




The Best Combination of Gas Sensor and Machine Learning Classification Algorithm in Detecting Mango (*Mangifera indica* L.) Quality

Joko Sumarsono¹ , Murad¹, Ida Ayu Widhiantari¹, Syahroni Hidayat²,
Ulfah Mediaty Arief², and Tatyantoro Andrasto²

¹ Department of Agricultural Engineering, Faculty of Food Technology and Agroindustries,
University of Mataram, Indonesia

² Departement of Electrical Engineering, Office E11 Sekaran Campus 50229, Universitas
Negeri Semarang, Indonesia
sumarsonoj@gmail.com

Abstract. Mango is a climacteric fruit with high transpiration activity when it reaches physiological maturity due to ethylene gas production. As a result, the quality of mangoes varies from day to day. Mango quality can be determined non-destructively by using gas sensors and machine learning to detect the gas produced. However, the classification accuracy remains low. Therefore, the aim of this study was to determine the type of gas sensor, the combination of gas sensors, and the combination of gas sensors and classification algorithms in determining the quality of mangoes. The gas sensors employed are TGS 2600, MQ3, MQ2, MQ4, and MQ8. While the classification algorithms are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). The results demonstrate that when paired with the SVM and KNN algorithms, the TGS 2600 sensor provided the best mango fruit quality classification results. Meanwhile, KNN's classification method outperforms SVM.

Keywords: Mango, Non-destructive, Machine learning.

1 INTRODUCTION

Mango (*Mangifera indica* L.) is a climacteric fruit. When it reaches a mature condition, the transpiration activity of mangoes increases every day. This certainly causes the quality of mangoes, when stored, to change every day [1], due to the increase in ethylene gas commonly produced by fruit, including mangoes [2], [3]. Numerous studies on determining climacteric fruit quality using non-destructive techniques have been conducted to determine and maintain fruit quality when kept. The quality of climacteric fruits such apples [2]–[4], bananas [5]–[7], tomatoes [6], and mangoes [1], [7], [8] has been the subject of numerous prior investigations. Non-destructive techniques include identifying the quality of fruit based on its physical characteristics using digital image

processing [3], [9]–[12] and based on its chemical characteristics using gas sensors [4], [13]–[15].

Fruit gas can be used to detect fruit quality faster, and it is also less expensive and easier to implement [6]. However, using gas produced by fruits has not provided better detection results than image processing [16]. The type of gas sensor used, the combination of gas sensor arrays, and the machine learning classification algorithm used can all contribute to this. Several types of gas sensors have been implemented, including ethylene gas sensors [13], the MQ-x family [4], and the TGS-x family [8], [15]. Artificial neural networks (ANN), Principal Component Analysis (PCA), and linear discriminant analysis (LDA) are the algorithms that have been used [15]. There are numerous other machine learning methods used in agriculture for classification, including Linear Regression, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Nave Bayes, K Nearest Neighbor (KNN), and Random Forest [17].

The purpose of this research is to determine the type of gas sensor, the combination of gas sensors, and the combination of gas sensors and classification algorithms in detecting the quality of mango fruit. The gas sensors used are from the MQ-x and TGS-x families.

2 PROPOSED METHOD

2.1 Method

The stages of this research were followed, as illustrated in Fig. 1. The subject of the investigation was 40 mangoes. Mangoes are kept at room temperature in an open location. The mango fruit was just halfway ripe on the first day of storage. The fruit is kept for a week. The fruit was examined in a lab on the eighth day to determine whether the pulp was good (positive, 1) or damaged (negative, 0). Every day, a gas sensor built inside an Arduino Mega 2560 is used to capture the gas data produced by each mango. The TGS 2600, MQ3, MQ2, MQ4, and MQ8 sensors are employed. The sensor and mango samples were placed in a customized container. This container is designed to hold mango fruit gas when it is filled and to ensure that clean air is only measured before the sensor is used again when it is empty. Fig. 2 depicts the tool series' organizational structure. Table 1 provides more information about the items and instruments utilized in this investigation.

