per tank observed every two weeks, the number of fish, and average fish weight per tank measured at the beginning of each sorting operation. This period takes two months on average, so we refer to this data collection cycle as two months. The detailed explanation of each attribute (variable) can be found in **Table 2**.

Table 2. Description for smart aquaculture data.

Data	Attribute (unit)	Interval
sensor data	dissolved oxygen, DO (mg/L)	one-minute
	water temperature (˚C)	one-minute
	pH	one-minute
	oxidation reduction potential, ORP (mV)	one-minute
	CO_2 (mg/l)	one-minute
	oxygen (L/m)	one-minute
	light (mA)	one-minute
feeding data	feed amount for a.m.	one-day
	feed amount for p.m.	one-day
	water amount for a.m.	one-day
	water amount for p.m.	one-day
growth data	fish weight (g)	two-weeks
	average fish weight per tank (g)	two-months
	number of fish per tank	two-months

2.3 Characteristics of smart aquaculture and the datasets

This section examines the characteristics of smart aquaculture dataset and provides brief description of how to deal with each characteristic. The first characteristic is that data is a mixture of data from IoT sensors and manual measurements. In traditional fish sampling methods, researchers manually collect fish samples from the water, which is labor-intensive and time-consuming and affects the fish population. In contrast, in smart aquaculture, IoT sensors are used to observe various ecological environmental factors at a high frequency, making data collection more cost-effective. However, the growth data such as the weight of the fish is still measured manually, resulting in mixed frequencies in the data set and a small number of weight data observations. Moreover, the difference in data collection methods can lead to overfitting in weight prediction models. While the amount of data collected by IoT sensors is large, the number of fish weight datasets is small, so the process of matching the frequency of the data results in a small number of data points. In addition, for the weight prediction problem, the sensor data and feeding data are set as input data and the weight data is set as output data. If the output data uses all the sensor data and feeding data of the observed

interval as input data, it becomes enormously high-dimensional data. The small number of data points with high-dimensionality can cause overfitting. Coincidentally, underwater sensors are susceptible to malfunctioning and require periodic replacements, leading to potential missing data. We should preprocess these periodic missing data.

The second characteristic is related to the growth process of the fish. Despite having the same ecological environment and feeding process, the growth of fish can vary greatly. This variability in growth can cause significant difficulties when trying to predict the growth of fish using point prediction methods such as deep neural networks or tree boosting methods. To solve this problem, researchers can create a model that returns a probability distribution. **Fig. 1** depicts a histogram and exponential kernel density estimation plot of fish weight for each date in tank 1. As illustrated in **Fig. 1**, the distribution of fish weight changes over time from a unimodal to a multimodal distribution. To consider this phenomenon. we assume that the weight distributions of fish over time are a mixture of several stochastic processes. The validity of this assumption will be tested in the experimental section. The histograms for other tanks are included in the appendix.

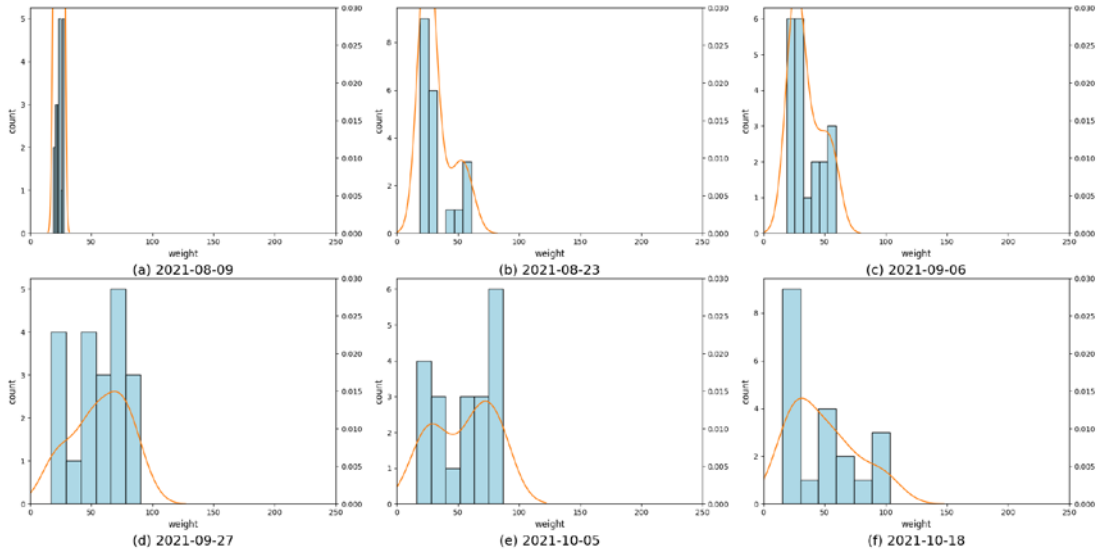


Fig. 1. Histogram and kernel density plot of fish weight of tank 1 for 2021/08 – 2021/10.

The last characteristic is the periodic grading and sorting of fish. As mentioned earlier, fish grow differently over time, so they need to be graded and sorted regularly [32]. **Fig. 2** illustrates the grading and sorting step of smart aquaculture when the number of tanks is 2. Periodically, the fish farmer grades and sorts the fish that have grown in each tank and places them back into new tanks. Grading and sorting have the advantage of preventing feed monopoly of mature fish and providing efficient feeding according to the size of the fish. However, this characteristic can lead to fragmented data along the time axis and missing data during sorting time. In addition, periodic replacement of underwater sensors can exacerbate the problem of missing data, making it difficult to accurately predict fish growth and behavior.

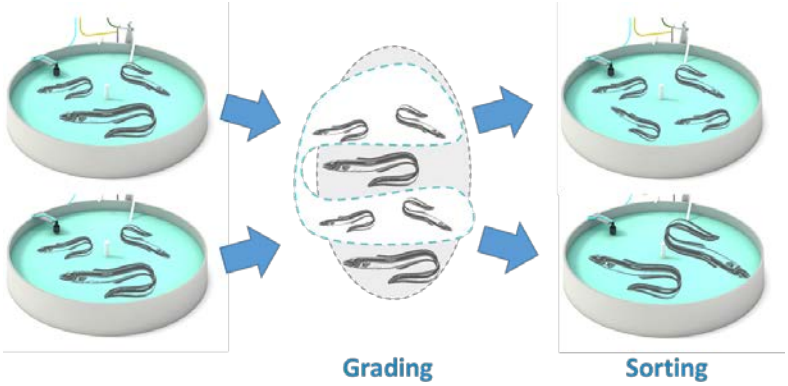


Fig. 2. Grading and sorting process in aquaculture

In summary, mixed sampling method (IoT sensor data and manual measurements) results in mixed frequency, lack of data, high-dimensionality and data missing. The growth process of fish demonstrates the need for a probabilistic model that returns a multimodal distribution, and grading and sorting produce a fragmented time series and missing data. The methods for addressing those difficulties through preprocessing and modeling are discussed in more detail in the following section.

3. Probabilistic modeling for fish growth

3.1 Preprocessing

This section describes the preprocessing step that converts daily feeding data, $\mathbf{X}_d$, and sensor data, $\mathbf{X}_s$, collected in minutes into an input dataset, $\mathbf{X}$. **Fig. 3** illustrates our algorithm.

