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/HNICEM51456.2020.9400140

Development of Sanitary Landfill's Groundwater Contamination Detection Model Based on 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* - This study describes the development of five machine learning models for the detection of groundwater contamination due to leachate leakage in a sanitary landfill. A prototype was constructed using Arduino Uno, Wi-Fi module, pH, electrical conductivity and temperature sensors. This prototype was used to gather data from the groundwater and leachate samples in the sanitary landfill. The sensors that were used in the study was calibrated prior to the actual data gathering in the sanitary landfill. Five machine learning model based on logistic regression, quadratic discriminant analysis, k-nearest neighbour, decision tree and support vector machine algorithm was trained and evaluated. Matlab software was used in this study for the development of each model. The accuracy of each model was then compared which results to a 97.8% accuracy for KNN, 97.7% for SVM and Decision Tree, 93.7% for quadratic discriminant and 92.6% for logistic regression model. Based on the results, KNN, SVM and decision tree based models provide the highest accuracy for the detection of leachate leakage on the groundwater located in a sanitary landfill.**

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

I. INTRODUCTION

For the past years, there had been an increasing number of sanitary landfills in the Philippines. As of 2016, there are already 118 sanitary landfills in the Philippines [1]. The Department of Natural Resources (DENR) estimates that 26,472 tons/day of solid wastes of the country were collected and delivered to its disposal sites. These solid wastes have a very diverse composition and according to the National Solid Waste Management Commission (NSWMC), these disposed waste are 52% biodegradable, 28% recyclable and 18% residuals. Due to its diverse composition, several byproducts are being introduced by these disposal sites which results to several health and environmental problems. Landfill gas and leachate are two of the primary byproducts of sanitary landfills. Leachate is a liquid that spread through the deposited solid wastes [2]. A huge number of pollutants and toxic substances can be found in the sanitary landfill leachates, some of this are toxic, and dangerous to the environment and human health [3]. This leachate may leak and contaminate the groundwater reservoir in the landfill area which possess a huge risk to people and animals consuming and using these waters.

Several related studies on groundwater quality detection and prediction by applying different technology and algorithms were done over the past years. A study about the surface water pollution detection using Internet of Things (IoT) were presented in [4] using a developed prototype that records water quality parameters of different water samples. Several machine learning models were also developed for the classification of water samples. Based on the experimental results of this study, deep neural network gives the highest accuracy of 93% [4]. Another related study about the application of different machine learning models in groundwater was presented in [5]. In [5], they compare the performance of ELM with other ML techniques to predict the fluoride content of groundwater. In [6] they developed a system that is incorporated with machine learning algorithm which has the ability to classify water quality using Support Vector Machine (SVM) algorithm. A device that monitors several groundwater parameters were also developed in [7] that is based on Internet of Things. This device provides a near to real time update on the pH level, electrical conductivity (EC) and groundwater's temperature in the sanitary landfill area. Lastly, in [8] different machine learning algorithms was used to predict the occurrence of rainfall in La Trinidad, Benguet which is based on various weather parameters. The results of the study show that KNN provides the highest accuracy in predicting the occurrence of rainfall.

Researches on the application of new technologies and different algorithms for the detection and prediction of water

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

quality was successfully done over the past years. But these researches focuses their scope on surface water which is one of the two classification of water. Groundwater that is located in areas such as landfills which are prone to possible contamination must be treated differently since groundwater pollution works differently from surface water pollution [9]. Some study also used machine learning algorithms for the classification of water, but these study focuses on a single algorithm such as Support Vector Machine (SVM) and Artificial Neural Network (ANN). Other machine learning classification algorithm can also be used for groundwater contamination detection such as K-Nearest Neighbor algorithm (KNN), Decision Trees, Quadratic Discriminant Analysis (QDA) and Logistic Regression algorithm. These algorithms might have a better accuracy if properly tuned.

The study generally aimed to developed a machine learning model for the detection of leachate leakage in groundwater. Specifically, the researcher aims to: 1) to develop a system that will collect groundwater parameters such as pH, temperature and electrical conductivity (EC); 2) to develop a machine learning model using Quadratic Discriminant Analysis, K-Nearest Neighbor, Decision Trees, Logistic Regression and Support Vector Machine; 3) to evaluate each machine learning models and select the best predictive models for groundwater contamination detection.