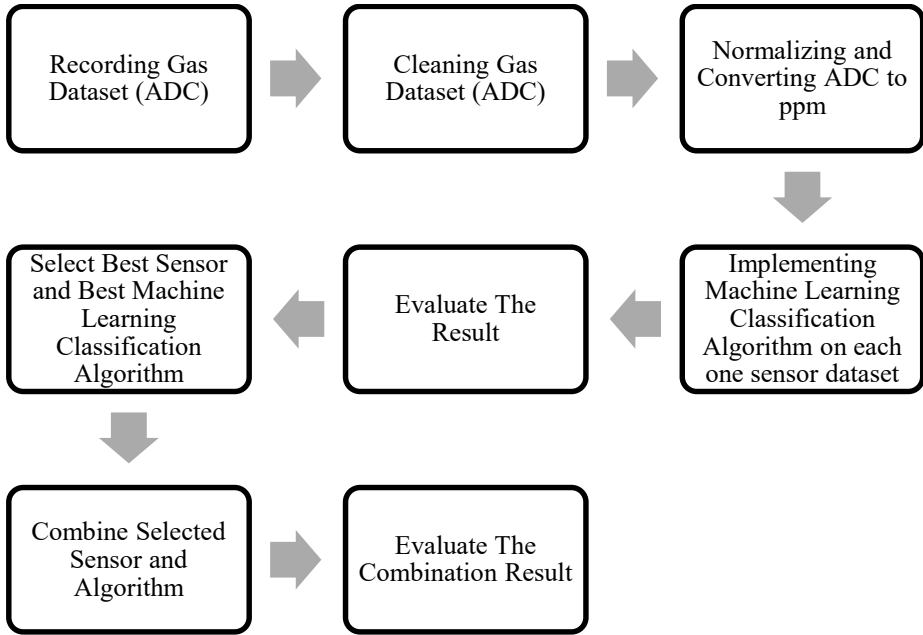


Fig. 1. Flowchart of proposed method.

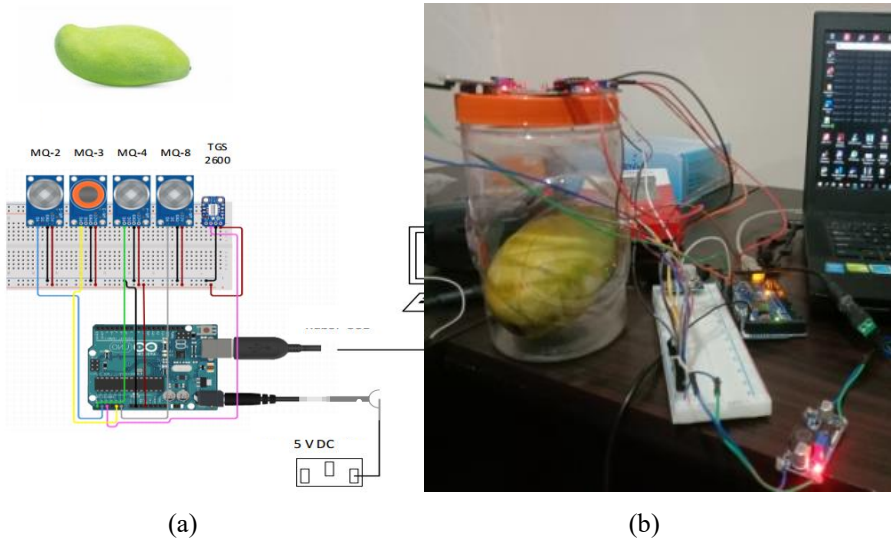


Fig. 2. (a) Schematic of research tool integration with five gas sensors and (b) Assembly Research tool with five sensors with gas trap container.

Table 1. Research tools and object

No	Name	Quantity	Specification
1	Gas Sensors:		
	TGS 2600	1	1 ~ 30 ppm (Hydrogen)
	MQ3	1	25 to 500 ppm (Ethanol)
	MQ4	1	300~10000 ppm (Methane)
	MQ2	1	300~10000 ppm (LPG, alcohol)
	MQ8	1	100~1000 ppm (Hydrogen)
2	Microcontroller	1	Arduino Mega 2560
3	Power source	1	5 V DC
4	PC Desktop	1	Intel core i5
5	Data logger	1	PuTTY
6	Programming Language		Python
7	Mango	40	Half-ripped

2.2 Recording Gas Dataset

Determine the gas sensor set point value is the first of two recording steps. The gas sensor measured the clean air in the gas trap container while it was empty for five minutes before recording the set point value. Out of the whole measuring period of five minutes, only one minute is used. The average set point value for each gas sensor is based on the steady state duration of 1 minute. Table 2 displays the data set points for each sensor.

