FoodAlytics: a Formalin Detection System Incorporating a Supervised Learning Approach

Swapnil Sayan Saha[1], Md. Sadman Siraj[1], Walid Bin Habib[2]
Department of Electrical and Electronic Engineering[1 2]
University of Dhaka[1 2]
Dhaka, Bangladesh
swapnilsayansaha@ieee.org[1], sadmansiraj.ss@gmail.com[1], walid.habib.1994@ieee.org[2]

***Abstract*— The rampant use of dangerous levels of formalin in food items has incited anxiety and concern amongst the Bangladeshi citizens. Sporadic monitoring, amalgamated with the paucity of well-grounded and inexpensive testing devices has created a need for the citizens to address the issue themselves. Impelled by promising results from a survey, FoodAlytics, a portable, economical and non-contact formalin detection system has been developed using locally sourced COTS components. Integrated with our own android application and bundled with the Grove VOC HCHO gas sensor, a recent innovation, our system has been able to detect up to 50 ppm of formalin. Polynomial regression, linear regression and the Levenberg-Marquardt algorithm, three supervised machine learning algorithms have been incorporated in our system to accurately predict the correct concentration of formalin at all temperatures. By applying logistic regression and support vector machines, our system is also able to correctly classify between artificially added and naturally formed formalin.**

Keywords—Supervised Learning; Machine Learning; Curve Fitting; Microcontrollers; Mobile Applications; Gas Detectors

I. Motivation and Market Analysis

The widespread adulteration of fresh fruits, vegetables and fish with harmful concentrations of formalin has become a severe issue in Bangladesh due to the health hazards associated with it [1][2]. Excessive concentrations of formalin has been found in almost all fruits available in the local market, thanks to lack of sufficient monitoring and raids [3]. Analogous to this issue, the authors decided to conduct a survey on 150 random subjects in and around Dhaka city to find out what our citizens think should be the impeccable solution to this problem. Figure 1 shows partial results of the survey.

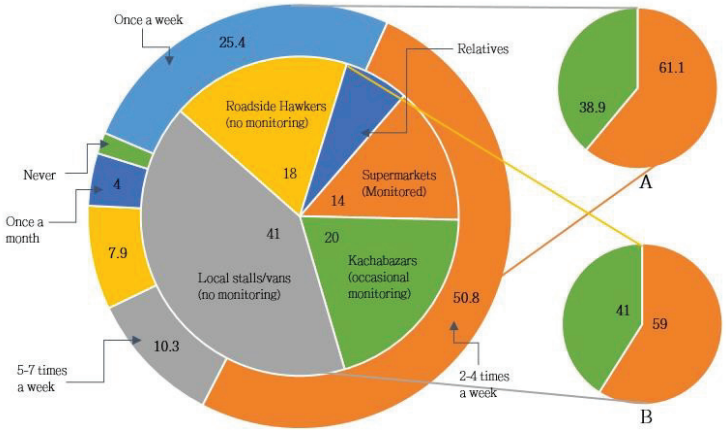


Figure 1. Partial results of the survey (quoted in percentages). The inner chart shows the sources of fruits and vegetables for the subjects while the outer chart shows the frequency of shopping of fresh fruits and vegetables. Chart A shows that 61.1% of our subjects shop regularly (more than 2-4 times a week) while Chart B shows that 59% of our subjects shop from sources with no monitoring for adulteration.

From Figure 1, it is clear that most of our citizens regularly shop for fruits and vegetables while more than half of them source their buys from roadside hawkers and local stalls/vans which are not monitored at all by our mobile courts. 66% of our subjects have confirmed that they think their buys are adulterated with toxic levels of formalin, while 10% of these subjects have checked for presence of formalin themselves and all of them found formalin in their fruits and vegetables.

The outcome of the survey was further intensified when our subjects stated that the core problem was the lack of a dedicated formalin detection system that can be used by the general mass for detection of adulterated food items and reporting to concerned authorities. Furthermore, 91% of our subjects wanted a device in the market that the citizens themselves can use to check for adulterated food. Figure 2 outlines the specifications of the package as suggested by our subjects.

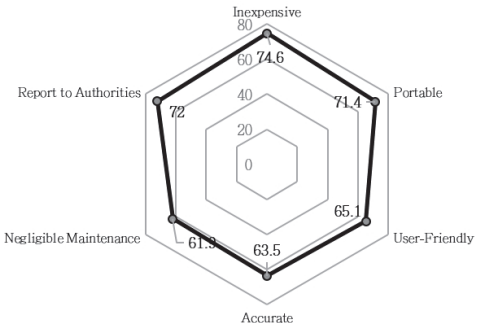


Figure 2. Specifications of the formalin detection system as suggested by our subjects. The figures are quoted in percentages.