The groundwater quality data that will be used in the study came from the pH, EC and temperature sensor that were installed in a monitoring well at the Sanitary Landfill of Tayabas City. Two sets of data gathering instruments will be used, one for the uncontaminated well and one for the leachate sample. Other groundwater quality parameters will not be consider in the study such as TDS, BOD, COD, ammoniacal nitrogen or nitrate, chloride and sulphate. This study will use Matlab software for the implementation of the different machine learning models.

II. Related Literature

Leachate is a dark wastewater that is produced due to precipitation, deposited waste moisture and water in a sanitary landfill. Leachate can penetrate through the ground and mix with groundwater and surface water especially during rainfalls. Even after the closure of a landfill, leachate may still possess possible threat of groundwater contamination. Due to the diverse composition of solid wastes being disposed into sanitary landfills, A huge number of pollutants and toxic substances can be found in the sanitary landfill leachate, some of this are toxic, and dangerous to the environment and human health [3]. The amount and characteristics of leachate generated in a sanitary landfill depends on several factors such as the amount of waste deposited, climate condition, hydrogeological structure of the area, operational conditions and age of the landfill [10]. Leachate is characterized to have high amount of dissolved organic and inorganic substances and small amount of suspended solids [11]. The composition of leachate undergoes continuous variation due to the new and emerging compound being discovered in the environment.

A. *Quadratic Discriminant Analysis*

QDA is an ML algorithm that Discriminant analysis is used to determine which variables distinguish the difference between the groups of datasets. Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA) are the most common type of discriminant analysis. In QDA, the datasets are assumed have the same covariance and to be normally distributed.

QDA is typically used when we have less number of predictor variables. This due to the association of the number of parameters the algorithm need to estimate. In QDA, a large number of predictor variables lead to high number of parameters to estimate which might result to high variance. In [12], a QDA based Global Navigation Satellite System (GNSS) signal quality monitoring method was developed. Using QDA based method, it shows an improved performance on detecting “evil waveforms” (EWFs) and provides an added capability to identify the failure types [12].

B. *Support Vector Machine*

Support Vector Machine (SVM) have been one of the most popular supervised machine learning algorithm for regression and classification problems today. 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.

This algorithm is very effective for high dimensional spaces and it is also memory efficient since it uses a subsets of training points in the decision function. On the other hand, this algorithm is not suitable for large data sets and with data sets that are overlapping.

C. *K-Nearest Neighbor (KNN)*

The k-nearest neighbors’ (KNN) algorithm is another supervised machine learning classification technique that uses the closeness as the basis for ‘sameness’. KNN algorithm is often used for classification problems. The algorithm takes several number of identified points and these points are used to train and learn on how to label the remaining points. To classify a new data, it looks at the previously classified data points that is nearest to that new data point.

The performance of this algorithm depends crucially on the technique used to recognize nearest neighbors [13]. This algorithm is simple and easy to implement, more flexible to noisy data, and effective if there is a large data set but the computation cost for this algorithm is typically high since it is necessary to compute the distance of each points on all the training data.

D. *Decision Trees*

Decision tree algorithm for classification and 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 [14], 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.

E. Logistic Regression

Logistic regression is another supervised learning classification algorithm that estimates the probability of a response variable. In this algorithm, the response variable should be in binary format which means that only two values should be assigned to the response variable. In logistic regression models, it calculates the probability of the response variable *y* as a function of the predictor or independent variables {xj} ,where j = 1…, n[15]. A better result of this algorithm is when we cleaned the data by removing unrelated variables as well as those variables with high correlation with each other. So it is important to cleaned and pre-process the datasets before training the model.

III. METHODOLOGY

The gathered data that was used in this study came from the sanitary landfill of Tayabas City as shown in Figure 1. The groundwater's quality data was collected from the well of the sanitary landfill by the EC, pH and temperature sensor. The sanitary landfill's leachate data was gathered from the leachate collection facility of the sanitary landfill by the EC, pH and temperature sensor. Machine learning classification algorithms was used to classify groundwater's quality according to the three input parameters.