Table 2. Average ADC measured as set point of sensor gasses

Variables	Sensors				
	TGS 2600	MQ3	MQ4	MQ2	MQ8
Average ADC _{measured} Value	70	120	190	200	95

The second step is the recording of mango fruit gas data. The mango fruit was kept in an open area for seven days straight during this method. For around three minutes, each mango sample was tested for gas. The gas recording log file is an ADC value.csv file that was created with the PuTTY program. Table III displays the findings from the mango fruit gas data recording log file. Incomplete or missing recordings at specific times are indications of a recording error in raw data. Both when the sensor has stopped recording and when it first detects gas by mango are during this condition. Consequently, a data cleaning procedure needs to be performed.

Table 3. Mango fruit gas measurement data log sample 35th day 7

No	===== PuTTY log 2021.09.24 14:58:08 =====
1	T MQ-8= 99
2	TGS2600= 74 MQ-3= 119 MQ-4= 185 MQ-2= 215 MQ-8= 99
3	TGS2600
4	TGS2600= 74 MQ-3= 118 MQ-4= 183 MQ-2= 215 MQ-8= 99
5	TGS2600= 74 MQ-3= 119 MQ-4= 185 MQ-2= 215 MQ-8= 99
6	TGS2600TGS2600= 73 MQ-3= 120 MQ-4= 188 MQ-2= 215 MQ-8= 101
7	TGS2600= 74 MQ-3= 121 MQ-4= 188 MQ-2= 216 MQ-8= 101
8	TGS2600= 72 MQ-3= 120 MQ-4= 187 MQ-2= 215 MQ-8= 100
.	...
.	...
.	...
302	TGS2600= 77 MQ-3= 143 MQ-4= 205 MQ-2= 232 MQ-8= 111
303	TGS2600= 77 MQ-3= 144 MQ-4= 206 MQ-2= 234 MQ-8= 113
304	TGS2600= 77 MQ-3= 143 MQ-4= 206 MQ-2= 234 MQ-8= 112
305	TGS2600= 77 MQ-3= 143 MQ-4= 205 MQ-2= 233 MQ-8= 112
306	TGS2600= 77 MQ-3= 144 MQ-4= 206 MQ-2= 234 MQ-8= 112
307	TGS2600= 0 MQ-3= 0 MQ-4= 0 MQ-2= 0 MQ-8= 0
308	TGS2600= 0 MQ-3= 0 MQ-4= 0 MQ-2= 0 MQ-8= 0
309	T

The mangoes were subjected to laboratory tests on the eighth day to assess their condition after seven days of storage. Mango is covered to reveal if it is in good (1) or damaged (0) . According to the lab test results, 19 mangoes were discovered to be in good (1) condition, while the remaining 21 were damaged (0). This mango appears to have significant damage that extends from the skin to the flesh.

2.3 Cleaning Gas Dataset

Reading the gas recording data log file is the first step in the data cleaning process. Next, the data is categorized according to the sensor, specifically the tgs26, mq3, mq4, mq2, and mq8. The result is a new table with dimensions M x N, where M is the number of data rows and N is the number of columns. Table 4 displays the outcomes of the sample data cleaning.

Table 4. ADC measurement data of cleaning results

Index	tgs26	mq3	mq4	mq2	mq8
1	96	271	277	199	166
2	98	302	312	200	208
3	96	271	277	199	166
4	98	302	312	200	208
.
.
.
314	260	512	549	284	390

Index	tgs26	mq3	mq4	mq2	mq8
315	259	507	548	283	389
316	259	506	547	284	390
317	260	507	551	284	392

Fig. 3 displays the graph of the measured ADC data following the cleaning procedure. There was a rise in the measured ADC data at the beginning of the recording, and then, after some time, it reached a steady state. The graph demonstrates that none of the sensors' measured ADC values start at zero at the beginning. Processes for normalization and data conversion are therefore required.

