ISSN 2442-6571

A comparison of machine learning methods for knowledge extraction model in a LoRa-based waste bin monitoring system

Aa Zezen Zaenal Abidin [a,b,1,*], Mohd Fairuz Iskandar Othman [b,2], Aslinda Hassan [b,3] , Yuli Murdianingsih [a,4], Usep Tatang Suryadi [a,5], Timbo Faritchan Siallagan [a,6]

[a] Informatics Departement, Universitas Mandiri, Jl.Marsinu no 5 Subang, Subang 41211 Jawa Barat, Indonesia
[b] Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[1] zezen@universitasmandiri.ac.id; [2] mohdfairuz@utem.edu.my; [3] aslindahassan@utem.edu.my; [4] yuli@universitasmandiri.ac.id; [5]usep@universitasmandiri.ac.id; [6] timbosiallagan@universitasmandiri.ac.id
* corresponding author

ARTICLE INFO

Article history
Received February 13, 2023
Revised July 19, 2023
Accepted September 30, 2023
Available online February 29, 2024

Keywords
IoT
Knowledge extraction model
LoRa
Machine Learning
Waste management

ABSTRACT

Knowledge Extraction Model (KEM) is a system that extracts knowledge through an IoT-based smart waste bin emptying scheduling classification. Classification is a difficult problem and requires an efficient classification method. This research contributes in the form of the KEM system in the classification of scheduling for emptying waste bins with the best performance of the Machine Learning method. The research aims to compare the performance of Machine Learning methods in the form of Decision Tree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, and Multi-Layer Perceptron, which will be recommended in the KEM system. Performance testing was performed on accuracy, recall, precision, F-Measure, and ROCS curves using the cross-validation method with ten observations. The experimental results show that the Decision Tree performs best for accuracy, recall, precision, and ROCS curve. In contrast, the K-NN method obtains the highest F-measure performance. KEM can be implemented to extract knowledge from data sets created in various other IoT-based systems.

1. Introduction

The Knowledge Extraction Model (KEM) is a system designed to monitor IoT-based waste bins using LoRa network media, facilitating sorting organic, inorganic, and metal waste. KEM enables the extraction of insights from waste management data to schedule waste disposal from intelligent bins efficiently. Constructed using an IoT-based framework utilizing LoRa network media, KEM employs Machine Learning techniques for data analysis, including Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) methods. KEM aims to optimize waste management in rural areas, mainly by enhancing knowledge extraction to schedule the emptying of smart waste bin points.

Waste continues to increase daily, per the population growth rate [1]–[4]. Waste is a problem in many countries. Therefore, a solution is required for collecting, sorting, and disposing waste [2]. Waste is not only a problem in urban areas but also in rural areas [5], [6]. Even rural communities tend to have poor, dangerous waste disposal practices such as burning waste and littering [7]. Internet of Things (IoT) technology based on LoRa network media could be an alternative solution to optimize waste management [6]. IoT using LoRa network media in a waste monitoring system and system modeling using the Machine Learning method is highly recommended [8], [9]. IoT has become an enabling technology in waste management, especially in distributing waste bins and distribution of waste carriers [10]. Adopting IoT technology with Machine Learning to properly dispose of waste [11]. Even Machine Learning methods in waste data management provide better classification accuracy [12].

The main task in Machine Learning is data classification, where classifying data is complex because it requires properly prepared data and efficient classification methods [13]. Khatiwada *et al.* [12] discuss the Machine Learning method to improve the accuracy of data classification. Several Machine Learning methods, such as DT, NB, SVM, K-NN, and MLP, are compared to obtain the best method for determining the best model in the Intrusion Detection System [14]. Even a combined Machine Learning (hybrid) method is used to clean data from damaged data, detect anomalies, and clean data [15].

Kontokosta *et al.* [16] involved many waste management data in residential areas. There are four critical variables in urban waste management research [17]. Monitoring the volume of waste in waste bins in the campus area [18], recently carried out for similar waste piles, Yusoff *et al.* [19] conducted research on smart waste bins, with one of its functions being waste monitoring. Waste data analysis in bins for retrieval optimization, using past data or historical data and viewing graphs on the cloud interface [20]. Prediction of the fullness of waste bins and evaluation techniques were carried out to optimize waste disposal [21]. In López *et al.* [22], several parameters were used, such as the type of truck, the location of the waste bin, travel time, and the number of employees. The decision to operate a smart waste bin uses rules but does not use past data or history to obtain rules [23].

IoT as an enabling technology has been used in previous works [1]–[3],[4]. It uses Machine Learning in data analysis but does not explicitly mention the method used and does not choose the method that performs best for the model. Also, Studies on IoT and machine learning approaches have been implemented in previous works [16]–[21], [24], and [22]. Furthermore, Binbusayyis *et al.* [14] compared machine learning methods for IoMT (Internet of Medical Things) system security. However, these studies did not address new knowledge generated from past data.

This research aims to compare Decision Trees (DT), Naive Bayes (NB), k-Nearest Neighbors (K-NN), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP) methods to determine the most effective Machine Learning approach for the Knowledge Extraction Method (KEM) system. The comparison is based on various performance metrics, including Accuracy, Recall, Precision, F-measure, and ROC curve. The KEM system is specifically applied to schedule waste disposal activities from IoT-based smart waste bins utilizing LoRa network media. The contribution of this research lies in two main aspects: firstly, identifying the method with the highest performance in extracting knowledge from IoT-generated data, particularly in the context of scheduling waste disposal from smart bins employing LoRa network technology. Secondly, developing a KEM system that is adept at extracting knowledge from IoT-based systems using Machine Learning reasoning methods, thereby advancing the capabilities of such systems in decision-making processes.

This paper contains Section 1, presenting an introduction and related works; Section 2, describing the KEM Design Architecture; Section 3, describing research methodology; Section 4, presenting the results and a discussion of the research activity; and Section 5, enclosing conclusions and future research steps.

2. Method

This research aims to compare Decision Trees (DT), Naive Bayes (NB), k-Nearest Neighbors (K-NN), Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP) methods to establish a waste disposal scheduling model using data from smart waste bins in rural areas. The methodology, illustrated in Fig. 1, involves seven steps: 1) Defining the extraction environment; 2) Acquiring waste volume data from smart waste bins; 3) Creating the dataset; 4) Designing the model; 5) Training the system model; 6) Conducting comparative testing of machine learning methods; and 7) Extracting knowledge in the form of scheduling rules for waste disposal from smart waste bins in rural regions.

