SENSOR TEST

The code that I wrote is available on my GitHub::
https://github.com/priya212000/Food-spoilage-detection/blob/main/Sensors.ipynb

ABSTRACT

(importance of good nutritious food 3-4 lines) (affordability 2 lines) This research work proposes a low-cost and effective food spoilage detection by using multi-sensory setup by using a list of gas sensors . In recent years there has been a rapid development of this method, especially in the area of food control. The work has to its credit a mechanism that performs dimensionality reduction based on decision boundary and logistic regression. The system trains different models using high quality training sets, test sets, and feature combinations for classification which serve as the basis for obtaining more accurate results. The classifier engine is responsible for classifying food items into one of the two classes: “spoiled” or “fresh”. The experimentation results clearly reveal the capability of the system to produce accurate results still maintaining low cost and versatility of usage.

INTRODUCTION:

One of the reasons for the spread of pandemic is the lack of stamina on people. Therefore, it is highly necessary that during such times we consume healthier food to build resistance against the virus. The proposed work provides a low-cost food spoilage identification system. The hardware comprises MQ series of analog gas sensors (MQ8, MQ135, MQ9, MQ4, MQ2, MQ3). The data frame is formed from the dataset which consists of 180 data points with labels that classify the food as spoilt or not. The data is manually collected from the Arduino Nano by using different fruits and vegetables which are spoilt and fresh and labeled correspondingly. Feature extraction is done based on decision boundary. CNN cannot be used as the dataset is not sufficient for the same. Therefore, logistic regression is used as the model for training and testing as overfitting occurs when any other complex model is used. The detailed implementation procedure and discussions on results obtained is presented in sections.

LITERATURE REVIEW:

For many years…..researchers have come up with various ideas to address the problem. The study of the literature reveals that many approaches like image classification, sensor test etc., have been adopted for assessing the quality of food items. The proposed system differs from the existing works in the learning methods, and the sensors used for the hardware design.
In terms of identifying food spoilage, a holistic approach of using both the images and the gases that are emitted while testing the food items is expected to give more accurate results.

IMAGE CLASSIFIER REVIEW:

[1] proposes a low cost high quality system for accessing quality of food samples by finding the presence of fungus by using Histogram of Oriented Gradients algorithm along with Support Vector Machine classifier. The features of the food samples captured in real time using a webcam were extracted using Histogram of Oriented Gradients algorithm. The extracted features were given to SVM classifier which compared these features with the trained one and displayed the quality of food samples. [2] discusses a solution for identifying food samples using CNN and compares it to the results from a generic classifier like an SVM. This work applies a convolutional neural network (CNN) to the tasks of detecting and recognizing food images. They apply CNN to the tasks of food detection and recognition through parameter optimization. CNN showed significantly higher accuracy than did traditional support-vector machine-based methods with handcrafted features of: Baseline 89.7 ± 0.73% CNN 93.8 ± 1.39%. Since this proves that CNN is a more viable option than SVM, it is used as the image classifier for this paper.

SENSOR TEST:

HARDWARE REVIEW:

We have chosen to identify spoilage using a low-cost setup, therefore prefer arduino which is an 8-bit microcontroller development board with a USB programming interface to connect to a computer and additional connection sockets to external electronics like sensors, motors speakers, diodes etc over the Raspberry pi which is a computer-based development board which runs on a Linux distribution referred to as Raspbian Linux.. [12] provides a comparison of Raspberry pi4 and Arduino for identifying gases. [13] provides a glimpse of the type of Arduino boards, working principles, software implementation and their applications. We have chosen arduino nano as it is breadboard friendly and goes in with our setup. **[14]** gives a review of different semiconductor-type gas sensors and the properties of each of them which would pave way for choosing the appropriate sensors for the hardware design.

FEATURE EXTRACTION AND MODEL REVIEW:

[3] proposes an Internet of Things (IoT) based system that allows users to know the groupings of gases in crude milk continuously. In [3], Microbial activity is determined using a gas sensor, high quality milk should have no salinity, so salinity of the milk is measured by using a salinity sensor and also level of the milk will be measured by using a level sensor. This method only helps in classifying milk, whereas our scope of research reaches out to fruits and vegetables as well. **[4]** proposes a solution whose objective is to design an electronic nose system that will be sensitive to the gases emitted by spoiled food samples namely banana, pechay, carrots and grapes operating in low level temperature particularly the refrigerator and then determine food spoilage using Principal Component Analysis – K Nearest Neighbors, however, it will not take any corrective actions. The system will gather readings from MQ gas sensors and will be subjected to PCA and KNN.

[5] aims to predict a past-due amount using traditional and machine learning techniques: Logistic Analysis, k-Nearest Neighbor (KNN) and Random Forest and compare the results of the models. [6] gives insights into which model to prefer for classification. Results state that KNN is comparatively slower than Logistic Regression (LR) and that LR can derive confidence level (about its prediction), whereas KNN can only output the labels. Therefore, we prefer LR

according to our dataset. Since outliers are minimal, we don't prefer SVM (Support Vector Machine) over LR. [7] proposes a gas classification method for an electronic nose (e-nose) system for combined features. [11] proposes a method for building a simple electronic nose based on commercially available sensors used to sniff in the market and identify spoiled/contaminated meat stocked for sale in butcher shops. Using a metal oxide semiconductor-based electronic nose, [11] measures the smell signature from two of the most common meat foods (beef and fish) stored at room temperature. Food samples were divided into two groups: fresh beef with decayed fish and fresh fish with decayed beef. The prime objective was to identify the decayed item using the developed electronic nose. Additionally, [11] tested the electronic nose using three pattern classification algorithms (artificial neural network, support vector machine and k-nearest neighbor), and compared them based on accuracy, sensitivity, and specificity.

DECISION BOUNDARY & LOGISTIC REGRESSION ALGORITHMS:

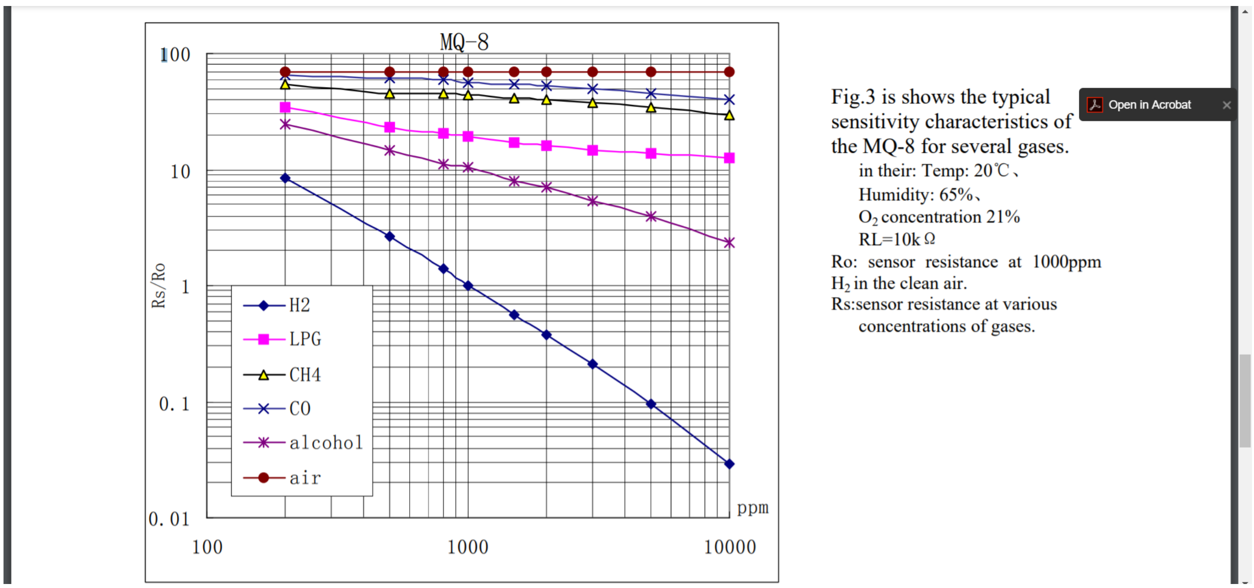
[15] proposes a variable selection method for a nonparametric framework, in which they consider the problem of classifying a categorical response Y whose distribution depends on a vector of predictors X, where the coordinates Xj of X may be continuous, discrete, or categorical which is called Categorical Adaptive Tube Covariate Hunting. [8] proposes a gradient descent DBFE method (GDDBFE) that shows a substantial improvement in processing time. Based on this GDDBFE, they then propose an incremental gradient descent decision boundary feature extraction method (IGDDBFE). The proposed IGDDBFE method consists of two steps: updating the decision boundaries and adding discriminately informative features with newly added samples and then updating the feature vectors by incremental eigenvector updates. [9] proposes a novel approach to feature extraction for classification for non-parametric classifiers is proposed based directly on the decision boundaries. [9] claims that feature extraction is equivalent to retaining informative features or eliminating redundant features; thus, the terms “discriminately information feature” and “discriminately redundant feature” are first defined relative to feature extraction for classification. [9] shows how discriminately redundant features and discriminantly informative features are related to decision boundaries. [10] demonstrates the preferred pattern for the application of logistic methods with an illustration of logistic regression applied to a data set in testing a research hypothesis. It also offers recommendations for appropriate reporting formats based on logistic regression results and the minimum observation-to-predictor ratio.

SUMMARY OF THE LITERATURE REVIEW:

From the literature survey, one can clearly understand the need for an Arduino Nano over Raspberry pi for the hardware setup due to the effective data that the former provides in a low-budget scenario. The obtained results are further processed using Machine learning methods. For feature extraction, decision boundary based algorithm is preferred due to the non-parametric nature of the data. The model structure of nonparametric models is not specified *a* priori (as mentioned later in the doc) but is instead determined from data. The term nonparametric is not meant to imply that such models completely lack parameters, but rather that the number and nature of the parameters are flexible and not fixed in advance. Finally, the scheme uses a logistic regression model for classification due to the effective handling of lesser training data without overfitting for achieving faster results and ensuring minimal outliers.

