TOWARDS DEVELOPING A CLASSIFICATION MODEL FOR WATER POTABILITY IN PHILIPPINE RURAL AREAS

Melchizedek Ibarrientos Alipio

Electronics and Communications Engineering Department, Gokongwei College of Engineering,
De La Salle University, Manila, Philippines, Tel: +63 2 524 4611, e-mail:
melchizedek.alipio@dlsu.edu.ph

Received Date: October 3, 2019; Revised Date: March 1, 2020; Acceptance Date: April 8, 2020

Abstract

In the Philippines, access to safe and sustainable water source is a major problem especially in rural areas. Thus, water monitoring in different water resources has been practiced to ensure safe drinking water. However, manual monitoring of safe drinking water is known to be inconvenient since it requires high operational and transportation costs, and time consuming. This study develops a data-driven water classification model for rural household areas using sensor nodes and machine learning algorithm. Sensor nodes are installed in several water sources in different rural areas to collect water parameters such as pH, turbidity, total dissolved solids, and temperature which are wirelessly transmitted to a base station. The collected sensor data is used to build and train the model to classify water potability using a hard-voting method in ensemble learning. The ensemble learning combined three machine learning algorithms namely k-nearest Neighbor, Naive Bayes, and Classification and Regression Tree. Finally, data are sent to a cloud for data storage and remote monitoring. Results show that the voting classifier model achieves an accuracy of 97% compared with other stand-alone classification algorithms. Furthermore, the model achieves 90% match with conventional industrial laboratory test.

Keywords: Classification, Ensemble learning, Hard-voting, Water potability, Sensors

Introduction

Water is considered as a necessity to all life on Earth. In the Philippines, access to safe and sustainable water source is a major problem. Out of 101 million Filipinos, 9 million rely on unimproved, unsafe, and unsustainable water sources and about 19 million lack access to improved sanitation [1], [2]. Therefore, water monitoring of different sources has been practiced to ensure the quality of drinking water. However, manual monitoring is known to be inconvenient since it requires high operational and transportation costs. Water quality monitoring is mainly developed using Internet of Things (IoT) in this era where each network device can collect and sense the data, then share it across the internet. It is a network embedded with sensors, actuators, and software which allows interaction and exchange of data between devices [3]. The use of IoT in water monitoring have made an evident progress for water source inspection and operation [4]. Moreover, an autonomous monitoring system is possible by applying machine learning to the IoT platform [5]. Machine learning is made of algorithms that analyze the data, learn from the data, and then apply the knowledge they have gained to make informed decisions and actions. Several works address the issue of devising a comprehensive methodology that analyzes and predicts water quality with the help of certain water quality parameters by suggesting a model based upon machine learning techniques [6].

Traditional water laboratory testing is available today but it requires transportation and labor costs, and it takes days or weeks before the result is released. In this work, we develop a system to monitor the quality of water and classify it as either potable or non-potable in real-time for rural household areas in the Philippines. Through ensemble learning, we build a classification model from sensor dataset obtained from different water sources located in rural areas. The model classifies the quality of water in terms of potability in which the result is disseminated in real-time using 2G/3G mobile communication. We design the system to be as less costly, greater efficiency, and portable as possible.

This study makes the following contributions: (1) develop a sensor node device capable to collect water parameters such as pH, turbidity, total dissolved solids and temperature; and (2) build a predictive model for water potability using voting method from different classification algorithms based on sensor dataset obtain from different water sources from several Philippine rural household areas.

Related Work

Several works involved the use of machine learning in monitoring and classifying the water quality.

A study develops a low-cost wireless sensor network for water quality monitoring integrated with sensing, data acquisition, communication, information decision, and application [7]. The system includes solar powered sensor node layer with built-in rechargeable battery which collects the water quality parameters such as pH, temperature, turbidity, dissolved oxygen, sulphate, ammonia, and nitrate. These are implemented on WaspMote that comprises of a slot for an external memory for scalability of the system and then transmits the data through WiFi to a base station. The system is simply used for notifying the changes in water quality to localities. It also updates the end user about the possible effects of the changes in the parameters to their health. However, it was used to provide necessary alerts to the concerned stakeholders and did not deploy machine learning helpful for environmental researchers to monitor and predict water quality. Since the data gathered from the base station still involves human interpretation on the analysis, discrepancies on the results may also occur.