Pertaining to the results of the survey, the authors have developed a low-cost, handy and comprehensible formalin detection package entitled "FoodAlytics" bundled with an android application. Supervised learning algorithms have been incorporated into our system to correctly classify between artificially added and naturally formed formalin in food items, as well as accurately infer the correct concentration of formalin present regardless of surrounding temperature.

II. Literature Review

Current handheld formalin detectors (Z-300) used by mobile courts in Bangladesh have a narrow concentration range (~1-30 ppm), practically importable, abstruse for everyday use, provides frequent erroneous readings and costs 1175 USD [4] [5]. In contrast, a chemical kit manufactured by Bangladesh

978-1-5386-2175-2/17/$31.00 ©2017 IEEE

The authors are grateful towards IEEE Bangladesh Section Humanitarian Activities Committee (IEEE BDS HAC) for providing necessary guidance and financially assisting in the making of this paper.

Council of Scientific and Industrial Research (BCSIR) [6] is also available in the market for 3 USD, however, this kit provides only a qualitative deduction on whether formalin is present or not, unsuitable for spontaneous use and can detect up to 5 ppm of formalin. Our designed system, built within 26 USD and integrated with our own android application can detect up to 50 ppm of formalin. The principles of electronic sensing of formaldehyde via gas sensors are closely related to a few independent studies [7] [8] [9]. However, our system, designed using a novel sensor and a different package, is also compensated for temperature via supervised machine learning algorithms. Furthermore, the package can also distinguish between naturally formed formaldehyde in food items and artificially added formaldehyde.

III. Materials and Methods

Figure 3 shows the block diagram of the designed formalin detection system.

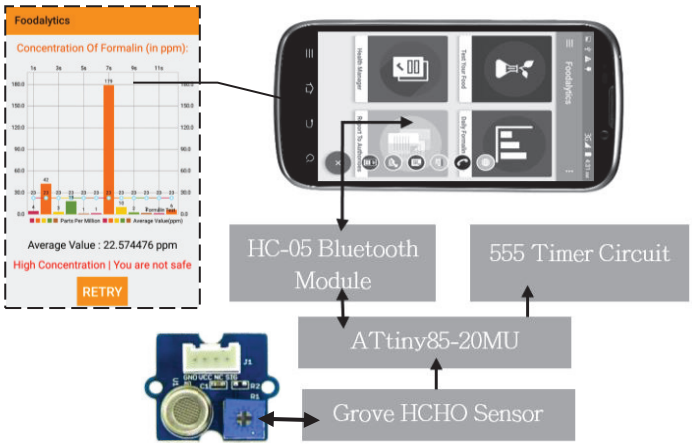


Figure 3. Block Diagram of the Entire System. The developed Android application in action and the Grove HCHO Sensor is also shown.

Figure 4 shows the partial schematic diagram of the system without the battery and charger circuit.

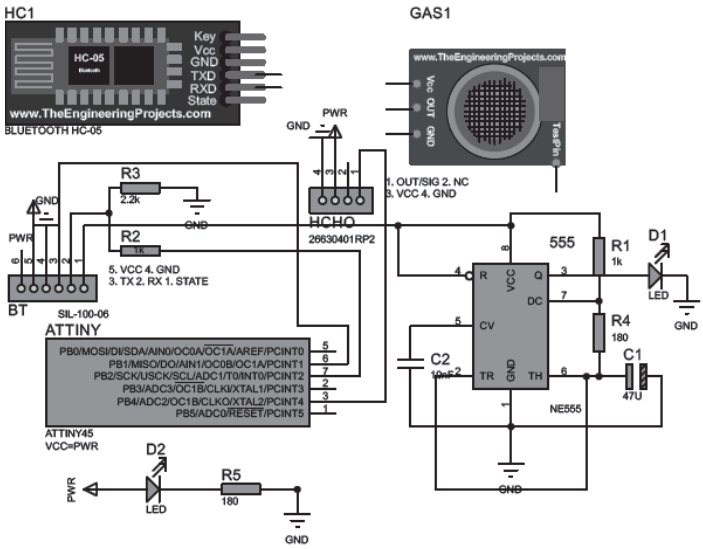


Figure 4. Schematic diagram of the entire system.