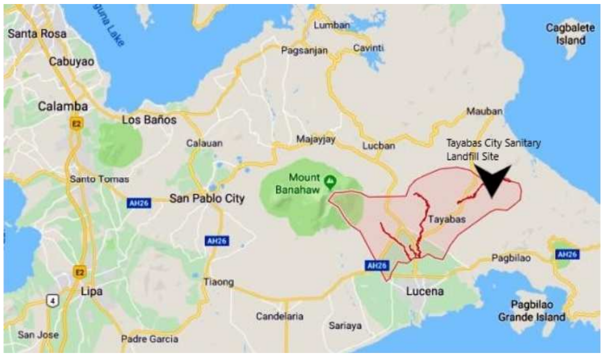

Fig. 1. Sanitary Landfill Map of Tayabas City

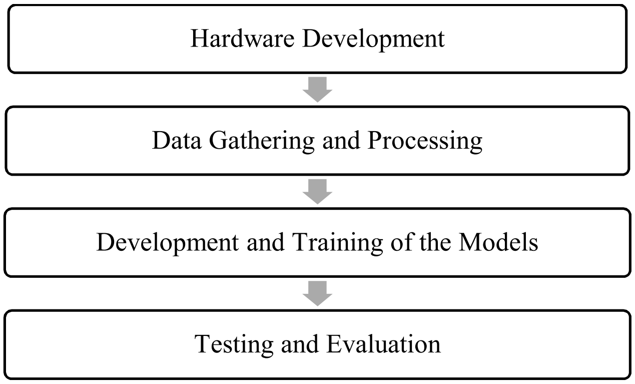
Hardware Development
Data Gathering and Processing
Development and Training of the Models
Testing and Evaluation

Fig. 2. Block Diagram of the Methodology

A. Hardware Development

A two set of prototype for the data gathering was constructed using Arduino Uno, Wi-Fi modules, pH sensors, EC sensors and water temperature sensors. The Arduino microcontroller received and processed all the raw data from each sensor nodes. Wi-Fi module was connected to the Arduino for the transmission of processed data to cloud. ESP8266 Wi-Fi Module was chosen for this study due to its availability, compatibility with Arduino and low cost.

Three different sensors for water quality were used in the prototype. Analog pH sensor, electrical conductivity sensor with industrial electrode were used to gather the pH values and electrical conductivity in mS/cm of the groundwater and leachate samples in the sanitary landfill. DS18B20 temperature sensor was then used to collect the groundwater's temperature and leachate samples temperature in degree Celsius. These sensors were chosen due to its low cost, compatibility with Arduino Uno and its availability in the Philippines.

Each sensor was calibrated using reference meter and different liquid samples to ensure accurate readings will be obtained from each sensor nodes as shown in Figure 3.

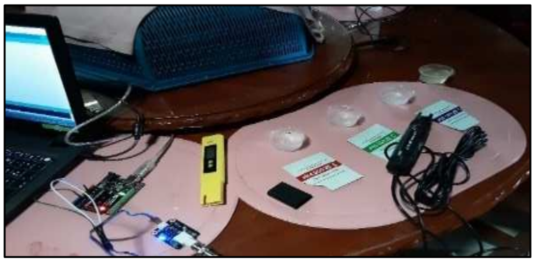

Fig. 3. Calibration Setup of pH and EC sensors

B. Data Gathering and Processing

Two sets of data gathering prototype were installed in the sanitary landfill. The first set was installed in a monitoring well adjacent to the actual landfill area while the second set was installed in the leachate collection facility.

The sensor readings coming from the pH, EC and temperature sensor served as the predictors for the different machine learning classification models in this study. The classification used for this study was whether the groundwater is contaminated or uncontaminated by leachate leakage. This study adopted supervised machine learning algorithms in which the researcher assigned a value of "1" for the sensor readings coming from the prototype installed in the leachate collection facility. A value of "0" was then assigned for the set of sensors reading obtained from the prototype installed in the clean or uncontaminated monitoring well.

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 a technique used as part of data pre-processing in machine learning. The outliers were identified as incomplete and incorrect values in the data set. The goal of normalization is to convert the values of the original data into a common scale, without altering the difference in span of values. Thru normalization, the data can only have values from 0 to 1.

C. *Development and Training of Models*

In this paper, five different machine learning (ML) supervised classification algorithms were used for the classification of groundwater quality from the dataset obtained from the prototype. The machine learning algorithms that were used were logistic regression, quadratic discriminant analysis (QDA), decision trees algorithm, support vector machine (SVM) and k – nearest neighbors (KNN). These algorithms were trained and implemented using Matlab software tools.