Firstly, we remove missing data or outlier in sensor data. Missing data and outlier came from sensor replacement, sensor failure, grading, and sorting. Sensor failure returns a nan value which indicates that the data is missing. We look for instances where a sensor was outputting an outlier due to failure or replacement and remove data corresponding to the date of output. Removing hourly data is based on the expert opinion; since all the causative processes take a few hours, we detected an outlier in minutes and then removed the sensor data from hours that included the outlier or missing data. We follow the Tukey's fences rule, which determines that the data is an outlier if it is greater than Q3 + 2(Q3-Q1) or less than Q1 – 2(Q3-Q1) where Q1 and Q3 are the lower and upper quartiles, respectively. **Fig. 4** depicts the pH sensor data collected over time. The detected outlier is highlighted in red in **Fig. 4** (a), while the data observed on the same hour as the detected outlier is colored in red in **Fig. 4** (b). Due to the presence of long-term missing data in addition to missing data caused by outliers, removing the missing data would be more appropriate for ensuring the accuracy of the model, rather than attempting to impute the missing data

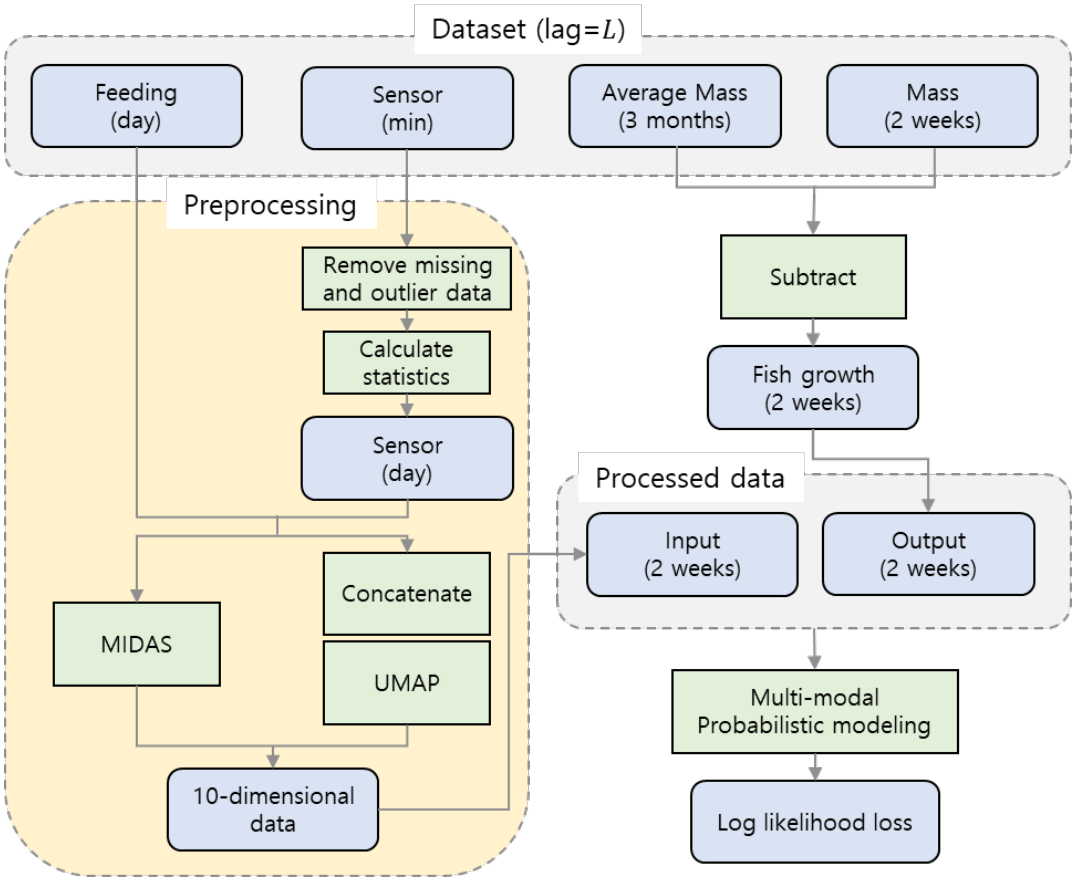
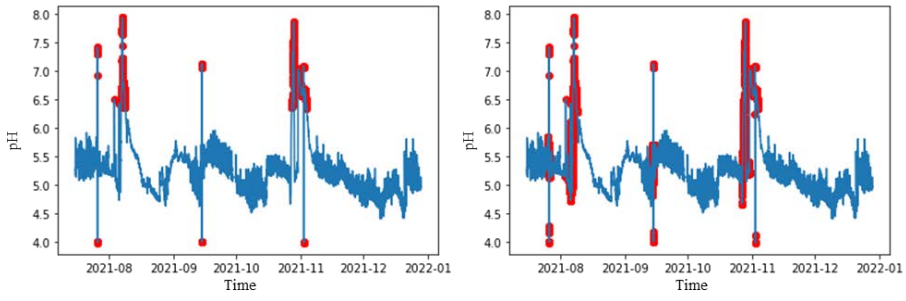


Fig. 3. Algorithm diagram of our model.

Subsequently, the sensor dataset per minute is transformed into daily sensor data collected from $t-(L-1)$ to t day, and then we get the L daily sensor and feeding data. As mentioned above, sensor data contains hours of missing data which can result in an erroneous model due to the low accuracy of imputation. To address this issue, we solve missing data and frequency transformation problems by calculating the statistics, obtaining mean and variance for each day and treating these values as daily sensor data. Even if there are some missing values, we can still calculate the statistics, and the frequency is successfully changed from minutes to days.



(a) Detecting outliers (b) Removing outlier neighborhoods

Fig. 4. Removing outliers from the data.

Finally, we transform L daily feeding and sensor data into one row data to match the frequency with our growth data. This problem is a mixed-frequency problem. We employ two methods, mixed data sampling (MIDAS) [34] and the uniform manifold approximation and projection (UMAP) [35], to solve this problem. MIDAS is the most popular and successful method to deal with these mixed-frequency data. Since MIDAS has a small number of parameters, it performs well in our situation where data is scarce, unlike recurrent neural network that typically have a large number of parameters. MIDAS consists of time-varying parameters and a regression model with those parameters. We use simple MIDAS regression formula with exponential Almon lag as follows:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 B\big(L, (\theta_1, \theta_2)\big)\mathbf{a}_t + \epsilon_t \\ B\big(L, (\theta_1, \theta_2)\big) &= \textstyle\sum_{l=1}^{L} b\big(l, (\theta_1, \theta_2)\big) D^l \\ b\big(l, (\theta_1, \theta_2)\big) &= \frac{e^{\theta_1 l + \theta_2 l^2}}{\sum_{l=1}^{L} e^{\theta_1 l + \theta_2 l^2}}, \end{aligned} \tag{1}$$

where ϵ_t denotes a Gaussian noise, $\mathbf{a}_t = (\mathbf{s}_t, \mathbf{d}_t) \in \mathbf{X}_s \times \mathbf{X}_d$ denotes daily input data consisting of daily sensor and feeding data at day t, D^k is a lag operator which means that $D^k(\mathbf{a}_t) = \mathbf{a}_{t-k}$, $\boldsymbol{\theta} = (\theta_1, \theta_2)$. We constrain θ_2 to be negative because the data has less impact over time. We treat $B\big(L, (\theta_1, \theta_2)\big)\mathbf{a}_t$ as input data, $\mathbf{x} \in \mathbf{X}$ for next step, probabilistic modeling. Even if this restricted functional space of MIDAS regression does not contain the optimal function, it helps to improve the model when the data is not too large [36]. This is because MIDAS is a good representative of a structure that loses influence over time with a small number of parameters. However, estimating MIDAS and the probabilistic model simultaneously requires training numerous parameters, which can be computationally expensive. Therefore, we optimize those parameters separately. Additionally, to confirm the suitability of MIDAS, we also implement another preprocessing method using UMAP as a baseline. We concatenate *L* daily data into one-row data and use UMAP to get the identical dimension with MIDAS. We compare two preprocessing strategies in Section 4.