PROPOSED SCHEME:

Image of the config of the MQ8 sensor:



Picture credit: HANWEI ELECTRONICS CO., LTD

As you can see in the picture for a sample sensor, a voltage drop is observed if the sensors detect the gas that they are sensitive to beyond the threshold value. The voltage is by the standard in an Arduino nano set between 0-5V.

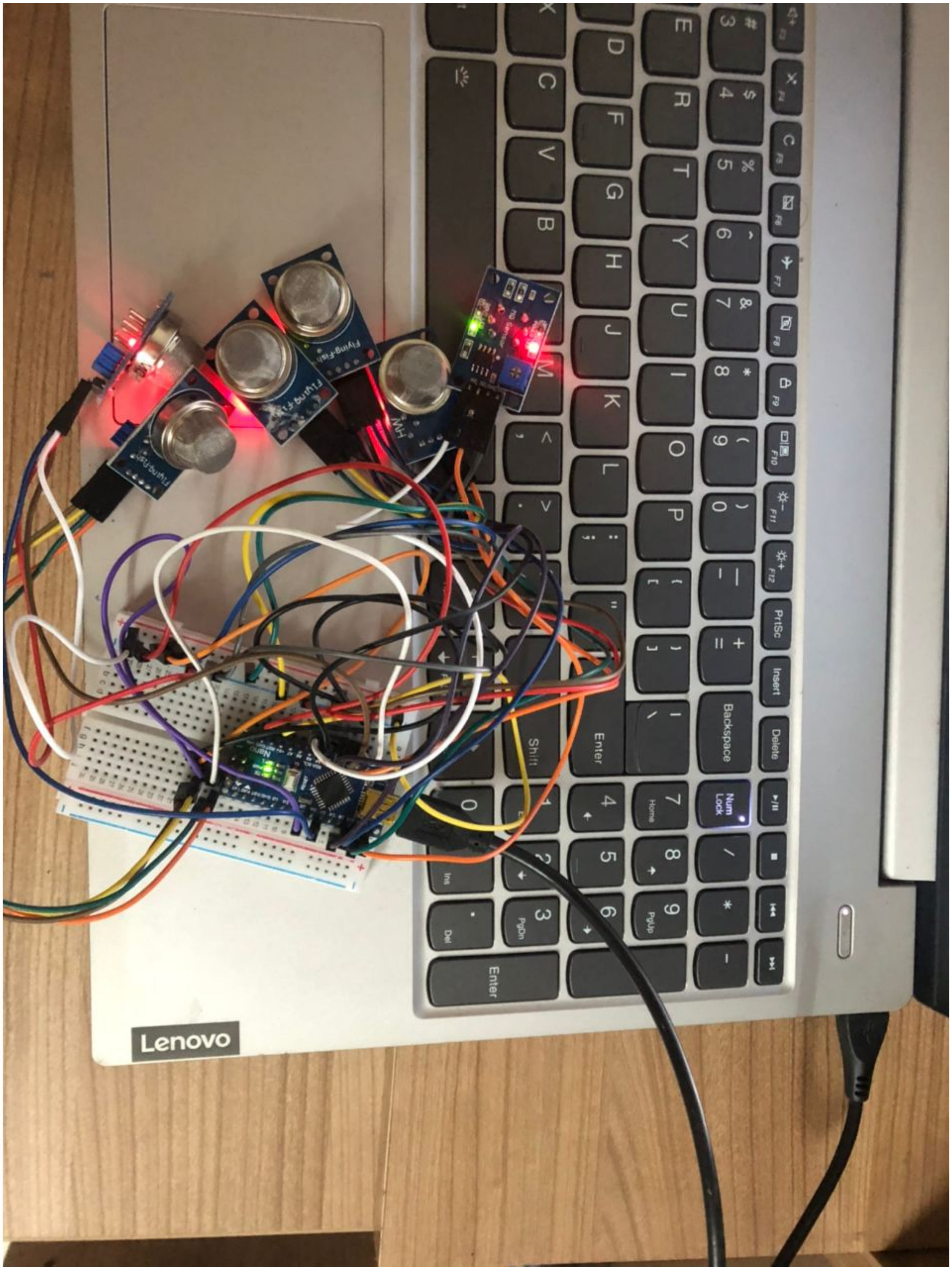
HARDWARE CONNECTION:

Code for programming the Arduino Nano:

Data screenshots

Setup:

- Breadboard
- Arduino Nano
- Male to Male jumper cables
- Female to Male jumper cables
- MQ sensors
- Computer
- USB cable


Lenovo

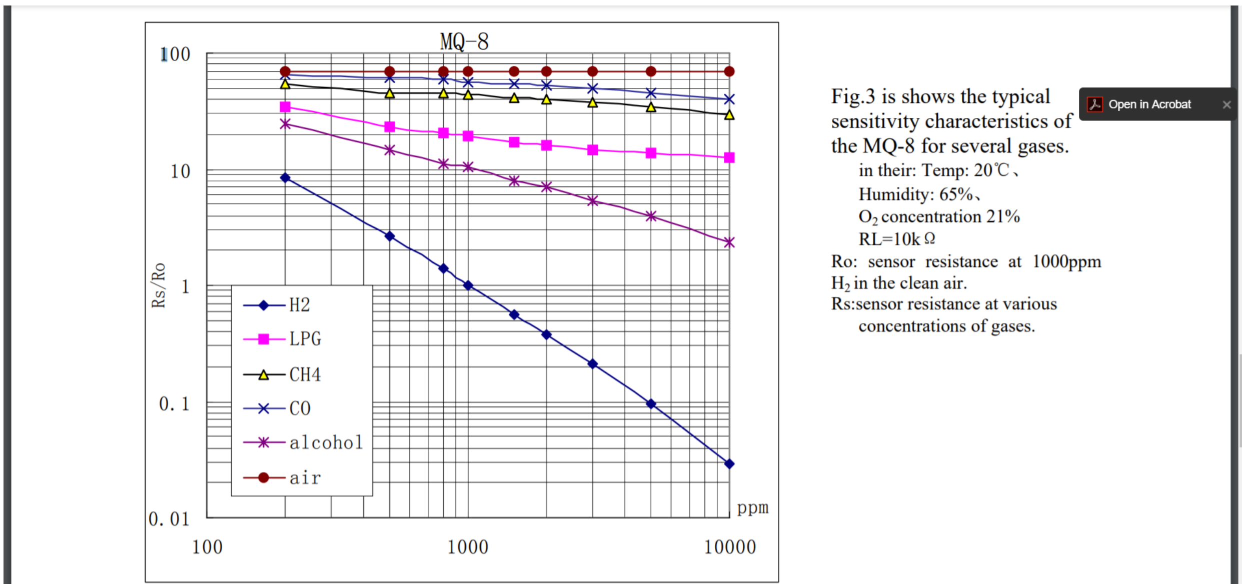
SENSOR PROPERTIES:(placed in hardware setup)

*The below diagrams belong to a company and represent how the voltage drops when the measure of gas exceeds the threshold.

Here are some features of the

1. MQ8 sensor:

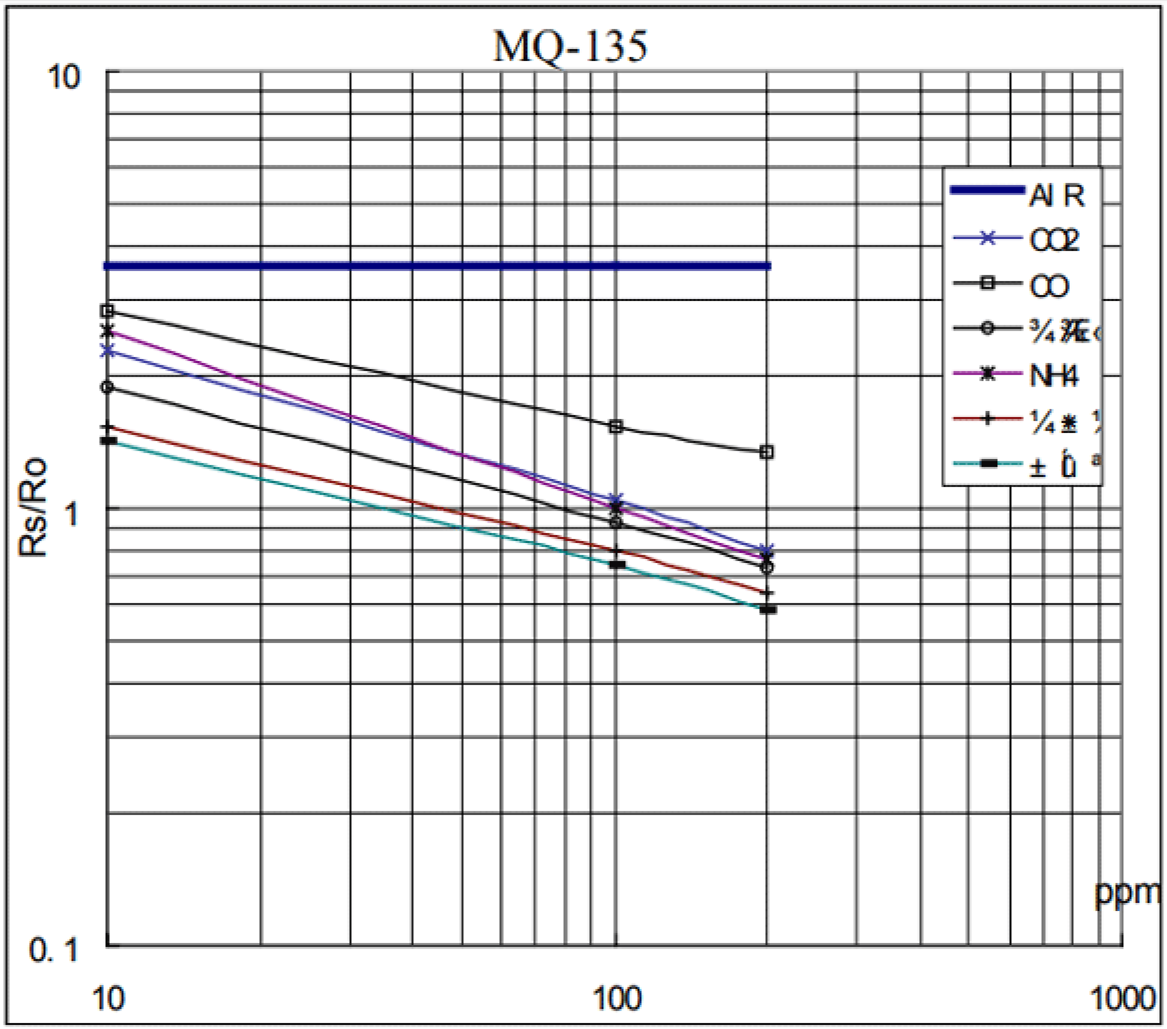
- High sensitivity to Hydrogen (H2)
- Small sensitivity to alcohol, LPG, cooking fumes