Another work develops a system that uses machine-to- machine (M2M) method which allows both wired and wireless systems to communicate, IoT process, and cloud computing [8]. The system uses 32-bit Intel Pentium Processor Galileo Gen 2, which serves as the interface of the device for controlling data acquisition, and performs processing and preprocessing of the data collected coming from the sensors used such as temperature, pH level, water level, and turbidity. These data are transmitted then received for monitoring purposes by the end user with the access of internet. The smart sensor device monitors the various parameters of water to verify its quality and monitor them for industrial application. However, the database maintained by the cloud computing only stores the received water parameters for future references. The system does not use machine learning which could be beneficial in monitoring as well as predicting the water quality in a certain water resource.

Covering a huge part of South Pacific Ocean, water around Fiji have been increasingly related to pollution and contamination. Another study develops a smart water quality monitoring system for Fiji seawater by measuring water quality parameters such as pH, oxidation and reduction potential (ORP), conductivity, and temperature using remote sensing technology [9]. The IoT system provides accurate and consistent data and enables real-time monitoring of water quality. The parameter values gathered from different water sources can be used in order to learn and develop an artificial system which can perform on its own in the form of neural network. The analysis involves error between the system output and the desired output which is fed back to the neural network. The system also used GSM technology to implement the sending of an alarm based on a reference parameter for immediate action to ensure water quality. However, the neural network created by the system only detects anomaly from the gathered data and does not make use of prediction of water quality.

Previous works in water quality monitoring involves the use of sensors which gather data from water sources and can be remotely access through the Internet. Some of these works applied machine learning to better understand the data from physico-chemical parameters of water and acquire useful information from them. As far as our knowledge, no study has yet to develop a classification model of water potability in rural areas in the Philippines. Classification is performed as it maximizes the use of machine learning and provides better understanding of sensor data wherein the end user can easily understand. This gap should be solved as the water quality classification and monitoring will remain tedious and unappreciated by the end user. Moreover, human interpretation might lead to inaccuracy in terms analysis of the results.

Building the Water Potability Classification Model

Building the model is comprised of two major parts: hardware and software. The hardware is consisting of the microcontroller and sensor devices to gather data about the water quality. The software includes data analytics using machine learning and server for data storage and predictive analytics for potability classification.

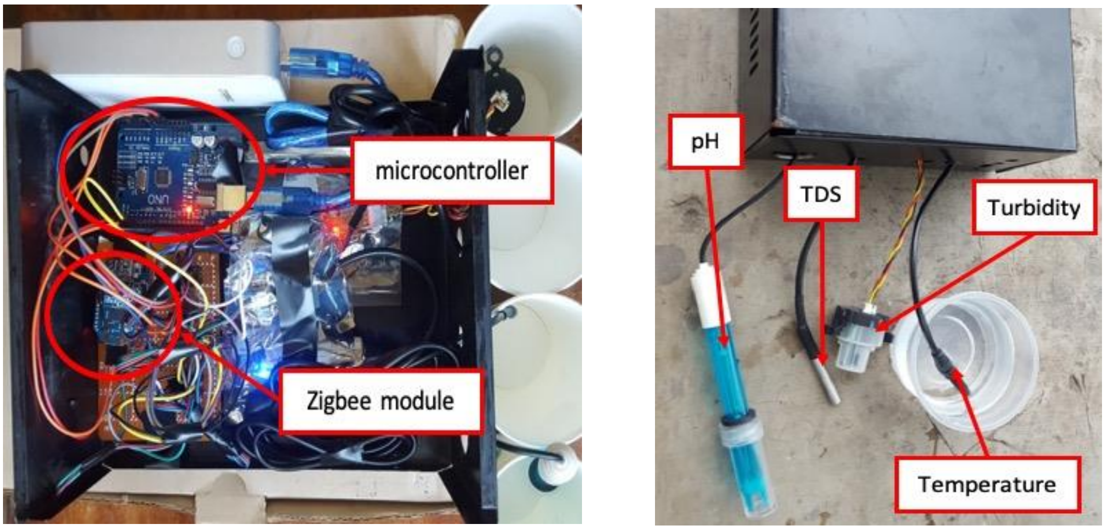
Design of Sensor Node Device

The sensor network comprises of sensor nodes composed of microcontroller, wireless module, and sensor. Different sensors measure the values of physico-chemical parameters from water namely temperature, turbidity, total dissolved solids (TDS), and potential hydrogen (pH). These sensors are waterproof so that each device can resist the corrosion once immersed in water. On the hand, Arduino [10] microcontroller was used to perform all the necessary computations which consist of processing the gathered data, integration of multiple data, managing the power source of the system, setting parameters for the sensors, and transmission to gateway.