3.2 Design output with a fragmented dataset

We mitigate a data shortage problem caused by the fragmented time series data**.** Since current fish weight is highly correlated to the past fish weight, it is important to add information for the history of fish weight into the input data. There are two types of collected fish weight data, 1) average fish weight measured every two months, 2) twenty fish weight collected by manual sampling every two weeks. If we utilize fish weight sampled two weeks ago, the time series data is not available at the initial stage, requiring the removal of initial data. Since our time series is fragmented—that is, the number of fragmented time series is not just one—we have to remove more data, which makes the insufficient data even more scarce. To preserve the number of data, we set the difference between the average fish weight measured every two months and the current fish weight as the output. The weight observed every two months is measured early in the fragmented time series, meaning it was measured immediately after the sorting process. Therefore, the difference represents the growth of the fish after the sorting process. We add the time taken from the initial point of fragmented timeseries to the data as a new variable.

3.3 Probabilistic modeling for multimodal dataset

We develop a multimodal probability model, specifically a multi-trajectory prediction model. These models can account for variability in growth by modeling multiple potential growth trajectories, which can lead to more accurate predictions. Aquaculture tanks group fish by their graded weight. This results in a unimodal weight distribution at first, but it changes to a multimodal one as fish grow at different rates over time. We point out that these aspects are seriously similar to the data association problem, modeling a mixture of multiple (continuous) stochastic processes. Consequently, we choose to implement overlapping mixture of Gaussian process (OMGP) [37], which is a Gaussian process (GP) [38] for data association problem.

In OMGP, the relationship between output $y_i \in \mathbf{y}$, and input $\mathbf{x}_i \in \mathbf{X}$ $(i = 1, \dots, n)$ is represented by the mixture process F where n denote the size of data (the number of data points). We assume Gaussian noise $\epsilon \sim N(0, \sigma^2)$ on output $\mathbf{y}$. The mixture process F consists of the K latent Gaussian process $\{f_j\}_{j=1}^{K}$ which help to model the K multi-modal distribution. The GP prior is expressed as $f_j \sim GP(\mathbf{0}, k_j)$ with zero mean $\mathbf{0}$ and kernel function k_j. OMGP assumes that each data $(\mathbf{x}_i, y_i)$ is generated from one of the latent functions. Let $\mathbf{Z}$ denote indicator matrix where *i*th row, j-th column entries, z_{ij} means that i-th data is generated from j-th latent Gaussian process. With these assumptions, the model's probability density functions are as follows:

$$\begin{aligned} P(\mathbf{y}|F, \mathbf{X}, \mathbf{Z}) &= \prod_{i=1}^{N}\prod_{j=1}^{K} N(y_i|f_i(\mathbf{x_i}), \sigma^2)^{z_{ij}} \\ P(\mathbf{Z}) &= \prod_{i=1}^{N}\prod_{j=1}^{K}(\Phi_{ij})^{z_{ij}}, \sum_{j=1}^{K}\Phi_{ij} = 1 \\ P(F|\mathbf{X}) &= \prod_{j=1}^{K} N\left(f_j(\mathbf{X})\middle|0, k_j(\mathbf{X}, \mathbf{X})\right), \end{aligned} \tag{2}$$

where $N(*\,|\mu, \sigma^2)$ denotes the probability density function of a normal distribution with mean μ and variance σ^2 with input $*$,and Φ_{ij} is the probability in which ith data is generated from f_j. Then the predictive distribution function, f_*, corresponding to a new input $\mathbf{x}_*$ is expressed as follows:

$$\begin{aligned} P(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &\approx \sum_{j=1}^{K} \Phi_{*j} N(f_*|\mu_*^j\ {\sigma_*^j}^2) \\ \mu_*^j &= k_j(\mathbf{x}_*, \mathbf{X})\left(k_j(\mathbf{X}, \mathbf{X}) + \mathbf{D}_j^{-\mathbf{1}}\right)^{-1}\mathbf{y} \\ \sigma_*^j &= \sigma^2 + k_j(\mathbf{x}_*, \mathbf{x}_*) \\ &\quad -k_j(\mathbf{x}_*, \mathbf{X})\left(k_j(\mathbf{X}, \mathbf{X}) + \mathbf{D}_j^{-\mathbf{1}}\right)^{-1}k_j(\mathbf{X}, \mathbf{x}_*), \end{aligned} \tag{3}$$

where $\mathbf{D}_j$ is a $n \times n$ diagonal matrix with i-th diagonal entry $\frac{\Phi_{ij}}{\sigma^2}$.

The biggest hurdle of OMGP prediction is how to set the proper prior Φ_{*j} for new input $\mathbf{x}_*$. Each trajectory of fish weight represents a distinct group of fish, and this group does not change significantly if sufficient time has passed since the sorting process. Therefore, our model approximates prior Φ_{*j} as the mean of Φ_{ij} over the nearby time points as follows:

$$\Phi_{*j} = \frac{\sum_{i \in N(x_*)} \Phi_{ij}}{||N(x_*)||}, \tag{4}$$

where $N(x_*)$ denotes the dataset observed in a period close to x_* and $||C(x_*)||$ means the number of elements of $N(x_*)$. For the tractable computation of the posterior distribution, $P(\mathbf{Z}, F(\mathbf{X})|\mathbf{X}, \mathbf{y})$, we adapt an improved variational bound for OMGP [37]. In the following sections, we use "OMGP" to denote the conventional model with uniform settings and "OMGP-SA" to denote our approach for smart aquaculture. We empirically determine the number of mixtures K according to the result of experiments. The setting of other model parameters will be explained in Section 4.

4. Experiments

4.1 Experimental settings

For our experiment, we implement two preprocessing methods: MIDAS and UMAP embedding. Additionally, we use four probabilistic models: GP, Mixture Density Network (MDN) [39], OMGP, and OMGP-SA. Our OMGP-SA is an extension of OMGP designed to handle multi-modal distribution and data association problems, with an appropriate prior for Φ_{*j}. GP serves as a basic probabilistic model, while MDN addresses multi-modal distribution, and OMGP tackles the data association problem. Notably, GP can be viewed as an MLP with infinite units in the hidden layer [40], and MDN is considered a special case of MLP. The features of each model are represented in **Table 3**.

