

Article

Predicting Restroom Dirtiness Based on Water Droplet Volume Using the LightGBM Algorithm

Sumio Kurose ¹, Hironori Moriwaki ², Tadao Matsunaga ¹ and Sang-Seok Lee ^{1,*} 
¹ School of Engineering, Tottori University, 4-101 Koyama-Minami, Tottori 680-8552, Japan; sumiokurose@yahoo.co.jp (S.K.)

² Asuka Bisou Co., Ltd., 296-1 Daigo, Kashihara 634-0072, Japan; hironori.moriwaki@asukabiso.co.jp

* Correspondence: sslee@tottori-u.ac.jp

Abstract: This study examines restroom cleanliness in public facilities, department stores, supermarkets, and schools by using water droplet volumes around washbowls as an indicator of usage. Rising cleaning costs due to labour shortages necessitate more efficient restroom maintenance. Quantifying water droplet accumulation and predicting cleaning schedules can help optimise cleaning frequency. To achieve this, water droplet volumes were measured at specific time intervals, with significant variations indicating increased restroom usage and potential dirt buildup. For real-world assessment, acrylic plates were placed on both sides of washbowls in public restrooms. These plates were collected every hour over five days and analysed using near-infrared photography to track changes in water droplet areas. The collected data informed the development of a prediction system based on the decision tree method, implemented via the LightGBM framework. This paper presents the developed prediction system, which utilises in situ water droplet volume measurements, and evaluates its accuracy in forecasting restroom cleaning needs.

Keywords: restroom dirtiness; LightGBM; data augmentation techniques; water droplet volume prediction; cleaning schedule prediction



Academic Editor: Jiachen Yang

Received: 28 December 2024

Revised: 23 February 2025

Accepted: 19 March 2025

Published: 30 March 2025

Citation: Kurose, S.; Moriwaki, H.; Matsunaga, T.; Lee, S.-S. Predicting Restroom Dirtiness Based on Water Droplet Volume Using the LightGBM Algorithm. *Sensors* **2025**, *25*, 2186. <https://doi.org/10.3390/s25072186>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Current Issues Faced by Cleaning Companies

Restrooms in public and commercial facilities, such as department stores, supermarkets, and schools, are typically maintained by cleaning companies. However, cleaning is often not considered a value-adding service, leading to cost compromises. Recently, rising labour costs have further increased cleaning expenses, prompting companies to focus on efficiency by optimizing cleaning frequency and duration.

Typically, commercial and public restrooms are cleaned before opening to the public, followed by periodic inspections. During these inspections, consumables such as toilet paper and hand soap are restocked, and any visible dirt or stains are addressed to maintain cleanliness [1–5]. This additional effort ensures that restrooms remain hygienic and well maintained.

1.2. Solution for Restroom Cleaning Issues

Historically, public restroom cleanliness has received minimal attention. However, restrooms in restaurants, department stores, and supermarkets are now more closely monitored, as their cleanliness directly impacts the facility's image. Recognizing this, businesses have increasingly prioritised restroom maintenance.

Figures 1 and 2 present survey responses from 2000 participants regarding restroom cleanliness in restaurants. Figure 1 illustrates responses to the question ‘How do you feel about clean restaurant restrooms?’. Nearly 80% of respondents stated that clean restrooms improve a restaurant’s image. Figure 2 shows responses to the question ‘How do you feel about dirty restaurant restrooms?’. Here, 77% of respondents felt that unclean restrooms negatively impact a restaurant’s image, indicating a lack of consideration for customers. These findings suggest that restroom cleanliness significantly influences public perception, not just in restaurants but also in department stores and supermarkets.

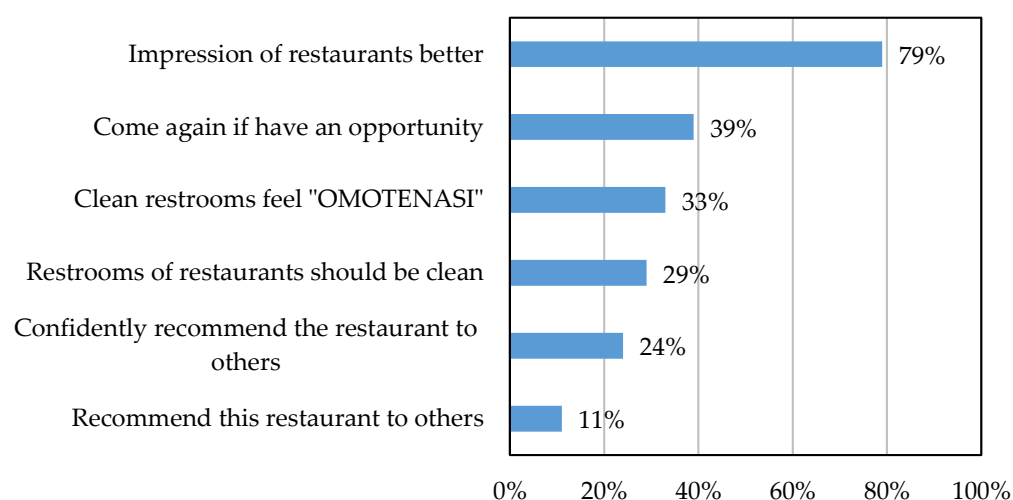


Figure 1. Responses from 2000 participants to the survey question: ‘How do you feel about clean restrooms in restaurants?’ [6].

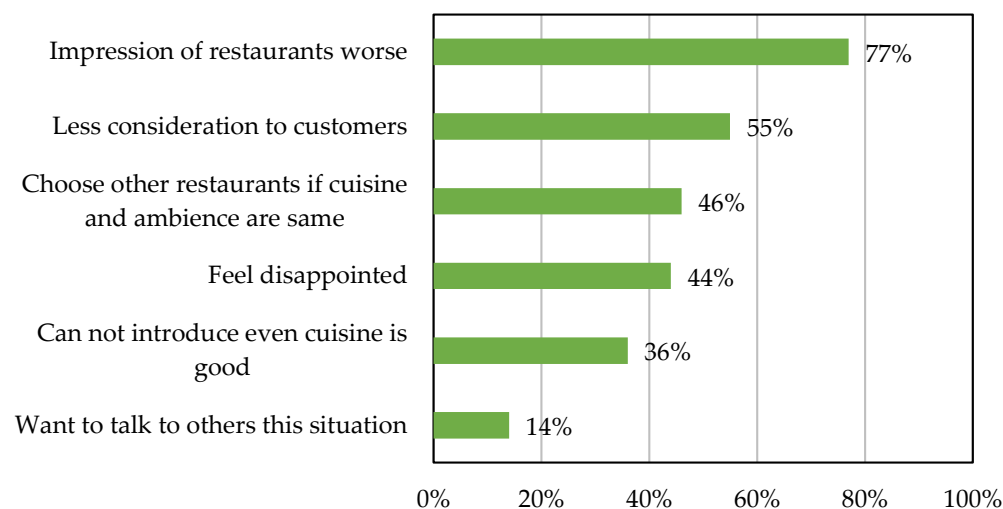


Figure 2. Responses from 2000 participants to the survey question ‘How do you feel about dirty restrooms in restaurants?’ [6].

Figure 3 presents survey responses to the question ‘How comfortable do you feel using public restrooms?’. Participants, aged 20 to 60, included 80 respondents per sex per age group (totalling 800 respondents). Multiple answers were allowed [7].

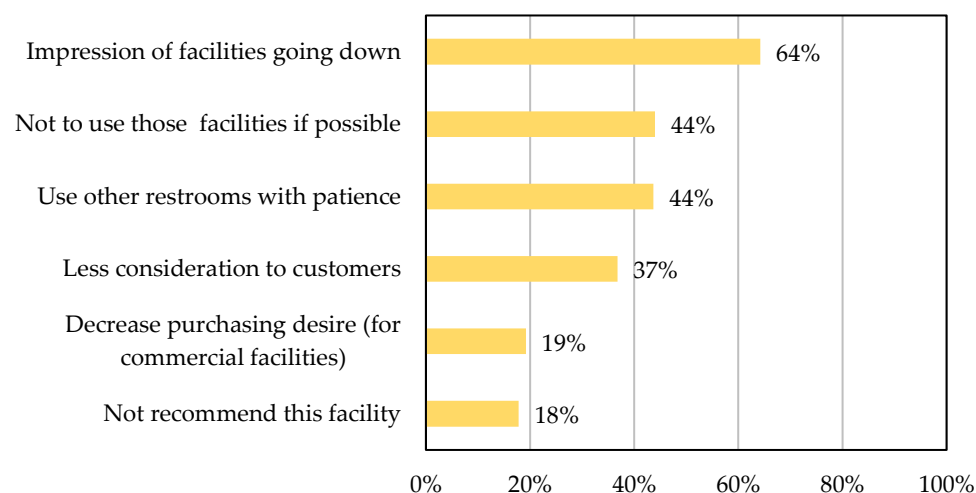


Figure 3. Responses of participants to the survey question ‘How comfortable do you feel using public restrooms?’ [7].

Current cleaning routines are insufficient to ensure restroom cleanliness at all times. Setting an appropriate cleaning frequency and timing is essential [8,9]. Furthermore, if cleaning duration can be accurately predicted, restrooms can remain clean while minimizing cleaning costs.

As previously mentioned, assessing restroom dirtiness enables the determination of an optimal cleaning frequency and duration, allowing for a more flexible and efficient cleaning strategy. To achieve this, this study develops a prediction system and establishes a methodology for effective restroom maintenance.

2. Dirtiness Quantification Method

2.1. Definition of Dirtiness

Restrooms contain various sources of dirt, such as water droplets, hairs, and paper waste. While all types of dirt should be removed during cleaning, quantifying these different types of dirt presents challenges. Specifically, paper waste, dirt around urinals, and hairs in washbowls are difficult to measure precisely. To simplify the process, we defined the presence of water droplets around washbowls as an indicator of overall dirtiness.

Figure 4 displays images of water droplets placed using a digital pipette. The relationship between the number of pixels representing the water droplets and the corresponding volume was determined before the experiments, as shown in Figure 5. This relationship is approximated as linear.