From the sensor node, Zigbee module was used to wirelessly transmit the sensor data to a local server. XBee S2C is a Zigbee RF module typically used for wireless communication that provides an established wireless connectivity to end- point devices for data exchange at low data rates and medium range [11]. The power of the system is an external power bank that provides higher voltage enough to supply the whole system, and has an advantage of portability. The sensor node is enclosed in a metal casing to prevent the microcontroller and other modules from being exposed to external environment. The actual sensor node is shown in Figure 1 and the logical flow structure of data collection and transmission is shown in Figure 2.

Deployment and Data Collection

The target areas of the deployment focus on rural areas wherein lack of water monitoring for drinking water is apparent as per World Health Organization (WHO) [2]. In this study, we chose certain areas in Southern Luzon region (CALABARZON) which is composed of five provinces namely Cavite, Laguna, Batangas, Rizal and Quezon. Water samples are taken from 23 rural areas within the region which are identified to have poor safe drinking water as per the local government records. Figure 3 shows the geographical area of the region and the respective number of water samples obtained in a span of 90 days. Moreover, the locations are randomly selected among all the active areas in the region wherein no water and wastewater service providers are available.



(a) Interior components

(b) Exterior casing and sensors

Figure 1. Actual sensor node

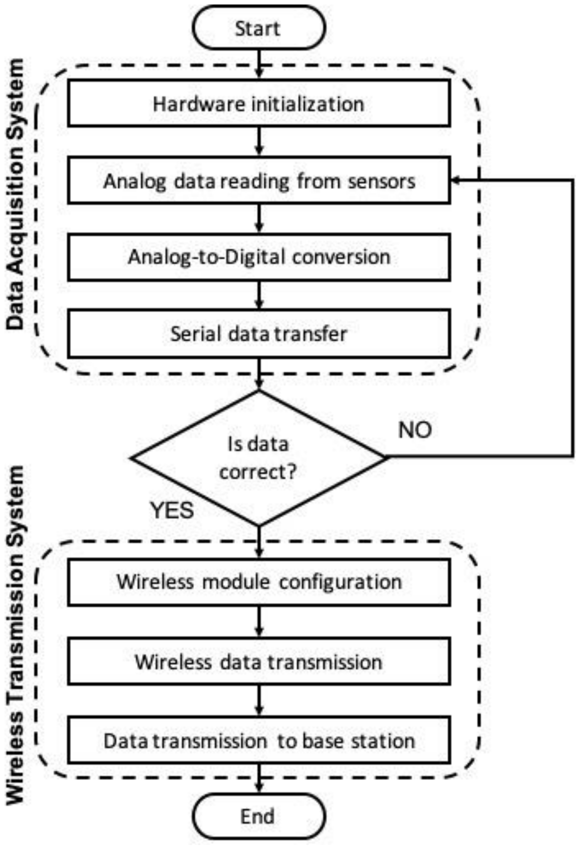

Start
Data Acquisition System
Hardware initialization
Analog data reading from sensors
Analog-to-Digital conversion
Serial data transfer
Is data correct?
NO
YES
Wireless Transmission System
Wireless module configuration
Wireless data transmission
Data transmission to base station
End