Table 3. Feature of each implemented models

	Probabilistic modeling	**Multi-modal distribution**	**Data association problem**	**Proper prior for Φ_{*j}**
GP	O	X	X	-
MDN	O	O	X	-
OMGP	O	O	O	X
OMGP-SA (Ours)	O	O	O	O

Table 3 denotes instances where each model handles a specific feature with an "O", cases where it does not with an "X", and situations where the feature is not applicable with a "-".

We implement GP, OMGP, and OMGP-SA on Python with GPy package [41], and MDN with PyTorch package [42]. Each method trains with input data from two different embedding techniques, MIDAS and UMAP. In the models based on GP, we set prior mean and covariance to be zero and radial basis function (RBF) kernel, and the prior distribution P(Z) to be the Dirichlet distribution. We apply the same prior to GP. OMGP uses the L-BFGS-B optimizer (scipy implementation [43]), while GP and MDN use the Adam optimizer. The learning rate and number of iterations are 0.01 and 1000, respectively. Dataset is split into training and test datasets according to date. To evaluate the performance of our model, we designate the data observed on 2021-12-13 as our test set, while using the remaining data as our training set. To select the dataset observed in a period close to x_*, we consider $N(x_*)$ as the set of data observed on 2021-11-29 and 2021-12-27.

For quantitative analysis, we adopt two evaluation metrics: mean squared of error (MSE) and mean standardized log-likelihood (MSLL). To obtain MSE, we must have at least one

prediction value for each input. Since our model is probabilistic, we choose the mean as a representative value for each prediction. Also, when using multiple mixture elements (such as MDN or OMGP) for each data point, we calculate a weighted sum of the density values and MSE values using the mean of each mixture element. We assign weights to each mixture element based on the probabilities that the data point belongs to each mixture element. MSLL is also calculated in a similar way, and the formula is given as follows:

$$\mathrm{SE}(\mathbf{y},\mathbf{y}') = (\mathbf{y}-\mathbf{y}')^{\mathrm{T}}(\mathbf{y}-\mathbf{y}'),$$

$$\mathrm{MSE}(\mathbf{y}_*, \mathrm{f}_*) = \begin{cases} \frac{1}{\mathrm{n}}\mathrm{SE}(\mathbf{y}_*, \boldsymbol{\mu}_*) & \text{if f is GP} \\ \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{K}} \Phi_{*\mathrm{j}}\mathrm{SE}\left(\mathbf{y}_*, \boldsymbol{\mu}_*^{\mathrm{j}}\right) & \text{if f consists of K mixtures,} \end{cases}$$

$$\mathrm{SLL}(\mathrm{y},\sigma,\mu) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(\mathrm{y}-\mu)^2}{2\sigma^2}, \tag{5}$$

$$\mathrm{MSLL}(\mathbf{y}_*, \mathrm{f}_*) = \begin{cases} \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{SLL}(\mathrm{y}_{*\mathrm{i}}, \sigma_{*\mathrm{i}}, \mu_{*\mathrm{i}}) & \text{if f is GP} \\ \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{K}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \Phi_{*\mathrm{j}}\mathrm{SLL}\left(\mathrm{y}_{*\mathrm{i}}, \sigma_{*\mathrm{i}}^{\mathrm{j}}, \mu_{*\mathrm{i}}^{\mathrm{j}}\right) & \text{if f consists of K mixtures,} \end{cases}$$

where $\mathbf{y}_*$ is the fish weight, corresponding to a new input x_*. MSLL is an evaluation metric from a probabilistic view [38], so this metric is more suitable for assessing the probabilistic model than MSE. Indeed, the values of MSLL and MSE do not always align. Therefore, when we make comparisons between implemented probabilistic models, we say that the model with the better MSLL value outperforms unless otherwise noted. A lower score indicates a better fit of our model to the dataset in both metrics.

4.2 Experiment results

This section finds a suitable preprocessing method and model for our dataset. **Table 4** shows the quantitative comparison between our model and others. We highlight the cells that correspond to our model in yellow in the table. In each cell of **Table 4**, the first and second rows denote the MSLL and MSE values, respectively. The cells with the best MSLL value and the second best are highlighted in red and blue, respectively. From this table, we firstly see that all models trained with the MIDAS embedded data show better metric values than the UMAP averagely. Using MIDAS as the embedding method, GP achieves a better MSE value by 395.154 and a better MSLL value by 0.01425 than when using UMAP. We also compare the average per-expert values using MIDAS embedding with those using UMAP for MDN, OMGP, and OMGP-SA models, and find that MIDAS achieved better MSE values of 3143.05125, 68.37475, and 155.4 and better MSLL values of 0.04933, 0.09713, and 0.13381 than UMAP, respectively. The dominance of MIDAS is likely due to the fact that UMAP is generalized for all embedding situations, while MIDAS is specialized for mixed frequency data. UMAP should find the relationship between the same sensor values measured on different dates by itself, whereas MIDAS uses well-designed relationships. In addition, the MIDAS has very few parameters, which works well when the number of data is insufficient [36]. In the following, we will only compare models using MIDAS embedded data.

Table 4. Performance table for our smart aquaculture dataset.

First row: MSLL **Best MSLL value**
Second row: MSE **Second best MSLL**

Model	Embedding	Number of mixture (K)			
		2	**3**	**4**	**5**
GP	MIDAS	0.2799 3146.170			
	UMAP	0.29415 3541.324			
MDN	MIDAS	0.2746 3134.895	0.29055 2510.411	0.27775 3063.906	0.27685 2083.301
	UMAP	0.30645 4375.303	0.32075 6453.700	0.36785 6300.6219	0.322 6145.0931
OMGP	MIDAS	0.278 2338.572	0.29385 3003.035	0.36485 2563.207	0.3794 3542.594
	UMAP	0.34195 2670.375	0.3867 2712.878	0.33005 2545.094	0.65225 3792.56
OMGP-SA	MIDAS	**0.26515** 2342.421	**0.26855** 2545.400	0.3922 3193.065	0.28685 3287.255
	UMAP	0.40125 3170.205	0.2769 2338.575	0.523855 3193.706	0.546 3287.255

We confirm that our assumptions for data, fish weight distribution's multimodality, data association issues and approximating the prior Φ_{*j} by the average of near past Φ_{ij} is validate through experiments. MDN and OMGP exhibit better MSLL values than GP when K=2, and OMGP-SA outperform GP in terms of MSLL when K=2, 3. This demonstrates that modeling fish weight as multimodal distributions is appropriate for understanding smart aquaculture data, rather than assuming unimodal distributions. In addition, OMGP-SA when K=2, 3 outperforms MDN for all K. It suggests that considering the distribution of fish over time as a mixture of stochastic processes is a more effective approach for analyzing smart aquaculture data. All mixture models show the best MSLL values when K=2, indicating that two fish groups are enough for our smart aqua farm data. Additionally, the comparison between OMGP-SA and OMGP highlights the validity of our approximation method for the prior Φ_{*j}. Specifically, when K=2, OMGP had a worse MSLL value than MDN, indicating the limitations of traditional OMGP's prediction and emphasizing the importance of setting up a proper prior distribution, Φ_{*j}. Considering both the model and the number of mixtures, the best model in terms of MSLL is OMGP-SA with K=2.