Waste volume data in quantitative form from monitoring activities of smart waste bins and disposal decision data by officers are converted into historical data. Historical data in quantitative form from the waste bin level organic, in organic and metal parameters is made into a dataset. Model design, model training, and KEM model testing were done by comparing the method's performance using a dataset of waste disposal activities in rural areas. The performance of the Knowledge Extraction Model (KEM) model is compared to obtain the best method in KEM. The compared methods are DT, NB, K-NN, SVM, and MLP. The dataset used is data on the volume of waste in smart bins in the form of quantitative data with values ranging from 3 to 26 units high, and the decision to take waste from smart bins by garbage workers in the form of waste disposal and waste did not dispose as shown in Fig. 1.

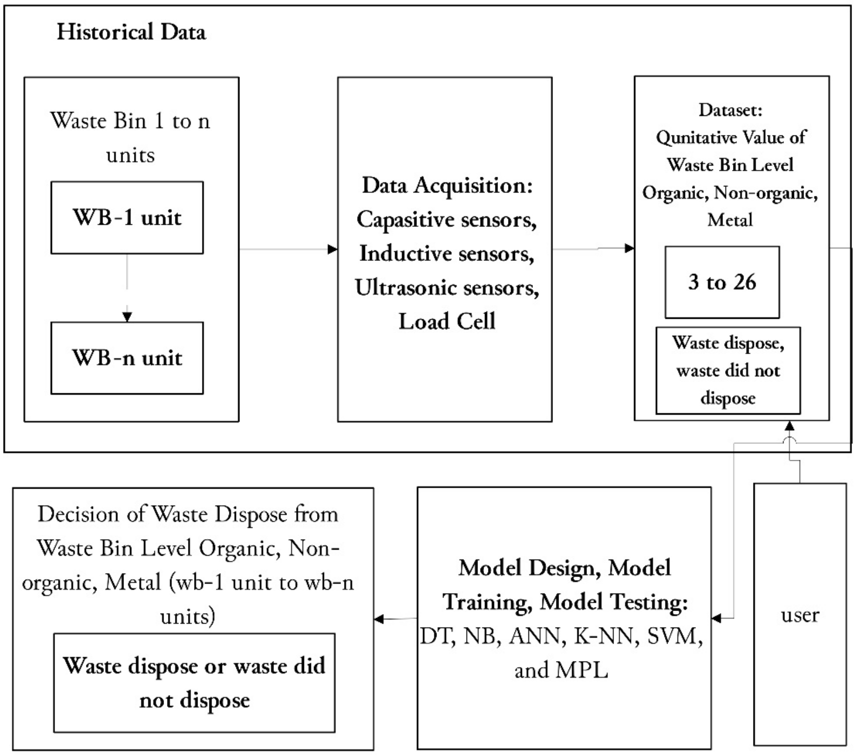

Fig. 1. Scheme of research methodology

2.1. Decision Tree Method

The decision tree (DT) method is used because it is reliable and can extract knowledge. DT is better than other methods of extracting knowledge. DT can generate rules and predict correctly [25]. In the DT method, we will perform two calculations, namely the Entropy and Gain quantities [26]. Deciding to collect garbage in this study was to do iterative calculations of Entropy and Gain on past data, namely the decision history of emptying garbage from smart bins by officers. In Bakhshi and Ahmed [27], the extraction of disease categorization, although not based on IoT, is the best categorical extraction from past data using the DT method. The Entropy using (1) and Gain formula using (2).

$$Entropy\ Total\ (S) = \sum_{i=1}^{n} -Pi \times log_2\ Pi \tag{1}$$

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{S} \times Entropy\ (Si) \tag{2}$$

2.2. Naïve Bayes Method

Based on its function in this session, the Naïve Bayes method is used for numerical data using the Gaussian distribution. The Gaussian distribution is the last step to obtain results from training data or a test data model by taking opportunities from training data. NB and DT in information extraction activities from past data are the most widely used [27], although DT generally performs better than NB. The maximum probability value obtained leads to one of the classes in the dataset for the class of a data record that does not yet have a value. Naïve Bayes Gaussian for numerical data is presented using (3).

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

2.3. K-NN Method

K-NN is a machine learning method that is carried out by calculating similarity. The similarity value is obtained by looking at the closest distance to the previously existing class value. Representation of knowledge extracted from banking data using K-NN [28], even in the NER system, K-NN is the best method for knowledge extraction [29]. The distance determination formula in K-NN using (4).

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2} \tag{4}$$

2.4. SVM Method

SVM, as a linear classifier, is a classification case that can be linearly separated, but SVM can also be used in non-linear problems. SVM can work on non-linear data by using a kernel approach to the initial data features of the data set. Research on factory activities [30] to extract knowledge in the form of sound. Even for precision performance, SVM is best for knowledge extraction activities in the form of tomato fruit textures [31]. Optimizing problem-solving with SVM using (5).

$$\min(w, b)\{\frac{1}{2}\|w\|^2 + C\sum_i \check{c}_i \tag{5}$$

2.5. MLP Method

MLP is the steps in reducing the error function defined on the learning set (x_i, d_i). Knowledge-enhanced Medical Relations Extraction (KemRE) for medical guidance uses the MLP method [32]. The formula follows.

Step 1: The X_i signal is received by each input $(X_i,\ i = 1, 2, 3, \dots, n)$, which then sends it to all of the neurons in the hidden layer.

Step 2: All of the weighted input signals are added together by each hidden layer neuron $(Z_j, j = 1, 2, 3, \dots, n)$, using (6).

$$z_{in-j} = \mu_{0j} + \sum_{i=1}^{n} x_i \mu_{ij} \tag{6}$$

The following equation can be used to determine how to apply the activation function to the output signal, using (7).

$$z_j = f(z_{in-j}) \tag{7}$$

All of the neurons in the output layer are then supplied the output signal that was applied to the activation function.

Step 3: Each neuron in the output layer $(Y_k, k = 1, 2, 3, \dots, m)$ sums up all the received weighted input signals using (8).

$$y_{in-k} = w_{0k} \sum_{j=1}^{k} z_j\, w_{jk} \tag{8}$$

Applying the activation function, which the equation can calculate, to the output signal, using (9)

$$y_k = f(y_{in-k}) \tag{9}$$

2.5. Define The Environment in which The Knowledge will be Extracted

At this stage, as presented in Fig. 2, we describe all the objects in the environment involved in the desired knowledge. As in Fig. 2, a waste object occupies one geographical point. The rural research area is marked on the Google Earth application. Waste bins along the streets are colored blue. Open areas for illegal waste dumping are marked with yellow dots. Open dumping areas are colored red.