1) Training and Model Development Using Logistic Regression

Logistic regression was the first model that was used in this study. This model is one of the simplest and easiest model to implement when modelling a categorical outcome variable. Logistic regression is also the most popular algorithm for analysis of binary response data. For this particular study a multivariate logistic regression was used. Three independent variables or predictor variables was used to estimate the probability of the dependent or response variable using equation 1.

$$\pi(x) = \frac{e^{\beta_o+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}{1+e^{\beta_o+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}} \qquad (1)$$

2) Training and Model Development Using Quadratic Discriminant Analysis (QDA)

Quadratic discriminant algorithm is a variant of linear discriminant analysis (LDA) in which in every class of observation, a covariance is estimated. QDA is particularly useful if the individual classes exhibit distinct covariance. This model is applicable for datasets that are not linearly separable. In this study the classifier assigned an X = x to the class for which the equation 2 is largest.

$$\delta_k(x) = \log\pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) \qquad (2)$$

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

Support Vector Machine (SVM) have been one of the most common supervised machine learning technique for regression and classification problems today. This algorithm 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 using kernels. In this study, SVM using Gaussian kernel was used. The kernel scale was set to $\frac{\sqrt{P}}{4}$ in which the P variable is the number of input variables. The Gaussian kernel that was used in this study is defined by equation 3.

$$k(x,y) = e^{-\frac{||x-y||^2}{2\sigma^2}} \qquad (3)$$

4) Training and Model Development Using K – Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) algorithm has wide applications in regression and classification problems. The algorithm takes several number of identified points and these points are used to train and learn on how to label the remaining points. To classify a new data, it looks at the previously classified data points that is nearest to that new data point. A dissimilarity function as shown in equation 4 is used to determine the closeness. In this study, fine KNN was used that makes finely-detailed dissimilarity between each class with neighbors' number set to 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \qquad (4)$$

5) Training and Model Development Using Decision Trees Algorithm

Decision tree algorithm for regression or classification problems is based on the tree structure. In this algorithm, the data sets were 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.

D. *Testing and Evaluation*

For the testing of each model, a fivefold cross-validation was used. This type of validation is typically used when we don't have large datasets. In this technique, the datasets are randomly split into 5 groups. The process starts by assigning one group as the test set and the other group as the training set. The test set will be used to determine the accuracy after it was trained on the training set. This process is repeated until each unique group has been used as the test set. The average accuracy was then calculated for each model. The accuracy will be the basis for the evaluation and comparison of the performance for each of the classifier model.

IV. Results and Discussion

This section provides the results obtained in this study. Table 1 shows the sample of the cleaned and normalized data from the different sensors that was installed in the sanitary landfill. Incorrect and missing values for some parameters were removed as part of the data cleaning process.

TABLE I. SAMPLE CLEANED AND NORMALIZED DATA

pH	EC	Temperature	y
0.500212	0.125686	0.543497	0
0.49675	0.127315	0.549509	0
0.496861	0.129759	0.55918	0
0.498313	0.129759	0.568982	0
0.493511	0.129759	0.57813	0
0.781213	0.253634	0.48528	1
0.558834	0.256773	0.465593	1
0.739555	0.212692	0.426394	1
0.646736	0.198464	0.400871	1
0.773703	0.434246	0.346065	1

Figure 4 presents the confusion matrix of the logistic regression model of this study. The number of true positive results is 1247 while for the true negative is 1256 which corresponds to 95 false positive and 104 false negative for the logistic regression model. This result to an overall accuracy of 92.6%.

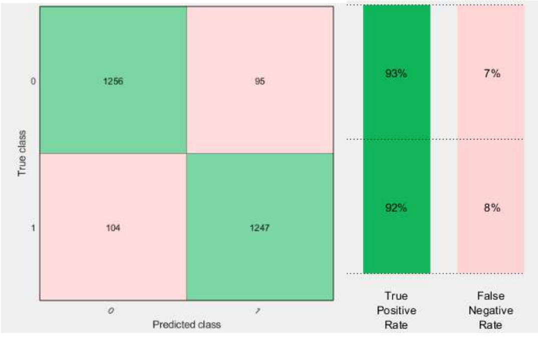

Fig. 4. Confusion Matrix for the Logistic Regression Model

Figure 5 shows the confusion matrix of the quadratic discriminant analysis model of this study. The lowest number of true positive results of 1237 was obtained in this while having a true negative of 1295 which corresponds to 56 false positive and 114 false negative. This result to an overall accuracy of 93.7% which is slightly higher than the logistic regression model.

