2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) | 978-1-6654-1971-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/HNICEM51456.2020.9400077

Development of Sanitary Landfill's Carbon Dioxide Concentration Models Using Machine Learning Algorithms

Zoren Mabunga
School of Graduate Studies, College of Engineering
Mapua University, Southern Luzon State University
Manila, Philippines
zorenmabunga@gmail.com

Jennifer Dela Cruz
School of EECE
Mapua University
Manila, Philippines
jennycdc69@gmail.com

Glenn Magwili
School of EECE
Mapua University
Manila, Philippines
gvmagwili@yahoo.com

Angelica Samortin
School of Graduate Studies
Mapua University
Manila, Philippines
angelicasamortin26@gmail.com

***Abstract* – Carbon dioxide is one of the major component of landfill gas being emitted by sanitary landfills. High concentration of this gas may cause several health condition. It is also one of the greenhouse gas that consistently contributes to climate change. Monitoring and assessing the carbon dioxide concentration in landfills is vital to ensure better living conditions. This study presents the development of carbon dioxide concentration model based on machine learning algorithms. A prototype was developed using Arduino Uno, Wi-Fi module, DHT11 temperature and humidity sensor, MQ4 and MQ135 gas sensors. This prototype was used to gather CO_2 and CH_4 concentrations, humidity and air temperature of the sanitary landfill. Five machine learning model based on linear regression, support vector machine, regression trees, boosted regression trees and neural network was trained and evaluated. Matlab software was used in this study for the development of each model. The R-square and MSE of each model was calculated and compared which results to an almost identical r-square value of 0.75 and 0.76. An MSE of 6.90857e-05 for the neural network model followed by SVM, Boosted Regression Trees, Regression Trees and Linear Regression with an MSE of 8.8168e-05, 9.0085e-05, 9.4227e-05 and 9.4652e-05 respectively was also obtained. Based on these results, it was concluded that the machine learning model based on neural network is the best algorithm for the carbon dioxide concentration modelling in sanitary landfills since it obtained the lowest MSE among the five models.**

Keywords -* *sensors, Arduino, machine learning, Matlab, landfill

I. Introduction

Rapid industrial development and exponential increase in human population contributes to the continuous increase in solid wastes being disposed in landfills. In return, these landfills tend to emit more landfill gas which had adverse effects to the environment and to human health. Landfill gas is one of the byproducts of sanitary landfills due to the decomposition of disposed organic waste. Landfill gas is primarily compose of methane gas and carbon dioxide [1]. These two gases are two of the major greenhouse gases in the atmosphere. Carbon dioxide is one the major components of landfill gas being emitted by sanitary landfills. Aside from its contribution to climate change, high concentration of carbon dioxide can also produce variety of health effects such as headaches and dizziness.

Several related studies on the monitoring and measurement of carbon dioxide concentration using different sensors and new technology had been done over the years. In [2], a carbon dioxide emission prediction based on machine learning was developed. In their study, a regression model was developed in which they use historical data of carbon dioxide emissions from 1960 to 2014. Another study was conducted based on a specific algorithm called support vector machine for the prediction of carbon dioxide emissions was developed in [3]. In their study, energy consumption such as electrical energy and burning coal was used as the input parameters for the prediction of carbon dioxide emissions. A methane and carbon dioxide gas emissions monitoring for sanitary landfills was also developed by Mabunga et al. in [4]. They developed an Internet of Things based monitoring using carbon dioxide sensor, methane sensor, temperature and humidity sensor. The data coming from these sensors are sent to cloud for storage and processing.

Previous researches on the application of new technologies and different algorithms for the monitoring and prediction of carbon dioxide concentration and emissions was successfully done over the past years. But these researches and developed model based on machine learning are not applicable to sanitary landfill's CO2 concentration. Several factors affect the variations of CO2 concentrations in sanitary landfills such as the diverse composition of waste, decomposition stage and changing environmental conditions in the sanitary landfill [5]. These factors were not considered on the previous study. In [4],

978-0-7381-1177-3/20/$31.00 ©2020 IEEE

their developed monitoring system for greenhouse gases in sanitary landfill uses carbon dioxide sensors that are very sensitive to changing atmospheric conditions that causes drifts in the sensor readings which results to frequent recalibration of CO2 sensors.