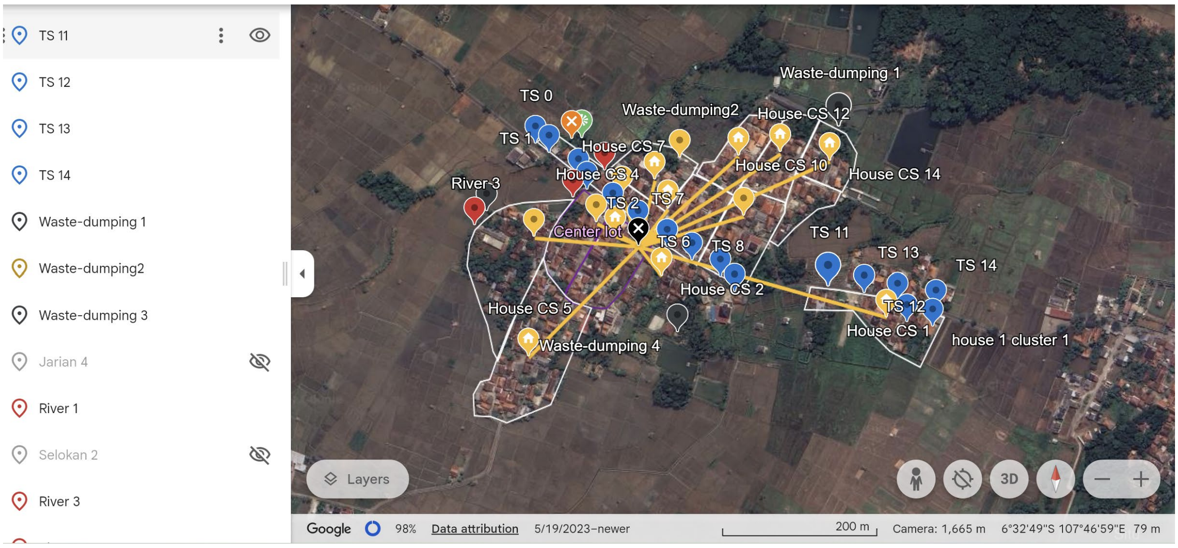

Fig. 2. Research location area [5]

The work area of the system heavily influences the KEM architecture. In this context, KEM is applied in a scheduling system for waste disposal from IoT-based smart bins using LoRa network media and Machine Learning methods. The system consists of an environment and acquisition system, a computer network system, and applications, as shown in Fig. 3. The environment and data acquisition system consists of a geographic environment, a waste bin with various sensors needed by the system, and a microcontroller. The computer network system consists of LoRa network media and internet network media. The application system comprises an IoT platform, a spreadsheet application, and a web browser interface. The slave section contains the sensors needed for waste sorting. On the master side, it is intended to identify users and send waste data via LoRa network media to cloud servers. Data on organic, inorganic, and metal waste volume will be collected on the cloud server from the Things board platform. Past monitoring and waste disposal activity data were obtained in MS Excel as time series data. Past data on waste volume can be saved to a local server and then analyzed to generate new knowledge for decision-makers to schedule waste disposal from smart bins.

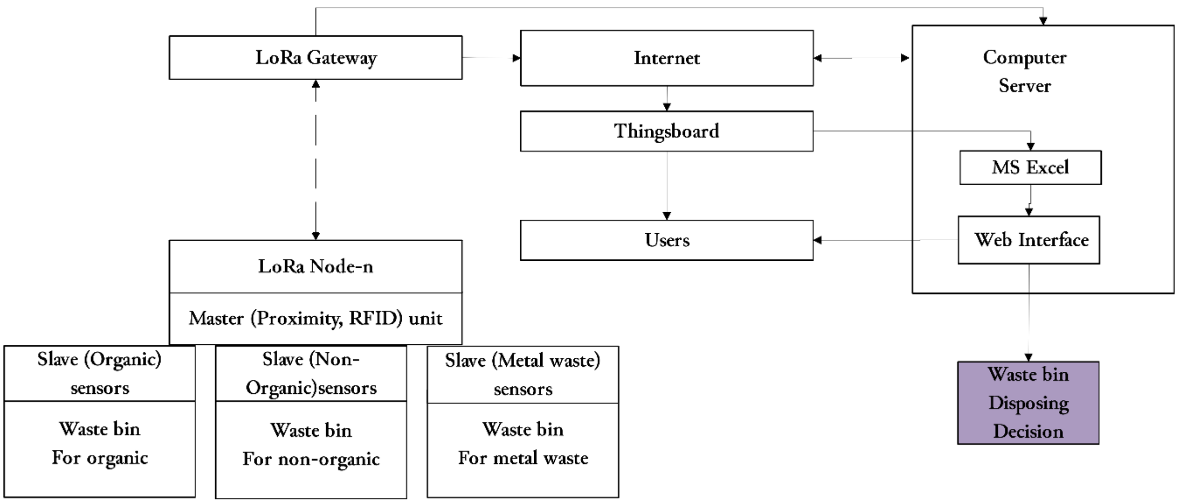

Fig. 3. KEM design architecture

2.6. Perform Waste Volume Data Acquisition from Smart Waste Bins

Monitoring the volume of waste in smart waste bins in this study is just one of the other activities, such as monitoring the weight and gas content of the waste in smart waste bins, as shown in Fig. 4 Monitoring waste bins using Ultrasonic sensors. Ultrasound sensors are placed on each lid of the waste bin. Monitoring is carried out by calculating the distance from the surface of the waste to the Ultrasonic sensor in distance units. The distance in question is the height from the bottom of the waste bin to the maximum point, where the highest distance unit indicates the waste bin is empty, the shortest unit is the distance closest to the waste to the Ultrasonic sensor, which shows the waste bin is full, and in between is the waste bin filled.

Fig. 4. A smart waste bin unit

2.7. Dataset Creation

The dataset consists of quantitative data on the level of waste in organic, inorganic, and metal type bins as attribute data. Data class is the officer's decision to empty the smart waste bin or leave the waste in the smart waste bin first. The data is obtained from acquiring data on the volume of waste in the waste via an Ultrasonic sensor.

The dataset describes 2 data variables: a waste volume variable and a waste collection or emptying decision variable. The waste volume variable consists of 3 parameters: organic, inorganic, and metal. The waste collection decision class is the variable of emptying the waste bin with one parameter.