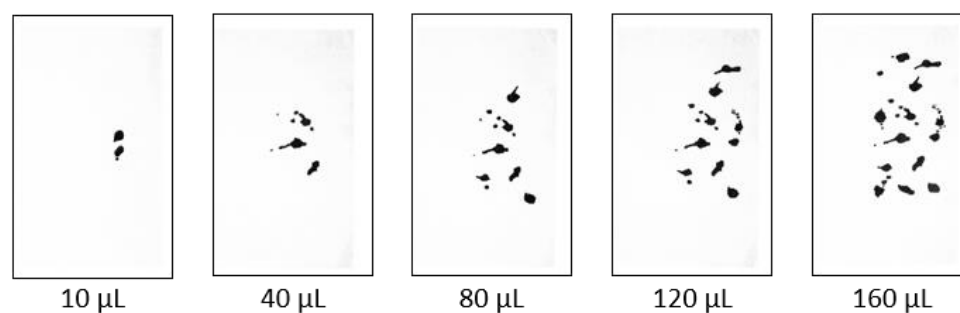


Figure 4. Relationship between pixel images of water droplets dropped using a digital pipette and the corresponding volumes.

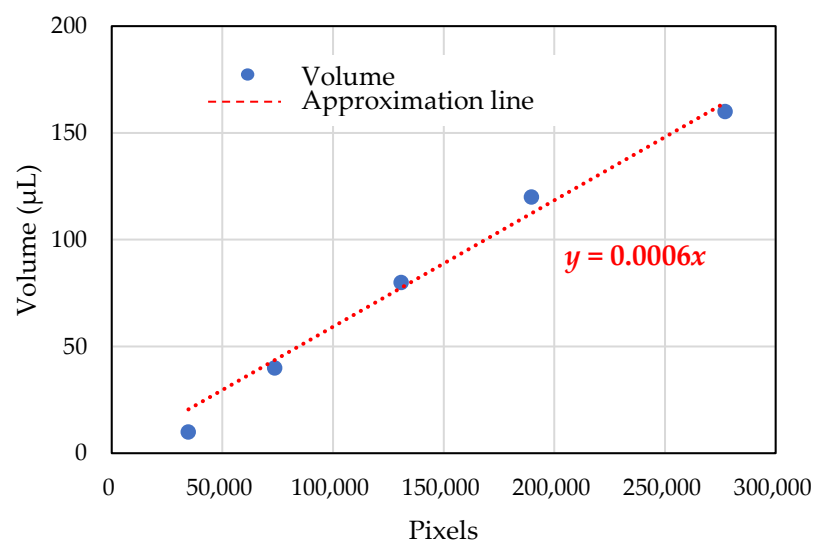


Figure 5. Conversion of pixel data to volume from the collected images.

For the in situ experiments, water droplets were collected on acrylic plates installed beside the washbowls. These droplets were then photographed, and the pixel-based images were converted to corresponding volumes using the method outlined in Section 2.2.

2.2. Overview of Method

The experiment followed a series of steps to determine the volume of water droplets around the washbowl. Water droplets were photographed using a near-infrared camera, as shown in Figure 6, where the droplets appear as black areas.

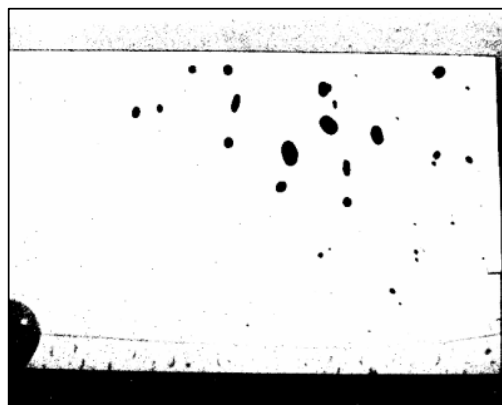


Figure 6. Example of a bitmap image of a water droplet captured by a near-infrared camera.

However, the image also contains extraneous black areas outside the acrylic plate, caused by background differences where no water droplets exist. To ensure accurate pixel counting, image-editing software (Photos, version 2024.11040.1002.0) was used to remove these irrelevant black areas. The cleaned-up bitmap image is shown in Figure 7. The number of pixels in the cleaned image (Figure 7) was then counted using the open-source software ImageJ (version 1.53). Finally, the pixel count was converted into the volume of water droplets using the established relationship shown in Figure 5.

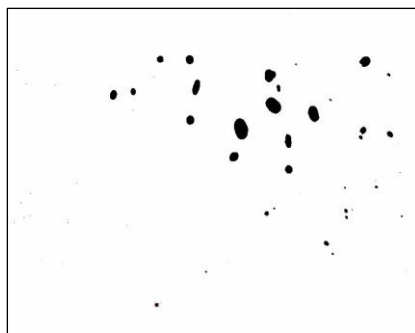


Figure 7. Bitmap image after removing the irrelevant black areas from Figure 6.

This method enables the quantification of restroom dirtiness based on the volume of water droplets around the washbowl. While it does not provide a perfectly accurate estimate due to factors such as overlapping droplets, the method offers a reliable means to assess water droplet amounts. Additionally, it allows for the establishment of a cleanliness threshold, which is crucial for determining optimal cleaning timing.

3. In Situ Experiment and Results

3.1. Preparation of Water Droplet Images

To capture time-varying images of water droplets, experiments were conducted in the restrooms of a language school with approximately 70 students. The number of students present on any given day varied, and most students were from Asian countries studying Japanese for college entrance exams or Japanese qualification exams.

To collect water droplets around the washbowl, acrylic plates were placed on either side of the washbowl and replaced every hour for analysis. Figure 8 shows the actual setup, with transparent acrylic plates used in the experiment. The plates were colour-coded blue in the figure for better visibility.



Figure 8. Photograph of the washbowl with acrylic plates placed on either side. Note that the acrylic plates, though transparent in the experiment, are shown in blue here for clarity.

Setting up a camera inside the restroom posed privacy concerns, and it was difficult to maintain the same camera position every hour. Moreover, distinguishing between male and female users was necessary. Consequently, the acrylic plates were collected every hour, and photographed in a separate room under consistent conditions, and the images were transferred to a computer connected to the camera. The camera lens was positioned 650 mm above the acrylic plates, as shown in Figure 9.

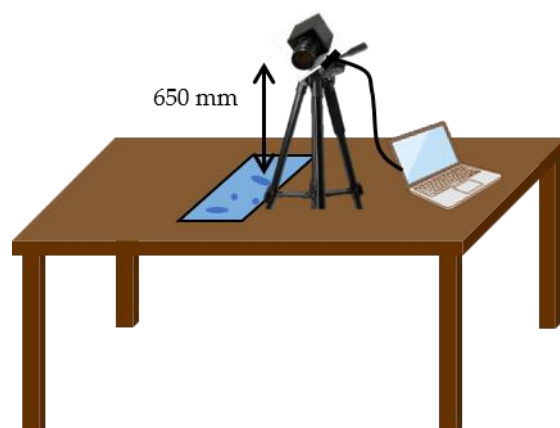


Figure 9. Setup of the camera was set up in a different room from the restroom to capture photographs of water droplets on the acrylic plates.

A near-infrared camera (NIRCam-640SN, Vision Sensing Co., Ltd., Osaka, Japan) was used for capturing the images, with specifications listed in Table 1. Since water absorbs at wavelengths of 1200, 1450, and 1940 nm in the near-infrared spectrum, a 1450 nm bandpass filter was attached to the camera to capture water droplets effectively. As a result, water droplets were depicted as black areas in the images.

Table 1. Specifications of the camera used in the experiments [10].

Image Sensor	InGaAs
Number of pixels	640 × 512 pixels
Wavelength sensitivity	0.9–1.7 μm
Frame rate	98 fps
Camera dimensions	61 × 59 × 81 mm (width × height × depth) (without the lens)

Representative photographs taken during the in situ experiment are shown in Figures 10–12. These images were captured at 10:30, 11:30, and 12:30 on the first day of the experiment at male restroom. The images on the left show the raw bitmap images, while the processed images on the right have had irrelevant black areas removed, as described in Section 2.2. Water droplet volumes were then estimated using the pre-determined relationship shown in Figure 5. These volume data were subsequently used as input for the prediction system.

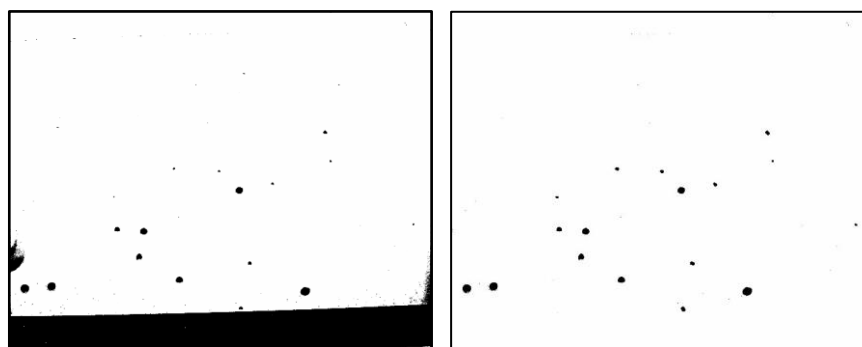


Figure 10. Captured bitmap image of water droplets taken at 10:30 on the first day of the experiment at male restroom (left) and the processed image after removing irrelevant black areas (right).

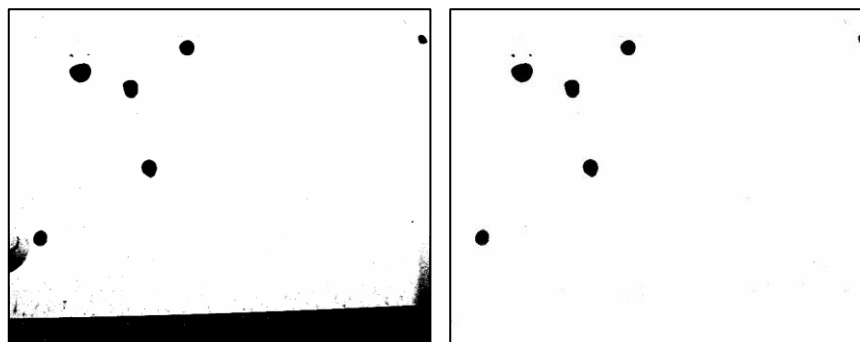


Figure 11. Captured bitmap image of water droplets taken at 11:30 on the first day of the experiment at male restroom (left) and the processed image after removing irrelevant black areas (right).