Other sensors:

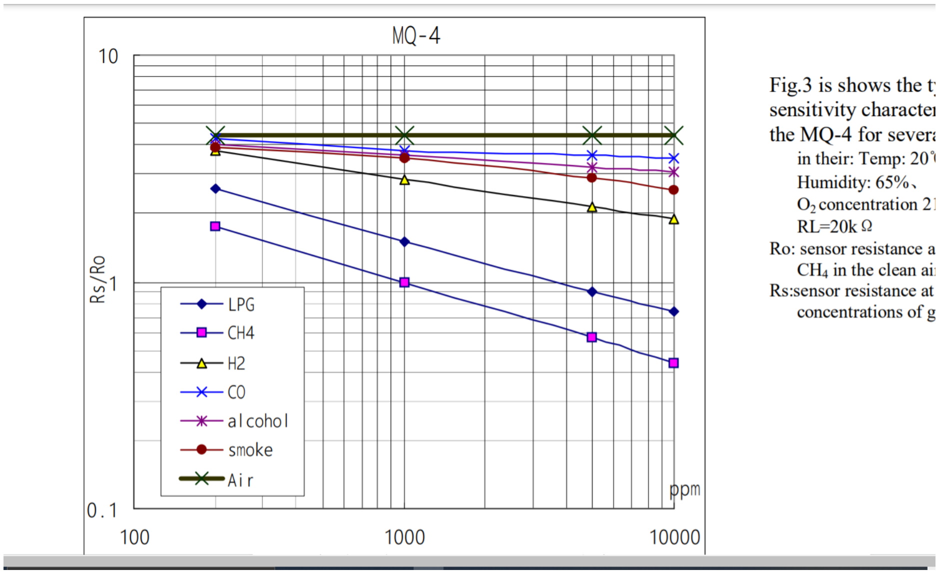
2. MQ135 sensor:

- Used to detect leakage/excess of gases like Ammonia, nitrogen oxide, alcohols, aromatic compounds, sulfide, and smoke.
- Air quality monitors.



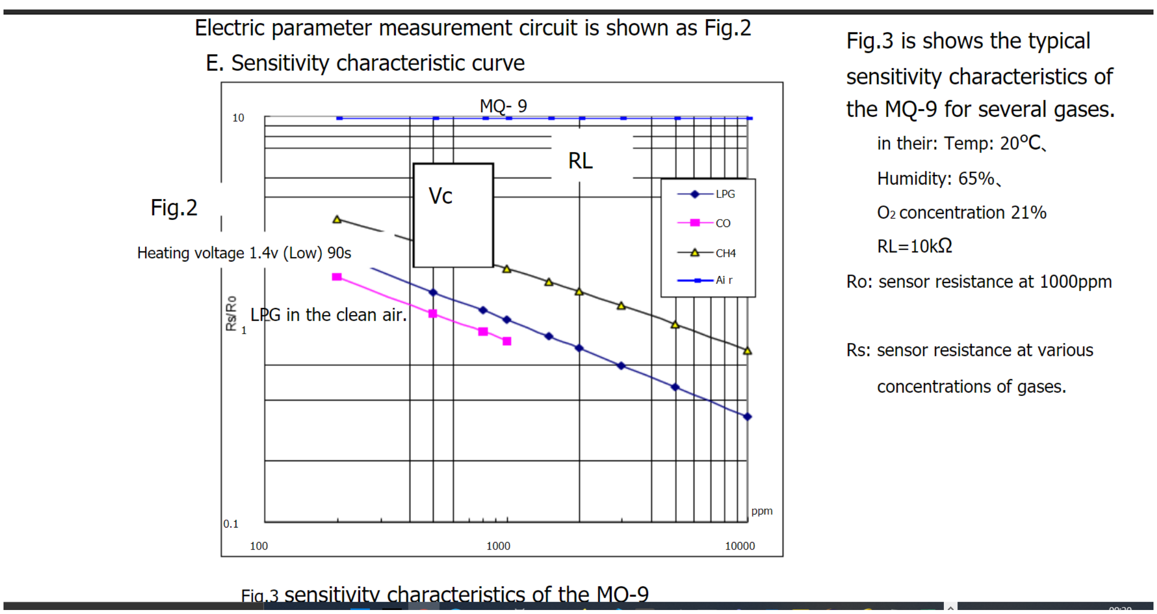
3. MQ4 sensor:

- High sensitivity to Methane, CNG Gas



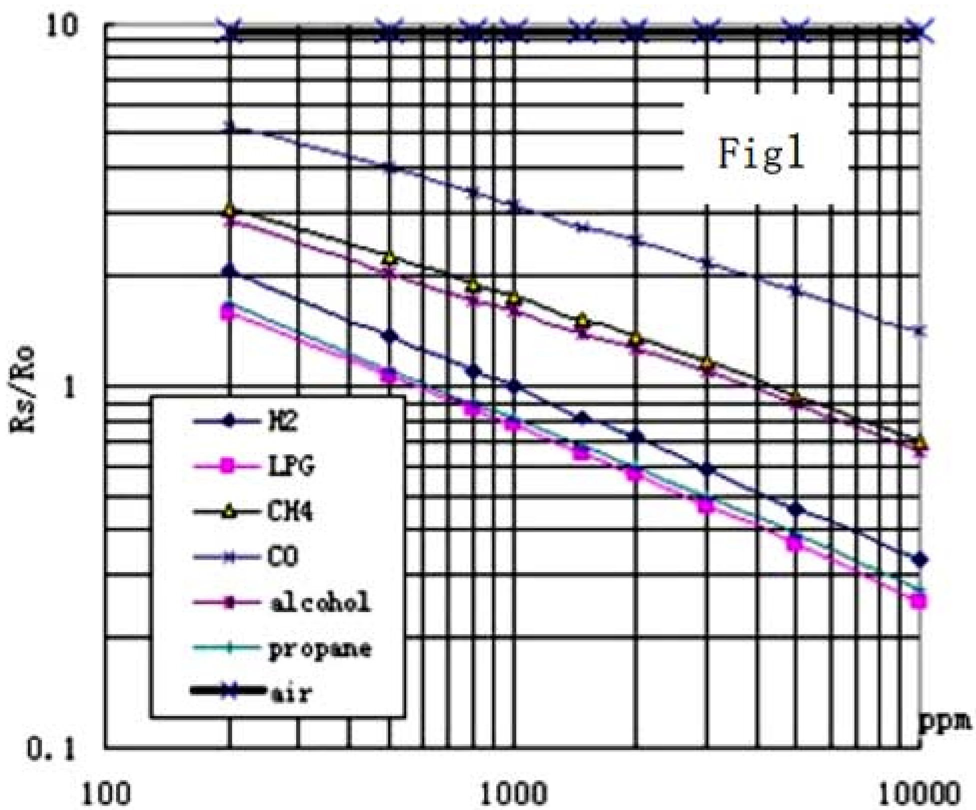
4. MQ9 sensor:

- High sensitivity to carbon monoxide and CH4，LPG



5. MQ2 sensor:

- Detects or measure Gases like LPG, Alcohol, Propane, CO, and even methane
- Air quality monitor



6. MQ3 sensor:

- Sensitive to Alcohol, Ethanol, Smoke

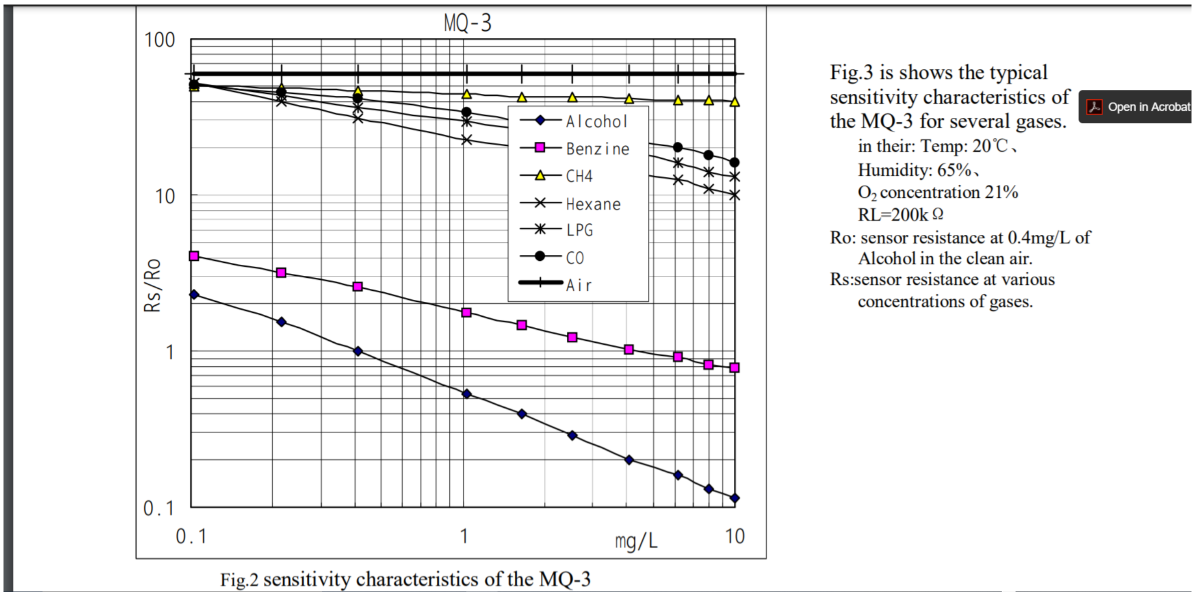


Fig.2 sensitivity characteristics of the MQ-3

DATASET:

- For example, MQ8A represents the analog value of sensor MQ8, and MQ8D represents the digital value for the same.