2.8. Design a Model

The experiment will compare Machine Learning methods in the form of DT, NB, K-NN, SVM, and MLP in the KEM model. The parameters for each method are shown in Table 1. For DT, it is the criterion in the form of Gini and gain. For Naïve Bayes, it is only Laplace Correction. For K-NN, it is a variation of the k value. For SVM, it is the kernel type and its C value. MLP includes the training cycle, the number of generations, and the number of ensembles. We hope to obtain optimal values from the test results with parameter values in design activities. Variations in parameter values include at least two different variable values. Hopefully, this design activity can provide primary results in experimental activities. So, which parameter value has a better effect on system performance.

Table 1. Parameter of comparision methods

Methods	Comparision	
	Parameter	*Range value*
Decision tree	Criterion	(Gini, gain)
Naïve bayes	Laplace Correction	
K-NN	k	(k=1, k=5)
SVM	Kernel type C	Dot (0,1)
MLP	Training cycles Number of generations Number of esemble	(10, 20) (10, 20) (4, 8)

2.9. Conduct Model Training

Model training uses DT, NB, K-NN, SVM, and MLP. The training process uses the k-fold cross-validation principle. That is, we will divide our research data into k subsets so that we will obtain k training data and k testing data for 164 in the existing dataset.

2.10. Model Testing

As a measure of performance, every time we research the nth training and testing data, we calculate performance quantities in the form of accuracy, recall, precision, and F-measure. Recap the performance values in each of the nth data in k amounts. We do a cumulative recap of the average performance value.

2.11. Comparative Testing of Machine Learning Methods

A comparative evaluation was carried out on several machine learning techniques, such as Decision Trees (DT), Naive Bayes (NB), Multi-Layer Perceptron (MLP), k-nearest Neighbors (K-NN), and Support Vector Machines (SVM). The evaluation focused on analyzing prior data models of trash volume from innovative waste bin units. The goal was to determine the most effective model to extract knowledge from the dataset, which contains historical data. The acquired knowledge can appear as rules or forecasts, guiding whether innovative waste units comprising organic, inorganic, and metal trash should be evacuated immediately or postponed.

3. Results and Discussion

In this study, the units of waste bins consisted of three types of waste bins. Each type of waste bin has a size of 90 liters. The first bin is for organic waste, the second is for inorganic waste, and the third is for metal waste, as shown in Fig. 5 for the front view, and the rear view is shown in Fig. 6. The smart waste bin that can sort waste in this research is a development of previous research [6]. The ability to sort waste in smart bins in this study uses capacitive and inductive proximity sensors. In previous studies, it only sorted two waste types: organic and inorganic. In this research, one type was added, namely the metal type.

Fig. 5. The front of the smart waste bin side

Fig. 6. The back of the smart waste bin side

Distance measurement in the waste bin uses an Ultrasonic sensor, where the highest distance unit is 26 distance units, and the lowest distance is three distance units. The lowest distance unit available is the distance detected by the Ultrasonic sensor, indicating the bin is full, while the highest distance indicates it is empty. Apart from full and empty conditions, another condition is that the waste bin is filled. Waste condition data is in the form of time series data obtained from the IoT platform interface using the Thingsboard Platform. The volume of waste is four units high for metal, 16 units high for inorganic, and 26 units high for organic waste. The waste volume data is the waste volume variable in the dataset. The dataset from smart waste units is data taken in June 2020. Data collection is only limited from 9 a.m. to 11 a.m. Time series data from MS Excel format, as in Fig. 7, the distance unit quantity, which shows the volume of waste from each type of organic, inorganic, and metal waste, is data in one unit of time from one waste bin unit. There are 14 smart waste units in the research area, as shown in Fig. 3.

	A	B	C	D	E	F	G	H
1	Timestamp	inductive	distance	capacity	type of waste	metal	non_organic	organic
2	6/17/2020 9:55		5					
3	6/17/2020 9:55					26		
4	6/17/2020 9:55							26
5	6/17/2020 9:55						13	
6	6/17/2020 9:55				Non-organic waste			
7	6/17/2020 10:00							
8	6/17/2020 10:00			1				
9	6/17/2020 10:00		2					
10	6/17/2020 10:00					25		
11	6/17/2020 10:00							23
12	6/17/2020 10:00						26	
13	6/17/2020 10:00				Organic waste			
14	6/17/2020 10:05							
15	6/17/2020 10:06							
16	6/17/2020 10:06							
17	6/17/2020 10:06							
18	6/17/2020 10:06							
19	6/17/2020 10:06		5					
20	6/17/2020 10:06					25		
21	6/17/2020 10:06							26
22	6/17/2020 10:06						26	
23	6/17/2020 10:06				Metal waste			
24	...	...	...	...	...	...	...	...
25	6/29/2020 10:09							
26	6/29/2020 10:09			1				
27	6/29/2020 10:09		5					
28	6/29/2020 10:09				Non-organic waste			
29								

New Timeseries table

Fig. 7. Sample of tabulation data downloaded from Thingboard

Each unit of waste bin occupies one cluster area. Each waste unit takes turns sending monitoring results every 5 minutes. One hundred sixty-four valid data were obtained from the total data, shown in Fig. 8.

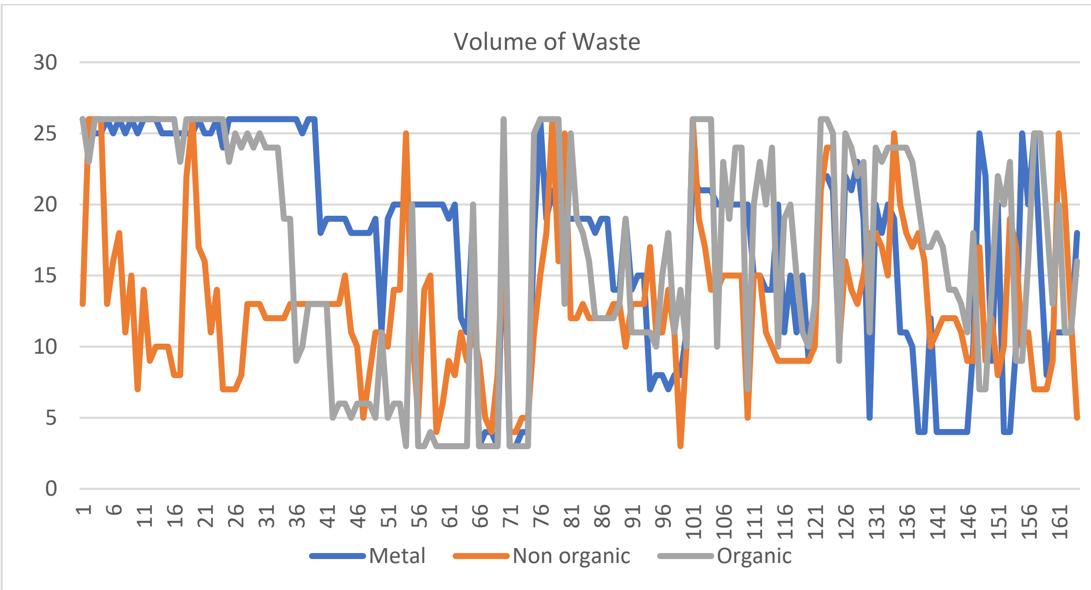

Fig. 8. The visualization of all valid data

The system then imports time series data in MS Excel format to be entered into the local database. The integration of data on garbage collection activities from smart waste bins by officers with time series data from the Thingsboard platform is stored in a local database into historical data. These old records are used today so that they can be extracted into current knowledge, as shown in Table 2, as a dataset in this research.