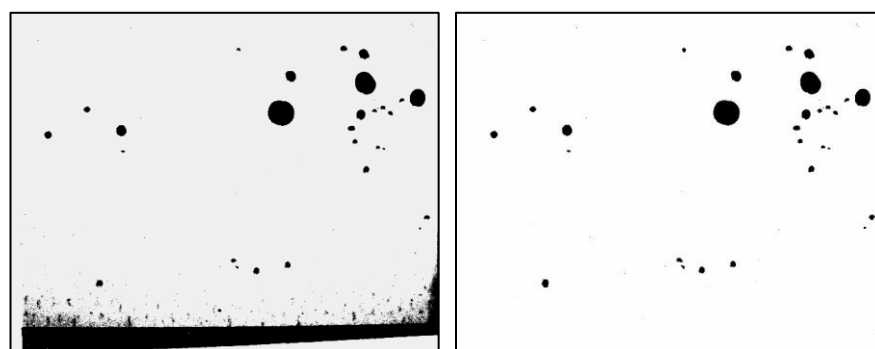


Figure 12. Captured bitmap image of water droplets taken at 12:30 on the first day of the experiment at male restroom (left) and the processed image after removing irrelevant black areas (right).

3.2. Data Collection

Photographs of water droplets on the acrylic plates were captured at 1 h intervals, and the data were accumulated into ‘1 h before’ datasets. These images were collected over five days from both male and female restrooms. The data are presented in Figure 13 and Table 2. The variables lag1030 through lag1730 represent data collected at hourly intervals from 10:30 to 17:30. Additionally, the average temperature and humidity were recorded, as shown in Table 2.

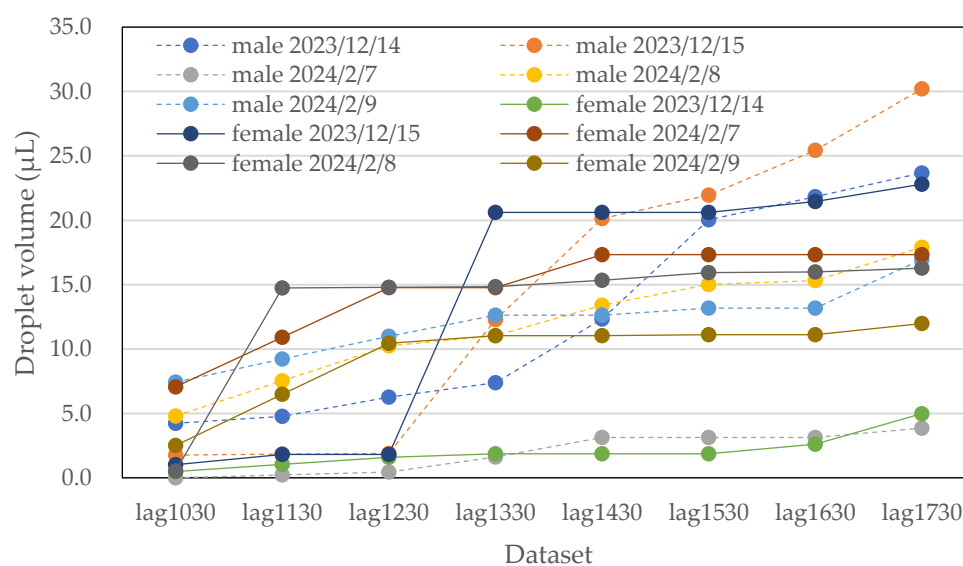


Figure 13. Five-day data from male (dashed lines) and female (solid lines) restrooms, converted to water droplet volume data.

Table 2. Five-day data collected from both male and female restrooms, converted to water droplet volumes, along with temperature and humidity data.

Sex	Temperature (°C)	Humidity (%)	lag1030 (μL)	lag1130 (μL)	lag1230 (μL)	lag1330 (μL)	lag1430 (μL)	lag1530 (μL)	lag1630 (μL)	lag1730 (μL)
Male	12.4	62.9	4.2	4.8	6.3	7.4	12.3	20.0	21.8	23.7
Male	16.3	82.9	1.8	1.8	1.9	12.3	20.1	22.0	25.4	30.2
Male	6.6	69.6	0.0	0.2	0.4	1.6	3.1	3.1	3.1	3.9
Male	6.3	60.6	4.8	7.5	10.2	11.0	13.4	15.0	15.3	17.9
Male	8.6	48.8	7.4	9.2	11.0	12.6	12.6	13.2	13.2	17.0
Female	12.4	62.9	0.5	1.0	1.6	1.9	1.9	1.9	2.6	5.0
Female	16.3	82.9	1.0	1.8	1.8	20.6	20.6	20.6	21.5	22.8
Female	6.6	69.6	7.1	10.9	14.8	14.8	17.3	17.3	17.3	17.3
Female	6.3	60.6	0.5	14.7	14.8	14.8	15.3	15.9	16.0	16.3
Female	8.6	48.8	2.5	6.5	10.5	11.0	11.0	11.1	11.1	12.0

The collected data on water droplet volumes exhibited significant variation, depending on individual usage, as shown in Figure 13. On some days, there were abrupt changes in droplet volumes in response to time variations. However, as the number of data points increased, this variation became less pronounced. Generally, the data indicated a tendency for the accumulated water droplet volume to increase with the number of restroom users. It is important to note that ventilation fans were continuously running in the restrooms, which led to the evaporation of water droplets over time. As a result, the observed water droplet volumes were slightly lower than the actual volumes, implying that the prediction may be somewhat underestimated.

3.3. Data Augmentation

As noted earlier, the data were collected for five days from both male and female restrooms but these were insufficient for training and prediction purposes. To overcome this limitation, a data augmentation method was applied, as long-term restroom experiments are constrained. Several data augmentation techniques have been reported in the literature, including the addition of random noise, data scaling, weighted dynamic time warping (DTW), dynamic time warping barycentric averaging (DBA) [11–13], moving block bootstrapping (MBB) [14–17], and the DoppelGANger framework based on generative adversarial networks (GAN) [18–20]. In anomaly detection, adding random noise is effective, and GANs can generate high-fidelity datasets. MBB generates pseudo-scattering through random sampling. However, our application required only a small portion of scattered datasets, and the datasets needed to be time-series. Therefore, we adopted the DBA method for our prediction model.

We primarily employed two methods for data augmentation:

1. Averaging and scaling method: This approach involved successively taking averages of two nearest neighbouring data points, two second-nearest neighbouring data points, and so on. These averages were then multiplied by 0.9 and 1.1, and the process was repeated. This method generated three sets of augmented data by expanding the margins of the minimum and maximum values.
2. Random number generation method: In this method, random numbers were generated between the maximum and minimum values for each dataset.

Based on the above, we primarily attempted the following two methods:

Details of the first method, involving averaging and scaling, are illustrated in Figure 14. This method resulted in a total of 60 datasets: five original datasets (data No. 1–5) and 55 augmented datasets (data No. 6–60). Each augmented dataset also included the average temperature and humidity for the day, obtained from a meteorological agency. In Figure 14,

data No. 6 represents the average of data No. 1 and data No. 2, including temperature and humidity averages. Similarly, data No. 11 is the average of data No. 1 and data No. 3, and so on. By successively taking averages with two nearest neighbours, two second-nearest neighbours, and so on, we generated 20 augmented data points from the original five datasets.