	MQ8A	MQ8D	MQ135A	MQ135D	MQ9A	MQ9D	MQ4A	MQ4D	MQ2A	MQ2D	MQ3A	MQ3D	output
0	583	0	268	1	203	1	462	1	241	1	660	0	1
1	589	0	272	1	207	1	462	1	247	1	659	0	1
2	594	0	275	1	211	1	464	1	250	1	659	0	1
3	598	0	277	1	215	1	468	1	253	1	659	0	1
4	603	0	279	1	219	1	473	1	256	1	658	0	1

RESULTS AND DISCUSSIONS:

DECISION BOUNDARY PLOTS:

The analog values are trained for each of the features. We validate the hypothesis to find which feature actually drives the results. Therefore, a decision boundary between positive and negative examples is observed and we see where the model fails for evaluating.

In the following plot images, different features were compared against each other on the same test and training data as per an 80-20 split.

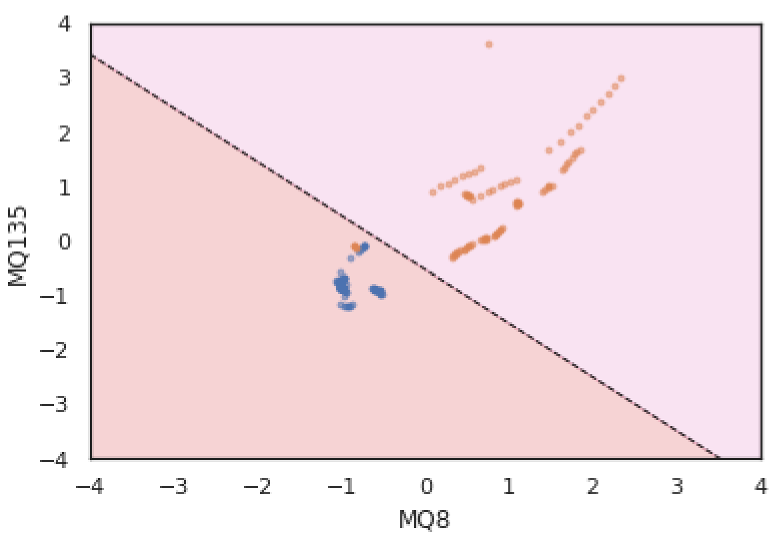
The below formula is used for finding the decision region where the hypothesis is neither positive nor negative.

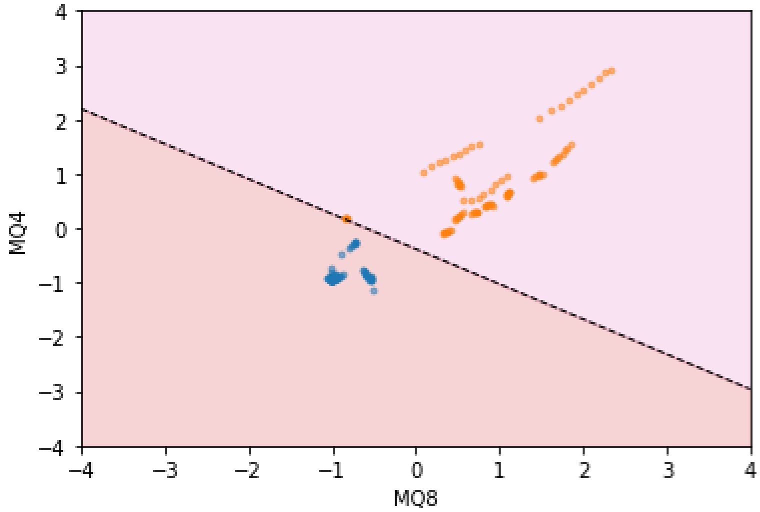
Where w0, w1, w2 are the parameters, and x1, x2 are the features, and h, the hypothesis. For example in the first plot, we are directly able to prove the hypothesis that if X1 (in the first case MQ8) >0 the food is classified as spoilt and if X1<0, the data points are classified as fresh. Therefore, MQ8 is a good feature.

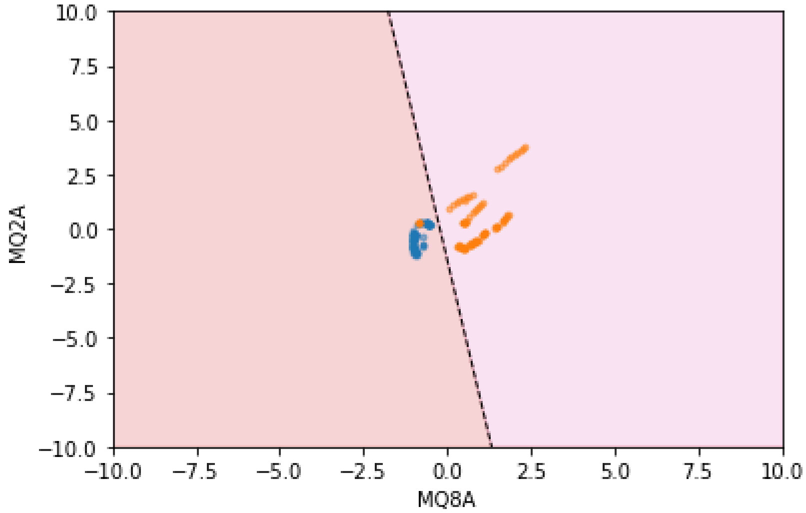
(Legend: The red dots are spoiled data points and blue are unspoiled and all the data points are normalized).

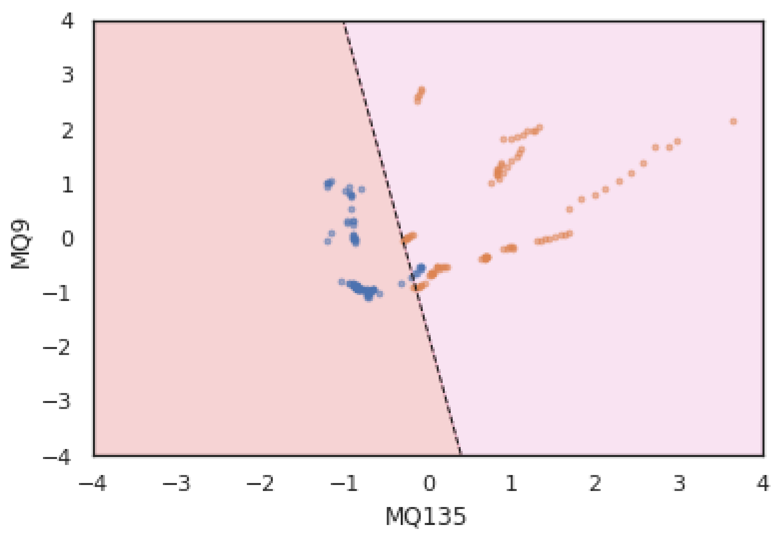
Similarly, we compare all the plots and look for the best features that contribute to the classification.

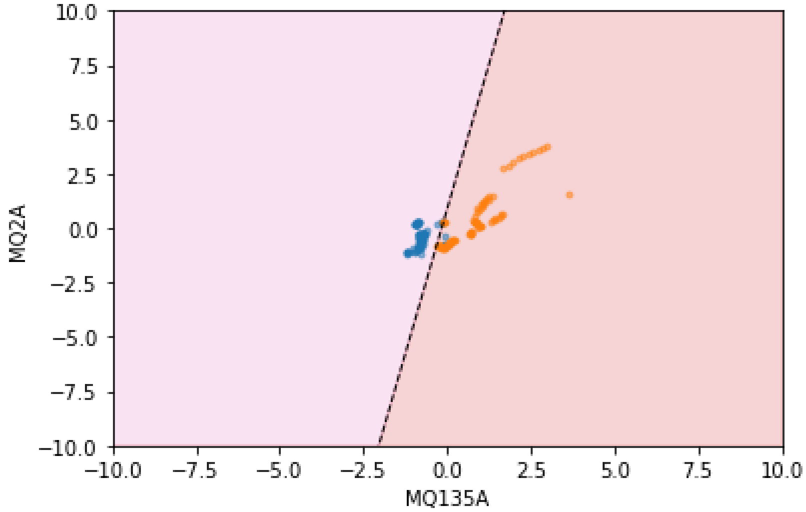
(figures, equations should be numbered and headings to be mentioned)