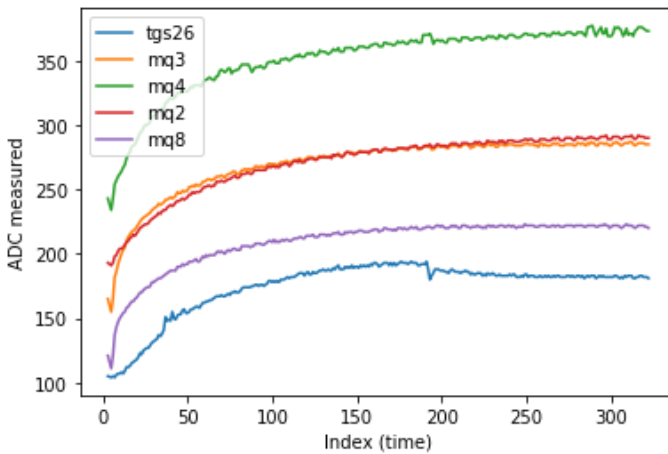


Fig. 3. Sample of ADC measured after data cleaning

2.4 Normalizing and Converting Gas Dataset

Normalization aims to make the initial value of each measured sensor is 0. Normalization is done by subtracting all measured ADC values with set point values using Equation (1). Then the calculation result of Equation (1) is reduced by its own minimum value using Equation (2).

$$dADC = ADC_t - SP_{SSR} \quad (1)$$

$$\Delta ADC = dADC - \min(dADC) \quad (2)$$

Data conversion is carried out to obtain the ppm value of each measured ADC value. The data needed to perform the conversion is the measurement range of each sensor, the set point, the measured ADC value, the ADC 1023 scale, the maximum input voltage V_{max} 5 V, and the DAC value [1][1]. The DAC value is obtained using the following equation (3):

$$DAC = \frac{ADC \text{ measured}}{1024} \times V_{maks} \quad (3)$$

The DAC value is then converted to ppm using Equation (4). The results of normalizing the measured ADC data and conversion to ppm values are shown in Fig. 4.

$$ppm_{ssr} = \frac{range_{max_ssr} - range_{min_ssr}}{V_{maks}} \times DAC \quad (4)$$

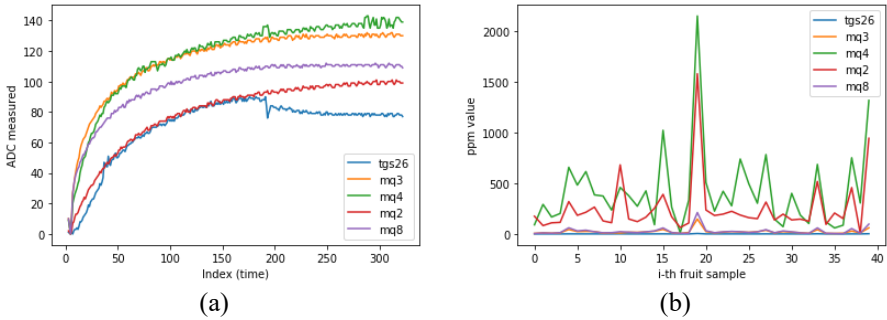


Fig. 4. (a) Sample of ADC measured normalized day 1, (b) Sample of ppm value conversion day 1

2.5 Machine Learning Algorithm

The machine learning classification techniques employed in this work are Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). In the learning process, 80% of the training dataset is used, the remaining 20% is used as the test dataset.

Logistic Regression (LR). Logistic regression is a data analysis technique that employs mathematics to determine the relationship between two data factors. Then, based on the other factors, use this relationship to predict the value of one of these factors. Logistic regression employs the logistic function, also known as the logit function in mathematics as the equation between x and y . The logit function converts y to x 's sigmoid function. This algorithm is powerful to be implemented in agriculture and horticulture research [18].