No.	temp	humidity	lag1030	lag1130	lag1230	lag1330	lag1430	lag1530	lag1630	lag1730	
1	12.4	62.9	4.2	4.8	6.3	7.4	12.3	20.0	21.8	23.7	Actual data
2	16.3	82.9	1.8	1.8	1.9	12.3	20.1	22.0	25.4	30.2	
3	6.6	69.6	Average	0.2	0.4	1.6	3.1	3.1	3.1	3.9	
4	6.3	60.6		7.5	10.2	11.0	13.4	15.0	15.3	17.9	
5	8.6	48.8	7.4	9.2	11.0	12.6	12.6	13.2	13.2	17.0	
6	14.4	72.9	3.0	3.3	4.1	9.8	16.2	21.0	23.6	26.9	Average nearest neighbor
7	11.5	76.3	0.9	Average	1.2	6.9	11.6	12.5	14.3	17.0	
8	6.5	65.1	2.4		5.3	6.3	8.3	9.1	9.2	10.9	
9	7.5	54.7	6.1	8.4	10.6	11.8	13.0	14.1	14.2	17.4	
10	11.5	60.9	5.2	6.3	7.5	11.2	14.4	17.1	18.4	21.9	
11	9.5	66.3	2.1	2.5	3.4	4.5	7.7	11.6	12.5	13.8	Average 2nd nearest neighbor
12	11.3	71.8	3.3	4.7	Average	1.7	16.8	18.5	20.4	24.1	
13	7.6	59.2	3.7	4.7		7.1	7.9	8.1	8.1	10.4	
14	10.3	66.8	3.9	5.4	7.2	10.4	14.8	18.0	19.5	22.4	
15	10.0	62.5	4.2	5.1	6.1	9.8	12.1	12.9	13.7	17.0	
16	9.4	61.8	4.5	6.2	8.3	9.2	12.9	17.5	19.6	20.8	Average 3rd nearest neighbor
17	12.5	65.9	4.6	5.5	6.4	12.4	15.4	17.6	19.3	23.6	
18	10.5	71.3	1.5	1.8	2.3	Average	9.7	12.1	13.4	15.4	
19	8.9	68.4	2.8	4.3	5.7	9.0	12.5	13.8	14.8	17.5	
20	7.5	57.0	4.9	6.5	8.2	9.5	10.4	11.1	11.2	13.9	
21	10.5	55.9	5.8	7.0	8.6	10.0	12.5	16.6	17.5	20.3	Average 4th nearest neighbor
22	15.3	77.9	2.4	2.6	3.0	11.0	18.2	21.5	24.5	28.6	
23	9.0	72.9	0.4	0.6	0.8	4.3	7.4	7.8	8.7	10.4	
24	6.4	62.9	3.6	5.7	7.8	8.7	10.8	12.1	12.3	14.4	
25	8.0	51.8	6.8	8.8	10.8	12.2	12.8	13.6	13.7	17.2	
26	13.6	69.2	4.7	5.3	6.9	8.1	13.6	22.0	24.0	26.0	Multiply by 1.1
27	17.9	91.2	1.9	2.0	2.1	13.5	22.2	24.2	28.0	33.2	
28	7.3	76.6	0.0	0.2	0.5	1.8	3.4	3.4	3.4	4.2	
29	6.9	66.7	5.3	8.3	11.3	12.1	14.8	16.5	17.1	17.7	
30	9.5	53.7	8.2	10.1	12.1	13.9	13.9	14.5	14.5	17.7	
31	11.2	56.6	3.8	4.3	5.6	6.6	11.1	18.0	19.7	21.3	Multiply by 0.9
32	14.7	74.6	1.6	1.6	1.7	11.0	18.1	19.8	22.9	27.2	
33	5.9	62.6	0.0	0.2	0.4	1.5	2.8	2.8	2.8	3.5	
34	5.7	54.5	4.3	6.8	9.2	9.9	12.1	Average	13.8	16.1	
35	7.7	43.9	6.7	8.3	9.9	11.4	11.4		11.9	15.3	
36	15.8	80.2	3.3	3.6	4.5	10.8	17.9	23.1	26.0	29.6	Average nearest neighbor of ×1.1 data
37	12.6	83.9	1.0	1.1	1.3	7.6	12.8	13.8	15.7	18.7	
38	7.1	71.6	2.6	4.3	5.9	7.0	9.1	10.0	10.1	12.0	
39	8.2	60.2	6.1	9.2	11.7	13.0	14.3	15.5	15.7	19.2	
40	10.3	55.1	6.0	7.2	8.9	10.3	12.5	16.3	17.1	20.0	
41	12.9	65.6	2.7	3.0	3.7	8.8	14.6	18.9	21.3	24.2	Average nearest neighbor of ×0.9 data
42	10.3	68.6	0.8	0.9	1.1	6.3	10.5	11.3	12.9	15.3	
43	5.8	58.6	2.2	3.1	4.8	5.7	7.4	8.2	8.3	9.8	
44	6.7	49.2	5.5	7.5	9.6	10.7	11.7	12.7	12.8	15.7	
45	11.8	62.1	5.0	6.0	7.2	11.1	14.6	17.5	18.9	22.4	
46	14.2	72.2	3.0	3.3	4.0	9.7	16.1	20.8	23.4	26.7	Multiply by 0.9 of averaged
47	11.3	75.5	0.9	1.0	1.2	6.9	11.5	12.4	14.1	16.9	
48	6.4	64.4	2.4	3.8	5.3	6.3	8.2	9.0	9.1	10.8	
49	7.4	54.2	6.1	8.3	10.1	1.7	12.9	14.0	14.1	17.3	
50	9.3	49.6	2.4	6.5	8.1	9.2	11.2	14.6	15.4	18.0	
51	11.6	59.0	2.4	2.7	3.3	8.0	13.2	17.0	19.1	21.8	Multiply by 1.1 of averaged
52	9.3	61.8	0.7	0.8	0.9	5.6	9.4	10.2	11.6	13.8	
53	5.2	52.7	1.9	3.1	4.3	5.1	6.7	7.4	7.5	8.8	
54	6.0	44.3	5.0	6.8	8.6	9.6	10.5	11.4	11.5	14.1	
55	10.6	55.8	4.5	5.4	6.5	10.0	13.1	15.7	17.0	20.2	
56	17.4	88.2	3.6	4.0	4.9	11.9	19.6	25.4	28.6	32.6	Multiply by 1.1 of averaged
57	13.9	92.3	1.1	1.2	1.4	8.4	14.1	15.2	17.3	20.6	
58	7.8	78.8	2.9	4.7	6.5	7.7	10.0	11.0	11.2	13.2	
59	9.0	66.2	7.4	10.1	12.9	14.3	15.8	17.1	17.2	21.1	
60	11.3	60.7	6.6	7.9	9.8	11.3	13.7	17.9	18.8	22.0	

Figure 14. Data augmentation results using the average and multiply method.

Next, data No. 26 and data No. 27 were obtained by multiplying data No. 1 and data No. 2 by 1.1, respectively. Similarly, data No. 31 and data No. 33 were multiplied by 0.9. This step provided an additional 10 augmented datasets. The process was repeated by

averaging the new data points, resulting in a total of 60 datasets. Figures 15 and 16 show the actual water droplet data from the male restroom (Figure 13) and the augmented data from the male restroom, respectively.

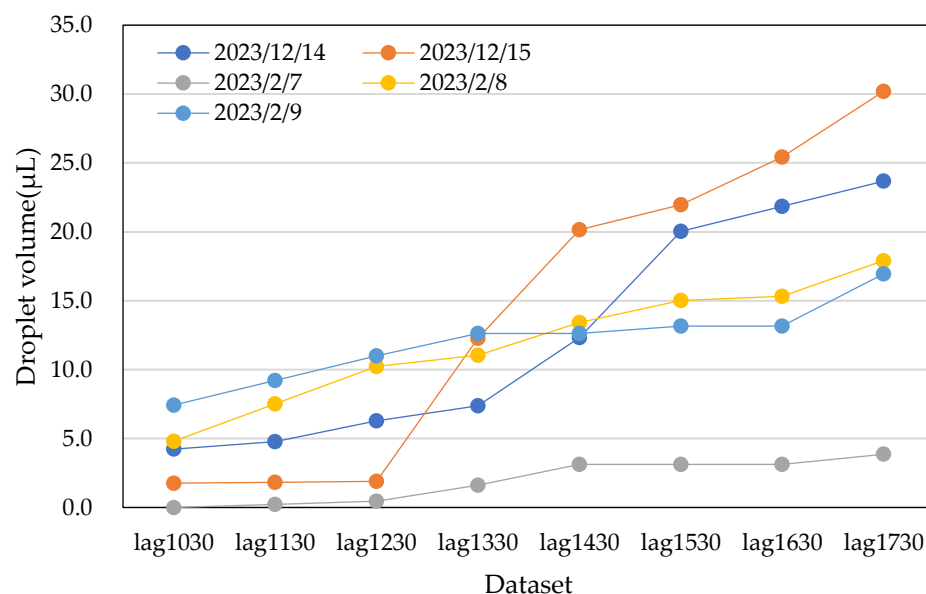


Figure 15. Actual water droplet volume data from the male restroom over five days, excerpted from Figure 13.

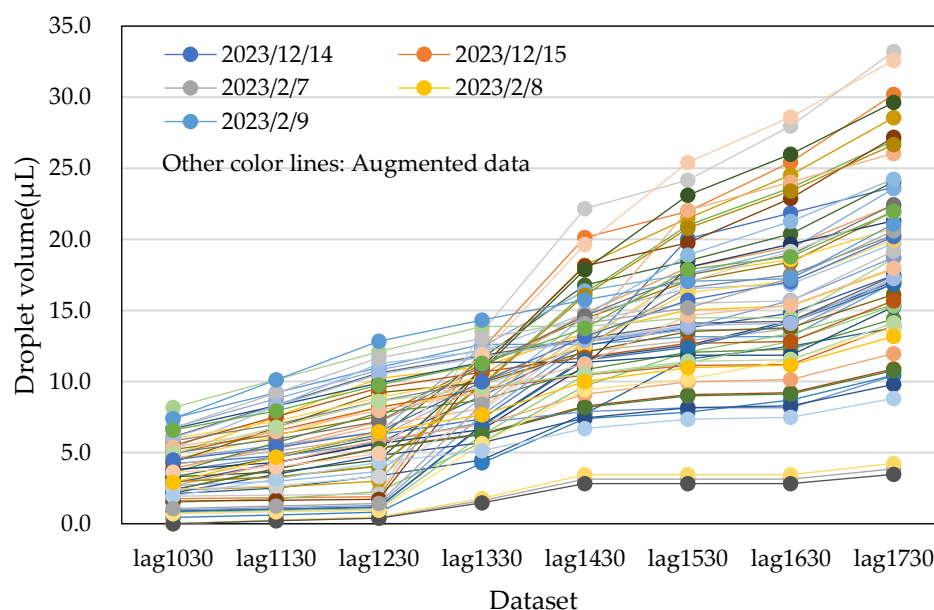


Figure 16. Augmented water droplet volume data from the male restroom using the averaging and scaling method. The total number of datasets is 60, including the five actual datasets.

The second method for data augmentation used random number generation. First, the maximum and minimum values from each hourly dataset were identified. The maximum value was multiplied by 1.1, and the minimum value by 0.9. Random numbers were then generated between these two values. This generated 55 random datasets, which, when combined with the five original datasets, resulted in a total of 60 datasets. To ensure consistency, the data were sorted in ascending order, as water droplet volumes could not decrease over time. Table 3 and Figure 17 display the augmented datasets generated by

this method. Temperature and humidity were not incorporated, as the random number generation process is not time- or date-dependent.

Table 3. Data augmentation results using the random number generation method. Dataset from No. 1 to 5 are actual data and others are augmented datasets.

No.	lag1030	lag1130	lag1230	lag1330	lag1430	lag1530	lag1630	lag1730
1	4.2	4.8	6.3	7.4	12.3	20.0	21.8	23.7
2	1.8	1.8	1.9	12.3	20.1	22.0	25.4	30.2
3	0.0	0.2	0.4	1.6	3.1	3.1	3.1	3.9
4	4.8	7.5	10.2	11.0	13.4	15.0	15.3	17.9
5	7.4	9.2	11.0	12.6	12.6	13.2	13.2	17.0
6	4.5	4.6	5.5	6.0	9.1	10.6	16.5	26.2
7	3.2	3.4	8.2	8.5	11.0	11.7	16.6	29.9
8	4.6	5.1	5.5	5.6	7.0	11.0	12.0	15.2
9	0.3	2.6	3.9	6.4	8.9	12.8	14.8	18.9
10	1.2	3.7	5.9	6.6	8.3	8.3	26.0	26.4
11	0.0	0.5	8.5	8.8	14.1	18.5	21.0	21.9
12	3.2	4.8	5.1	6.8	10.1	12.6	17.7	31.3
13	4.6	9.4	10.4	12.4	12.7	18.4	26.8	27.6
14	1.4	1.5	2.2	9.5	12.0	16.2	16.6	21.2
15	3.4	5.7	9.9	10.9	11.5	13.9	20.6	24.0
16	1.5	4.1	5.0	6.1	9.4	10.9	13.8	17.0
17	5.0	5.5	6.6	7.0	9.5	9.8	17.6	22.9
18	1.9	3.4	4.8	6.0	8.9	10.1	11.0	11.8
19	0.7	3.5	4.1	8.4	10.9	19.8	25.0	26.6
20	0.6	3.4	8.3	8.9	10.1	14.9	15.2	32.9
21	0.9	2.7	3.8	6.8	8.5	10.0	16.6	27.7
22	2.6	3.6	5.6	8.4	10.2	11.0	20.8	32.5
23	3.1	3.8	4.3	7.7	11.2	20.6	21.0	21.4
24	0.5	2.3	3.4	4.0	9.1	11.1	14.8	17.5
25	2.1	2.9	3.9	5.1	6.3	9.2	18.5	18.5
26	2.9	3.6	9.7	11.0	13.2	16.2	20.3	25.7
27	2.3	3.8	4.4	4.4	10.5	15.8	20.7	28.1
28	3.7	5.6	8.9	10.9	17.7	23.2	26.6	28.3
29	3.6	6.6	7.2	8.4	10.3	11.9	13.8	32.1
30	2.6	4.8	8.1	8.3	8.3	12.4	18.9	29.1
31	1.8	5.9	6.1	7.0	10.7	16.2	20.7	20.8
32	4.0	4.2	5.8	6.1	7.1	10.0	11.9	23.0
33	4.1	4.3	6.0	11.7	11.8	14.1	18.1	24.6
34	4.2	4.3	4.9	5.4	7.3	7.6	16.6	21.8
35	0.0	5.1	8.6	10.1	10.8	11.0	17.5	22.0
36	1.7	2.5	2.8	5.5	7.8	10.1	16.6	22.6
37	5.1	5.1	5.4	5.9	9.3	9.6	16.6	19.1
38	1.3	3.6	4.0	4.4	6.0	9.3	9.5	13.2

Table 3. Cont.