Historical datasets or past data on monitoring activities for smart waste bins are used as a reference for the collection schedule for waste disposal from every smart waste bin in rural areas. The waste volume data from Thingsboard is the waste volume data line with the waste dispose class as in the 138th data line in Table 2, where the officer decides to dispose of waste from the smart waste unit.

Table 2. The numerical value of the volume

Number	Table Column Head			
	Volume of metal	*Volume of Non Organic*	*Volume of Organic*	*Class*
1	26	13	26	Waste did not dospose
.....	....	....	....	....
138	4	16	26	Waste disposed
.....	....	....	....	....
164	18	5	16	Waste disposed

Implementation of model testing from the DT, NB, K-NN, SVM and MLP methods using Rapidminer tools, as shown in Fig. 9. The results of comparing Gini and gain criteria in the Decision Tree method obtained the most optimal value for gain, as shown in Table 3. Specifically for the method Naïve Bayes because it is a simple classification method for calculating a set of probabilities by adding up the frequencies and combinations of values from a dataset. In this study, the dataset of waste volume in smart bins and the decision to empty it is assumed to have a standard normal distribution.

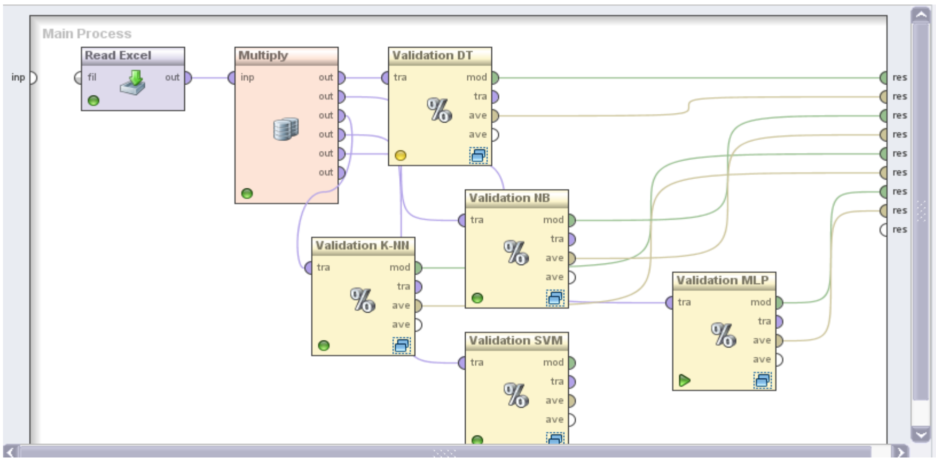

Fig. 9. Model in Rapidminer tool

As shown in Table 3, for the SVM method, the kernel type is chosen as a dot, and compared to a C value of zero and 1, the most optimal is at a C value of 1. For the MLP method, we can see from Table 3, that three parameters are being compared training cycles, number of generations and number of ensembles. The greater the value range of the three parameters in the MLP, the more optimal. So, this study's performance values of accuracy, recall, precision and F-measure are taken from the largest range of values. The method used in determining performance values is the confusion matrix method.

Table 3. Parameter of comparison methods

Methods	Comparision		
	Parameter	*Range value*	*Optimal*
Decision tree	Criterion	(Gini, gain)	Gain
Naïve bayes	Laplace Correction		
K-NN	K	(k=1, k=5)	K=5
SVM	Kernel type C	Dot (0,1)	1
MLP	Training cycles Number of generations Number of esemble	(10, 20) (10, 20) (4, 8)	20 20 8

The results of the Accuracy performance measurements are shown in Table 4. Accuracy, recall, precision and F1-measure values for each activity k to n $(k - n)$ are cumulative as the average. The total value is the average performance measure of 10 experiments based on the number of trials in [33], more than in [14], where only five were performed. The highest accuracy value, as shown in Table 4, is the accuracy value from the Decision Tree method of 88.13.

Table 4. Comparison accuracy values

K-fold	Methods				
	Decision Tree	*Naïve Bayes*	*K-NN*	*SVM*	*MLP*
X-1	59.38	81.25	78.12	81.25	80.00
X-2	85.62	68.75	76.25	68.75	78.75
X-3	81.25	67.50	68.75	70.00	72.50
X-4	91.10	81.00	79.10	43.75	90.95
X-5	100.00	100.00	98.67	100.00	96.67
X-6	94.38	43.75	90.62	43.75	60.00
X-7	95.00	97.50	93.75	97.50	93.75
X-8	97.33	88.67	85.33	90.00	92.67
X-9	83.75	93.75	93.75	86.88	84.38
X-10	93.50	71.00	70.00	67.00	78.00
Total	**88.13**	**81.11**	**83.43**	**74.88**	**82.76**

The results of measuring recall performance values for each k-n value are shown in Table 5. The largest average value is owned by the Decision Tree method, the largest average of recall is 86.89. The average recall value was obtained from 10 trials, as an implementation of the k-fold cross-validation evaluation method with $k = 10$.

Table 5. Comparision recall values

K-fold	Methods				
	Decision Tree	*Naïve Bayes*	*K-NN*	*SVM*	*MLP*
X-1	56.92	100.00	87.69	100.00	87.69
X-2	72.00	0.00	60.00	0.00	42.00
X-3	70.00	48.00	50.00	52.00	56.00
X-4	91.46	81.71	82.93	33.33	87.80
X-5	100.00	100.00	97.14	100.00	92.86
X-6	100.00	33.33	95.83	25.00	55.00
X-7	86.67	93.33	83.33	93.33	83.33
X-8	100.00	85.00	85.00	83.33	97.87
X-9	100.00	100.00	100.00	100.00	97.14
X-10	91.88	76.25	75..00	70.62	79.38
Total	**86.89**	**71.76**	**81.69**	**65.76**	**77.90**

Meanwhile, the precision values for each k-n value are shown in Table 6. The largest average precision value is 93.61. This value is owned by the Decision Tree method.

