The Study of Groundwater Source by using KNN Classification

Suziyanti Zaib[1], Muhammad Sharfi Najib[2], Suhaimi Mohd Daud[3], Muhammad Faruqi Zahari[4] and Mujahid Mohamad[5]

[1,2,3,4,5] Faculty of Manufacturing & Mechatronic Engineering Technology, University Malaysia Pahang, 26600 Pekan, Pahang, Malaysia

suziyantizaib@gmail.com

Abstract. This study was focused on assessing the groundwater as a source using odor by electronic nose (E-nose). Water is a finite resource that essential for humans and ecosystem existence. The suitable quality water resources need to be paid attention since it controlled by naturalistic activities such as geology, motion of groundwater, and water-rock interaction. In general, it is tasteless, odorless, and nearly colorless liquid but in other aspect, it also fulfills the need of minerals in human body up to a certain limit. The anthropogenic activities had caused an imbalance of these minerals in water that result in degradation of its quality. The aim of this study to apply an E-nose in classification of water and to identify odor pattern. It consists of sensor array which mimic the olfactory receptor in human nose that ability to sniff volatile odor that usually undetectable by human nose. K-Nearest Neighbor (KNN) is applied in performing the intelligent classification with mean feature data as an input. The finding results shows that the E-nose sensitivity, specificity and accuracy indicates at 100% for Euclidean distance.

Keywords: Groundwater, Tube well, E-Nose, Odor pattern, Mean data feature, Intelligent classification, K-Nearest Neighbors.

1 Introduction

Groundwater has become as main source of drinking water that meets the demand for all human life including animals and more than half of the world's population depends on ground water for survival [1]. It is susceptible to pollution because of varieties of manmade activities, an excessive usage of fertilizers, pesticides, rapid industrial growth, increased anthropogenic activities and pollution of ground water aquifers also make many of wells unfit for consumption [2, 3]. In fact, the groundwater quality is degrading day by day that become a serious matter of concern, as poor-quality water poses a threat to health and the cleanliness of living beings [4]. Hence it is needed authentication technologies and more study to improve the water quality and environmental sector.

In Malaysia, water sources are rivers and streams, which depend heavily on rainfall and groundwater with the quality of the raw water monitored by state water monitoring

and controlling authorities [5]. Meanwhile, the public water supply in Kelantan is operated by Air Kelantan Sdn. Bhd. (AKSB), through a concessionaire agreement with state government, responsible for the development, operation and maintenance of the ground water supply system in the state [6]. There are many research proposal based on water such as [7] study on groundwater measures or quality and its suitability for drinking and irrigation, and contaminant sources. Mubarra Noreen and Isma Younes make a research on appraisal of water quality measurements [8], [9] study on analysis of well water quality. The quality of raw water varies because of high amounts of microorganisms or industrial contaminants and these disturbances accompanied with unpleasant odor or taste that do generally not possesses a health risk, if the raw water is processed properly [10].

In recent years, application of E-nose in environmental monitoring has been classified into four main categories; water quality monitoring, air quality monitoring, process control, and pollution or odor control [11]. E-noses in advance provide low cost and portable electronic instruments that have rapid response, good sensitivity and precision [12]. The principle of E-nose is starts with odor molecules by the scent sampler is react with gas sensor array and converted into electrical signals [13]. Nowadays, it used selected combination of sensors in order to improve the accuracy [14].

The nonparametric analytical tools and concepts are needed to analyze such massive data to keep up with rapidly growing technology and which can also be used in the analysis of continuous streaming big data. [15] study on urban river (black odor river) with the efficacy of an E-nose with Linear Discriminant Analysis, Analysis of Variance for pH, chemical oxygen demand, total nitrogen content, and total phosphorous content. Meanwhile, [16] is focus on accuracy level of KNN and Support Vector Machine algorithm in classification water quality status based on river.