No.	lag1030	lag1130	lag1230	lag1330	lag1430	lag1530	lag1630	lag1730
39	1.6	3.7	7.0	8.4	10.1	11.0	12.6	14.3
40	1.0	1.2	5.1	5.9	10.8	13.8	16.3	26.2
41	0.8	1.6	3.2	3.4	3.7	8.2	11.8	11.8
42	4.4	6.6	7.9	11.8	13.1	18.8	20.2	27.1
43	1.2	4.0	6.7	8.8	8.9	17.5	18.1	18.2
44	6.0	6.3	6.6	7.8	11.4	12.7	15.5	17.3
45	4.3	4.4	4.9	6.4	10.5	10.7	20.1	31.3
46	1.4	6.4	7.4	9.8	13.3	19.4	21.5	31.5
47	2.0	2.3	2.5	6.2	7.2	22.0	25.9	27.0
48	3.9	5.6	6.8	10.2	11.3	16.7	18.5	20.2
49	5.9	6.7	7.1	7.3	7.4	7.4	9.6	10.3
50	5.6	6.2	8.9	10.5	12.4	13.3	16.9	20.3
51	5.5	6.0	8.0	8.3	9.9	10.9	20.4	20.6
52	2.9	5.3	7.5	8.3	8.9	13.1	16.9	31.2
53	0.4	6.9	7.4	12.6	13.4	20.2	20.5	25.3
54	1.1	2.7	3.3	9.0	12.7	16.0	16.8	19.0
55	2.0	4.2	5.3	5.7	6.5	6.5	7.2	8.5
56	0.2	2.5	2.9	4.1	7.1	7.7	8.9	15.4
57	1.7	3.2	3.8	5.2	5.2	10.2	14.3	24.0
58	1.0	3.5	4.2	6.1	7.0	7.6	10.0	15.7
59	2.6	5.3	6.7	8.0	8.1	10.8	14.4	27.9
60	0.7	6.3	6.5	6.9	8.7	16.4	22.6	24.7

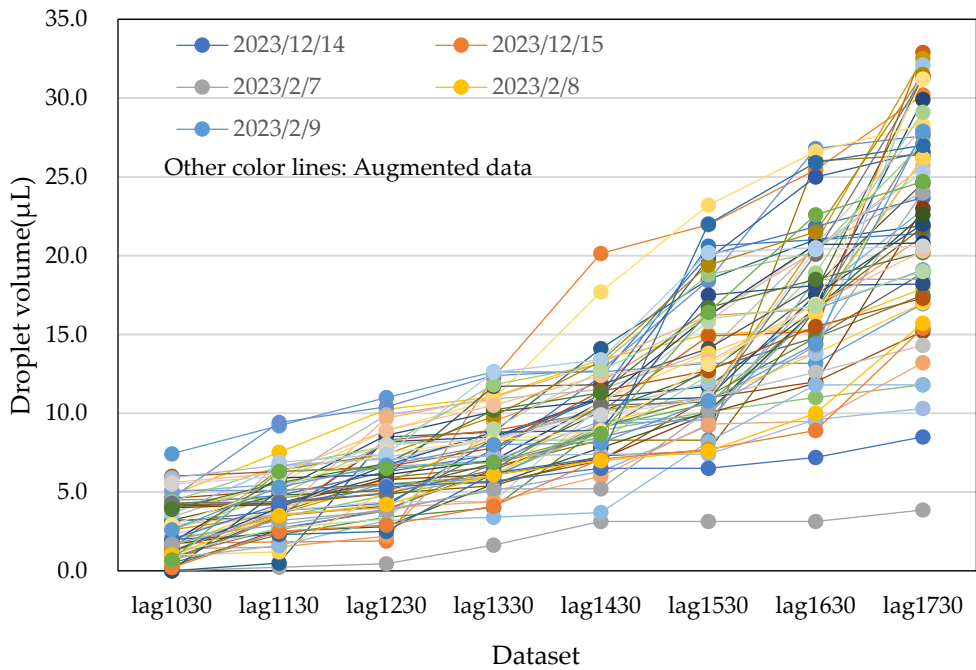


Figure 17. Augmented water droplet volume data from the male restroom using the random number method. The total number of datasets is 60, including five actual datasets.

Ultimately, three augmented datasets were prepared: (1) the averaging and scaling method, including temperature and humidity, (2) the averaging and scaling method without temperature and humidity, and (3) the random number generation method. These datasets were then used to predict water droplet volumes, and the accuracy of the prediction methods was evaluated.

3.4. Water Droplet Volume Prediction

We aimed to predict water droplet volumes at 17:30 based on data collected from 10:30 to 16:30. Specifically, our goal was to predict the water droplet volume for the hour following data collection. Although predicting the volume for 2 or 3 h intervals is more practical in real-world cleaning scenarios, this study focused on constructing and evaluating a model. Therefore, we initially chose to predict the volume for just 1 h after the data collection period.

For our predictions, we used the light gradient boosting machine (LightGBM) method [21–26], which operates within a decision tree framework. Specifically, we employed the DecisionTreeRegressor algorithm. Several other methods are available for time-series predictions, including recurrent neural networks (RNN) [27–29] and long short-term memory (LSTM) networks [30–35]. Both RNNs and LSTMs are based on neural networks: in RNNs, past time-series data are stored in hidden layers, whereas LSTMs store them in additional gates (forget, input, and output gates). These stored values are weighted and added to the input data.

We first evaluated the applicability and validity of the LightGBM method for our prediction task. While RNNs and LSTMs are widely used, they present challenges. RNNs are susceptible to the vanishing gradient problem, which occurs due to the activation function, and the exploding gradient problem, which arises from matrix multiplication. The data we work with are highly variable and depend on individual behaviours, making them inherently random. Although LSTMs attempt to handle this randomness by incorporating weighted past data through their gates, designing and stabilizing the weights for such unpredictable data can be difficult.

Given these challenges, we decided to first assess the LightGBM (version 4.4) framework for our prediction task. In the LightGBM model, the data were split as follows: 70% for training (42 datasets out of 60) and 30% for testing (18 out of 60).

4. Results and Discussion

4.1. Prediction Using Actual Data

First, we present the results of predictions made using only the actual datasets obtained from the in situ experiment. There were five datasets for each sex, but since this number is too small to construct a reliable prediction system with the LightGBM framework, we combined the datasets without considering gender in order to evaluate the accuracy of the prediction system. In this case, the hyperparameters were set as follows: 50% of the data (5 out of 10 datasets) was used for training, while the remaining 50% (5 out of 10) was used for testing. The decision tree depth (*max_depth*) was set to 9, and the *random_seed* was fixed. Figure 18 presents a comparison between the predicted and actual values. The results show an R^2 value of 0.656 and a root mean squared error (RMSE) of 5.137, indicating relatively poor prediction accuracy due to the limited dataset. To improve accuracy, it was clear that augmented datasets, as described in Section 3.3, would be necessary.

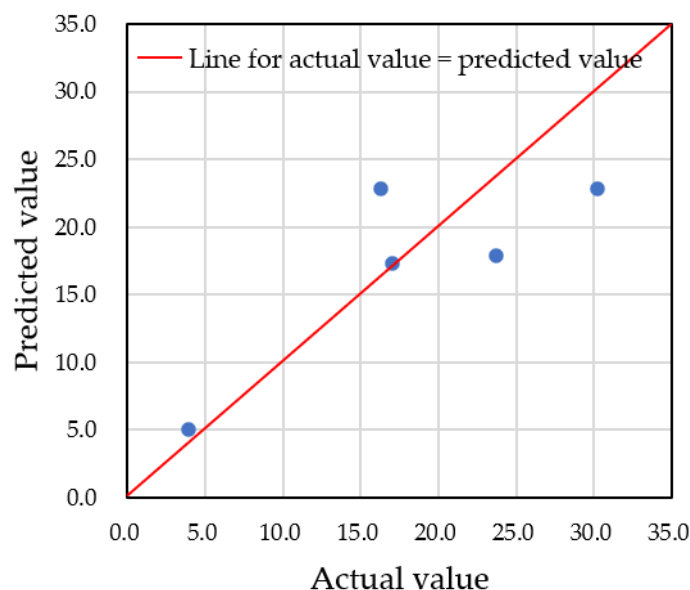


Figure 18. Comparison of predicted values and actual values when only the actual datasets were used for prediction.

4.2. Setting Hyperparameters of LightGBM

For accurate predictions of water droplet volume using augmented datasets, it is critical to carefully set the hyperparameters of the model. In this study, we focused on adjusting the depth of the decision tree, controlled by the ‘*max_depth*’ parameter. If the number of nodes is too small, multiple predicted data points will converge into one node, resulting in inaccurate predictions. Therefore, an adequate number of nodes is needed to properly distribute the data. In our case, we had a total of 60 datasets, with predictions made on 18 testing datasets. To classify each testing dataset into one node, at least five nodes were required. We examined three different *max_depth* values—3, 5, and 9—and compared the predicted values to the actual data. Figures 19–21 show the results for *max_depth* values of 3, 5, and 9, respectively.