The study generally aimed to developed a machine learning model for the measurement of carbon dioxide (CO_2) concentrations in sanitary landfill. Specifically, the researcher aims to: 1) to develop a prototype that will gather concentrations of CO_2 and CH_4, air temperature and humidity of the landfill; 2) to develop a machine learning (ML) regression models using Linear Regression, Support Vector Machine (SVM), Regression Trees, Neural Network (NN) and Boosted Trees algorithm; 3) to evaluate each machine learning models and select the best model for carbon dioxide concentration measurement.

The data that will be used in this study came from the CO2, CH4, temperature and humidity sensors that were installed in the gas vents of the sanitary landfills. Other parameters such as atmospheric pressure and wind velocity will not be consider in this study. This study will use Matlab software for the implementation of the different machine learning models.

II. Related Literature

A. Landfill Gas

Landfill gas is one of the main contributors of greenhouse gases in the atmosphere. Landfill gas is produce during the decomposition phase of municipal solid wastes in a landfill. The amount and composition of landfill gas varies between different sanitary landfills due to the complex composition of the solid wastes being disposed on these disposal sites. According to [1], carbon dioxide is estimated to be 35% in volume of the landfill gas. The continuous increase of landfill gas is one of the main contributors to global warming. Ngwabie et al., observed that there are huge variations in landfill gas emissions likely as a result of complex nature of the waste, decomposition stage and changing conditions of the environment within the waste[5]. According to the NSWC, these gases should be monitored accurately at least four times a year.

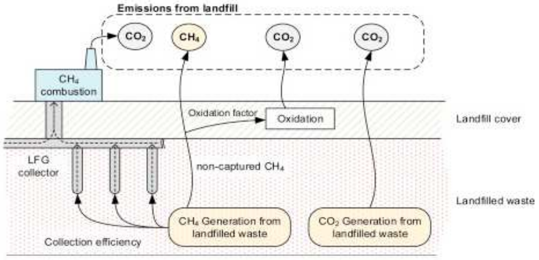


Fig. 1. Landfill Gas emissions in landfill

B. Neural Network

Neural Network (NN) is currently the most used tool in machine learning applications. NNs function just like how the neural systems of human works [6]. NN is composed of several layer connected neural nets. These layers are consisting of nodes that are connected to every other node on the adjacent layer. A relative weight if assigned into each neurons in which it accepts several inputs from the input datasets. These weights affects the impact of the input on the predicted output of the neural network [6]. Neural Network can be used to solve linear and non-linear datasets which makes it to be an effective algorithm in many complex systems [7]. In [8], five different machine learning models was developed and neural network was selected as the best model for predicting outdoor air quality.

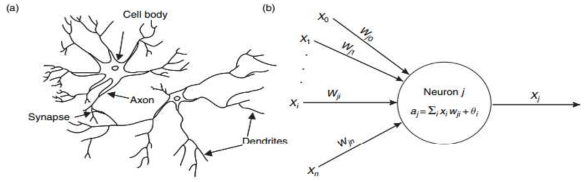


Fig. 2. (a) biological neuron and (b) artificial neuron

C. Support Vector Machine

Support Vector Machine is one of the most popular techniques in machine learning for regression and classification problems at present. SVM have been used in several applications such as image recognition and classification, text categorization, bioinformatics and database marketing. SVM uses two hyperplanes that separate the two class of data as much as possible. SVM can separate linearly separable dataset and even non-linear datasets. Same concept is being used for regression problems. The objective of this algorithm for regression problems is to consider all the points within the decision boundary line and fit a line with these points. This line is called the hyperplane that contains the maximum number of points inside the decision boundary.