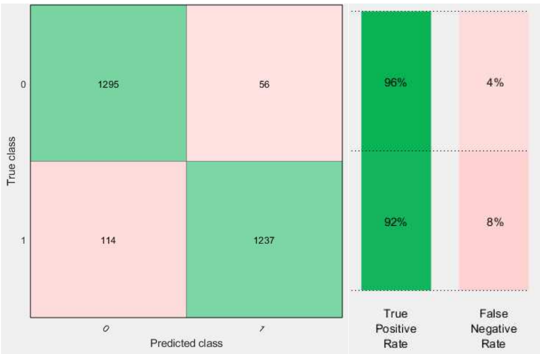

Fig. 5. Confusion Matrix for the Quadratic Discriminant Analysis Model

Figure 6 demonstrates the confusion matrix for the support vector machine model of this study. SVM model was able to achieve an overall accuracy of 97.7% with true positive results of 1308 and true negative of 1333 which corresponds to 18 false positive and 43 false negative. This result to an overall accuracy of 97.7%.

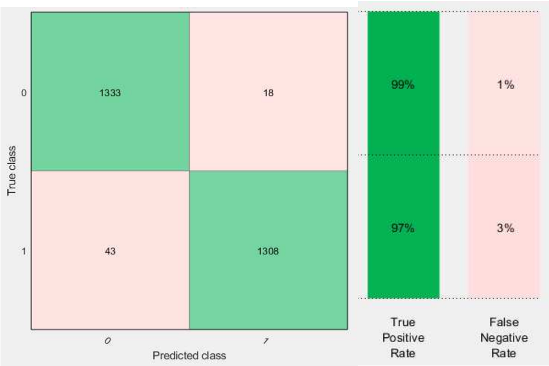

Fig. 6. Confusion Matrix for the SVM Model

Figure 7 illustrates the confusion matrix for the KNN model of this study. In this model, the number of true positive results is 1304 while for the true negative is 1339 which corresponds to 12 false positive and 47 false negative. This result to the highest accuracy among the five models with 97.8%.

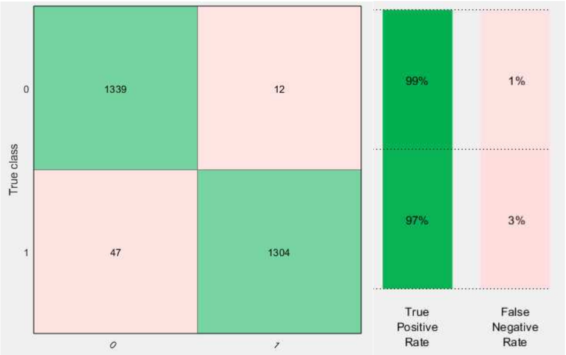

Fig. 7. Confusion Matrix for the KNN Model

The confusion matrix for the decision tree model of this study is shown in figure 8. An almost identical results with the SVM model for the number of true positive results and true negative was obtained in the decision tree model which is 1334 and 1306 respectively which corresponds to 17 false positive and 45 false negative. This result to an equal overall accuracy of 97.7% with the SVM model.

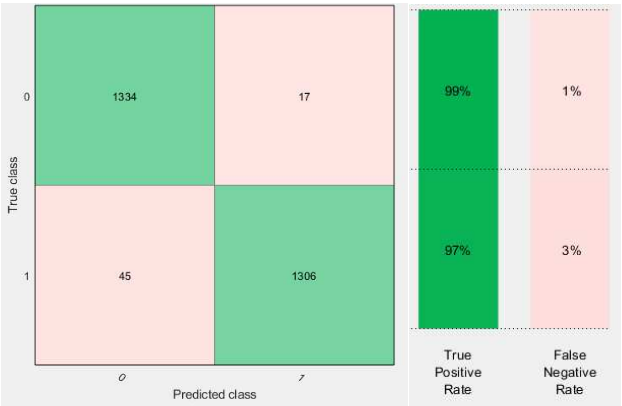

Fig. 8. Confusion Matrix for the Decision Tree Model

Table 2 presents the accuracy obtained after training and cross-validation for the five classification models.

TABLE II. SUMMARY OF MACHINE LEARNING CLASSIFICATION MODEL RESULTS

Model	Accuracy
Logistic Regression	92.6%
Quadratic Discriminant	93.7%
Support Vector Machine	97.7%
K-Nearest Neighbor	97.8%
Decision Tree	97.7%

According to the results, the k-nearest neighbor model provide the highest accuracy among the five classification models with an accuracy of 97.8%. It was followed by Decision tree and SVM with Gaussian Kernel model with and accuracy of 97.7%. Logistic regression and quadratic discriminant algorithm were able to obtained accuracy of 92.6% and 93.7% respectively.