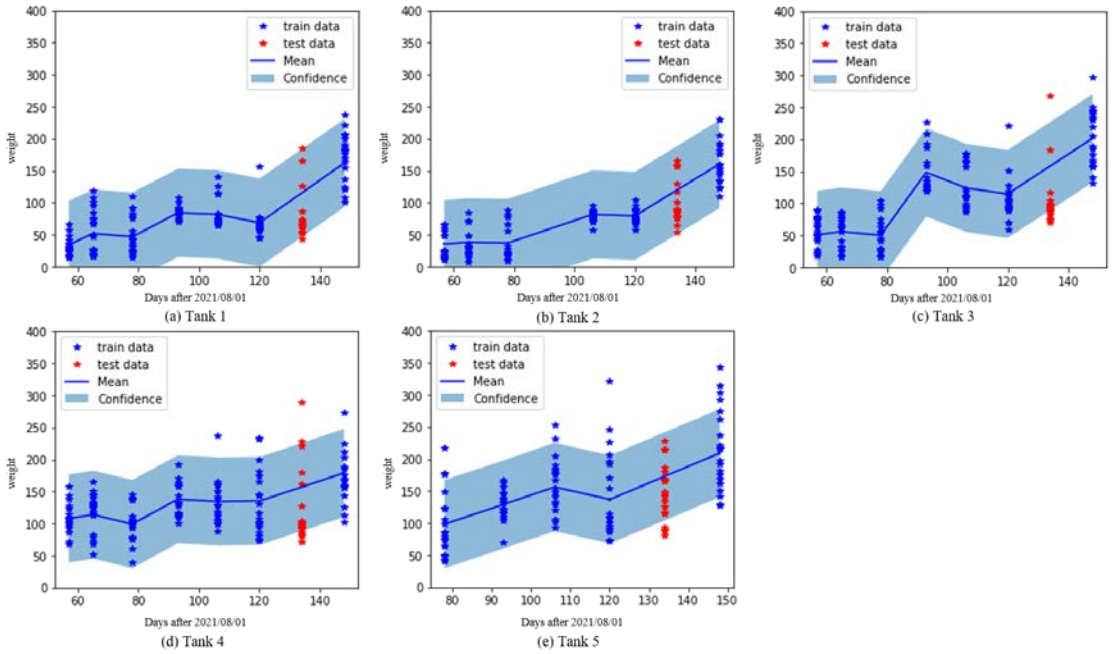
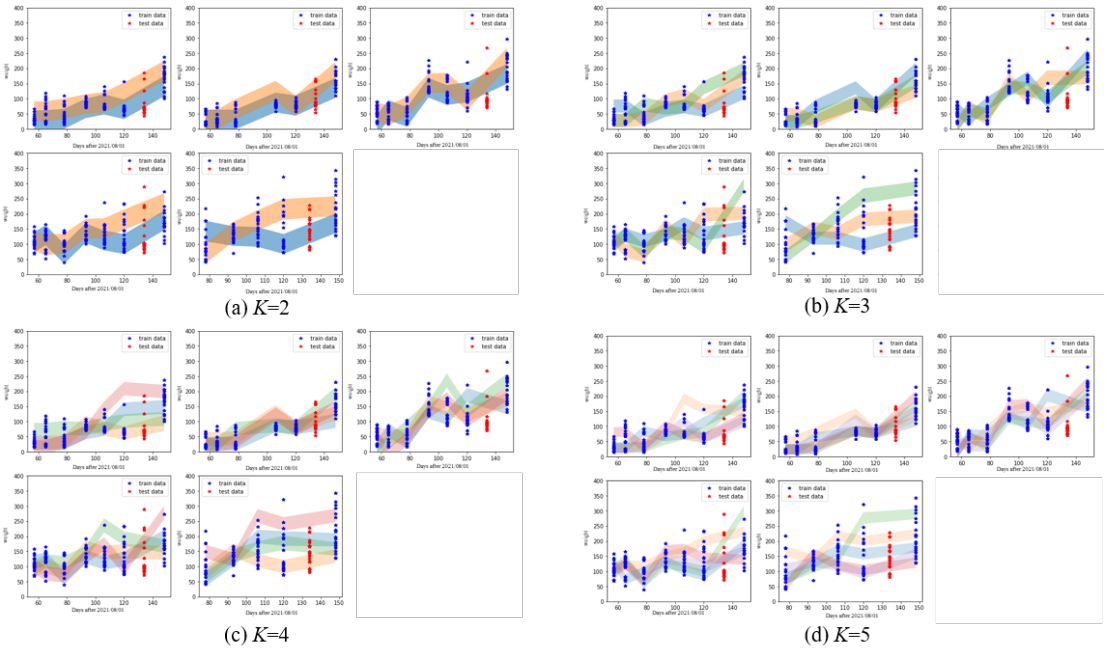


Fig. 5. Confidence interval plots of GP.

Fig. 5 and **6** show how our data exhibits the characteristics of a probabilistic multimodal distribution and how each mixture distinguishes different groups of fish. **Fig. 5** depicts the confidence intervals for GPs for each tank. The fish in each tank are plotted in different subfigures because they are not mixed, but we train the model with the entire data. The x-axis represents the number of days since 2021-08-01, one of days of sorting and grading, and the y-axis represents the fish weight. The train data was scatter-plotted in blue and the test data in red. The confidence interval is $[\mu - 2\sigma, \mu + 2\sigma]$ where and are the posterior mean μ and σ^2 are the posterior mean and variance, respectively.

Fig. 6 shows the confidence intervals of OMGP-SA for each number of mixtures, and each subplot has the same format as **Fig. 5**. Grading and sorting occur once, 87 days after 2021-08-01. As illustrated in **Fig. 5** and **6**, GP requires a larger confidence interval than the proposed model does to cover entire data due to its unimodality. This is interpreted as a result of trying to represent the entire data with a single mixture, resulting in large confidence intervals. On the other hand, we can see that our model has a smaller confidence interval but a mean value that is slightly closer to the actual data. This feature helps our model represent the data more accurately, as the results in the **Table 4** explain. As shown in **Fig. 6**, fish groups exhibit constantly changing patterns. Fish weight distribution follows a unimodal distribution at x=87, and then it diverges into various mixtures. We speculate that each fish group represents a distinct set of fish that diverges as fish grow. We numerically confirm that it is appropriate to divide the data into groups, but the interpretation of the groups remains an open question. When K=2 (the optimal number of clusters) each group has a distinct weight distribution with minimal overlap. So, we can say that the groups are divided based on differences in fish growth. However, for K=3, 4, and 5, the weight distributions of the fish in each group become less distinct over time and intersect with each other. For K=3, 4, and 5, it is possible that the model's interpretive power is weak, leading to incorrect results.



(a) K=2 (b) K=3

(c) K=4 (d) K=5

Fig. 6. Confidence interval plots of OMGP-SA for each number of mixtures, K.