For non-linear SVM problems, a kernel is introduced in the algorithm. Kernel is function that can make non-linear SVM problems to linear. There are different kernels for SVM such as polynomial kernel, linear kernel, non-linear kernel, Gaussian kernel, radial basis function (RBF), and sigmoid that improves the performance of the model[9].

D. Linear Regression

Linear regression is currently the simplest and easiest to implement type of regression model in machine learning. The basic linear regression model is basically a linear function of a predictor/independent variable x. For regression problems that involves more than one predictor, a variation of linear regression called multiple linear regression must be used. In this model, every value of the predictor parameters is associated with the value of the response parameter [10].

E. Regression Trees

Decision tree algorithm for classification or regression is based on the tree structure. In this algorithm, the data sets are broken down into smaller subsets. This type of classification method is applicable in handling diverse data as well as missing data. Some of the common applications of decision trees algorithm are in market segmentation, anomaly and fraud detection and in medical diagnosis. In[11], decision tree algorithm was used in the identification of breast cancer. The results of their study shows that the decision tree algorithm gives a 90% accuracy in the identification of breast cancer whether it is malignant or benign.

F. Boosted Regression Trees Algorithm

Boosted Regression Trees algorithm is a machine learning algorithm for regression applications. This algorithm is based on the building of several regression trees in a step-wise fashion. A loss function is also used in this model to calculate the error in each step and automatically correct these errors in the next step. In simpler terms, this algorithm is the ensemble or combination of weaker prediction models.

III. METHODOLOGY

The data that was used in this study was collected from the sanitary landfill in Tayabas City. The temperature, humidity, methane and carbon dioxide sensor were place near the gas vents of the sanitary landfill for two weeks of continuous data gathering. Five machine learning regression algorithms was used to model the carbon dioxide concentration of the sanitary landfill using the three inputs namely methane concentration, temperature and humidity.

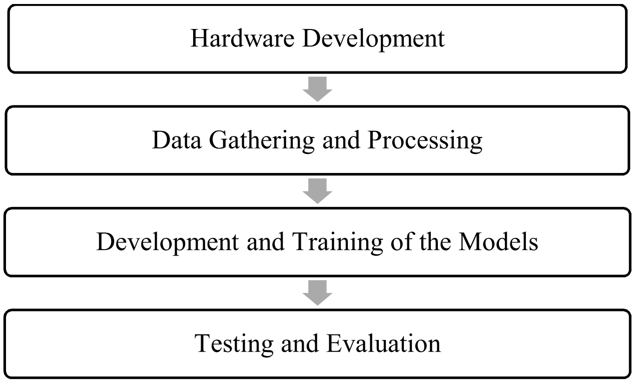

Hardware Development
Data Gathering and Processing
Development and Training of the Models
Testing and Evaluation

Fig. 3. General Methodology

A. Hardware Development

A prototype consisting of Arduino microcontroller, ESP8266 Wi-Fi module, temperature, humidity, carbon dioxide gas and methane gas sensor was developed to gather the needed data. The sensors were calibrated prior to deployment in the sanitary landfill to ensure that accurate readings will be obtain. The Arduino Uno received and processed all the raw data from the sensor nodes. Wi-Fi module was combine with Arduino Uno for the transmission of data to the cloud. DHT11 sensor was used for the temperature and humidity measurement, MQ4 gas sensor for the methane measurement and MQ135 gas sensor for the carbon dioxide measurement. These sensors were chosen for this particular study due to their availability in the Philippines and its low cost.

B. Data Gathering and Processing

The prototype was installed near the gas vents of the landfill. This prototype gathered the methane gas (CH_4) concentration in parts per million (ppm), carbon dioxide (CO_2) gas concentration in parts per million (ppm), humidity in percent (%) and temperature in degree Celsius (°C) of the sanitary landfill for two weeks. The humidity, temperature and CH4 concentration were used as the predictors variables of the study while the CO2 concentration will be the output or response variable.