KNN is a simplest memory-based algorithm that widely used in predictive analysis by observations in the training set to find the most similar properties[17, 18]. It also selected as one of the top-10 data mining algorithms that assigns a class label to an unlabeled object based on the class labels of its *k* nearest neighbors [19]. The aim of this method to classify new objects based on training data that closest to the object [20, 21]. In determine the *k*-variable this issue had been consider in this research with the smaller *k* will result a higher noise sensitivity, while the larger *k* smoother decision boundaries and lower noise sensitivity [22]. So, it recommended to choose the number of *k* is an odd sequence, and must not be a multiple of the number of classes in order to avoid any possible tie [23, 24]. As for KNN classifier, the distance function and classify rule were explored to obtain the optimum output in identify differences between samples with one portion of the first data is used as training and the remaining is for testing purposed [25].

The depth of well will depend the concentrations of arsenic [26]. Tube wells is classified into three depth categories, shallow wells (< 140 feet), intermediate-depth wells (140-300 feet) and deep wells (≥ 300 feet)[27]. Nonparametric methods have capabilities to yield transformational changes in research based on the data generated by researchers and also needed to analyze data to keep up with rapidly growing technology or used in the analysis of continuous streaming big data [28]. This study is focus on

classification of groundwater sources with the samples taken from tube wells with KNN as a classifier and to identify the odor pattern.

2 Materials and Methodology

2.1 Preparation of Sample, Measured the Data and E-nose Setup

The experiment setup in this research is beginning with preparation of ground water sample via tube wells and E-nose setup that shows in Figure 1. The data and samples used in this study are collected in Kelantan from four tube well with depth in the range of 150 to 180 feet (intermediate depth wells). The sample of ground water that taken on the site is stored in bottle and been labelled. Meanwhile one sample of ground water that had undergo treatment process (sample E) is collected at water plant Kg. Puteh as a reference. The samples from tube well are labelled as sample A, B, C and D. One set of E-nose that consist of four MOS sensor is setup and connect to laptop (software) to read and save the raw data of all samples. Before the measurement of sample is taken, it is run first without samples in order to warm up and testing the E-nose. The measurement of E-nose is based on different response of sensor and odorant molecule by observing the change of electrical resistance that later is stored as raw data.

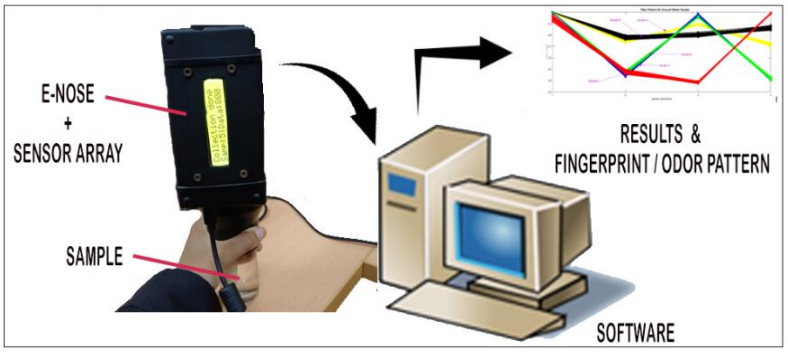

Fig. 1 E-nose diagram and preparation of data measurement for groundwater.

To perform this study for software analysis, a software code was developed in MATLAB. The process of data measurement is starts by collect 200 data x 4 sensors x 5 times for each sample. Next, E-nose is in a rest condition for a few minutes before proceeded to the other samples as to free it from odorant molecule of the previous sample. In order to get optimal sense of sensor, the purging process or air cleaning is done for every next cycle data collection to prevent sensor blockage of the previous odorant samples. So, the total output of raw data from E-nose for sample A, B, C, D and E is 1000 data x 4 sensor for each (matrix numeric number). In data pre-processing, normalization and mean calculation technique is applied. For normalization, it used to normalize the raw data to get only in the range (0 - 1) for each sensor so in order to reduce error and normalizing the range of the data. From the normalize data, the mean

calculation is applied to minimize the size of data set from 200 data x 5 x 4 sensor to 200 data x 1 x 4 sensor for each sample. The processing phase is very crucial in data analysis starts with raw data, normalize the raw data measurement and mean data features that interprets by odor pattern graph.