5. Conclusion

This paper proposes a preprocessing and probabilistic fish growth model for smart aquaculture systems. Initially, we examine the characteristics of our dataset, which is a blend of IoT data and human-collected data. Four primary characteristics of the dataset present analytical challenges: mixed frequencies, varied farming methods, periodic missing data, and fragmented time series.

To address these characteristics, we employ a statistical preprocessing method that addresses missing data and adjusts frequencies using MIDAS. We note that fish share common traits that shift as a group over time. We incorporate this phenomenon into the model using OMGP and name it 'data association.' However, OMGP has limitations in setting a proper prior distribution for data belonging to a mixture element during prediction. Therefore, we modify OMGP to better suit our problem by designing a prior distribution with the assumption that a fish's group does not change over time. Our model outperforms, in terms of both MSLL and MSE, those models that do not consider the specific characteristics of the dataset. We validate our model using eel data from an actual aquaculture system.

Due to a lack of data, we estimate the parameters of the preprocessing model and probabilistic modeling separately. Should more data become available in the future, we can conduct a more comprehensive analysis of the fish growth model. Furthermore, this study exclusively utilizes tabular data. In the future, we aim to incorporate image and video data available through monitoring cameras, which could enhance data quality and reduce reliance on human-collected data.

6. Appendix

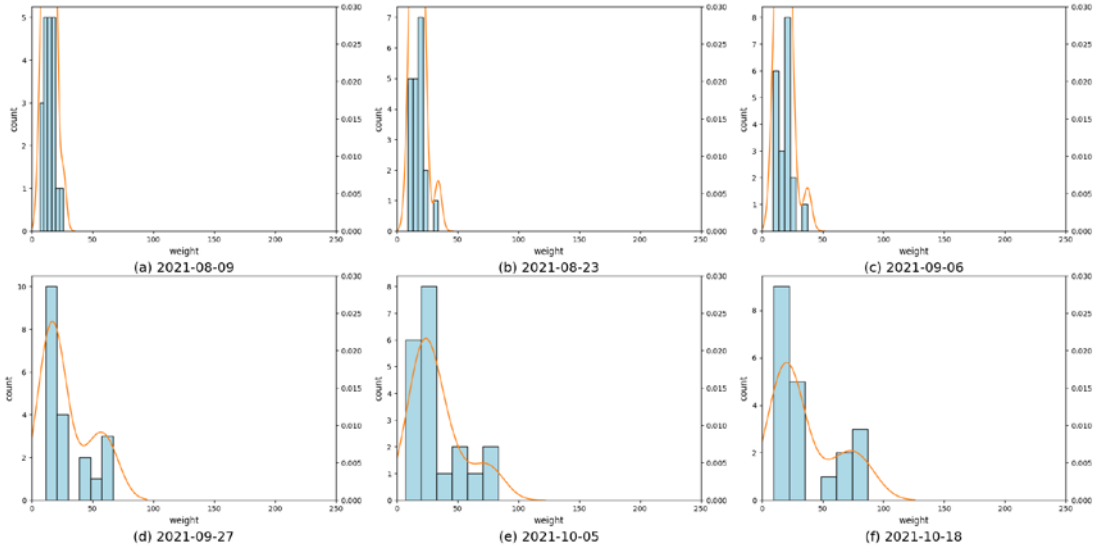


Fig. A.1. Histogram and kernel density plot of fish weight of tank 2 for 2021/08 – 2021/10.

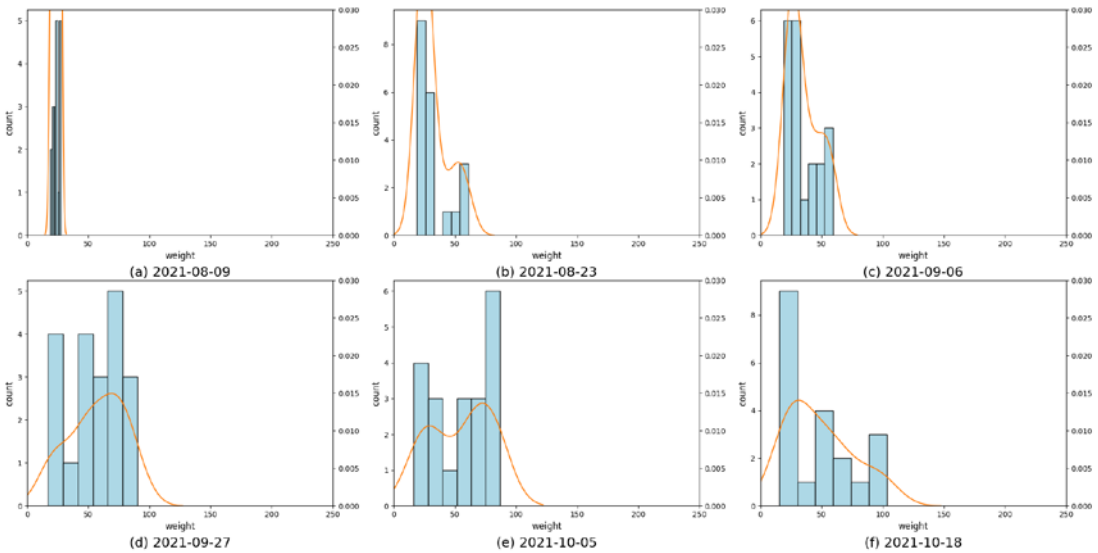


Fig. A.2. Histogram and kernel density plot of fish weight of tank 3 for 2021/08 – 2021/10.

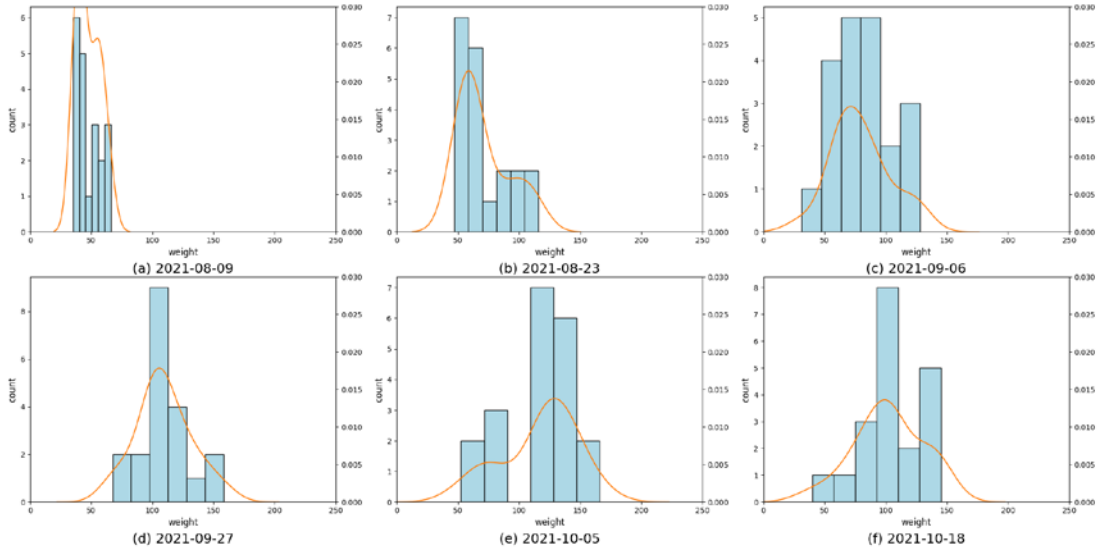


Fig. A.3. Histogram and kernel density plot of fish weight of tank 4 for 2021/08 – 2021/10.