After the data gathering phase, the datasets undergo data cleaning and normalization. Data cleaning was done by eliminating outliers in the data sets while normalization is another technique in the pre-processing of datasets that was used in the development of a certain machine learning model. Some of the outliers in the data set were remove such as those with incomplete values and incorrect sensor readings. Normalization was done to convert the raw data in the dataset to a similar scale, without changing the differences in their values. Thru normalization, the data can only have values from 0 to 1. One formula to achieve normalization is presented in equation 1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

C. Development and Training of Models

In this paper, supervised training was implemented. Five different ML regression algorithms were used for the carbon dioxide concentration modelling. The algorithms that were used are linear regression, support vector machine (SVM), boosted regression trees, neural network and regression trees algorithm. These algorithms were trained and implemented using Matlab software tools.

1) Training and Model Development Using Linear Regression

Linear regression was the first regression model that was used in this study. This model is one of the simplest and easiest model to implement when modelling a quantitative outcome variable. This model calculates the probability of occurrence of a binary response variable (y) with respect to the input or predictor variables (x) [12]. For this particular study, a linear regression with interactions terms was used as shown in equation 2. Interactions terms were used in this study since the effect of the predictor variables in the response/target variable changes depending on the values of other predictor variables.

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_1X_2 + b_5X_1X_3 + b_6X_2X_3 + b_7X_1X_2X_3 \quad (2)$$

2) Training and Model Development Using Support Vector Machine (SVM)

SVM algorithm have been one of the most popular supervised machine learning algorithm for regression and classification problems today. In this study, the concept of regression SVM was used. The objective of this algorithm for regression problems is to consider all the points within the decision boundary line and fit a line with these points. This line is called the hyperplane that contains the maximum number of points inside the decision boundary. An SVM with quadratic kernel was used in this particular study as presented in equation 3. Higher order polynomial kernel was not used in this study since it may over fit the training data and may not perform well on new data sets.

$$K = (1 + x \cdot y)^2 \qquad (3)$$

3) Training and Model Development Using Regression Tree

Regression tree algorithm for classification or regression is based on the tree structure. The objective of this algorithm is to create a model thru supervised training which can be used to predict the class or value of the response variables based on the rules learned from the previous datasets. In this algorithm, the data sets are broken down into smaller subsets. In this study a decision tree algorithm with a maximum number of splits of 20 was used. This type of decision tree algorithm is also called medium decision tree algorithm.

4) Training and Model Development Using Boosted Regression Trees

Boosted regression trees is another technique in machine learning that tends to increase the performance of one model by combining several models. Boosted regression trees collects the advantages of regression tree algorithm by handling different types of independent variables and supplying missing datasets. This algorithm was also used in this study since it is applicable to complex nonlinear relationships and it also automatically handle the interaction between the independent variables[13]. This algorithm is also applicable for large training data since it consumes relatively little memory compare to the other ensemble machine learning algorithms.

5) Training and Model Development Using Neural Network

The last model that was used in this study is the neural network. This algorithm was implemented using the neural net fitting tool of Matlab software. In this study, the model was based on neural network that has 3 neurons for the input layer, 10 neurons for the hidden layer and 1 neuron for the output layer. This model was train using Levenberg-Marquardt training algorithm since this algorithm does not require huge amount of computational memory and the training automatically stops when the cost function is at its minimum value[14].

D. Testing and Evaluation

For the testing and validation of each model, hold-out validation was used with 30% held out. This type of validation is typically used when there is a large amount of data sets. In this type of testing and validation, the data set was divided into two different sets, called the training set and the testing set. The training set was used solely for the training of each model and the test set was used to test the model. To compare the performance of each model, the mean square error and the r-squared of each model was used as shown in equation 4 and 5 respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2 \qquad (4)$$

$$R^2 = \left(\frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}\right)^2 \qquad (5)$$

IV. Results and Discussion

This section provides the results generated in this study. Table 1 shows the sample of the sample data prior to normalization. These data are the data that was directly sent by the prototype to the cloud.