2.2 K-Nearest Neighbors (KNN)

The mean features data is used as input in KNN model for intelligent classification phase. KNN had been chosen for this study as it one of the simplest machine learning algorithms for classification, easy to run and the mostly used by others researcher in their project. About 200 mean data splits into two sections, training and testing data by specific ratio (X_1:Y_1 to X_i: Y_i) with the selection ratio is divided by 10 datasets of sample. This selection is repeated by increase the ratio 10% until it reaches X_i: Y_i is 90:10 where i=10 that show in the Table 1. As for example of 10:90, the dataset has been divided into 10% training and 90% testing data.

Table 1 KNN data splitting ratio

Splitting Ratio (X_i, Y_i)	Training	Testing data
X_1:Y_1	TR_1	TE_1
X_2:Y_2	TR_2	TE_2
…	…	…
…	…	…
X_i: Y_i	TR_m	TE_i

In this case, the training dataset is for validation while the testing dataset is used for predicting. Next, the similarity function is obtained using Euclidean distance with implication of three different rule (nearest, random and consensus). This process in done by *knnclassify* function in MATLAB with parameter of *k*=1,2,3,6 and 9 based on nearest neighbors (represent in confusion matrix). Lastly, the result from KNN classification is display in term of percentage for sensitivity, specificity and accuracy.

3 Results and Analysis

3.1 Data Analysis

The results of normalized data for all sample (5 sample x 1000 data) were plot with yellow plus sign represent for sample A from tube well 1, black cross mark represent for sample B from tube well 2, blue square mark represent for sample C from tube well 3, green circle mark represent for sample D from tube well 4 and red star mark represent for water sample from well that been treated. Figure 2 shows visualization pattern with 1000 of data for each sample in each sensor (S1, S2, S3 and S4). It shows that S1 is the highest response of sensor to odorant sample with the value in the range of 0.8 to 1.0.

The other three sensor less sensitive towards samples because of wide range of sensor responds with average minimum responds is 0.3. However, it can be concluded that the significant variation responds produce different sensitivity of sensor toward sample. Vertical axis shows the quantitative value of the resistance reading in whilst horizontal axis shows the sensor attribution or sensor array.

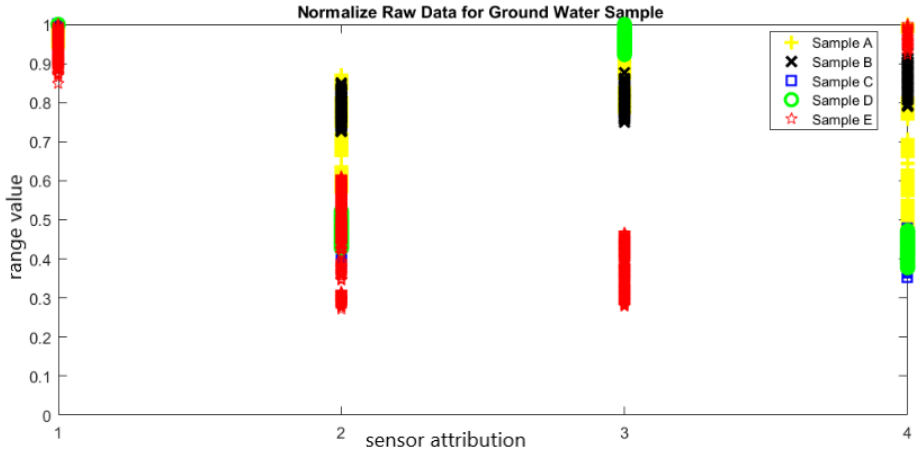

Fig. 2 Normalize Raw Data for Sample A, B, C, D and E.