Decision Tree (DT). Decision Tree algorithm is a supervised learning algorithm. The decision tree technique, in contrast to other supervised learning methods, is capable of handling both classification and regression issues. By learning straightforward decision rules derived from previous data, a Decision Tree is used to build a training model that

may be used to predict the class or value of the target variable (training data). In decision trees, we begin at the tree's root when anticipating a record's class label. We contrast the root attribute's values with that of the attribute on the record. We follow the branch that corresponds to that value and go on to the next node based on the comparison[19].

Random Forest (RT). A series of decision trees are produced by Random Forest classifiers, each of which is constructed using a vector that is generated randomly but uniformly across all trees. After producing a sizable number of trees, each one votes for the class that it thinks is the most popular, and the final model classifies according to the class that received the most votes. This classifier's ability to prevent overfitting due to the large numbers law is an intriguing feature[19], [20].

Support Vector Machine (SVM). A binary classifier called SVM looks for linear hyperplanes that maximize class separation. By transforming the data into a higher-dimensional space and locating a separable hyperplane across classes, SVM replicates decision boundaries between classes. SVM is a binary classifier, hence, to use it for multiclass classifications, multiple classifiers must be constructed and combined. We used the LIBSVM library provided by python [11].

K-Nearest Neighbours (KNN). KNN is a supervised algorithm that classifies the results of a new query instance based on the majority of the categories in the KNN. The goal of this algorithm is to classify new objects based on attributes and training samples. KNN keeps track of all the cases and uses a similarity metric to classify new cases [21].

2.6 Evaluation

System performance indicators including accuracy, precision, sensitivity, and specificity are employed for evaluation since it is crucial. The matrix, which is a confusion matrix, generates four indications: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). According to [8], [15], TP represents the proportion of normal samples that were correctly detected, TN represents the proportion of defective samples that were correctly detected, FP represents the proportion of normal samples that were incorrectly detected, and FN represents the proportion of faulty samples that were incorrectly detected. Fig. 5 shows the confusion matrix's shape.

		Actual Value	
		Positive (1)	Negative (0)
Predictive Value	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 5. Confusion matrix

Each value in the confusion matrix above is used to determine the following Accuracy (*Acc*), Precision (*Pr*), Sensitivity (*Se*), and Specificity (*Sp*) values:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$Pr = \frac{TP}{TP + FP} \quad (6)$$

$$Se = \frac{TP}{TP + FN} \quad (7)$$

$$Sp = \frac{TN}{FP + TN} \quad (8)$$

In this study *Acc* shows the overall performance of the model. *Pr*, is the ratio of predicted good fruit compared to the overall yield predicted as good fruit. *Se*, refers to the performance of the model to correctly detect good fruit samples. *Sp*, refers to the performance of the model to correctly detect samples of damaged fruit.

2.7 Selecting Sensor Combination

After examining the results of mango categorization based on the average accuracy value, the optimal sensor and algorithm are chosen. In the following phase, the most accurate sensors and algorithms will be employed. While the algorithm's accuracy limit is the best two, the sensor's accuracy limit is the best three. The two best classification algorithms are also used to reclassify a group of sensors that includes two and three sensors.

3 RESULT AND DISCUSSION

Table 5 displays the test results for the classification of mango quality using Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN) classifiers.

Table 5. Evaluation of all sensors

No	Classifier	Accuracy (Acc. %)					Average Classifier Acc.
		TGS	MQ3	MQ4	MQ2	MQ8	
1	LR	62.5	75.0	62.5	62.5	75.0	67.5
2	DT	62.5	75.0	62.5	62.5	75.0	67.5
3	RF	75.0	75.0	62.5	50.0	62.5	65.0
4	SVM	100	75.0	75.0	75.0	75.0	80.0
5	KNN	87.5	75.0	87.5	87.5	75.0	82.5
Av. Sensors Acc.		77.5	75.0	70.0	67.5	72.5	

The average classification accuracy performance of the TGS sensor is 77.5%, followed by MQ3 at 75%, MQ8 at 72.5%, MQ4 at 70%, and MQ2 at 67.5%. In the meantime, KNN, SVM, LR, DT, and RF all received 67.5% average accuracy for the classification algorithm, with KNN achieving the greatest average accuracy at 82.5%. As a result, the KNN and SVM classifiers will be employed in conjunction with the TGS 2600, MQ3, and MQ8 sensors for the following stage.