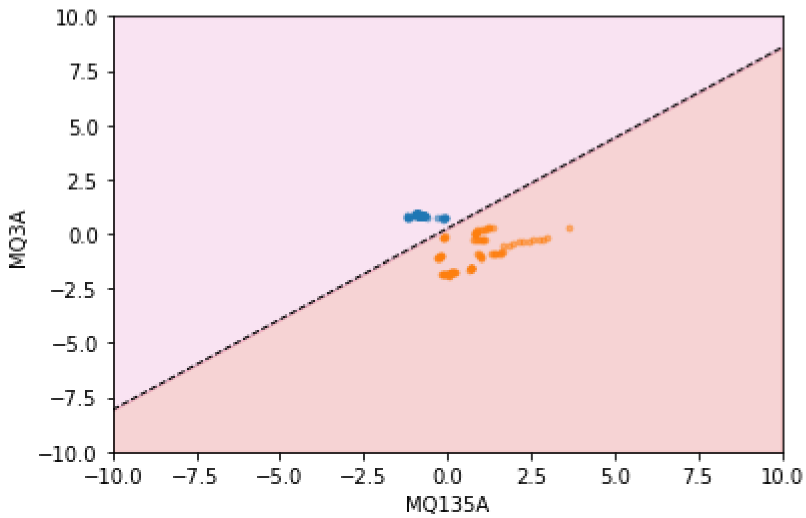


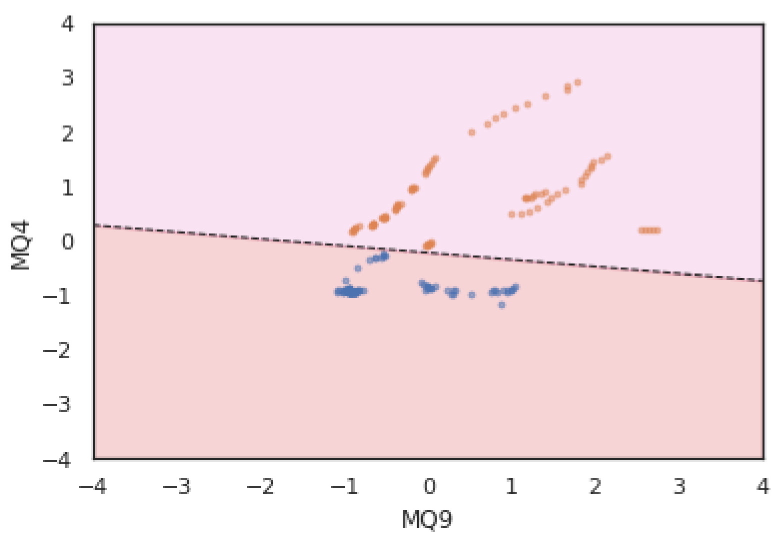

MQ4
MQ8

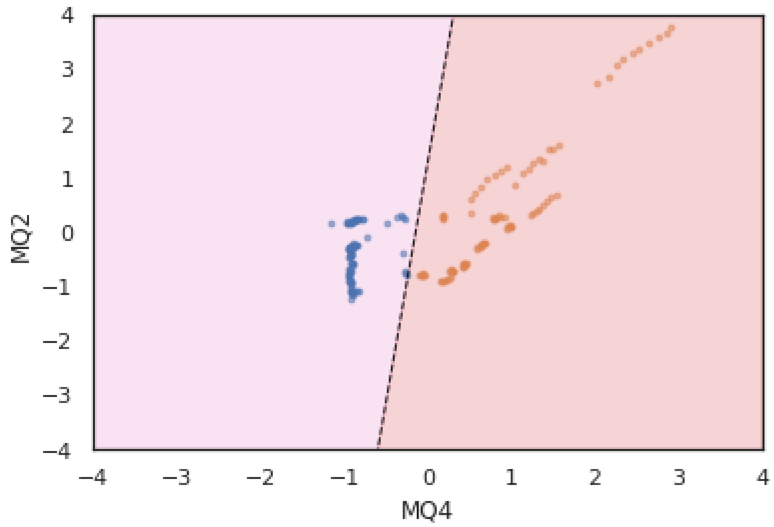

MQ2A
MQ8A

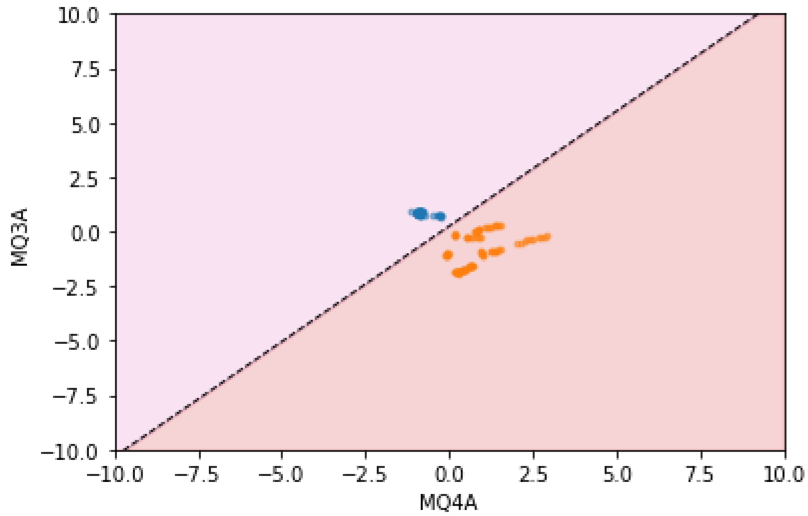

MQ9
MQ135

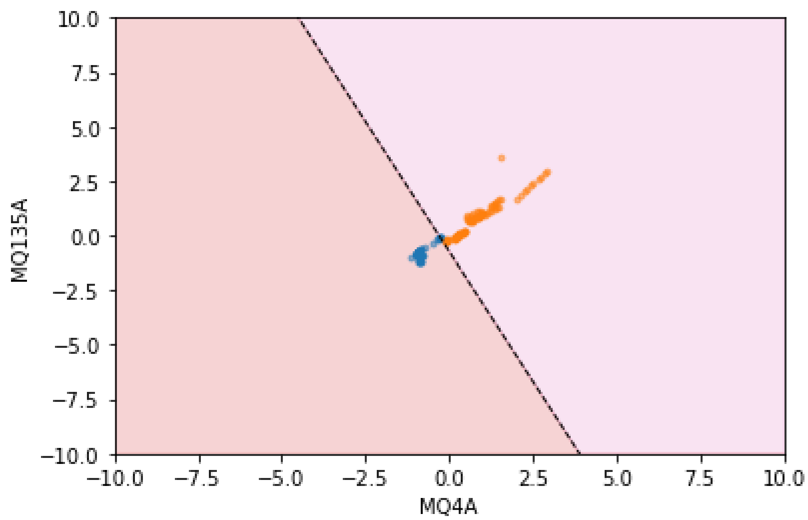

MQ2A
MQ135A

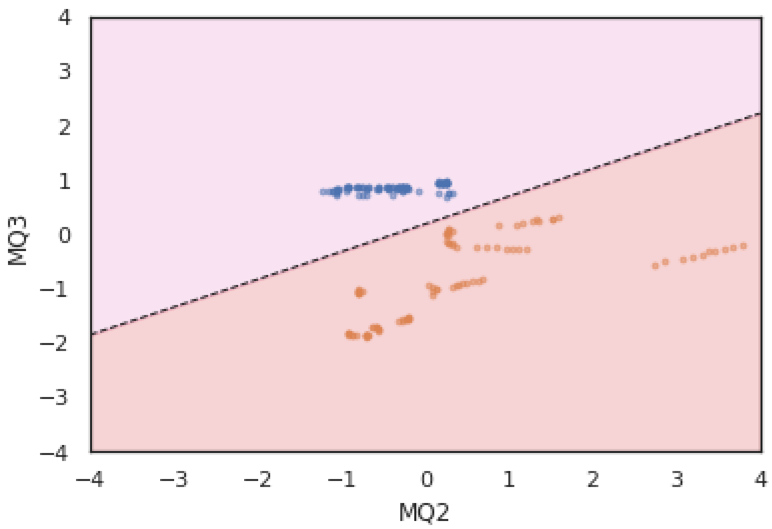

MQ3A
MQ135A


MQ4
MQ9


MQ2
MQ4


MQ3A
MQ4A


MQ135A
MQ4A

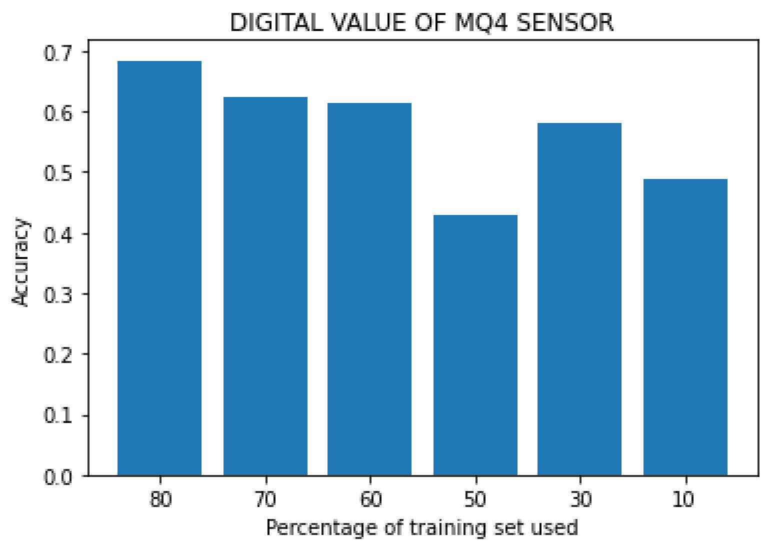


FURTHER PLOTS FOR EVALUATION:

It can be observed from the above plots that MQ8 and MQ4 sensors (next closest will be MQ3; therefore, we filter out the other sensors) classify the data points most accurately between the two classes. Since we do not know apriori on which features we need to go with, we experiment with different features individually (or sometimes in combination with some features; trying out all permutations and combinations) by training different models (with different train and test data) and look for the one with maximum accuracy.