In order to make the graph pattern easier to see by naked eyes, the solid line is drawn between sensor's response. The previous 1000 data for each sensor now become 200 data after the size of dataset in reduce by mean calculation technique. In Figure 3 represent the most similar responds in between sample C and D. In general, S1 and S2 shows the response a bit similar for each sensor with S1 higher than S2 but not implement to S3 and S4. From the mean data features, S1 still produced the highest response in the range 0.9 to 1.0. Meanwhile S2 had two different range which is 0.7 to 0.9 represent for sample A and B and 0.4 to 0.6 for sample C, D and E. The most difference response of 5 samples clearly can be shown in S4.

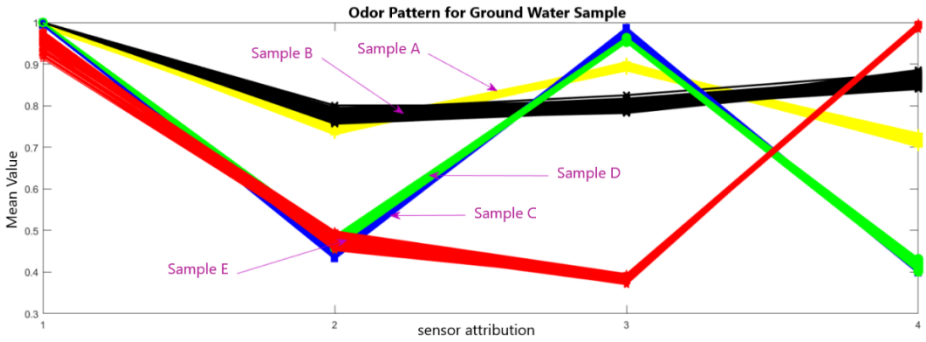

Fig. 3 Normalize Raw Data for Sample A, B, C, D and E.K-Nearest Neighbors (KNN)

The result of KNN voting based on nearest neighbors has been produced in confusion matrix to evaluate the KNN algorithm classification and demonstrates the correct classification and misclassification of data. Parameter of k=1, 2, 3,6 and 9 is used in this project as a controlling variable parameter for KNN classifier.

Fig. **4** represent for the Euclidean distance with nearest rule implication. All data are in the correct class indicate that KNN has ability to classify all samples for all splitting ratio with the nearest rule.

	k=1					k=2					k=3					k=6					k=9				
	Training : 10 , Testing : 90 [10 : 90]																								
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
A	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0
B	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0
C	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0
D	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0
E	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180
	Training : 20 , Testing : 80 [20 : 80]																								
A	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0
B	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0
C	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0
D	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0
E	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160
	Training : 30 , Testing : 70 [30 : 70]																								
A	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0
B	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0
C	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0
D	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0
E	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140
	Training : 40 , Testing : 60 [40 : 60]																								
A	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0
B	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0
C	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0
D	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0
E	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120
	Training : 50 , Testing : 50 [50 : 50]																								
A	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0
B	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0
C	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0
D	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0
E	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	Training : 60 , Testing : 40 [60 : 40]																								
A	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0
B	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0
C	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0
D	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0
E	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80
	Training : 70 , Testing : 30 [70 : 30]																								
A	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0
B	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0
C	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0
D	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0
E	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60

Fig. 4(a) Confusion Matrix of KNN for Euclidean distance with nearest rule.

	Training : 80 , Testing : 20 [80 : 20]																								
A	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0
B	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0
C	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0
D	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0
E	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40
	Training : 90 , Testing : 10 [90 : 10]																								
A	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0
B	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0
C	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0
D	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0
E	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20

Fig. 4(b) Confusion Matrix of KNN for Euclidean distance with nearest rule.