V. CONCLUSION

In this study, the researcher was able to develop a prototype that will collect and transmit data to the cloud using Arduino Uno, Wi-Fi module, EC, pH and water temperature sensor. Each sensor was properly calibrated using reference meters to avoid sensor drifts during the data gathering phase. The data that was gathered from the landfill were pre-processed by removing outliers and thru normalization because uncleaned data results to inaccurate prediction which tend to decrease the prediction accuracy of each model. These data were then used in the training and testing of the five different machine learning models. Among the five machine learning classification models that was developed, k-nearest neighbor (KNN) is the best machine learning model for the detection of contamination of groundwater in the landfill area since it provides the highest accuracy of 97.8%.

For future works, it is recommended to increase the number of predictor variables by adding other biological and chemical properties of groundwater. The degree of contamination can also be added as an additional classification of groundwater quality.

ACKNOWLEDGMENT

The researcher would like to express his gratitude to the Sanitary Landfill's personnel of Tayabas City for assisting him during the data gathering phase of this study.

REFERENCES

[1] Senate Economic Planning Office, "Philippine Solid Wastes," *Philipp. Solid Wastes A Glance*, vol. AG-17-01, no. 01, pp. 1–4, 2017, [Online]. Available: https://www.senate.gov.ph/publications/SEPO/AAG_Philippine Solid Wastes_Nov2017.pdf.

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

[3] T. Eggen, M. Moeder, and A. Arukwe, "Municipal landfill leachates: A significant source for new and emerging pollutants," *Sci. Total Environ.*, vol. 408, no. 21, pp. 5147–5157, 2010, doi: 10.1016/j.scitotenv.2010.07.049.

[4] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar, and H. Khurshid, "Surface Water Pollution Detection using Internet of Things," *2018 15th Int. Conf. Smart Cities Improv. Qual. Life Using ICT IoT, HONET-ICT 2018*, pp. 92–96, 2018, doi: 10.1109/HONET.2018.8551341.

[5] R. Barzegar, A. Asghari Moghaddam, J. Adamowski, and E. Fijani, "Comparison of machine learning models for predicting fluoride contamination in groundwater," *Stoch. Environ. Res. Risk Assess.*, vol. 31, no. 10, pp. 2705–2718, 2017, doi: 10.1007/s00477-016-1338-z.

[6] R. P. N. Budiarti, S. Sukaridhoto, M. Hariadi, and M. H. Purnomo, "Big Data Technologies using SVM (Case Study: Surface Water Classification on Regional Water Utility Company in Surabaya)," *Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019*, pp. 94–101, 2019, doi: 10.1109/ICOMITEE.2019.8920823.

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

[8] R. E. N. MacAbiog and J. C. Dela Cruz, "Rainfall Predictive Approach for la Trinidad, Benguet using Machine Learning Classification," *2019 IEEE 11th Int. Conf. Humanoid, Nanotechnology, Inf. Technol. Commun. Control. Environ. Manag. HNICEM 2019*, 2019, doi: 10.1109/HNICEM48295.2019.9072761.

[9] T. Harter, "Groundwater Quality and Groundwater Pollution," *Groundw. Qual. Groundw. Pollut.*, 2003, doi: 10.3733/ucanr.8084.

[10] M. El-Fadel, E. Bou-Zeid, W. Chahine, and B. Alayli, "Temporal variation of leachate quality from pre-sorted and baled municipal solid waste with high organic and moisture contentNeugebauer, M., & So, P. (2017). The use of green waste to overcome the dif fi culty in small-scale composting of organic househol," *Waste Manag.*, vol. 22, pp. 269–282, 2002, doi: 10.1016/S0956-053X(01)00040-X.

[11] R. C. Contrera, M. J. Lucero Culi, D. M. Morita, J. A. D. Rodrigues, M. Zaiat, and V. Schalch, "Biomass growth and its mobility in an AnSBBR treating landfill leachate," *Waste Manag.*, vol. 82, pp. 37–50, 2018, doi: 10.1016/j.wasman.2018.10.006.

[12] C. Zhuang, H. Zhao, C. Sun, and W. Feng, "Detection and Classification of GNSS Signal Distortions Based on Quadratic Discriminant Analysis," *IEEE Access*, vol. 8, pp. 25221–25236, 2020, doi: 10.1109/ACCESS.2020.2965617.

[13] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, 2010, vol. 1, no. Fskd, pp. 91–94, doi: 10.1109/FSKD.2010.5569740.

[14] 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.

[15] 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.