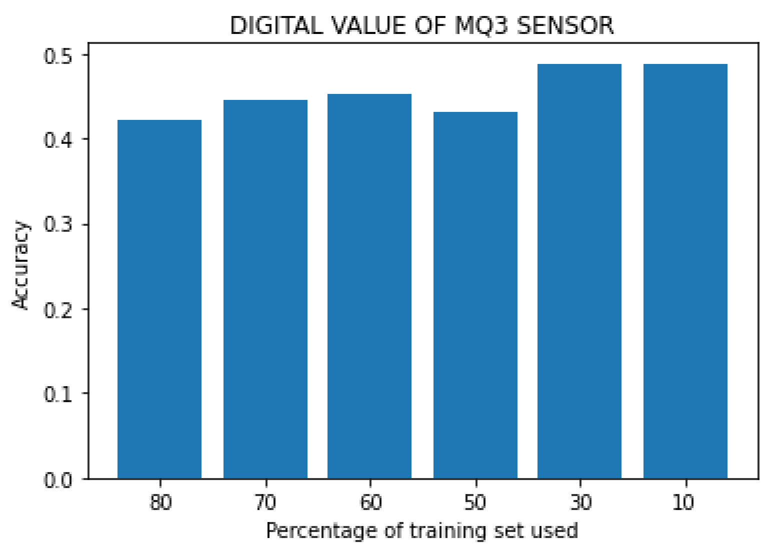
1. The digital value of the MQ4 sensor alone is taken as a feature and compared against the percentage of the training set used.
   *The accuracy is scaled down to lie between 0 and 1.

[0.6842105263157895, 0.625, 0.6133333333333333, 0.43010752688172044, 0.5801526717557252, 0.4880952380952381]

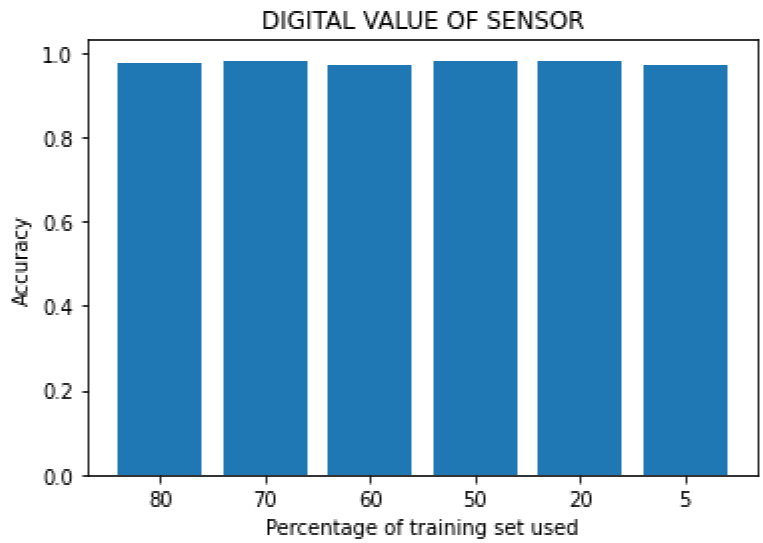


2. Similarly, the process is repeated (look at the labels for info)

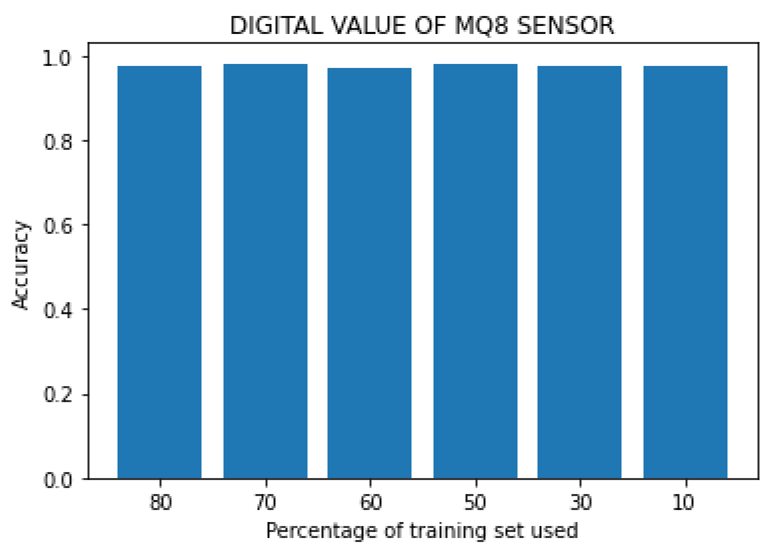
[0.42105263157894735, 0.44642857142857145, 0.4533333333333333, 0.43010752688172044, 0.48854961832061067, 0.4880952380952381]



[0.9736842105263158, 0.9821428571428571, 0.9733333333333334, 0.978494623655914, 0.9770992366412213, 0.9761904761904762]

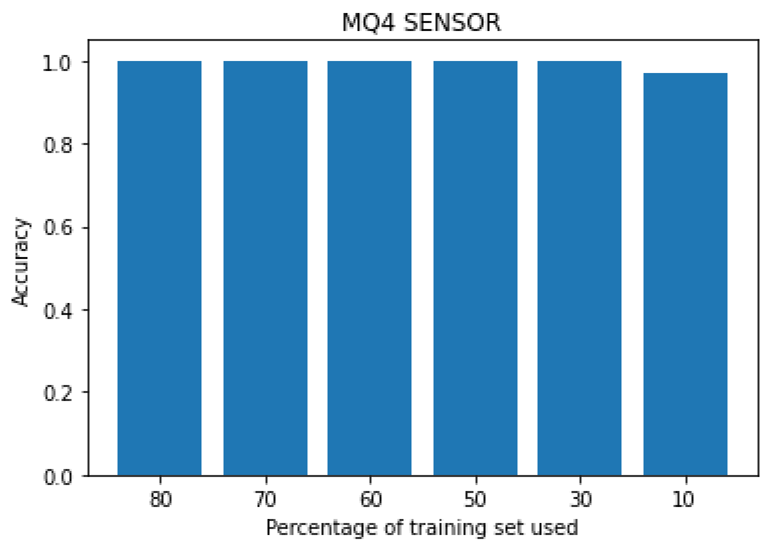


If all the sensors are used as features to the model we observe an accuracy of about 97%. But we still need to look for the best features for the model.

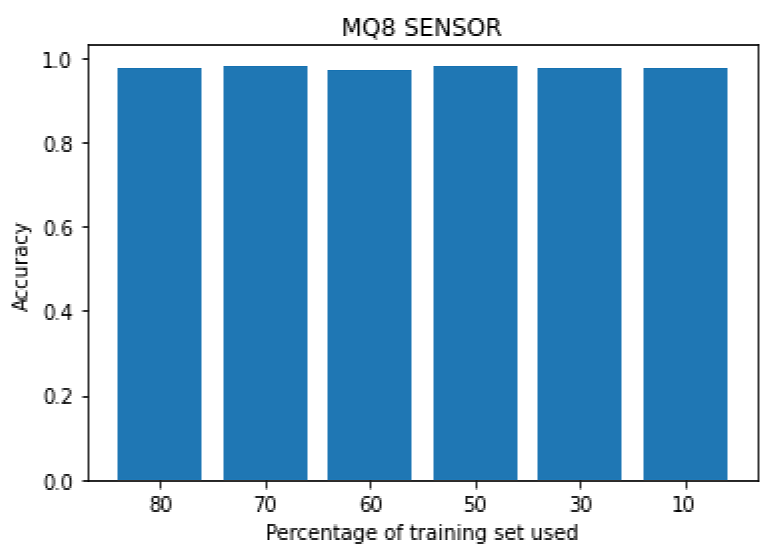


We observe that MQ8 has the highest accuracy within its digital value followed by MQ4.
Now we compare the analog values of the features similarly.

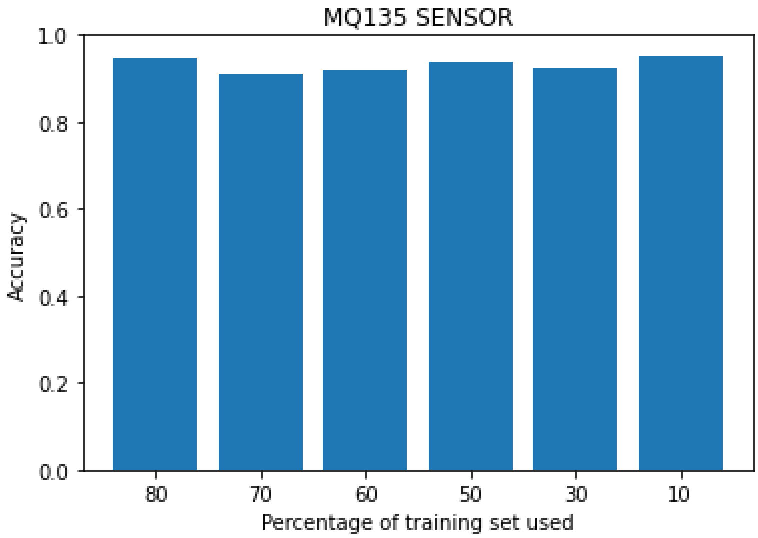
[1.0, 1.0, 1.0, 1.0, 1.0, 0.9702380952380952]