TABLE I. SAMPLE RAW DATA

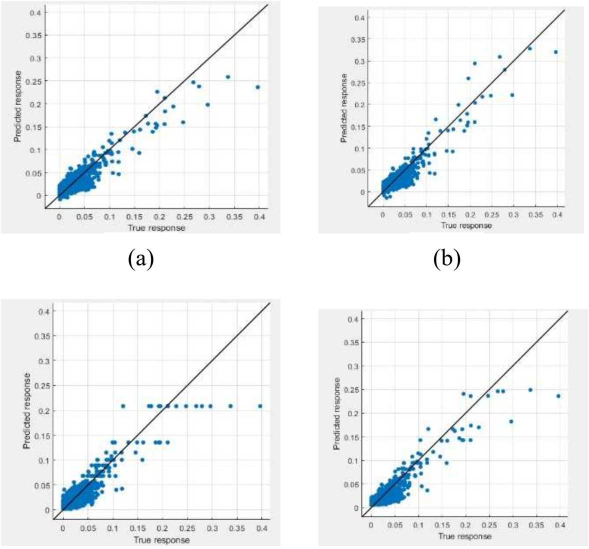
CO2 Concentration (ppm)	**CH4 Concentration (ppm)**	**Humidity (%)**	**Air Temperature (Celsius)**
587.69	149.6	99.8	26.6
587.67	158.67	99.5	26.6
404.42	131.97	97.8	27
434.17	148.23	79.2	30.9
416.09	140.63	77.8	30.8
415.73	140.87	79	30.6
429	139.52	80.2	30.6
437.2	140.65	80.4	30.6
578.18	210.3	80.9	30.7
454.25	149.16	81.1	30.7

Table 2 presents the cleaned and normalized data from the different sensors that was installed in the sanitary landfill. Incorrect and missing values for some parameters are removed as part of the data cleaning process. A total of 17,397 data sets were used in the study for the training, validation and testing of each model.

TABLE II. SAMPLE CLEANED AND NORMALIZED DATA

CO2 Concentration (ppm)	**CH4 Concentration (ppm)**	**Humidity (%)**	**Air Temperature (Celsius)**
0.039381	0.014121	0.997555	0.403101
0.039377	0.017193	0.990220	0.403101
0.009010	0.008151	0.948655	0.434109
0.013940	0.013657	0.493888	0.736434
0.010944	0.011084	0.459658	0.728682
0.010884	0.011165	0.488998	0.713178
0.013083	0.010708	0.518337	0.713178
0.014442	0.011090	0.523227	0.713178
0.037805	0.034676	0.535452	0.720930
0.039381	0.014121	0.997555	0.403101

Figure 4 presents the predicted vs true response of the linear regression, SVM, regression trees and boosted regression trees model. The four graphs show how well these regression models perform for different response values. The error in these model is represented by the vertical distance from the diagonal line to each data point. For these four model, an MSE of 9.4652e-05, 8.8168e-05, 9.4277e-05, 9.0085e-05 were obtained for the linear regression, SVM, regression trees and boosted regression trees model respectively. The r-square value for the four model are almost identical with value of 0.75 for the linear regression and regression trees and 0.76 for the SVM and boosted regression trees model.



(a)

(b)

(c)

(d)

Fig. 4. Predicted vs True Response of (a) Linear Regression, (b) Support Vector Machine, (c) Regression Trees, (d) Boosted Regression Trees

Figure 8 shows regression plot for the neural network model that was developed using Matlab Neural Network tool. Based on the results the R value of the model are 0.86214 during the training phase, 0.88831 during the validation phase and 0.84479 on the testing phase. This results to an overall R value of 0.8647 which is equivalent to an R-squared of 0.75.