Table 6. Comparision precision values

K-fold	Methods				
	Decision Tree	*Naïve Bayes*	*K-NN*	*SVM*	*MLP*
X-1	89.16	81.25	85.71	81.25	87.69
X-2	80.00	0.00	62.50	0.00	80.77
X-3	100.00	100.00	100.00	100.00	100.00
X-4	92.69	82.75	80.00	80.00	84.27
X-5	100.00	100.00	100.00	100.00	100.00
X-6	93.02	80.00	92.00	100.00	86.84
X-7	100.00	100.00	100.00	100.00	100.00
X-8	93.75	86.44	79.69	90.91	90.15
X-9	87.50	87.50	87.50	76.92	74.73
X-10	100.00	85.92	85.71	85.61	92.03
Total	**93.61**	**72.26**	**87.31**	**81.46**	**89.64**

The results of measuring the F-Measure performance values for each k-n value are shown in Table 7. The highest value is 83.49, and the average f-measure value of the K-NN method. There is a difference between the mean value of the Decision Tree and K-NN of 0.27.

Table 7. Comparison of F1-Score values

K-fold	Methods				
	Decision Tree	*Naïve Bayes*	*K-NN*	*SVM*	*MLP*
X-1	69.48	89.65	86.68	89.65	87.69
X-2	75.78	0	61.22	0	55.26
X-3	82.35	64.86	66.66	68.42	71.79
X-4	92.07	82.22	81.43	47.05	85.99
X-5	100	100	98.54	100	96.29
X-6	96.38	47.05	93.87	40	67.34
X-7	92.85	96.54	90.90	96.54	90.90
X-8	96.77	85.71	82.25	86.95	93.85
X-9	93.33	93.33	93.33	86.95	84.47
X-10	95.76	80.79	79.98	77.39	85.23
Total	**83.22**	**74.01**	**83.49**	**69.29**	**81.88**

The next system test was carried out using the RapidMiner tool, as shown in Fig. 10.

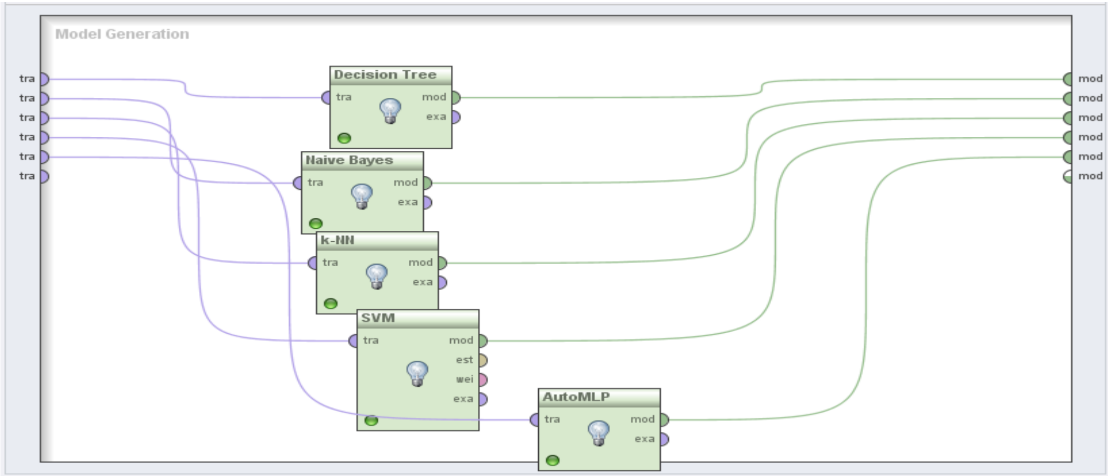

Fig. 10. Model comparison methods

The performance of the ROC curve of each method is shown in Fig. 11. The highest ROC graph marked in green is the ROC graph for the Decision Tree.

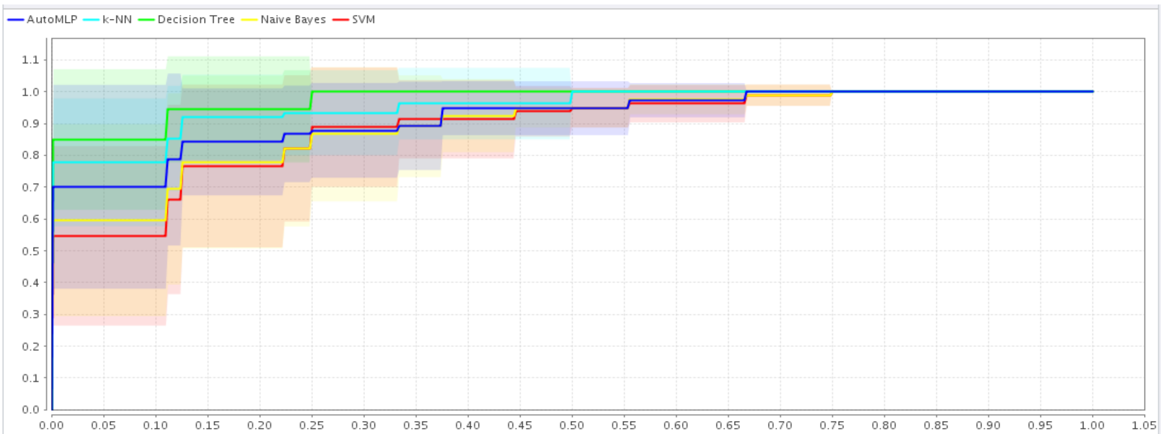

Fig. 11. ROCs Grafict comparison

Historical data on waste management activities is in the form of quantitative quantities as in Table 2 above. The DT method was chosen to implement the KEM system. Quantitative waste volume data is converted into categorical quantity data. Conversion to categorical magnitude uses a Likert scale as shown in Table 8. Where the decision to empty the waste bin or not refers to the rules extracted by the system. Rules derived from historical data on waste management in rural areas are today's new knowledge based on historical data in the past.