Next, the results of *knnclassify* function with Euclidean distance and random rule shown in Figure 5 are in correct class that also has ability to classify all samples for all splitting ratios.

	k=1					k=2					k=3					k=6					k=9				
	Training : 10 , Testing : 90 [10 : 90]																								
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
A	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0
B	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0
C	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0
D	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0
E	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180
	Training : 20 , Testing : 80 [20 : 80]																								
A	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0
B	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0
C	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0
D	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0
E	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160
	Training : 30 , Testing : 70 [30 : 70]																								
A	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0
B	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0
C	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0
D	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0
E	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140
	Training : 40 , Testing : 60 [40 : 60]																								
A	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0
B	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0
C	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0
D	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0
E	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120
	Training : 50 , Testing : 50 [50 : 50]																								
A	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0
B	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0
C	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0
D	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0
E	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	Training : 60 , Testing : 40 [60 : 40]																								
A	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0
B	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0
C	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0
D	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0
E	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80

Fig. 5(a) Confusion Matrix of KNN for Euclidean distance with random rule.

	Training : 70 , Testing : 30 [70 : 30]																								
A	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0
B	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0
C	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0
D	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0
E	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60
	Training : 80 , Testing : 20 [80 : 20]																								
A	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0
B	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0
C	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0
D	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0
E	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40
	Training : 90 , Testing : 10 [90 : 10]																								
A	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0
B	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0
C	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0
D	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0
E	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20

Fig. 5(b) Confusion Matrix of KNN for Euclidean distance with random rule.

Meanwhile, KNN with consensus rule is represent in Figure 6 also give the same results as nearest and random rule. For example, the total actual sample is same with the total predicted samples with = 180, 160, 140,120, 100 for the data splitting ratio 10:90, 20:80, 30:70, 40:60 and 50:50 The data is classified by using 1, 2, 3, 6 and 9 nearest neighbors. In summary, the different rule and parameter in this project produce the same results.

	k=1					k=2					k=3					k=6					k=9				
	Training : 10 , Testing : 90 [10 : 90]																								
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
A	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0
B	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0
C	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0
D	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0
E	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180	0	0	0	0	180
	Training : 20 , Testing : 80 [20 : 80]																								
A	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0
B	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0
C	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0
D	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0
E	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160	0	0	0	0	160
	Training : 30 , Testing : 70 [30 : 70]																								
A	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0
B	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0
C	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0
D	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0
E	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140	0	0	0	0	140
	Training : 40 , Testing : 60 [40 : 60]																								
A	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0
B	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0
C	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0
D	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0
E	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120	0	0	0	0	120

Fig. 6(a) Confusion Matrix of KNN for Euclidean distance with consensus rule.

	Training : 50 , Testing : 50 [50 : 50]																								
A	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0
B	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0
C	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0
D	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0
E	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0	0	100
	Training : 60 , Testing : 40 [60 : 40]																								
A	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0
B	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0
C	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0
D	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0
E	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80	0	0	0	0	80
	Training : 70 , Testing : 30 [70 : 30]																								
A	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0
B	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0
C	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0
D	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0
E	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60	0	0	0	0	60
	Training : 80 , Testing : 20 [80 : 20]																								
A	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0
B	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0
C	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0
D	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0
E	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40	0	0	0	0	40
	Training : 90 , Testing : 10 [90 : 10]																								
A	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0
B	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0
C	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0
D	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0
E	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	0	0	20

Fig. 6(b) Confusion Matrix of KNN for Euclidean distance with consensus rule.

Based on confusion matrix of KNN classifier, the performance measured is evaluated by statistical analysis in term of percentage of accuracy, specificity and sensitivity. Accuracy is the sum of true positive (TP) and false negative (FP) of sample A, B, C, D and E divided by the total number of the samples. Meanwhile specificity is the number of TN divided by the total data of FP and TN. Sensitivity is the result of TP divided by the sum of true positive false negative result. The classification performance measure of ground water sample based on accuracy, sensitivity and specificity is 100% as show in Figure 7.

k	1			2			3			6			9		
TRANING					10		TESTING							90	
	Accu	Sen	spec	Accu	Sen	spec	Accu	Sen	spec	Accu	Sen	spec	Accu	Sen	spec
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					20		TESTING							80	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					30		TESTING							70	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					40		TESTING							60	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					50		TESTING							50	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					60		TESTING							40	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					70		TESTING							30	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					80		TESTING							20	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TRAINING					90		TESTING							10	
Nearest	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Random	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Consensus	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Fig. 7 Performance Measure.