The HCHO sensor [10] by Grove was designed to detect formaldehyde gas. However, since formalin is a volatile 40% aqueous solution of formaldehyde [1], the sensor is able to accurately detect presence of formalin as well. The voltage output of the HCHO sensor is exponentially proportional [10] to the concentration of formaldehyde gas. The ATtiny85 microcontroller feeds the output voltage to the Bluetooth module, which transmits the data to the android application. The Timer Circuit is used to flash an indicator LED once the system is connected to the application via the Bluetooth Module.

IV. Results

A. Experimental Setup and Results for Raw Formalin

The goal of the first experiment was to find out the response of the system for different concentrations of formalin. The setup is shown in Figure 5.

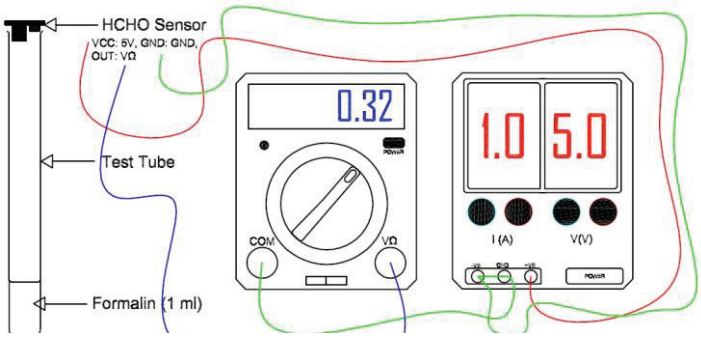


Figure 5. Computer Aided Drawing of the experimental setup. The multimeter was set to 2V DC and was used to measure the output voltage of the sensor when exposed to different concentrations of formalin. The DC power supply was used to supply 5V DC to the sensor.

Figure 6 outlines the results of the experiment. 10 trials were conducted at constant temperature (301 K) and constant relative humidity (48 % ± 2%) to find out the system response at 8 different concentrations of a constant volume (1 ml) of formalin. Using the Levenberg-Marquardt algorithm [11],

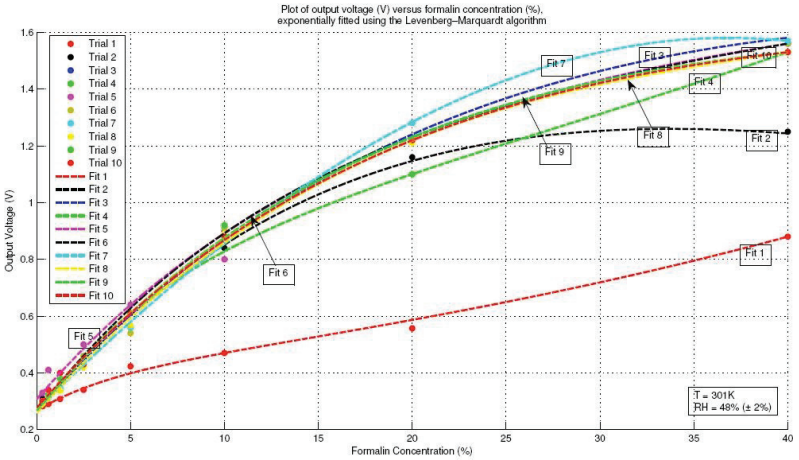


Figure 6. System response (V) versus concentration of formalin (%) for 10 trials at T = 301K and RH = 48 % ± 2%. A family of exponential curves was generated for the trials using the Levenberg-Marquardt algorithm.

which is an advanced and elegant optimization algorithm, a family of exponential curves were fitted to the results.

From the curves shown in Figure 6, the empirical mathematical relation between the output voltage (V) and formalin concentration (x) is given by:

$$V = ae^{bx} + ce^{dx} \quad (1)$$

where a, b, c, d are constants; $a, b, c, d \in \mathbb{R}$ and $|b| << 1$ and $|d| << 1$; $0 \leq x \leq 40$

Trial 1 is clearly anomalous and is thus omitted. Using data from Trials 2-10 as training set, regularized polynomial regression (with $\lambda = 1$) (a popular supervised learning algorithm) [12] and the Levenberg-Marquardt algorithm was applied to the training set. The results are shown in Figure 7.