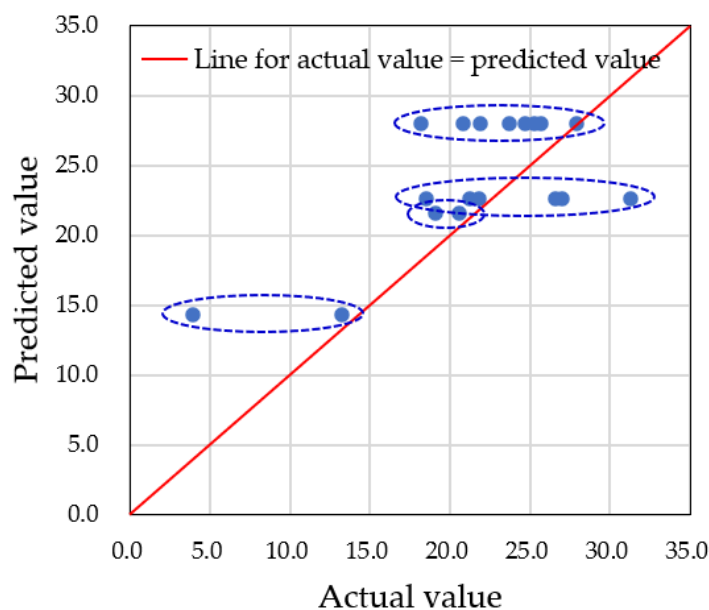


Figure 19. Four multiple predictions were obtained when the *max_depth* was set to 3. Points having the same predicted value for different actual values are grouped with dotted line circle.

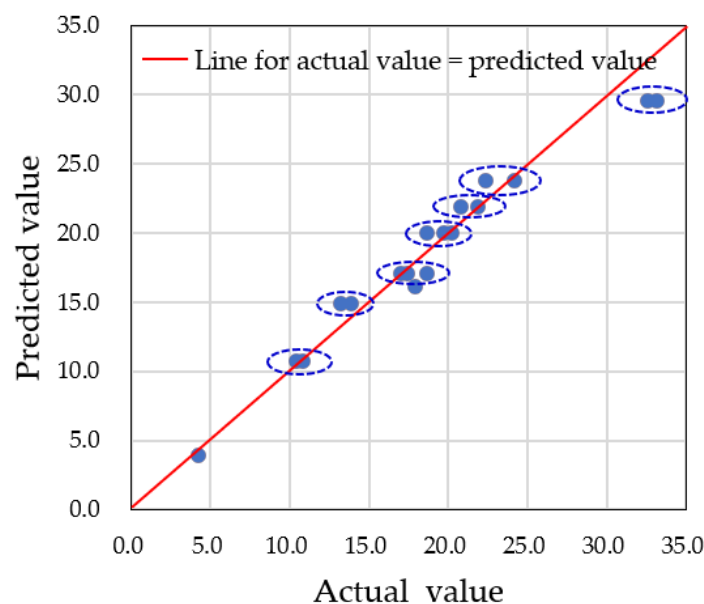


Figure 20. Seven multiple predictions were obtained when the *max_depth* was set to 5. Points having the same predicted value for different actual values are grouped with dotted line circle.

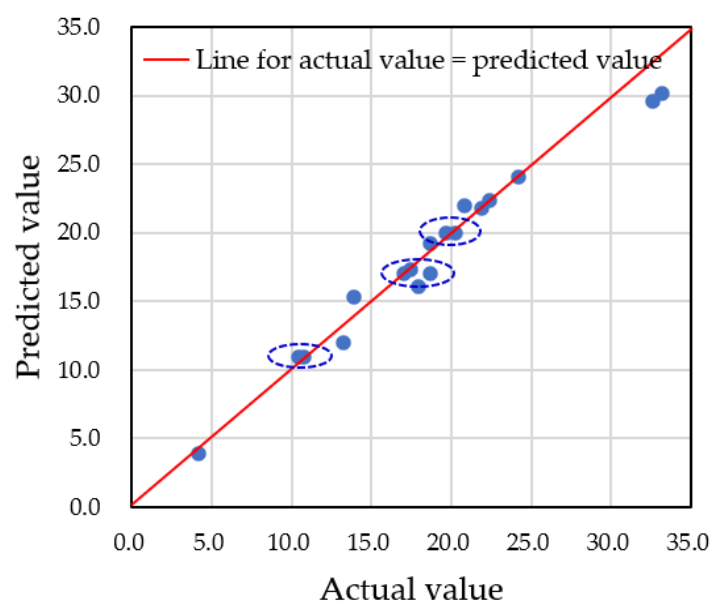
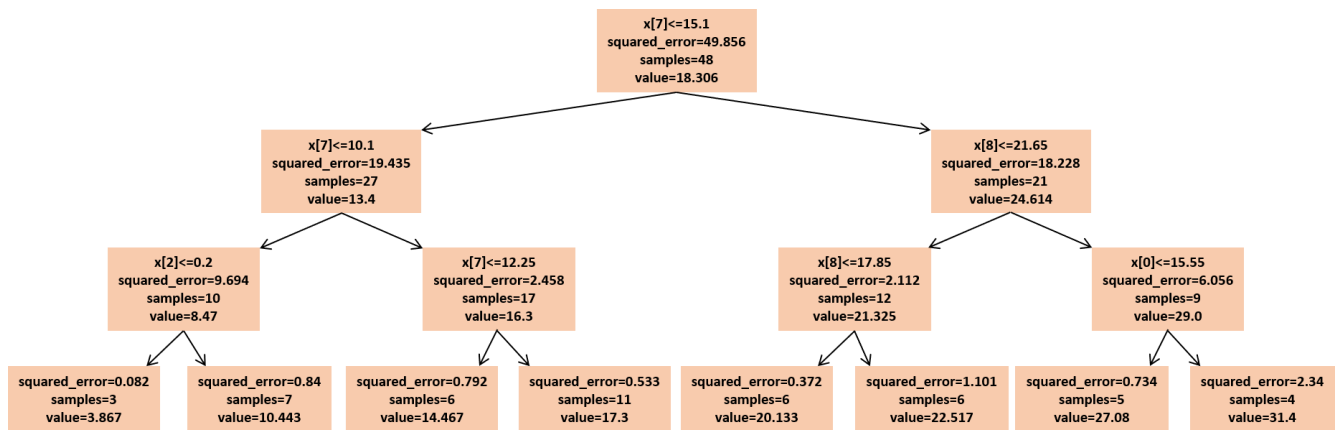
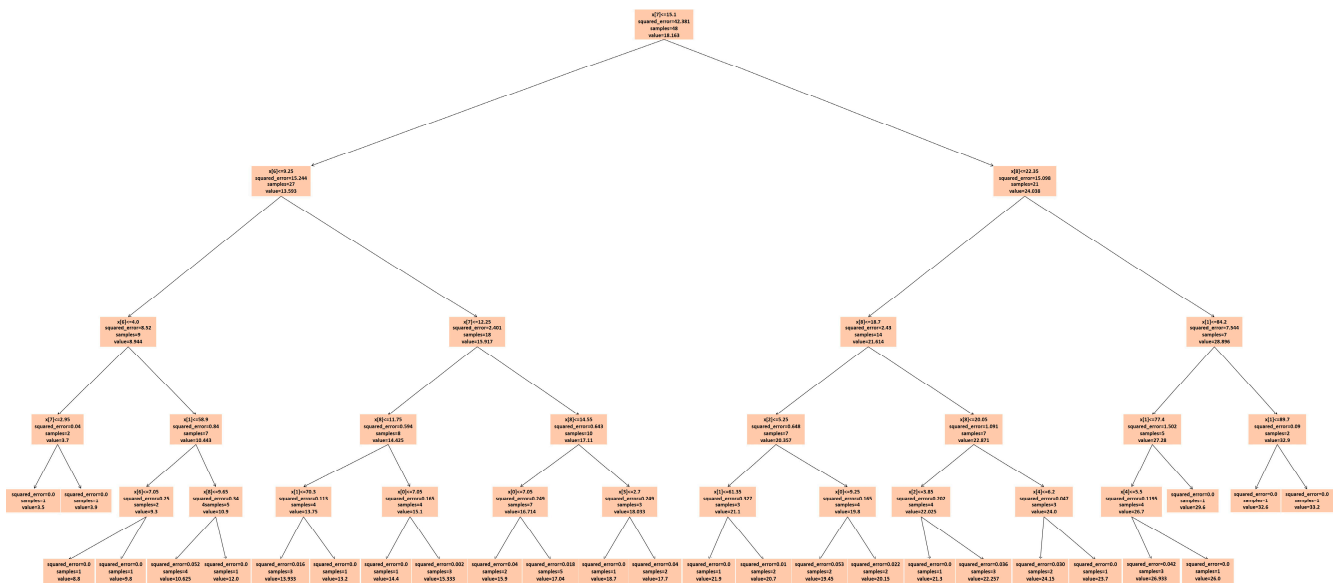
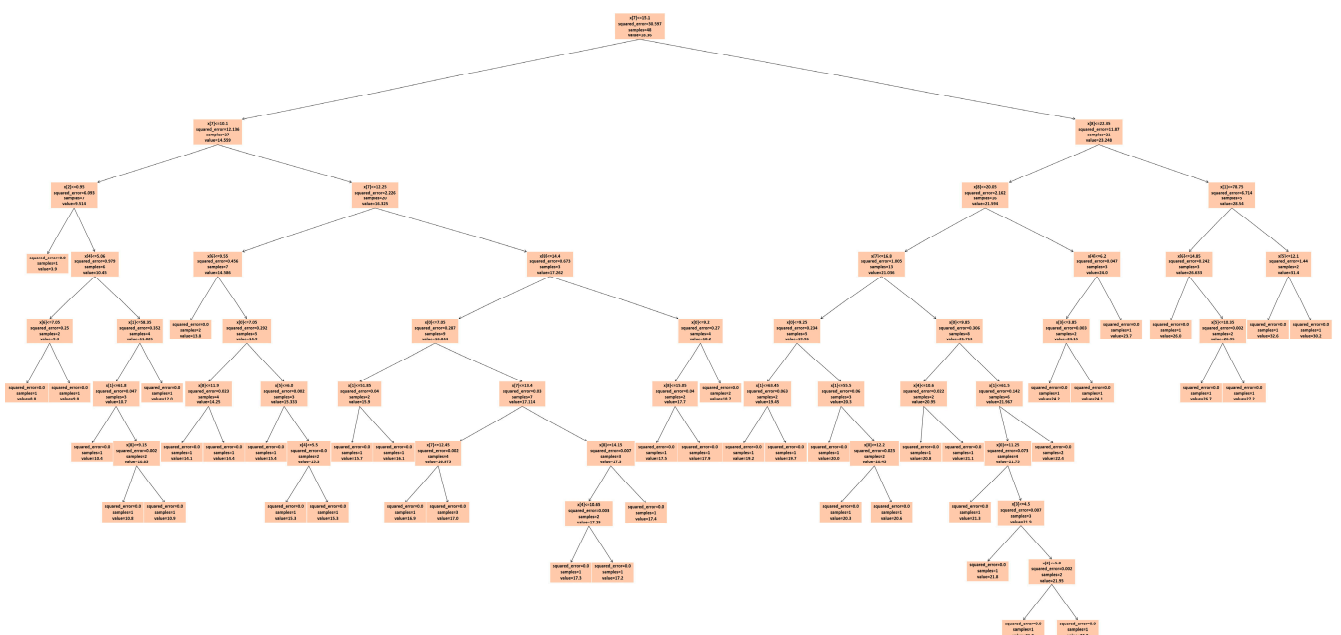


Figure 21. Three multiple predictions were obtained when the *max_depth* was set to 9. Points having the same predicted value for different actual values are grouped with dotted line circle.