4 Conclusions

In conclusion, odor pattern for samples A, B, C, D and E that represent for ground water samples were established. E-nose as a measurement tool in environmental field especially in water had been apply with its capability to classify 5 different samples. The

classification of ground water is successful with the accuracy, sensitivity and specificity approximately 100%.

Acknowledgements

This work was funded by UMP Postgraduate Research Grants Scheme (PGRS200346). We would like to thank and appreciate all staff and students of Faculty Electrical and Electronic Engineering, Faculty of Manufacturing & Mechatronic Engineering Technology, Universiti Malaysia Pahang (UMP) that involved in this project and also the staff at water's laboratory at AKSB for giving contribution in providing the sample.

References

1. Yisa, J., Jimoh, T.O., Oyibo, O.M.: Underground Water Assessment using Water Quality Index. Leonardo J. Sci. 33–42 (2012).
2. Navis Karthika, I., Thara, K., Dheenadayalan, M.S.: Physico-Chemical Study of the Ground Water Quality at Selected Locations in Periyakulam, Theni district, Tamilnadu, India. Mater. Today Proc. 5, 422–428 (2018).
3. Kumar, K.S., Prrasad, K.H., Rajesh, B., Prasad, R.S., Venkatesh, T.: Assessment of Ground Water Quality Using Water Quality Index. 2, 103–108 (2015).
4. Priyadarshi, H., Pradesh, U., Priya, S., Nagar, M., Alvi, S.H., Jain, A., Pradesh, U.: Physico-Chemical Analysis of Selected Groundwater Samples From Sikandra Rao Town and Its Adjoining Villages (Hathras District) Uttar. 10, 355–368 (2019).
5. Rahmanian, N., Ali, S.H.B., Homayoonfard, M., Ali, N.J., Rehan, M., Sadef, Y., Nizami, A.S.: Analysis of physiochemical parameters to evaluate the drinking water quality in the state of perak, Malaysia. J. Chem. 2015, 1-10, (2015).
6. Department of Irrigation and Drainage (DID): Review of the National Water Resources Study (2000-2050) and Formulation of National Water Resources Policy: Final Report, August 2011, Volume 10, Kelantan. (2011).
7. Adimalla, N., Li, P., Venkatayogi, S.: Hydrogeochemical Evaluation of Groundwater Quality for Drinking and Irrigation Purposes and Integrated Interpretation with Water Quality Index Studies. Environ. Process. 5, 363–383 (2018).
8. Apriliani, I.M., Purba, N.P., Dewanti, L.P., Herawati, H., Faizal, I.: Open access Open access. Citizen-Based Mar. Debris Collect. Train. Study case Pangandaran. 2, 56–61 (2021).
9. Hashim, M., Nor, S.S.M., Nayan, N., Mahat, H., Saleh, Y., See, K.L., Norkhaidi, S.B.: Analysis of Well Water Quality in the District of Pasir Puteh, Kelantan, Malaysia. IOP Conf. Ser. Earth Environ. Sci. 286, 0–10 (2019).
10. Vagin, M.Y., Eriksson, M., Winquist, F.: Section 2 : The electronic tongue Manuscript 27 : Drinking water analysis using an electronic tongues. In: Electronic Noses and Tongues in Food Science. pp. 255–263 (2016).
11. Bieganowski, A., Józefaciuk, G., Bandura, L., Guz, Ł., Łagód, G., Franus, W.: Evaluation of hydrocarbon soil pollution using e-nose. Sensors (Switzerland). 18, (2018).
12. Alphus Dan Wilson, A.D.: Recent progress in the design and clinical development of