The sensor and classifier combination consists of one sensor, two sensors, and three sensors. As a result, the sensor combinations will be TGS, MQ3, MQ8, TGS+MQ3, TGS+MQ8, MQ3+MQ8, and TGS+MQ3+MQ8. Table 6 shows the results of the learning evaluation of each of these combinations as a confusion matrix, and table 7 shows the performance evaluation.

Table 6. The confusion matrix of combination of selected sensors and classifier

Classifier		TGS		MQ3		MQ8		TGS+MQ3		TGS+MQ8		MQ3+MQ8		TGS+MQ3+MQ8	
		P	N	P	N	P	N	P	N	P	N	P	N	P	N
		SVM	P	4	0	2	2	2	2	2	2	2	2	2	2
	N	0	4	0	4	0	4	0	4	0	4	0	4	0	4
KNN	P	3	1	3	1	2	2	3	1	2	2	3	1	3	1
	N	0	4	1	3	0	4	1	3	0	4	0	4	0	4

It can be concluded that the combination of a single sensor, the TGS 2600, and the two selected classification algorithms, SVM and KNN, provides the highest accuracy, by 100% and 87.5%, respectively. Meanwhile, the MQ3 and MQ8 sensors both have a 75% accuracy. The MQ3 + MQ8 sensor and the KNN classifier achieve the highest accuracy of 87.5% in classifying the quality of mango fruit using a combination of two sensors. The combination of two other sensors, which use SVM and KNN, provides an

accuracy of 75%. Finally, by combining three sensors, the KNN classifier outperforms the SVM classifier by 87.5%. As a result, when used independently, the TGS 2600 sensor achieves the best results. In terms of classification, KNN outperforms SVM in classifying mango quality.

The Precision (Pr) system demonstrates the ability of a combination of sensors and classifiers to detect ripe (good) fruit. The combination of TGS and SVM provides the best precision of 100%, while TGS and KNN provides a precision of 75%. When comparing the average precision performance of all sensor combinations with classifiers, KNN still outperforms SVM. In terms of the system's Sensitivity (Se) performance in detecting good condition fruit, the SVM classifier outperforms the KNN for all sensor combinations. Finally, in terms of detecting the damaged condition, Specificity (Sp), the performance of the combination of TGS with SVM outperforms that of TGS with KNN. However, when the average performance of the Specificity (Sp) of all sensor combinations with classifiers is considered, KNN outperforms SVM.

Table 7. Performance evaluation of combination of selected sensors and classifier

Classifier	Criteria	Sensors Performance (%)							Av. Classifier Performance
		TGS	MQ3	MQ8	TGS+MQ3	TGS+MQ8	MQ3+MQ8	TGS+MQ3+MQ8	
SVM	Acc	100	75.0	75.0	75.0	75.0	75.0	75.0	78.6
	Pr	100	50.0	50.0	50.0	50.0	50.0	50.0	57.1
	Se	100	100	100	100	100	100	100	100
	Sp	100	66.7	66.7	66.7	66.7	66.7	66.7	71.4
KNN	Acc	87.5	75.0	75.0	75.0	75.0	87.5	87.5	80.4
	Pr	75.0	75.0	50.0	75.0	50.0	75.0	75.0	67.9
	Se	100	75.0	100	75.0	100	100	100	92.9
	Sp	80.0	75.0	66.7	75.0	66.7	80.0	80.0	74.8

4 CONCLUSION

According to the findings of the research, only the TGS 2600 sensor combined with the SVM and KNN classifiers can provide the best mango quality classification accuracy. Meanwhile, KNN outperforms SVM in terms of classifier performance. By combining sensors and increasing the number of sensors, the overall system performance in detecting mango fruit quality can be reduced.