Table 8. Likert scale for waste bin volume range

Waste bin volume	Category and Value		
	Result of Likert Scale Value	*Value*	*Category*
Volume of metal	Highest= 26= 100% Lowest= 3/26 * 100%= 11.54% Range= 100%- 11.54%= 88.46% Category= 3 Interval= 88.46%/3= 29.49% Then interval: Interval= 29.49% * 26= 7.67 Assesment criteria: 100 %- 29.49%= 70.51% * 26= 18.33	> 18 11-18 < 11	empty filled full
Volume of Non Organic	Highest= 26 Lowest= 3 With the same manner as volume of metal	> 18 11-18 < 11	empty filled full
Volume of Organic	Highest= 26 Lowest= 3 With the same manner as volume of metal	> 18 11-18 <11	empty filled full

As in Table 9, empty categories for data with more than 18 distance units. The categories are filled for the amount of data starting from 11 to 18 distance units. The category of full bins is the distance units that are smaller than 11 distance units. Quantitative historical data pieces are shown in Table 2. Historical data in the form of categorical data are shown in Table 3.

Table 9. Categorical value of the volume

Number	Table Column Head			
	Volume of metal	*Volume of Non Organic*	*Volume of Organic*	*Class*
1	empty	Filled	empty	Waste did not dospose
...	...	...	...	...
138	full	Filled	empty	Waste disposed
...	...	...	...	...
164	filled	Full	filled	Waste disposed

The decision tree-generated rule code, depicted in Fig. 12, is utilized to ascertain whether waste should be discarded from the smart waste bin. The categorical value within the rule denotes the present state of the smart waste volume. Among the seven potential decisions, three entail waste disposal, while the remaining four pertain to retaining the waste within the bin.

1. IF volume of nonorganic = empty THEN 'waste did not dispose'
2. IF volume of non organic = full THEN 'waste disposed'
3. IF volume of non organic = filled AND volume of organic= full THEN 'waste disposed'
4. IF volume of non organic = filled AND volume of organic= empty THEN 'waste did not dispose'
5. IF volume of non organic = filled AND volume of organic= filled AND volume of metal = filled THEN 'waste did not dispose'
6. IF volume of non organic = filled AND volume of organic= filled AND volume of metal = empty THEN 'waste did not dispose'
7. IF volume of non organic = filled AND volume of organic= filled AND volume of metal = full THEN 'waste disposed'

Fig. 12. The seven rules represent new knowledge in scheduling the emptying of the waste

4. Conclusion

Scheduling from IoT-based smart waste bins using LoRa network media was explored in this study. It was found that higher parameter values led to more optimal performance. Specifically, the Decision Tree (DT) method demonstrated the highest average performance values for accuracy (88.13%), recall (86.89%), and precision (93.61%). Additionally, the K-Nearest Neighbors (K-NN) method exhibited the highest F-measure performance value (83.49%), which was only slightly lower than that of DT (0.27 difference). Thus, the Knowledge Extraction Method (KEM) system, particularly utilizing DT, proved most effective for waste disposal scheduling in rural areas. Future research should focus on applying KEM to larger datasets and developing adaptive algorithms for waste disposal scheduling. Furthermore, exploring the management of additional waste types by smart waste bins could lead to more complex decision-making processes, emphasizing the importance of leveraging KEM to extract knowledge from various domains and complex data variations. Integrating data from diverse IoT systems can enhance decision-making capabilities both presently and in the future.

Acknowledgment

The authors would like to thank to Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) and Informatics Engineering Department, Universitas Mandiri, Subang, Jawa Barat, Indonesia for their incredible support for this research.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors has received any funding from any institution or source.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] H. M. K. K. M. B. Herath and M. Mittal, “Adoption of artificial intelligence in smart cities: A comprehensive review,” *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 1, p. 100076, 2022, doi: 10.1016/j.jjimei.2022.100076.

[2] G. Sanathkumar, K. J. Nagesh, G. Hadimani, Laxman, B. R. Charanraj, and P. B. Honnavalli, “Smart Waste Segregation,” *IEEE Reg. 10 Humanit. Technol. Conf. R10-HTC*, vol. 2021-Septe, pp. 1–6, 2021, doi: 10.1109/R10-HTC53172.2021.9641526.

[3] M. A. Ramírez-Moreno *et al.*, “Sensors for sustainable smart cities: A review,” *Appl. Sci.*, vol. 11, no. 17, pp. 1–29, 2021, doi: 10.3390/app11178198.

[4] P. Marques *et al.*, “An IoT-based smart cities infrastructure architecture applied to a waste management scenario,” *Ad Hoc Networks*, vol. 87, pp. 200–208, 2019, doi: 10.1016/j.adhoc.2018.12.009.

[5] A. Z. Zaenal Abidin, M. F. I. Othman, A. Hassan, Y. Murdianingsih, U. T. Suryadi, and Z. Muslim, “Verifying Waste Disposal Practice Problems of Rural Areas in Indonesia Using the Apriori Algorithm,” *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, 2021, doi: 10.1109/ICIC54025.2021.9632987.

[6] A. Z. Z. Abidin, M. F. I. Othman, A. Hassan, Y. Murdianingsih, U. T. Suryadi, and M. Faizal, “LoRa-Based Smart Waste Bins Placement using Clustering Method in Rural Areas of Indonesia,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 14, no. 3, pp. 105–123, 2022, doi: 10.15849/IJASCA.221128.08.

[7] G. Vinti and M. Vaccari, “Solid Waste Management in Rural Communities of Developing Countries: An Overview of Challenges and Opportunities,” *Clean Technol.*, vol. 4, no. 4, pp. 1138–1151, 2022, doi: 10.3390/cleantechnol4040069.

[8] R. Hajovsky *et al.*, “Design of an IoT-Based Monitoring System as a Part of Prevention of Thermal Events in Mining and Landfill Waste Disposal Sites: A Pilot Case Study,” *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023, doi: 10.1109/TIM.2022.3225046.

[9] P. Jiang, Y. Van Fan, J. Zhou, M. Zheng, X. Liu, and J. J. Klemeš, “Data-driven analytical framework for waste-dumping behaviour analysis to facilitate policy regulations,” *Waste Manag.*, vol. 103, pp. 285–295, 2020, doi: 10.1016/j.wasman.2019.12.041.

[10] T. Anagnostopoulos *et al.*, "A stochastic multi-agent system for Internet of Things-enabled waste management in smart cities," *Waste Manag. Res.*, vol. 36, no. 11, pp. 1113–1121, 2018, doi: 10.1177/0734242X18783843.

[11] D. Alwis, P. Munasinghe, S. Rajapaksha, B. Ranawaka, J. Krishara, and W. Tissera, "A Smart Waste Disposal System: To Encourage Proper Waste Disposal," *4th Int. Conf. Adv. Comput. ICAC 2022 - Proceeding*, pp. 441–446, 2022, doi: 10.1109/ICAC57685.2022.10025307.

[12] B. Khatiwada, R. Jariyaboon, and K. Techato, "E-waste management in Nepal: A case study overcoming challenges and opportunities," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 4, no. April, p. 100155, 2023, doi: 10.1016/j.prime.2023.100155.