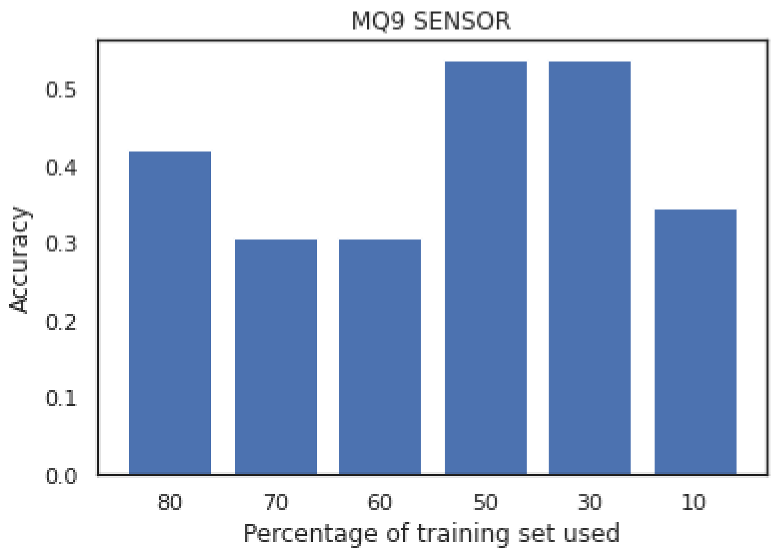


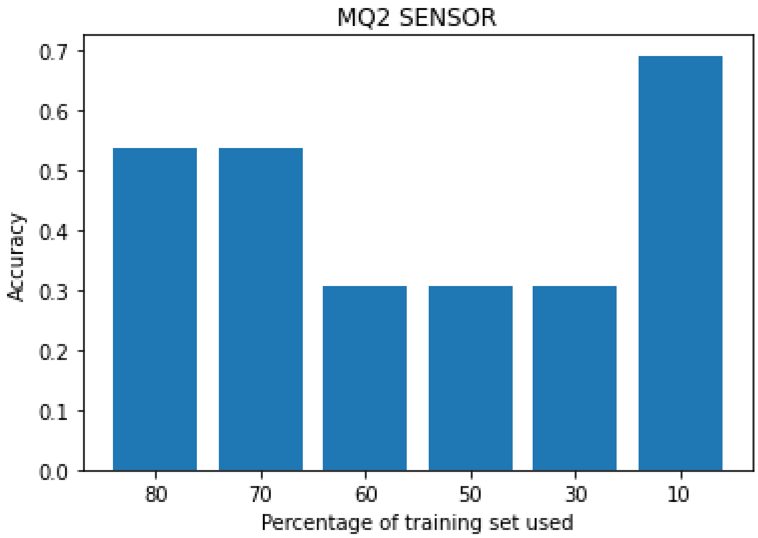
[0.9736842105263158, 0.9821428571428571, 0.9733333333333334, 0.978494623655914, 0.9770992366412213, 0.9761904761904762]



[0.9473684210526315, 0.9107142857142857, 0.92, 0.9354838709677419, 0.9236641221374046, 0.9523809523809523]

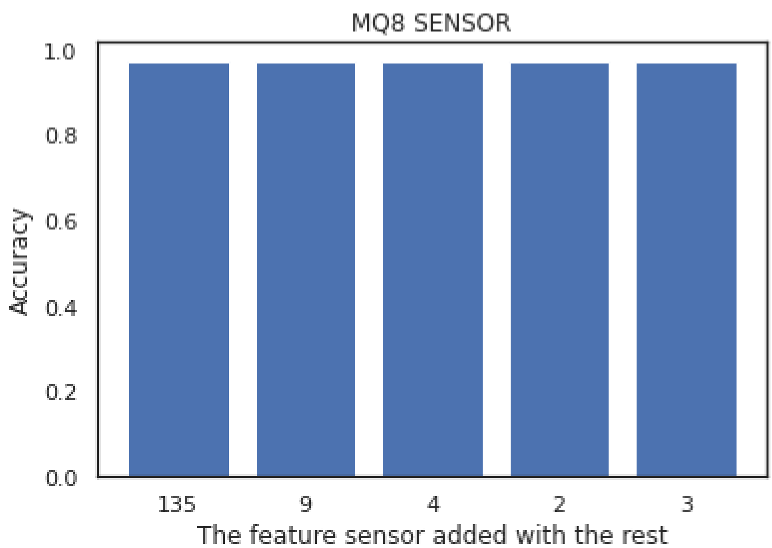

MQ135 SENSOR
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
80
70
60
50
30
10
Percentage of training set used

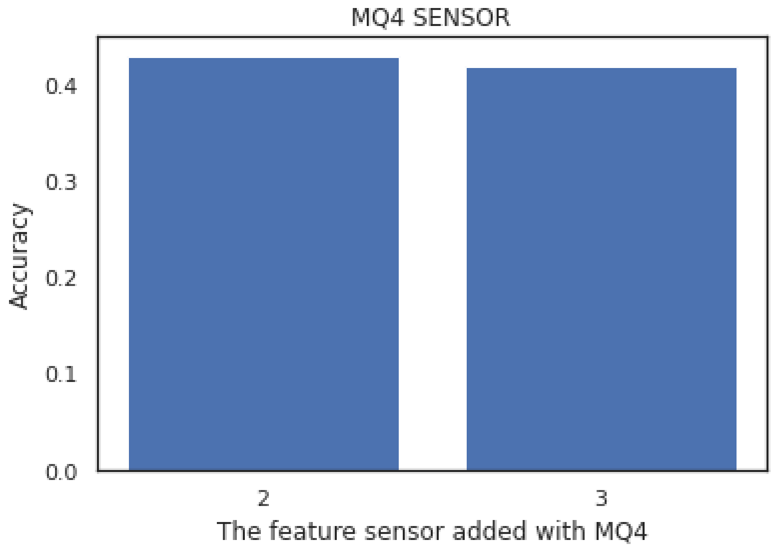

MQ9 SENSOR
0.5
0.4
0.3
0.2
0.1
0.0
Accuracy
80
70
60
50
30
10
Percentage of training set used

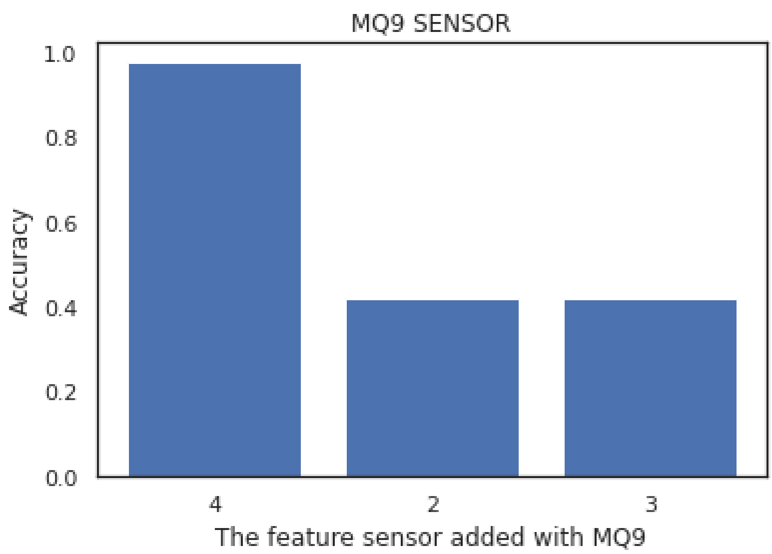


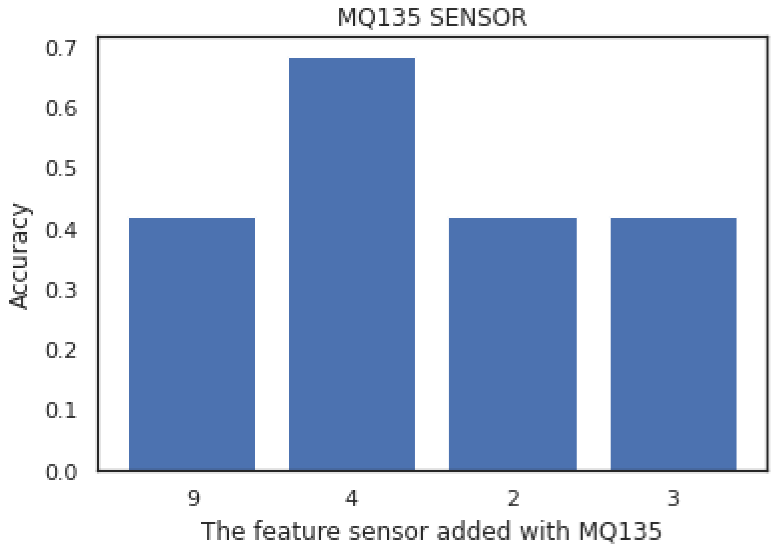
We similarly find that MQ4 followed by MQ8 has the highest accuracy.

To get conclusive evidence, we try adding each feature in combination with another feature. The x-axis shows the features (sensor no; eg: if x-axis value= 4 it means MQ4 sensor is taken, and so on..) that are added independently with the feature in the heading of the plots.