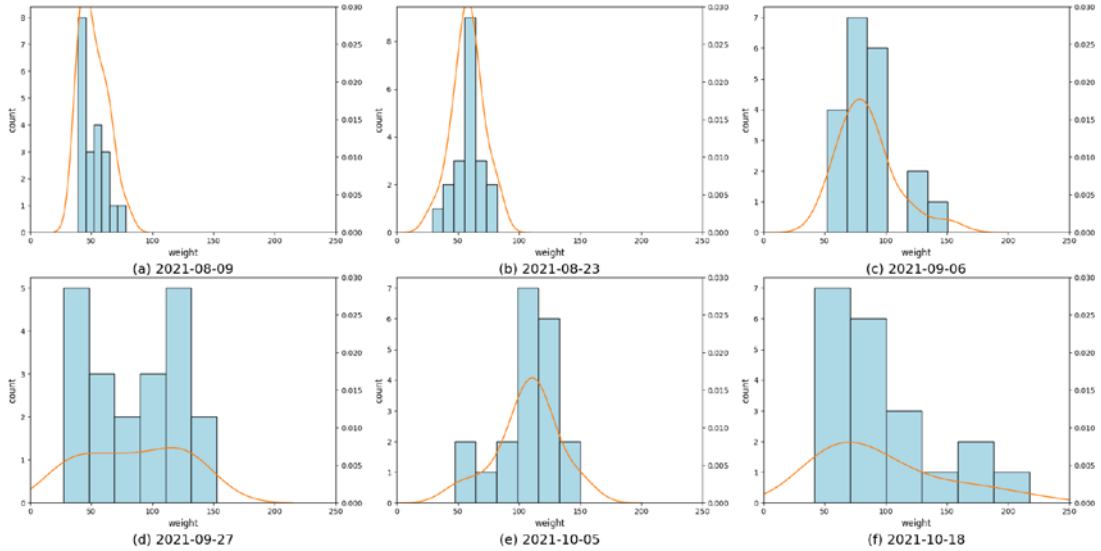


Fig. A.4. Histogram and kernel density plot of fish weight of tank 5 for 2021/08 – 2021/10.

Acknowledgement

We are grateful to the KETI researchers for providing us with essential materials and facilities for this research. We also appreciate the insightful academic discussions with the POSTECH LST Lab. researchers.

References

[1] FAO, “The State of World Fisheries and Aquaculture 2022,” *Towards Blue Transformation, Technical Report, FAO (Food and Agriculture Organization of the United States)*, 2022. Article (CrossRef Link)

[2] L. Parra, S. Sendra, J. Lloret, and J. J. Rodrigues, “Design and deployment of a smart system for data gathering in aquaculture tanks using wireless sensor networks,” *International Journal of Communication Systems*, vol. 30, 2017. Article (CrossRef Link)

[3] M.-C. Chiu, W.-M. Yan, S. A. Bhat, and N.-F. Huang, “Development of smart aquaculture farm management system using iot and ai-based surrogate models,” *Journal of Agriculture and Food Research*, vol. 9, 2022. Article (CrossRef Link)

[4] R. L. Naylor, R. W. Hardy, A. H. Buschmann, S. R. Bush, L. Cao, D. H. Klinger, D. C. Little, J. Lubchenco, S. E. Shumway, and M. Troell, “A 20-year retrospective review of global aquaculture,” *Nature*, vol. 593, pp. 551–563, 2021. Article (CrossRef Link)

[5] Choi, Wook, Yong Lee, and Sang-Chul Kim, “A reporting interval adaptive, sensor control platform for energy-saving data gathering in wireless sensor networks," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 5, no. 2, pp. 247-268, 2011. Article (CrossRef Link)

[6] C. Wang, Z. Li, T. Wang, X. Xu, X. Zhang, and D. Li, “Intelligent fish farm—the future of aquaculture,” *Aquaculture International*, vol. 29, pp. 2681–2711, 2021.Article (CrossRef Link)

[7] I. Ullah and D. Kim, “An optimization scheme for water pump control in smart fish farm with efficient energy consumption,” *Processes*, vol. 6, 2018. Article (CrossRef Link).

[8] Z. Hu, R. Li, X. Xia, C. Yu, X. Fan, and Y. Zhao, “A method overview in smart aquaculture,” *Environmental Monitoring and Assessment*, vol. 192, pp. 1–25, 2020. Article (CrossRef Link)

[9] X. Yang, S. Zhang, J. Liu, Q. Gao, S. Dong, and C. Zhou, “Deep learning for smart fish farming: applications, opportunities and challenges,” *Reviews in Aquaculture*, vol. 13, pp. 66–90, 2021. Article (CrossRef Link)

[10] Q. Kong, R. Du, Q. Duan, Y. Zhang, Y. Chen, D. Li, C. Xu, W. Li, and C. Liu, “A recurrent network based on active learning for the assessment of fish feeding status,” *Computers and Electronics in Agriculture*, vol. 198, 2022. Article (CrossRef Link)

[11] L. Du, Z. Lu, and D. Li, “Broodstock breeding behaviour recognition based on resnet50-lstm with cbam attention mechanism,” *Computers and Electronics in Agriculture*, vol. 202, 2022. Article (CrossRef Link)

[12] J.-C. Jang, Y.-R. Kim, S. Bak, S.-W. Jang, and J.-M. Kim, “Abnormal behaviour in rock bream (oplegnathus fasciatus) detected using deep learning-based image analysis,” *Fisheries and Aquatic Sciences*, vol. 25, pp. 151–157, 2022. Article (CrossRef Link)

[13] H. Wang, S. Zhang, S. Zhao, Q. Wang, D. Li, and R. Zhao, “Real-time detection and tracking of fish abnormal behavior based on improved yolov5 and siamrpn++,” *Computers and Electronics in Agriculture*, vol. 192, 2022. Article (CrossRef Link)

[14] Y. Mei, B. Sun, D. Li, H. Yu, H. Qin, H. Liu, N. Yan, and Y. Chen, “Recent advances of target tracking applications in aquaculture with emphasis on fish,” *Computers and Electronics in Agriculture*, vol. 201, 2022. Article (CrossRef Link)

[15] M. Saberioon and P. C ́ısa ̌r, “Automated within tank fish mass estimation using infrared reflection system,” *Computers and electronics in agriculture*, vol. 150, pp. 484–492, 2018. Article (CrossRef Link)

[16] C. Shi, R. Zhao, C. Liu, and D. Li, “Underwater fish mass estimation using pattern matching based on binocular system,” *Aquacultural Engineering*, vol. 99, 2022. Article (CrossRef Link)

[17] Aliyu, I., Gana, K. J., Musa, A. A., Adegboye, M. A., and Lim, C. G, "Incorporating recognition in catfish counting algorithm using artificial neural network and geometry," *KSII Transactions on Internet and Information Systems (TIIS)*, vol.14, no.12, pp. 4866-4888, 2020. Article (CrossRef Link)