For *max_depth* = 3 (Figure 19), the model grouped multiple actual data points into a single predicted data point, resulting in four incorrect predictions. The decision tree diagram for this case, with only three nodes, is shown in Figure 22. When *max_depth* was increased to 5 (Figure 20), seven incorrect predictions occurred, indicating insufficient classification. With *max_depth* = 9 (Figure 21), only three incorrect predictions were made, with a more suitable distribution of nodes. Decision tree diagrams for *max_depth* values of 5 and 9 are shown in Figures 23 and 24, respectively. Based on these results, we determined that a *max_depth* of 9 provided the most accurate predictions and should be adopted depending on the number of test cases.

Figure 22. Decision tree diagram for $max_depth = 3$.Figure 23. Decision tree diagram for $max_depth = 5$.Figure 24. Decision tree diagram for $max_depth = 9$.

4.3. Differences in Prediction Accuracy with Augmented Data

The use of only five real-world data points was insufficient for accurate predictions. Additionally, continuous data collection beyond these five days was not feasible due to privacy concerns in the restroom facilities. To address this limitation, we augmented the data using two methods, as described earlier. Figures 25–27 display the prediction results using augmented datasets generated by the averaging and scaling method (including temperature and humidity), the averaging and scaling method (without temperature and humidity), and the random number generation method, respectively. In these cases, the *random_seed* was fixed to ensure consistency in the training and testing data across all methods. Similarly, *max_depth* was fixed at 9 for all comparisons.

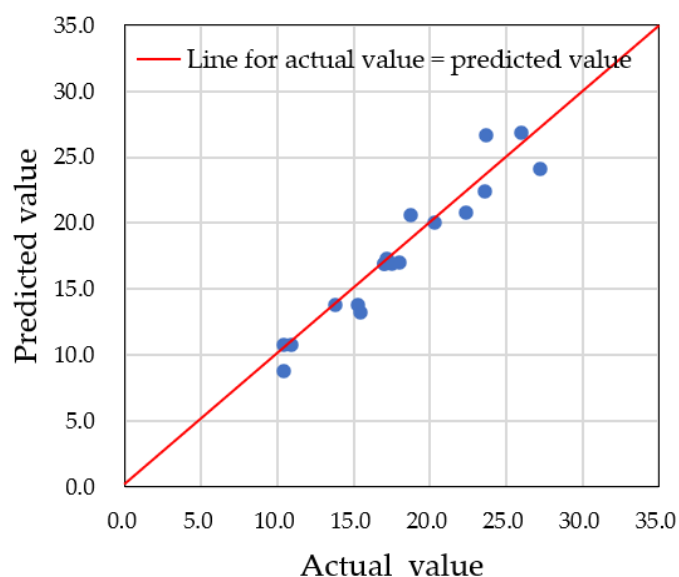


Figure 25. Comparison of predicted values and actual values using augmented datasets generated by the averaging and scaling method, including temperature and humidity.

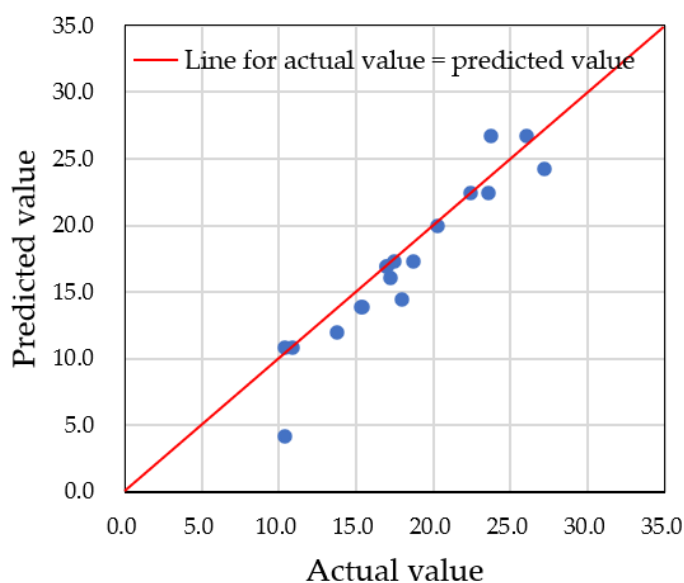


Figure 26. Comparison of predicted values and actual values using augmented datasets generated by the averaging and scaling method, excluding temperature and humidity.

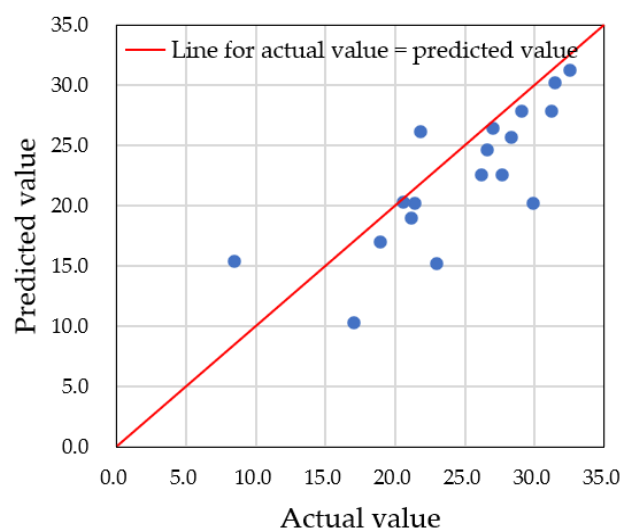


Figure 27. Comparison of predicted values and actual values using augmented datasets generated by the random number generation method.

Using the augmented datasets generated by the averaging and scaling method (including temperature and humidity), the model achieved an R^2 value of 0.913 and an RMSE of 1.459 (Figure 25). When the augmented datasets did not include temperature and humidity data, the R^2 value decreased to 0.812, with an RMSE of 2.138 (Figure 26). Finally, using the random number generation method resulted in an R^2 of 0.462 and an RMSE of 4.358, as shown in Figure 27. From Figures 25 and 26, we observed that omitting temperature and humidity slightly decreased prediction accuracy, suggesting that these weather variables are important for more precise predictions. However, for a simplified prediction system, omitting these variables may be acceptable, as their influence on water droplet volume is relatively minor, as shown in Figure 28. This figure compares actual water droplet data with temperature and humidity measurements taken from a male restroom. Figure 29 further illustrates the contribution of various feature variables to the prediction model, showing that temperature and humidity are less significant contributors.

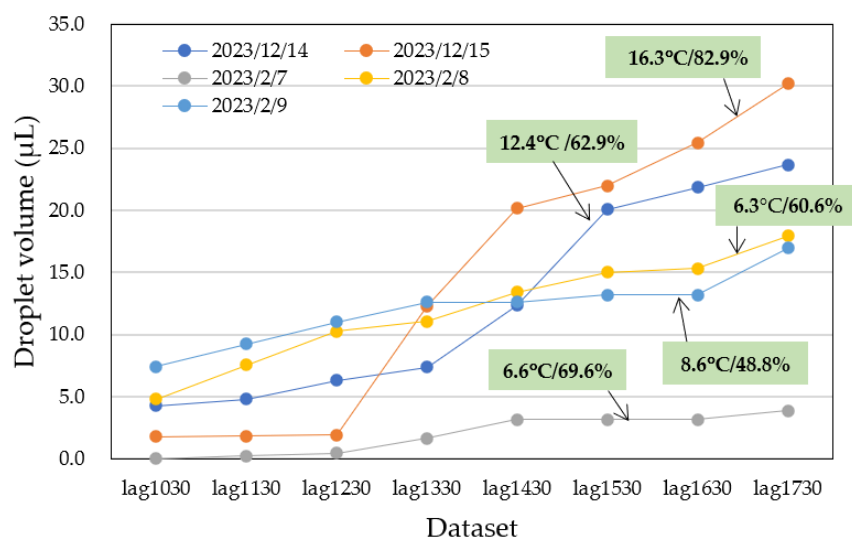


Figure 28. Actual water droplet volume data collected from the male restroom, along with temperature and humidity measurements.

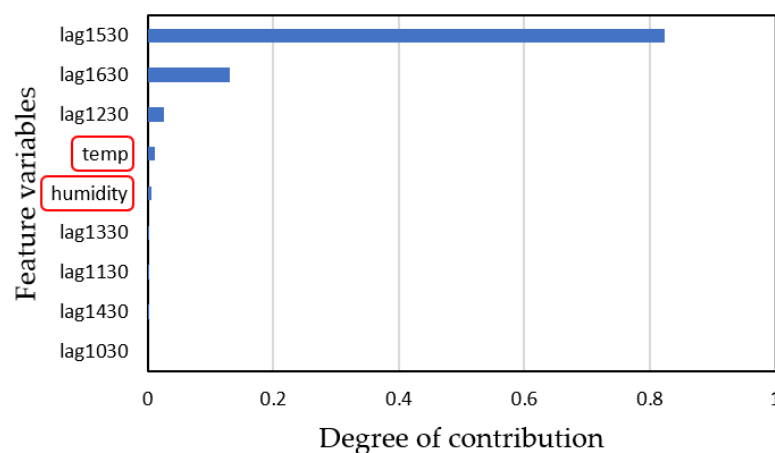


Figure 29. Contribution of feature variables in predicting water droplet volume. Temperature and humidity red boxed in are less significant contributors among feature variables.