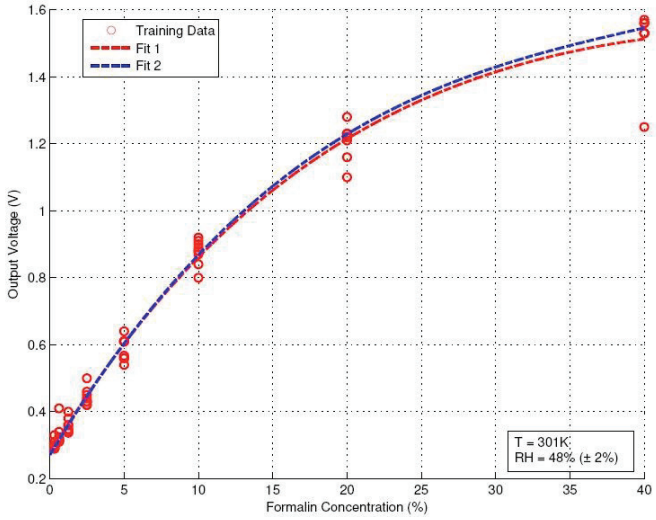


Figure 7. Plot of output voltage (V) versus formalin concentration (%) after applying regularized polynomial regression (blue line) and the Levenberg-Marquardt algorithm (red line), after 400 iterations.

The empirical mathematical relation obtained from regularized polynomial regression is given by:

$$V = ax^3 + bx^2 + cx + d \quad (2)$$

where, $a = 1.276e\text{-}05$, $b = -0.001567$, $c = 0.07415$, $d = 0.2702$

The empirical mathematical relation obtained from Levenberg-Marquardt algorithm is of the form shown in equation (1). Since equation 1 is mathematically implicit, there is no definite mathematical rule to predict formalin concentration for a given sensor voltage. Since regularized polynomial regression provides a good approximation, we may use equation (2) instead of equation (1).

$$x = \{q + [q^2 + (r\text{-}p^2)^3]^{1/2}\}^{1/3} + \{q - [q^2 + (r\text{-}p^2)^3]^{1/2}\}^{1/3} + p \quad (3)$$

where, $p = -b/(3a)$, $r = c/(3a)$, $q = p^3 + (bc\text{-}3a(d\text{-}V))/(6a^2)$

Equation (3) can be used to predict the concentration of formalin (in percentage ranging from 0-40) in a given sample from the sensor's output voltage at 28°C.

B. *Experimental Results for Raw Formalin at Different Temperatures*

Using a 60W incandescent light bulb as a heater and a digital thermometer to monitor temperatures, the setup in Figure 5 was used to find out the effects of change in temperature on the system's response for a fixed concentration. of formalin. The output voltage of the system was recorded for three different fixed concentrations (10%, 20%, 40%) of 1 ml formalin from 300K (27°C) to 323K (50°C) at constant RH = 48 % ± 2%. The results are outlined in Figure 8.

Applying Cardano's Formula [13] to equation (2):