[13] P. Jiraamphorn and E. Pacharawongsakda, *DB2SM: An Efficient Resampling Technique for Imbalanced Data Classification Pharnuphon*, vol. 251, no. Ic2it. pp. 1-12, 2021, doi: 10.1007/978-3-030-79757-7_1.

[14] A. Binbusayyis, H. Alaskar, T. Vaiyapuri, and M. Dinesh, "An investigation and comparison of machine learning approaches for intrusion detection in IoMT network," *J. Supercomput.*, vol. 78, no. 15, pp. 17403–17422, 2022, doi: 10.1007/s11227-022-04568-3.

[15] A. Parvathy, G. L. Kasyap, D. V. Abhinav, A. N. V. S. Sai, R. Sriranjani, and N. Hemavathi, "Hybrid Machine Learning based False Data Injection Attack Detection and Mitigation Model for Waste Water Treatment Plant," *Int. Conf. Autom. Comput. Renew. Syst. ICACRS 2022 - Proc.*, no. Icacrs, pp. 674–680, 2022, doi: 10.1109/ICACRS55517.2022.10029198.

[16] C. E. Kontokosta, B. Hong, N. E. Johnson, and D. Starobin, "Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities," *Comput. Environ. Urban Syst.*, vol. 70, no. March, pp. 151–162, 2018, doi: 10.1016/j.compenvurbsys.2018.03.004.

[17] M. A. Valencia Diaz *et al.*, "Machine Learning Techniques to the Prediction of Variables of the Urban Solid Waste Collection Process," *2022 IEEE Colomb. Conf. Appl. Comput. Intell. ColCACI 2022 - Proc.*, pp. 1–6, 2022, doi: 10.1109/ColCACI56938.2022.9905363.

[18] T. A. Khoa *et al.*, "Waste Management System Using IoT-Based Machine Learning in University," *Wirel. Commun. Mob. Comput.*, vol. 2020, Arti, p. 13, 2020, doi: 10.1155/2020/6138637.

[19] S. H. Yusoff, U. N. K. A. Din, H. Mansor, N. S. Midi, and S. A. Zaini, "Design of smart waste bin and prediction algorithm for waste management in household area," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 2, pp. 748–758, 2018, doi: 10.11591/ijeecs.v12.i2.pp748-758.

[20] C. J. Baby, H. Singh, A. Srivastava, R. Dhawan, and P. Mahalakshmi, "Smart bin: An intelligent waste alert and prediction system using machine learning approach," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 771–774, 2018, doi: 10.1109/WiSPNET.2017.8299865.

[21] G. Lahcen, E. Mohamed, G. Mohammed, H. Hanaa, and A. Abdelmoula, "Waste solid management using Machine learning approch," *in 2022 8th Int. Conf. Optimiz. App. (ICOA),* pp. 1–5, 2022, doi: 10.1109/icoa55659.2022.9934356.

[22] J. C. Rodríguez López, M. I. Rodríguez-García, J. A. Moscoso Lopez, J. J. Ruíz Aguilar, J. M. Alcántara Pérez, and I. J. Turias Domínguez, "Prediction of container filling for the selective waste collection in Algeciras (Spain)," *Transp. Res. Procedia*, vol. 58, pp. 583–590, 2021, doi: 10.1016/j.trpro.2021.11.077.

[23] M. A. Al Mamun, M. A. Hannan, A. Hussain, and H. Basri, "Theoretical model and implementation of a real time intelligent bin status monitoring system using rule based decision algorithms," *Expert Syst. Appl.*, vol. 48, pp. 76–88, 2016, doi: 10.1016/j.eswa.2015.11.025.

[24] S. Squire, S. Ur Rehman, and M. A. Khan, "Investigating Efficient Municipal Solid Waste Collection Through Technology," *Proc. 2021 IEEE Work. Microw. Theory Tech. Wirel. Commun. MTTW 2021*, pp. 211–216, 2021, doi: 10.1109/MTTW53539.2021.9607088.

[25] A. Zhandos and J. Guo, "An approach based on decision tree for analysis of behavior with combined cycle power plant," *Proc. 2017 Int. Conf. Prog. Informatics Comput. PIC 2017*, pp. 415–419, 2017, doi: 10.1109/PIC.2017.8359583.

[26] Z. Gustiana and A. N. Sari, "Combination of C 4.5 Algorithm and Profile Matching for Determining University Students Graduation," in *2021 International Conference on Computer Science and Engineering (IC2SE)*, IEEE, pp. 1–6, 2022, doi: 10.1109/ic2se52832.2021.9791954.

[27] T. Bakhshi and M. Ahmed, "IoT-Enabled Smart City Waste Management using Machine Learning Analytics," *ICECE 2018 - 2018 2nd Int. Conf. Energy Conserv. Effic. Proc.*, vol. 7, pp. 66–71, 2018, doi: 10.1109/ECE.2018.8554985.

[28] M. Gao *et al.*, "Identification Method of Electrical Load for Electrical Appliances Based on K-Means ++ and GCN," *IEEE Access*, vol. 9, pp. 27026–27037, 2021, doi: 10.1109/ACCESS.2021.3057722.

[29] F. M. Kamau, K. O. Ogada, and C. W. Kipruto, "Analysis of Machine-Based Learning Algorithm used in Named Entity Recognition," *Informing Sci. Int. J. an Emerg. Transdiscipl.,* vol. 26, pp. 69–84, 2023, doi: 10.28945/5073.

[30] R. Ramly, A. Abu, B. Sajak, and M. Rashid, "IoT recycle management system to support green city initiatives," *Indones. J. Electr. Eng. Comput. Sci.,* vol. 15, no. 2, pp. 1037–1045, 2020, doi: 10.11591/ijeecs.v15.i2.pp1037-1045.

[31] G. Pattnaik and K. Parvathi, "Machine learning-based approaches for tomato pest classification," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 20, no. 2, pp. 321–328, 2022, doi: 10.12928/TELKOMNIKA.v20i2.19740https://doi.org/10.1016/j.jbi.2021.103834.

[32] T. Qi *et al.*, "KeMRE: Knowledge-enhanced medical relation extraction for Chinese medicine instructions," *J. Biomed. Inform.*, vol. 120, p. 103834, 2021, doi: 10.1016/j.jbi.2021.103834.

[33] O. A. Ebrahim and G. Derbew, "Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021," *Sci. Rep.*, vol. 13, no. 1, pp. 1–12, 2023, doi: 10.1038/s41598-023-34906-1.