References

1. Gianguzzi, G., Farina, V., Inglese, P., Rodrigo, M. G. L.: Effect of Harvest Date on Mango (*Mangifera indica* L. Cultivar Osteen) Fruit's Qualitative Development, Shelf Life and Consumer Acceptance. *Agronomy* 11(4), (2021).
2. Janssen, S., Schmitt, K., Blanke, M., Bauersfeld, M. L., Wöllenstein, J., Lang, W.: Ethylene detection in fruit supply chains. *Philosophical Transactions of the Royal Society A*:

- Mathematical, Physical and Engineering Sciences 372(2017), (2014).
3. Bratu, A. M., Popa, C., Bojan, M., Logofatu, P. C., Petrus, M.: Non-destructive methods for fruit quality evaluation. *Scientific reports* 11(1), 1–15, (2021).
 4. Liu, S., Bai, L., Hu, Y., Wang, H.: Image captioning based on deep neural networks. *MATEC web of conferences* 232, 1–6 (2018).
 5. Maduwanthi, S. D. T., Marapana, R. A. U. J.: Induced ripening agents and their effect on fruit quality of banana. *International Journal of Food Science* 2019, (2019).
 6. Nair, K., Sekhani, B., Shah, K., Karamchandani, S.: Expiry Prediction and Reducing Food Wastage Using IoT and ML. *International Journal of Electrical and Computer Engineering Systems* 12(3), 155-162 (2021).
 7. Srividhya, V., Sujatha, K., Jayachitra, N.: *International Journal of Pure and Applied Mathematics* 118(18), 3191–3207 (2018).
 8. Murad, M., Sukmawaty, S., Ansar, A., Sabani, R., Hidayat, S.: Mango damage detection system using gas sensor with DCS - LCA method. *JTIM J. Teknol. Inf. dan Multimed.* 3(4), 186–194 (2021).
 9. Dhiman, B., Kumar, Y., Kumar, M.: Fruit quality evaluation using machine learning techniques: review, motivation and future perspectives. *Multimedia Tools and Applications* 81(12), 16255-16277 (2022).
 10. Priya, P. S., Jyoshna, N., Amaraneni, S., Swamy, J.: Real time fruits quality detection with the help of artificial intelligence. *Materials Today: Proceedings* 33, 4900-4906 (2020).
 11. Worasawate, D., Sakunasinha, P., Chiangga, S.: Automatic classification of the ripeness stage of mango fruit using a machine learning approach. *AgriEngineering* 4(1), 32-47 (2022).
 12. Hasanzadeh, B., Abbaspour-Gilandeh, Y., Soltani-Nazarloo, A., Hernández-Hernández, M., Gallardo-Bernal, I., Hernández-Hernández, J. L.: Non-Destructive Detection of Fruit Quality Parameters Using Hyperspectral Imaging, Multiple Regression Analysis and Artificial Intelligence. *Horticulturae* 8(7), (2022).
 13. Geethapriya, M. N., Praveena, S. M.: Evaluation of fruit ripeness using electronic nose. *Evaluation* 6(5), 1-5 (2017).
 14. Baietto, M., Wilson, A. D.: Electronic-nose applications for fruit identification, ripeness and quality grading. *Sensors* 15(1), 899-931 (2015).
 15. Aghilinategh, N., Dalvand, M. J., Anvar, A.: (2020). Detection of ripeness grades of berries using an electronic nose. *Food Science & Nutrition* 8(9), 4919-4928 (2020).
 16. Mavani, N. R., Ali, J. M., Othman, S., Hussain, M. A., Hashim, H., Rahman, N. A.: Application of artificial intelligence in food industry—a guideline. *Food Engineering Reviews* 14(1), 134-175 (2022).
 17. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: Machine learning in agriculture: A review. *Sensors* 18(8), (2018).
 18. Marini, R. P., Lavelly, E. K., Baugher, T. A., Crassweller, R., Schupp, J. R.: Using logistic regression to predict the probability that individual ‘Honeycrisp’ apples will develop bitter pit. *HortScience* 57(3), 391-399 (2022).
 19. Mercol, J. P., Gambini, M. J., Santos, J. M.: Automatic classification of oranges using image processing and data mining techniques. In: XIV Congreso Argentino de Ciencias de la Computación, pp. 1–12. Chilecito (2008).
 20. Zawbaa, H. M., Hazman, M., Abbass, M., Hassanien, A. E.: Automatic fruit classification using random forest algorithm. In: 14th International Conference on Hybrid Intelligent Systems (HIS 2014), pp. 164–168. IEEE Xplore, Kuwait (2014).
 21. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research* 5, 1-16 (2020).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