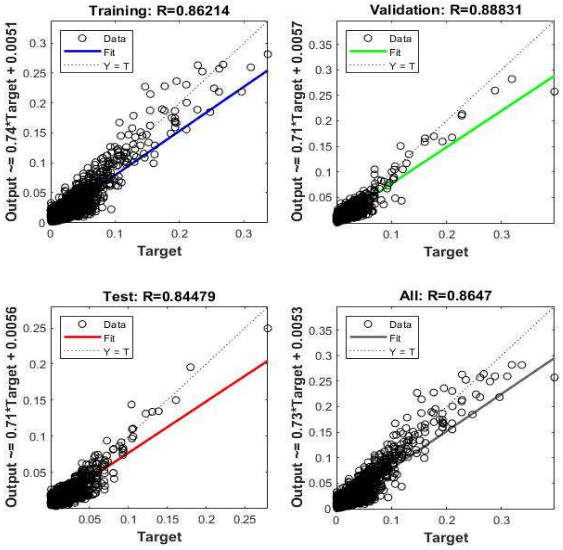


Fig. 5. Regression Plot of the Neural Network Model

Figure 9 shows the error histogram of the neural network model. For this particular model, a mean-square error of 6.90857e-05 was calculated on the test set.

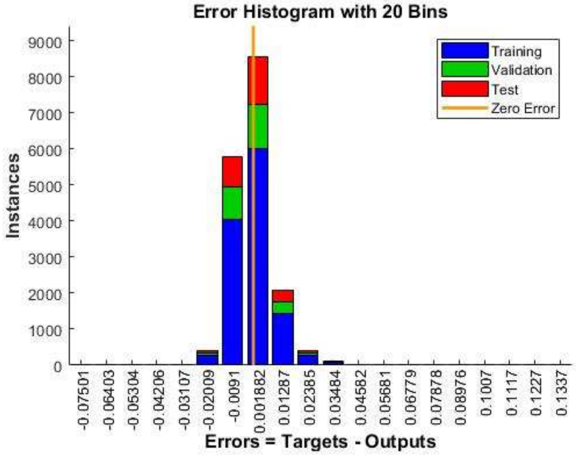


Fig. 6. Error Histogram of the Neural Network Model

Table 3 presents the mean-square error and R-squared obtained after training and hold-out validation for the five regression models.

TABLE III. SUMMARY OF MACHINE LEARNING REGRESSION MODEL RESULTS

Model	MSE	R-Squared
Linear Regression	9.4652e-05	0.75
Support Vector Machine	8.8168e-05	0.76
Regression Trees	9.4277e-05	0.75
Boosted Regression Trees	9.0085e-05	0.76
Neural Network	6.90857e-05	0.75

According to the results, the neural network model provides the lowest mean-square error among the five regression models with an MSE of 6.90857e-05. It was followed by SVM, Boosted Regression Trees, Regression Trees and Linear Regression with an MSE of 8.8168e-05, 9.0085e-05, 9.4227e-05 and 9.4652e-05 respectively. While all the five regression models provide almost identical R-squared value of 0.75 or 0.76.

V. CONCLUSION

In this study, the researcher was able to develop an electronic measuring device consisting of three sensors that will collect the carbon dioxide concentration, methane concentration, humidity and temperature of the sanitary landfill. The sensors were calibrated since these sensors are prone to sensor drift due to changing environmental conditions in the sanitary landfill. The data that was gathered are pre-processed by removing incomplete data and thru normalization since uncleaned and not normalized data may cause high bias and variance for each model. The processed data were then used for the training, validation and testing of five different machine learning regression models. The MSE and R-squared of each model was recorded and compared. In terms of the R-squared, an almost identical value was obtained among the five models which is 0.75 and 0.76 but in terms of MSE, the neural network model provides the lowest MSE of 6.908577e-05.

For future works, the developed model on this paper can be integrated with IoT sensors to improve the existing carbon dioxide gas sensors performance. Other parameters such as wind velocity, amount of rainfall, atmospheric pressure and the amount of waste being disposed in the landfill can also be consider as inputs to the model. Other data pre-processing techniques can also be used for future study such as the method presented in [15].