[18] J. Liu, C. Yu, Z. Hu, Y. Zhao, Y. Bai, M. Xie, and J. Luo, “Accurate prediction scheme of water quality in smart mariculture with deep bi-s-sru learning network,” *IEEE Access*, vol. 8, pp. 24784–24798, 2020. Article (CrossRef Link)

[19] Q. Ren, X. Wang, W. Li, Y. Wei, and D. An, “Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network,” *Aquacultural Engineering*, vol. 90, 2020. Article (CrossRef Link)
[20] Jones, R. E., R. J. Petrell, and D. Pauly, "Using modified length–weight relationships to assess the condition of fish," *Aquacultural engineering*, vol. 20(4), pp. 261-276, 1999. Article (CrossRef Link)
[21] D. Pauly, "Gill size and temperature as governing factors in fish growth: a generalization of von Bertalanffy's growth formula," 1979. Article (CrossRef Link)
[22] Chambers, M. S., Sidhu, L. A., O’Neill, B., and Sibanda, N, "Flexible von Bertalanffy growth models incorporating Bayesian splines," *Ecological Modelling*, vol. 355, p. 1-11, 2017. Article (CrossRef Link)
[23] Føre, Martin, et al., "Modelling growth performance and feeding behaviour of Atlantic salmon (Salmo salar L.) in commercial-size aquaculture net pens: Model details and validation through full-scale experiments," *Aquaculture*, vol. 464, pp. 268-278, 2016.Article (CrossRef Link)
[24] Siegfried, Kate I., and Bruno Sansó., "Two Bayesian methods for estimating parameters of the von Bertalanffy growth equation," *Environmental Biology of Fishes*, vol. 77, pp. 301-308, 2006. Article (CrossRef Link)
[25] Hamre, J., Johnsen, E., and Hamre, K, "A new model for simulating growth in fish," *PeerJ*, 2, e244, 2014. Article (CrossRef Link)
[26] Lopez Quintero, Freddy Omar, et al. "Flexible Bayesian analysis of the von Bertalanffy growth function with the use of a log-skew-t distribution," *Fishery Bulletin*, vol. 115, pp. 13-26, 2017. Article (CrossRef Link)
[27] He, J. X. and J. R. Bence, "Modeling annual growth variation using a hierarchical Bayesian approach and the von Bertalanffy growth function, with application to lake trout in southern Lake Huron," *Transactions of the American Fisheries Society*, vol. 136, pp. 318-330, 2007. Article (CrossRef Link)
[28] Schnute, Jon T. and Laura J. Richards, "A unified approach to the analysis of fish growth, maturity, and survivorship data," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 47, pp. 24-40, 1990. Article (CrossRef Link)
[29] Gamito, Sofia, "Growth models and their use in ecological modelling: an application to a fish population," *Ecological modelling*, vol. 113, pp. 83-40, 1998. Article (CrossRef Link)
[30] Katsanevakis, Stelios, "Modelling fish growth: model selection, multi-model inference and model selection uncertainty," *Fisheries research*, vol. 81, pp. 229-235, 2006. Article (CrossRef Link)
[31] Chiu, Min-Chie, et al., "Development of smart aquaculture farm management system using IoT and AI-based surrogate models," *journal of Agriculture and Food Research*, vol. 9, p. 100357, 2022. Article (CrossRef Link)
[32] A. M. Kelly, D. Heikes, et al., “Sorting and grading warmwater fish, Southern Regional Aquaculture Center Stoneville,” *Mississippi*, 2013.
[33] Neal, Radford M., *Bayesian learning for neural networks*, Vol. 118, Springer Science & Business Media, 2012.
[34] E. Ghysels, A. Sinko, and R. Valkanov, “MIDAS regressions: Further results and new directions,” *Econometric reviews*, vol. 26, pp. 53–90, 2006. Article (CrossRef Link)
[35] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint*, 2018. Article (CrossRef Link)
[36] E. Ghysels, V. Kvedaras, and V. Zemlys, “Mixed frequency data sampling regression models: the R package midasr,” *Journal of statistical software*, vol. 72, pp. 1–35, 2016. Article (CrossRef Link)
[37] M. L´azaro-Gredilla, S. Van Vaerenbergh, and N. D. Lawrence, “Overlapping mixtures of gaussian processes for the data association problem,” *Pattern recognition*, vol. 45, pp. 1386–1395, 2012. Article (CrossRef Link)
[38] C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning*, ch. 2, MIT Press, 2006.
[39] C. M. Bishop, *Mixture density networks*, Technical report, Aston University, unpublished, 1994.
[40] C. Rasmussen, Z. Ghahramani, “Infinite mixtures of gaussian process experts,” *Advances in neural information processing systems*, vol.14, 2001.

[41] GPy, GPy: A gaussian process framework in python. [Online]. Available: http://github.com/SheffieldML/GPy,
[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol.32, 2019.
[43] E. Jones, T. Oliphant, P. Peterson, et al., "SciPy: Open source scientific tools for Python," 2001. [Online]. Available: http://www.scipy.org/

Jongwon Kim received his B.S. and M.S. degrees from the Dept. of Industrial and Management Engineering, Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2020 and 2022, respectively. He is currently a Ph.D. candidate at Dept. of Industrial and Management Engineering, POSTECH, Pohang, South Korea, under the supervision of Prof. Young Myoung Ko. His research interest includes probabilistic modeling, Gaussian process, timeseries model, and unsupervised learning.

Eunbi park received B.S. from Dept. of Mathematics, Catholic University of Korea, South Korea, in 2022. She is currently a Master candidate at Dept. of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, South Korea, under the supervision of Prof. Young Myoung Ko. Her research interests are machine learning, active learning, and probabilistic modeling.

Sungyoon Cho received the B.S., M. S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2006, 2008, and 2013, respectively. From 2013 to 2020, he was with Samsung Electronics, Korea, as a Staff Engineer to research cellular communication systems and develop 4G and 5G modem chipset. Since 2020, he has been with Korea Electronics Technology Institute (KETI) as a Principal Researcher to develop the advanced technologies for embedded network. His research interests are in the fundamental aspects of wireless communication and signal processing, and learning algorithms for practical application.

Kiwon Kwon received B.S. and M.S. degrees in computer engineering from Kwangwoon University, Korea, in 1997 and 1999, He also received the Ph.D. degree in the School of Electrical & Electronics Engineering from Chung-Ang University, Korea, in 2011. In 1999, he joined in KETI, Korea, where he is currently a Group Leader with Oceans and Fisheries ICT Group. His research interests are in the area of advanced broadcasting/communication system, digital twin, oceans and fisheries ICT.

Young Myoung Ko is an associate professor in the Department of Industrial and Management Engineering at Pohang University of Science and Technology (POSTECH), Pohang, South Korea. He received the B.S. and M.S. degrees in industrial engineering from Seoul National University, Seoul, South Korea, and the Ph.D. degree in industrial engineering from Texas A&M University, College Station, TX, USA, in 1998, 2000, and 2011, respectively. His research interests include, but are not limited to, simulation and optimization of stochastic systems, such as telecommunication networks, ICT infrastructure, and renewable energy systems.