MQ4 SENSOR
Accuracy
0.0
0.1
0.2
0.3
0.4
2
3
The feature sensor added with MQ4

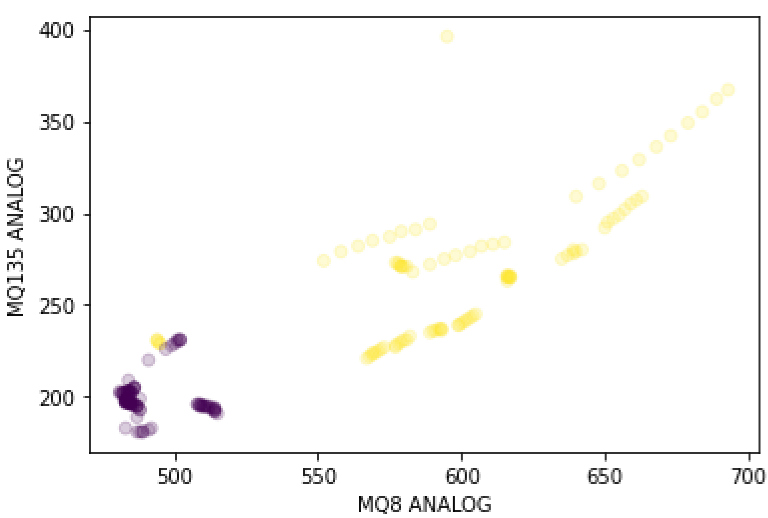

MQ9 SENSOR
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
4
2
3
The feature sensor added with MQ9

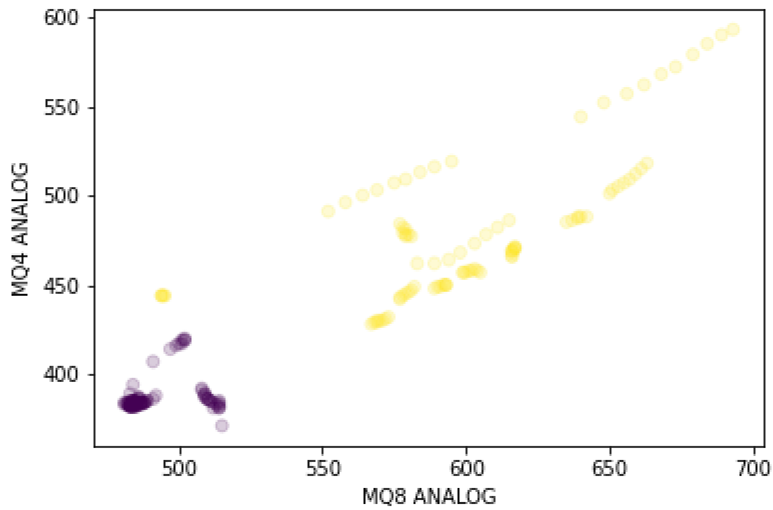


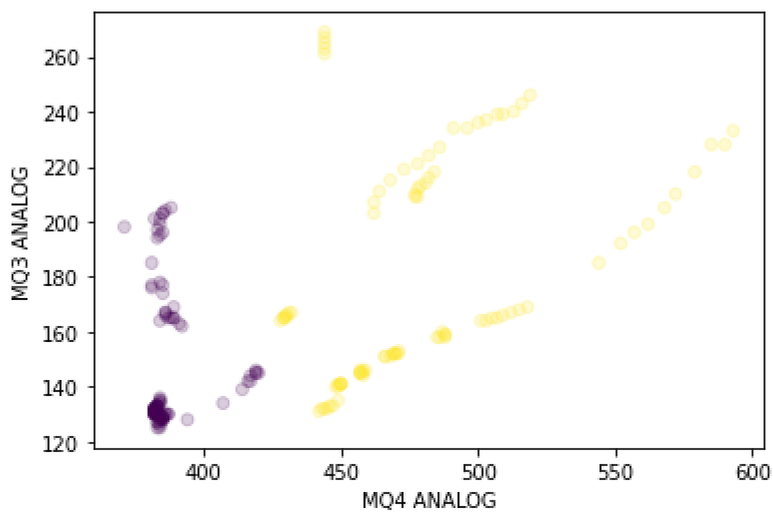
FURTHER SCATTER PLOTS:

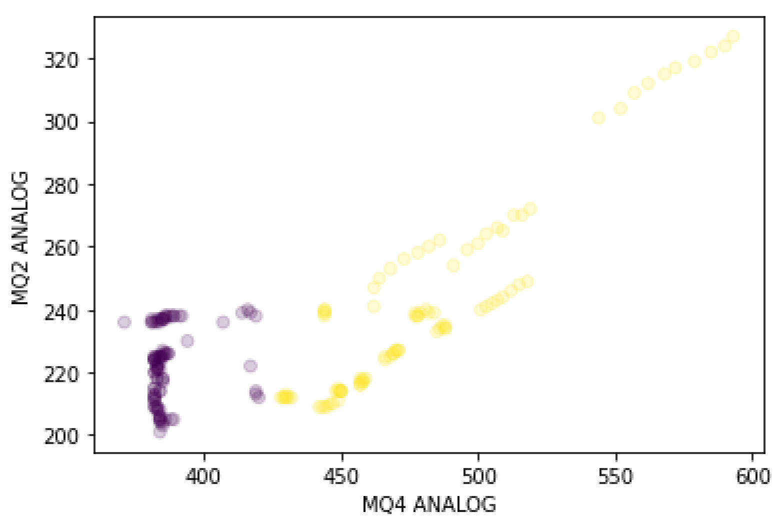
These are some scatter plots for further visualization and the underlining of the conclusion.

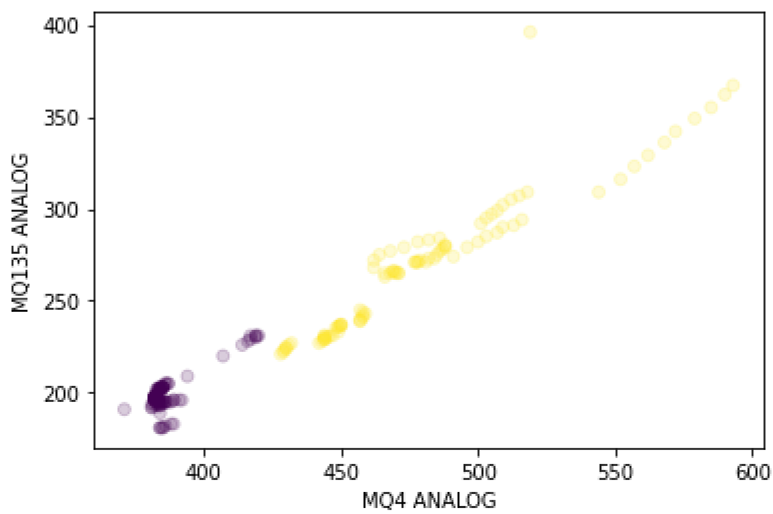
With the plots, we can comfortably conclude that the MQ8 sensor contributes the highest to the identification of the spoilage of food as the accuracy is the highest when it is added as a feature to the model. The next most important feature for the model would be the MQ4 sensor, which from the above plots brings in reasonable accuracy when added. MQ9 and MQ2 sensors contribute the least and to the classification, and MQ135 contributes reasonably to the identification of the spoilage. Now, a better understanding of what each sensor does is necessary.




MQ4 ANALOG
MQ8 ANALOG


MQ3 ANALOG
MQ4 ANALOG


MQ2 ANALOG
MQ4 ANALOG



CONCLUSION:

A thorough experimentation was conducted using the proposed hardware setup on a dataset with …..The results obtained revealed that if a food is spoiled a majority of hydrogen, followed by good amounts of methane, and CNG, traces of ammonia, nitrogen oxide, aromatic compounds, sulfide, and smoke get emitted in addition to the air quality level going down. The work could be further extended to include sensors capable of working under certain humidity, temperature, and diff voltages for better results.

REFERENCES:

1. S. Jayanthy, S. Valli Ayswaria, M. Vaishali, and P. Udhaya Pooran: A HIGH QUALITY EMBEDDED SYSTEM FOR ASSESSING FOOD QUALITY USING HISTOGRAM OF ORIENTED GRADIENTS
2. Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa: https://www.researchgate.net/profile/Kiyoharu_Aizawa/publication/266357771_Food_Detection_and_Recognition_Using_Convolutional_Neural_Network/links/542d52930cf29bbc126d2897/Food-Detection-and-Recognition-Using-Convolutional-Neural-Network.pdf?origin=publication_detail

3. Dr. G. Rajakumar, Dr. T.Ananth Kumar, Dr. T.S. Arun Samuel, Dr. E.Muthu Kumaran: IoT BASED MILK MONITORING SYSTEM FOR DETECTION OF MILK ADULTERATION
4. Roselle Marian M. Lafuente, Mark B. Malonzo, Wen-Yaw Chung: https://ieeexplore.ieee.org/document/9072715
5. Jie Hao and Jennifer Lewis Priestley: https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1002&context=dataphdgreylit

6. Towards data science: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:~:text=KNN%20is%20a%20non%2Dparametric%20model%2C%20where%20LR%20is%20a,can%20only%20output%20the%20labels.
7. Sang-Il Choi, Taekyu Eom, and Gu-Min Jeong: Gas Classification Using Combined Features Based on a Discriminant Analysis for an Electronic Nose
8. Seongyoun Woo and Chulhee Lee: Incremental feature extraction based on decision boundaries | Pattern Recognition
9. Chulhee Lee and David A. Landgrebe: Feature Extraction Based on Decision Boundaries
10. CHAO-YING JOANNE, PENG KUK LIDA LEE, and GARY M. INGERSOLL: (PDF) An Introduction to Logistic Regression Analysis and Reporting
11. Najam Ul Hasan , Naveed Ejaz · Waleed Ejaz and Hyung Seok Kim: https://www.mdpi.com/1424-8220/12/11/15542-
12. Samson Otieno Ooko : (PDF) A Comparison of Arduino, Raspberry Pi and ESP8266 Boards
13. Leo Louis : (PDF) WORKING PRINCIPLE OF ARDUINO AND USING IT AS A TOOL FOR STUDY AND RESEARCH
14. Anand Nayyar, Vikram Puri, Dac-Nhuong Le: A Comprehensive Review of Semiconductor-Type Gas Sensors for Environmental Monitoring
15. Shijie Tang, Lisha Chenb, Kam-Wah Tsuic, Kjell Doksumd: http://www.stat.yale.edu/~lc436/Tang_Chen_etal_2012_CSDA.pdf

*The research papers are accessed using sci hub.