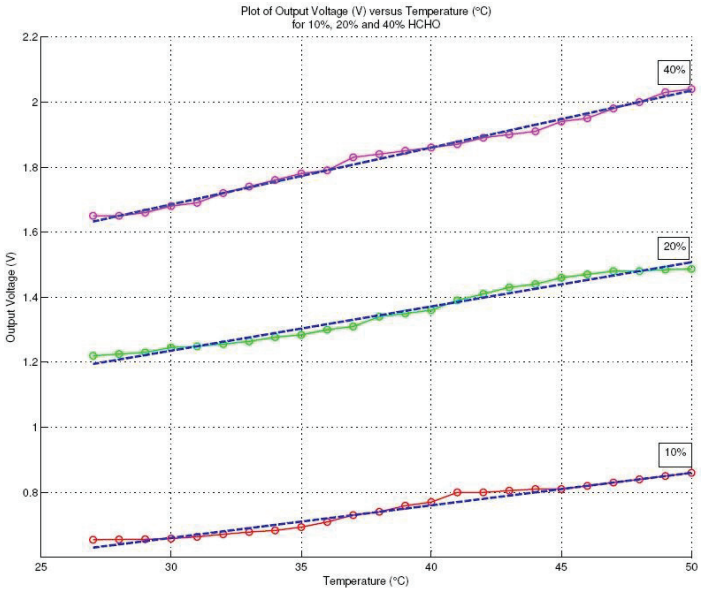


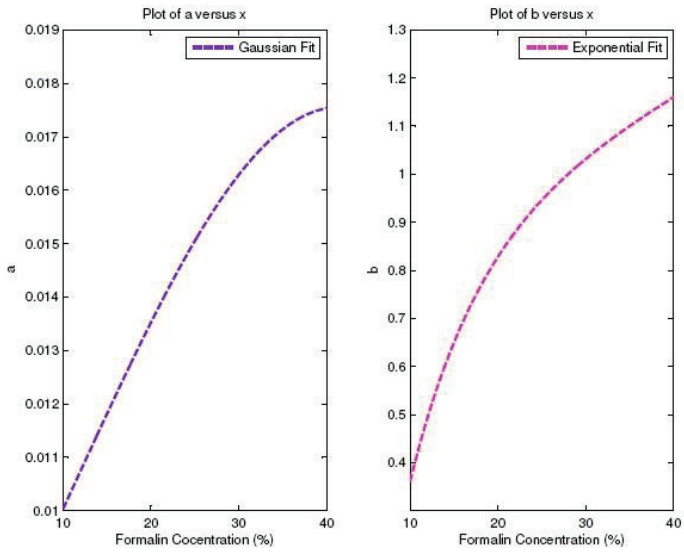
Figure 8. Plot of output voltage (V) versus temperature (°C) for three fixed concentrations of formalin (10%, 20%, 40%) from 300K (27°C) to 323K (50°C) at constant RH = 48 % ± 2%. The blue lines result from applying linear regression [14].

From Figure 8, we can conclude that the output voltage of the system increases slightly with increase in temperature. The empirical mathematical relation between V and t is given by:

$$V = at + b \quad (4)$$

where a and b are constants; a, b $\in \mathbb{R}$; t is temperature in °C and V is output voltage in volts.

It is also evident that change in voltage (a) is greater for higher concentrations (x) of formalin. By applying Gaussian process for machine learning by normalizing x [15] and the Levenberg-Marquardt algorithm, the curves obtained for rate of change of voltage (a) and b versus x are shown in Figure 9. The curves in Figure 9 can be used to compensate for temperature changes for high concentrations of formalin (>10%).



C. Using Logistic Regression and Support Vector Machines to obtain decision boundaries between natural and artificial formalin

Using BCSIR's chemical kit as reference, the response of the system for various fruits were noted using the setup in Figure 5. The results are outlined in Table I. The proponents marked with asterisk (*) were spiked artificially with formalin. The kit turns red if artificial formalin is present [7].

TABLE I. RESULTS FOR EXPERIMENT C.

Fruit	Output Voltage (V)	Chemical Kit	Artificial (1) Natural (0)
Orange	0.240	Yellow	0
Mango	0.242	Yellow	0
Mango	0.244	Yellow	0
*Raw Formalin (40%)	1.840	Red	1
Pure Water	0.200	Yellow	0
*Raw Formalin (20%)	1.180	Red	1
Banana	0.212	Yellow	0
Apple (Red)	0.188	Yellow	0
Apple (Green)	0.189	Yellow	0
*Mango	0.913	Yellow (failed to detect)	1
*Apple (Red)	0.700	Red	1
*Banana	0.890	Red	1
*Orange	0.750	Red	1
Orange	0.490	Yellow	0
Longan	0.180	Yellow	0
*Longan	1.03	Red	1

Logistic Regression [16] and Support Vector Machine (SVM) [17] were applied to the data in Table I and the decision boundaries shown in Figure 10 were obtained. Either of the boundaries can be used to distinguish between natural and artificial formalin in fruits. In our android application, we have used the boundary obtained from logistic regression.

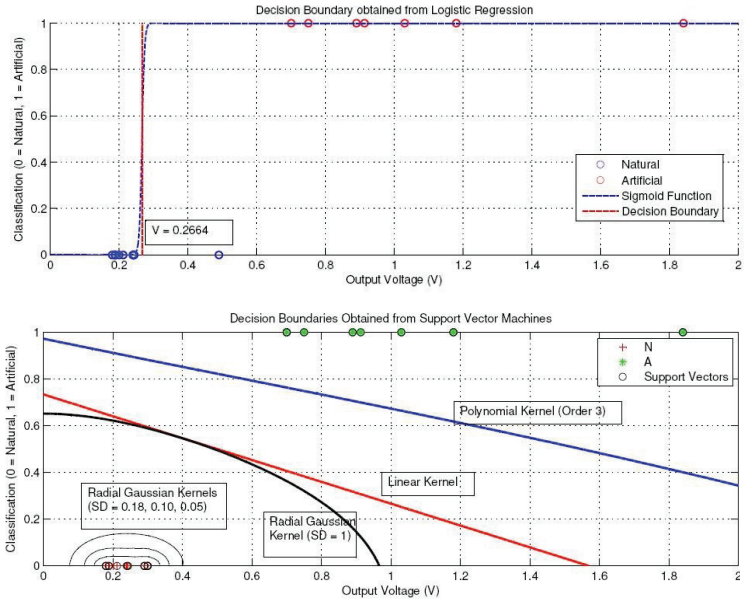


Figure 10. (Top) Decision boundary obtained by applying logistic regression. (Bottom) Decision boundary obtained by applying Support Vector Machines (SVM) for various kernels.