The prediction using the random number generation method resulted in unnatural data, as seen in Figure 27, and produced the least accurate predictions. Therefore, the random number generation method is not suitable for this application.

In this study, we used the DecisionTreeRegressor framework within the LightGBM method for predicting water droplet volumes. LightGBM, which utilises decision trees, is well-suited for time-series data and is commonly used for regression tasks, such as the one in this study.

The main focus of our prediction system was to determine when water droplet volumes would exceed a predefined threshold, indicating maximum dirtiness. This system quantifies dirtiness using water droplet volumes around the washbowl and is designed for practical use in cleaning environments. However, due to the limited amount of real-world data, data augmentation was essential to improve prediction accuracy. Ultimately, our augmented dataset-based model allowed for reliable predictions of water droplet volumes at specific times based on earlier measurements in restrooms.

5. Conclusions

In this study, we predicted restroom cleanliness by evaluating the volume of water droplets outside washbowls as an indicator of dirtiness. Using a near-infrared camera, we captured photographs of the water droplets, converted the image data into volume measurements, and accumulated the data over a 1 h period. These measurements were collected over five days through experiments, but the limited amount of data was insufficient to develop an effective learning and prediction algorithm. To address this, we augmented the dataset using two methods: the first method involved averaging and scaling the data, while the second used random number generation. The random number method was found unsuitable, as it caused the water droplet data to fluctuate unpredictably. In contrast, the averaging and scaling method proved effective for our application.

The LightGBM-based prediction system, with optimised hyperparameters and data augmentation, is nearly practical for real-world use. Specifically, this methodology is suitable for restrooms with limited user traffic, such as those in offices and schools. By predicting future dirtiness levels, the system can determine optimal cleaning times or adjust cleaning frequency, ultimately leading to more efficient and future-oriented cleaning practices.

6. Future Study

Our system successfully predicts water droplet volume 1 h ahead, but it could also predict 2 or 3 h future volumes. The process is sequential: we first predicted the 1 h ahead droplet volume, used this as actual data, and then predicted the 2 h ahead volume, followed by 3 h prediction. However, the accuracy of predictions for the 2 and 3 h intervals is expected to be lower than for the 1 h prediction. To improve the accuracy of longer-term predictions, additional data collection is necessary.

Currently, our prediction system relies on photographs of water droplets, which may not be practical in all real-world scenarios. Future work will explore methods that eliminate the need for photographs, making the system more feasible for actual cleaning operations.

Although the LightGBM method treats data as an array, not a time-series, other techniques could be integrated into our system. LSTM networks, which are designed to handle time-series data and retain past information, could enhance the prediction accuracy. Furthermore, the GRU [36,37] offers a simpler and computationally less intensive alternative to LSTM. RNNs are another promising approach. We plan to evaluate these three methods—LSTM, GRU, and RNN—to determine the most optimal solution for our application.

Author Contributions: Conceptualization, S.-S.L.; Methodology, S.K., H.M. and S.-S.L.; Formal analysis, S.K. and S.-S.L.; Investigation, S.K., H.M., T.M. and S.-S.L.; Writing—original draft, S.K.; Writing—review & editing, S.-S.L.; Supervision, S.-S.L.; Project administration, S.-S.L.; Funding acquisition, S.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Hironori Moriwaki was employed by the company Asuka Bisou Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Kim, H.; Bachman, J.R. Examining Customer Perceptions of Restaurant Restroom Cleanliness and Their Impact on Satisfaction and Intent to Return. *J. Foodserv. Bus. Res.* **2019**, *22*, 191–208. [CrossRef]
- Afacan, Y.; Gurel, M.O. Public Toilets: An Exploratory Study on the Demands, Needs, and Expectations in Turkey. *Environ. Plan. B Plan. Des.* **2015**, *42*, 242–262. [CrossRef]
- Hartigan, S.M.; Bonnet, K.; Chisholm, L.; Kowalik, C.; Dmochowski, R.R.; Schlundt, D.; Reynolds, W.S. Why Do Women Not Use the Bathroom? Women's Attitudes and Beliefs on Using Public Restrooms. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2053. [CrossRef] [PubMed]
- Jayasinghe, L.; Wijerathne, N.; Yuen, C.; Zhang, M. Feature Learning and Analysis for Cleanliness Classification in Restrooms. *IEEE Access* **2019**, *7*, 14871–14882. [CrossRef]
- Lewkowitz, S.; Gilliland, J. A Feminist Critical Analysis of Public Toilets and Gender: A Systematic Review. *Urban Aff. Rev.* **2024**, *61*, 282–309. [CrossRef]
- TOTO LTD. Customer Awareness Survey of Restaurant Restrooms. Available online: <https://jp.toto.com/products/machinaka/restaurantquestionnaires/> (accessed on 8 December 2024).
- Aica Kogyo Company, Limited. Survey on Awareness About Comfort of Restrooms. Available online: https://www.aica.co.jp/news/detail/post_143.html (accessed on 8 December 2024).
- Lokman, A.; Ramasamy, R.K.; Ting, C.Y. Scheduling and Predictive Maintenance for Smart Toilet. *IEEE Access* **2023**, *11*, 17983–17999. [CrossRef]
- Raendran, V.; Ramasamy, R.K.; Rosdi, I.S.; Ab Razak, R.; Fauzi, N.M. IoT Technology for Facilities Management: Understanding end User Perception of the Smart Toilet. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 353–359. [CrossRef]

10. Vision Sensing Co., Ltd. InGaAs Near Infrared Camera NIR640SN. Available online: https://www.vision-sensing.jp/en/catalog/catalog_all_en.pdf (accessed on 8 December 2024).
11. Forestier, G.; Petitjean, F.; Dau, H.A.; Webb, G.I.; Keogh, E. Generating Synthetic Time Series to Augment Sparse Datasets. In Proceedings of the IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 17–21 November 2017; pp. 865–870. [\[CrossRef\]](#)
12. Morel, M.; Achard, C.; Kulpa, R.; Dubuisson, S. Time-Series Averaging Using Constrained Dynamic Time Warping Tolerance. *Pattern Recognit.* **2018**, *74*, 77–89. [\[CrossRef\]](#)
13. Petitjean, F.; Gançarski, K. A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering. *Pattern Recognit.* **2011**, *44*, 678–693. [\[CrossRef\]](#)
14. Künsch, H.R. The Jackknife and the Bootstrap for General Stationary Observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [\[CrossRef\]](#)
15. Vogel, R.M.; Shallcross, A.L. The Moving Blocks Bootstrap Versus Parametric Time Series Models. *Water Resour. Res.* **1996**, *32*, 1875–1882. [\[CrossRef\]](#)
16. Hidalgo, J. An Alternative Bootstrap to Moving Blocks for Time Series Regression Models. *J. Econom.* **2003**, *117*, 369–399. [\[CrossRef\]](#)
17. Radovanov, B.; Marcikić, A. A Comparison of Four Different Block Bootstrap Methods. *Croat. Oper. Res. Rev.* **2014**, *5*, 189–202. [\[CrossRef\]](#)
18. Lin, Z.; Jain, A.; Wang, C.; Fanti, G.; Sekar, V. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In Proceedings of the IMC '20: Proceedings of the ACM Internet Measurement Conference, Virtual, 27–29 October 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 464–483. [\[CrossRef\]](#)
19. Brophy, E.; Wang, Z.; She, Q.; Ward, T. Generative Adversarial Networks in Time Series: A Systematic Literature Review. *ACM Comput. Surv.* **2023**, *55*, 1–31. [\[CrossRef\]](#)
20. Lin, Z.; Jain, A.; Wang, C.; Fanti, G.; Sekar, V. Generating High-Fidelity, Synthetic Time Series Datasets with Doppelganger. *arXiv* **2019**, arXiv:1909.13403.
21. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1–9.
22. Yan, J.; Xu, Y.; Cheng, Q.; Jiang, S.; Wang, Q.; Xiao, Y.; Ma, C.; Yan, J.; Wang, X. LightGBM: Accelerated Genomically Designed Crop Breeding through Ensemble Learning. *Genome Biol.* **2021**, *22*, 1–24. [\[CrossRef\]](#)
23. Cao, Q.; Wu, Y.; Yang, J.; Yin, J. Greenhouse Temperature Prediction Based on Time-Series Features and LightGBM. *Appl. Sci.* **2023**, *13*, 1610. [\[CrossRef\]](#)
24. Khan, A.A.; Chaudhari, O.; Chandra, R. A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778. [\[CrossRef\]](#)
25. Zhang, D.; Gong, Y. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. *IEEE Access* **2020**, *8*, 220990–221003. [\[CrossRef\]](#)
26. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
27. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [\[CrossRef\]](#)
28. Amalou, I.; Mouhni, N.; Abdali, A. Multivariate Time Series Prediction by RNN Architectures for Energy Consumption Forecasting. *Energy Rep.* **2022**, *8*, 1084–1091. [\[CrossRef\]](#)
29. Selvin, S.; Vinayakumar, R.; Gopalakrishnan, E.A.; Menon, V.K.; Soman, K.P. Stock Price Prediction Using LSTM, RNN and CNN-Sliding Window Model. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Udipi, India, 13–16 September 2017; pp. 1643–1647. [\[CrossRef\]](#)
30. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[PubMed\]](#)
31. Van Houdt, G.; Mosquera, C.; Nápoles, G. A Review on the Long Short-Term Memory Model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955.
32. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM—A Tutorial into Long Short-Term Memory Recurrent Neural Networks. *arXiv* **2019**, arXiv:1909.09586.
33. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [\[CrossRef\]](#)
34. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* **2000**, *12*, 2451–2471. [\[CrossRef\]](#)
35. Nelson, D.M.; Pereira, A.C.; De Oliveira, R.A. Stock Market's Price Movement Prediction with LSTM Neural Networks. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 1419–1426.

36. Dey, R.; Salem, F.M. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks. In Proceedings of the IEEE 60th International Midwest Symposium on Circuits and Systems, Medford, MA, USA, 6–9 August 2017; pp. 1597–1600. [\[CrossRef\]](#)
37. Fu, R.; Zhang, Z.; Li, L. Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction. In Proceedings of the 31st Youth Academic Annual Conference of Chinese Association of Automation, Wuhan, China, 1–13 November 2016; pp. 324–328.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.