Acknowledgment

The researcher would like to extend its sincerest gratitude to the local government unit of Tayabas City for allowing him to conduct the testing of this research on their controlled sanitary landfill and to the Department of Science and Techonology – Engineering Research and Development for Technology for the scholarship grant to the author.

References

[1] Environmental Protection Agency, *LANDFILL MANUALS LANDFILL MONITORING 2nd Edition*. 2003.

[2] P. Kadam and S. Vijayumar, "Prediction Model: CO 2 Emission Using Machine Learning," *2018 3rd Int. Conf. Converg. Technol. I2CT 2018*, pp. 1–3, 2018, doi: 10.1109/I2CT.2018.8529498.

[3] C. Saleh, N. R. Dzakiyullah, and J. B. Nugroho, "Carbon dioxide emission prediction using support vector machine," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 114, no. 1, 2016, doi: 10.1088/1757-899X/114/1/012148.

[4] Z. Mabunga and G. Magwili, "Greenhouse Gas Emissions and Groundwater Leachate Leakage Monitoring of Sanitary Landfill," 2019, doi: 10.1109/HNICEM48295.2019.9072872.

[5] N. M. Ngwabie, Y. L. Wirlen, G. S. Yinda, and A. C. VanderZaag, "Quantifying greenhouse gas emissions from municipal solid waste dumpsites in Cameroon," *Waste Manag.*, vol. 87, pp. 947–953, 2019, doi: 10.1016/j.wasman.2018.02.048.

[6] S. Lek and Y. S. Park, "Artificial Neural Networks," S. E. Jørgensen and B. D. B. T.-E. of E. Fath, Eds. Oxford: Academic Press, 2008, pp. 237–245.

[7] L. H. Hassan, M. Moghavvemi, H. A. F. Almurib, and O. Steinmayer, "Current state of neural networks applications in power system monitoring and control," *Int. J. Electr. Power Energy Syst.*, vol. 51, pp. 134–144, 2013, doi: 10.1016/j.ijepes.2013.03.007.

[8] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2018-October, no. October, pp. 668–672, 2019, doi: 10.1109/TENCON.2018.8650518.

[9] P. R. Meris *et al.*, "IOT Based - Automated Indoor Air Quality and LPG Leak Detection Control System using Support Vector Machine," *2020 11th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2020 - Proc.*, no. August, pp. 231–235, 2020, doi: 10.1109/ICSGRC49013.2020.9232472.

[10] S. Il Pak and T. H. Oh, "Correlation and simple linear regression," *J. Vet. Clin.*, vol. 27, no. 4, pp. 427–434, 2010, doi: 10.1007/978-3-319-89993-0_6.

[11] P. Sathiyanarayanan, S. Pavithra, M. Sai Saranya, and M. Makeswari, "Identification of breast cancer using the decision tree algorithm," *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 2–7, 2019, doi: 10.1109/ICSCAN.2019.8878757.

[12] A. Urso, A. Fiannaca, M. La Rosa, V. Ravì, and R. Rizzo, "Data mining: Prediction methods," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, pp. 413–430, 2018, doi: 10.1016/B978-0-12-809633-8.20462-7.

[13] X. Du, F. Zeng, G. Shi, and Y. Feng, "Smart pollution source tracing via gradient tree boosting regression," *Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019*, pp. 341–344, 2019, doi: 10.1109/MLBDBI48998.2019.00077.

[14] L. Sun, J. Hu, Y. Liu, L. Liu, and S. Hu, "A comparative study on neural network-based prediction of smart community energy consumption," *2017 IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trust. Comput. Scalable Comput. Commun. Cloud Big Data Comput. Internet People Smart City Innov. SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 -* , pp. 1–8, 2018, doi: 10.1109/UIC-ATC.2017.8397441.

[15] C. Sun and P. Guo, "Data preprocessing of wind turbine based on least squares support vector machine and neighbor model," *Proc. 29th Chinese Control Decis. Conf. CCDC 2017*, pp. 1441–1446, 2017, doi: 10.1109/CCDC.2017.7978744.