electronic-nose technologies. Nanobiosensors Dis. Diagnosis. 15 (2016).
13. Xu, L., Yu, X., Liu, L., Zhang, R.: A novel method for qualitative analysis of edible oil oxidation using an electronic nose. Food Chem. 202, 229–235 (2016).
14. Le Maout, P., Wojkiewicz, J.L., Redon, N., Lahuec, C., Seguin, F., Dupont, L., Mikhaylov, S., Noskov, Y., Ogurtsov, N., Pud, A.: Polyaniline nanocomposites based sensor array for breath ammonia analysis. Portable e-nose approach to non-invasive diagnosis of chronic kidney disease. Sensors Actuators, B Chem. 274, 616–626 (2018).
15. Qiu, S., Hou, P., Huang, J., Han, W., Kang, Z.: The Monitoring of Black-Odor River by Electronic Nose with Chemometrics for pH , COD , TN , and TP. Chemosens. 2021. 9, 1–12 (2021).
16. Danades, A., Pratama, D., Anggraini, D., Anggriani, D.: Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status. 2016 6th Int. Conf. Syst. Eng. Technol. 137–141 (2017).
17. Barua, S., Ahmed, M.U., Begum, S.: Towards intelligent data analytics: A case study in driver cognitive load classification. Brain Sci. 10, 1–19 (2020).
18. Cherif, W.: Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. Procedia Comput. Sci. 127, 293–299 (2018).
19. Li, L., Yu, Y., Bai, S., Hou, Y., Chen, X.: An Effective Two-Step Intrusion Detection Approach Based on Binary Classification and k-NN. IEEE Access. 6, 12060–12073 (2017).
20. Rojik, A., Endroyono, Irfansyah, A.N.: Water Pipe Leak Detection using the k-Nearest Neighbor Method. Proc. - 2019 Int. Semin. Intell. Technol. Its Appl. ISITIA 2019. 393–398 (2019).
21. Nikhath, A.K., Subrahmanyam, K., Vasavi, R.: Building a K-Nearest Neighbor Classifier for Text Categorization. Int. J. Comput. Sci. Inf. Technol. 7, 254–256 (2016).
22. Ertuğrul, Ö.F., Tağluk, M.E.: A novel version of k nearest neighbor: Dependent nearest neighbor. Appl. Soft Comput. J. 55, 480–490 (2017).
23. Rahman, M.M., Charoenlarpnopparut, C., Suksompong, P., Toochinda, P., Taparugssanagorn, A.: A false alarm reduction method for a gas sensor based electronic nose. Sensors (Switzerland). 17, 1–19 (2017).
24. Sha'abani, M.N.A.H., Fuad, N., Jamal, N., Ismail, M.F.: kNN and SVM Classification for EEG: A Review. In: Lecture Notes in Electrical Engineering. Volume 632. pp. 555–565. Springer (2020).
25. Zahed, N., Najib, M.S., Tajuddin, S.N.: Categorization of Gelam, Acacia and Tualang Honey Odor-Profile Using K-Nearest Neighbors. Int. J. Softw. Eng. Comput. Syst. 4, 15–28 (2018).
26. World Health Organization, Addendum, F., Third, T.O., WHO, WHO, Edition, S.: Guidelines for drinking-water quality. World Health Organization (1997).
27. Wu, J., Yunus, M., Streatfield, P., Van Geen, A., Escamilla, V., Akita, Y., Serre, M., Emch, M.: Impact of tubewell access and tubewell depth on childhood diarrhea in Matlab, Bangladesh. Environ. Heal. 10, 1–12 (2011).
28. Mathur, S.: Nonparametric data science: Testing hypotheses in large complex data. Handb. Stat. 44, 201–231 (2021).