V. CONCLUSION AND FUTURE WORK

The models obtained in Figure 7 had a coefficient of determination (R^2) of 0.998 (the Levenberg-Marquardt algorithm) and 0.9952 (Polynomial Regression) with a sum of squared errors of prediction (SSE) of 0.02633 and 0.0631 respectively. Temperature compensation curves in Figure 9 yielded a mean R^2 of 0.9802 with a mean SSE of 0.00365. Differentiation between natural and artificial formaldehyde via logistic regression (Figure 10 – top) resulted in a R^2 of 1.00 and SSE of 0.00002234, while the differentiation via SVM yielded R^2 of 1.00 and SSE of 0. Thus, we can conclude that application of machine learning algorithms is certainly validated, increasing the capability of the system, which is the first of its kind capable of differentiating between natural and artificial formalin at all temperatures; all within a portable, inexpensive, accurate and user-friendly package.

Some topics of future work include:

- Study the effects of varying humidity on the system.
- Repeat Experiment C for more specimens.
- Explore the possibility of implementing a neural network instead of using the different algorithms.
- Explore the possibility of commercially developing a photonic or MEMS based formalin sensor and using it in our designed system [18] [19].

REFERENCES

[1] Md. Kamruzzaman, "Formalin Crime in Bangladesh: A Case Study", European Journal of Clinical and Biomedical Sciences. Vol. 2, No. 5, 2016, pp. 39-44

[2] Headlines and Global News, "Deadly Formalin-Laced Fruits In Bangladesh Could Cause Slow Poison Mass Killing", 9th Jun 2014.

[3] The Daily Star, "Formalin in fruits", 12th Jun 2013.

[4] The Daily Star, "Find proper formalin detection kit", 25th Nov 2014.

[5] Calright Instruments, "Formaldehyde Meter Model Z-300", Z-300 Datasheet, Feb 2010 [Revised Feb 2015].

[6] Riaz Uddin, et al. "Detection of formalin in fish samples collected from Dhaka City, Bangladesh." Stamford Journal of Pharmaceutical Sciences 4.1 (2011): 49-52.

[7] Abdullah Al Mamun and Aklima Akhi, "Microcontroller Based Formalin Detector: Hardware Design & Implementation", BSc. Thesis, Dept. of Elec. and Electronic Eng., University of Dhaka , Dhaka, 2017.

[8] Crystalynne D. Cortez, et al. "Development of formaldehyde detector." International Journal of Information and Electronics Engineering 5.5 (2015): 385.

[9] Crystalynne D. Cortez and Jennifer L. Santos. "Evaluation of the Developed Formaldehyde Detector." International Journal of Environmental Science and Development 7.6 (2016): 449.

[10] "Grove - HCHO Sensor - Seeed Wiki", wiki.seed.cc, 2017. [Online]. Available: http://wiki.seeed.cc/Grove-HCHO_Sensor. [Accessed: 14 Jun 2017].

[11] Jorge J Moré, "The Levenberg-Marquardt Algorithm: Implementation and Theory." Numerical analysis. Springer, Berlin, Heidelberg, 1978. 105-116.

[12] Ethem Alpaydin, "Introduction to Machine Learning." MIT press, 2014.

[13] Allan Clark, "Elements of Abstract Algebra." Courier Corporation, 1984.

[14] Christopher M. Bishop, "Pattern Recognition and Machine Learning." Springer, 2006.

[15] Carl Edward Rasmussen and Christopher KI Williams, "Gaussian Processes for Machine Learning." Vol. 1. Cambridge: MIT press, 2006.

[16] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant, " Applied Logistic Regression." Vol. 398. John Wiley & Sons, 2013.

[17] Bernhard Scholkopf and Alexander J. Smola, " Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, 2001.

[18] Md Faizul Huq Arif, et al. "Simulation based analysis of formalin detection through photonic crystal fiber." Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on. IEEE, 2016.

[19] Umme Hafsa Himi, et al. " MEMS Based Formalin Gas Sensor: A noncontact Sensing Approach." Informatics, Electronics and Vision (SCIEV), 2016 Student Conference on. CNSER, 2016.