Figure 2. Logical flow structure of data collection and transmission

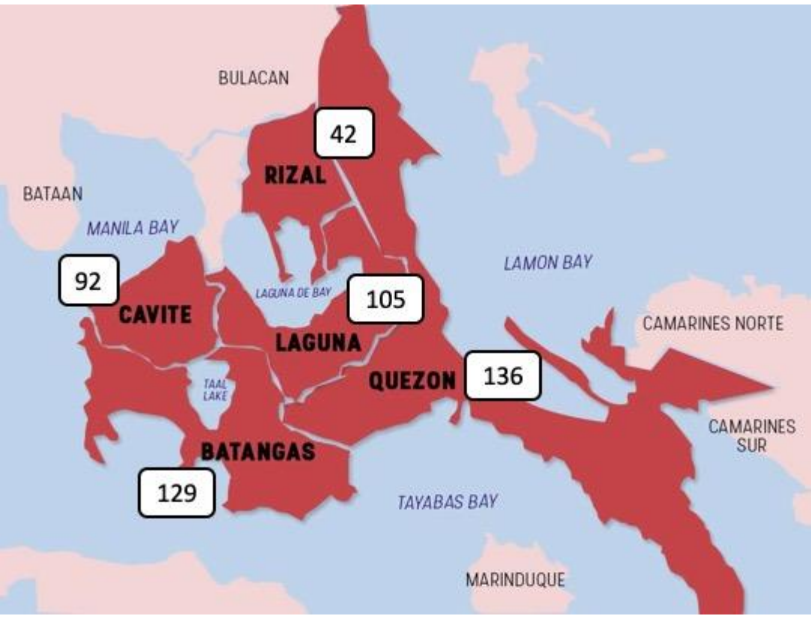
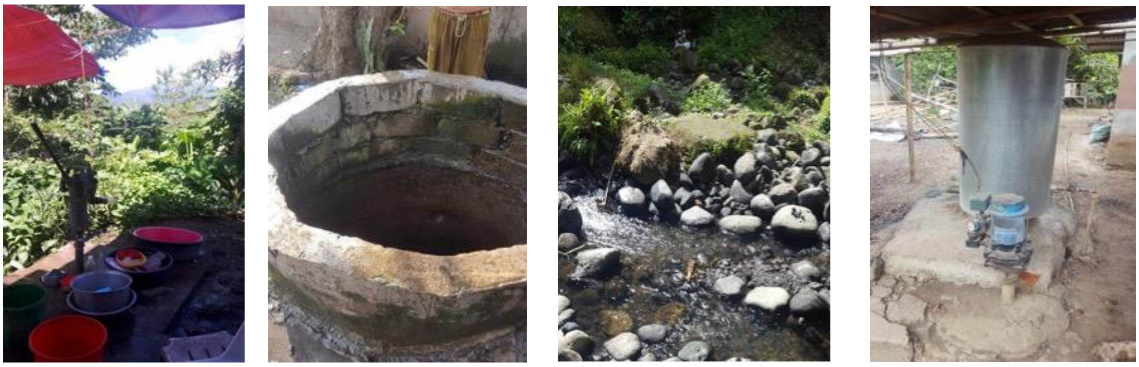


Figure 3. Water sample collection in CALABARZON region

A total of 500 samples was collected from different water sources such as artesian wells, deep wells, springs and cisterns/water tanks near the residential rural areas as shown in Figure 4. Table 1 shows the water sample dataset format consists of the following features: date, time, location, type of water source, and the terrain type. These features were selected based from the physico-chemical properties of potable water according to the Philippine National Standards for Drinking Water (PNSDW).

Table 1. Water Sample Dataset Format

Date	Time	Location	Source	pH	Turbidity	TDS	Temp	Terrain
04/18	13:06	Dayap, Calauan	Deep well	7.09	4.02	187	29.62	Residential
05/18	16:03	Balibago, Sta Rosa	Cistern	7.17	3.69	860	30.81	Residential
05/18	12:28	San Jose, Tagaytay	Spring	7.55	3.89	339	27.12	Mountain



(a) Artesian well (b) Deep well (c) Spring (d) Cisterns

Figure 4. Water source at rural areas

Water Potability Classification and Ensemble Learning

In building the model to classify the water potability, ensemble learning method was used as shown in Figure 5. Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance [12].

This method helps to minimize noise, bias and variance from the model. These methods are designed to improve the stability and the accuracy of machine learning algorithms. In this work, we used hard voting method of ensemble learning to combine three different machine learning algorithms namely Naive Bayes, K-Nearest Neighbor (kNN), and Classification and Regression Tree (CART). The first step is to create multiple classification models using some training dataset. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method. The samples obtained from different water sources served as the dataset for building the classification model for water potability. In this work, 80% (400 samples) of the total sample was used for training and 20% (100 samples) was used for testing. The 80-20 dataset splitting is the most used and conventional distribution used in machine learning binary classification modeling. The generated model classifies either the water is potable or not potable.

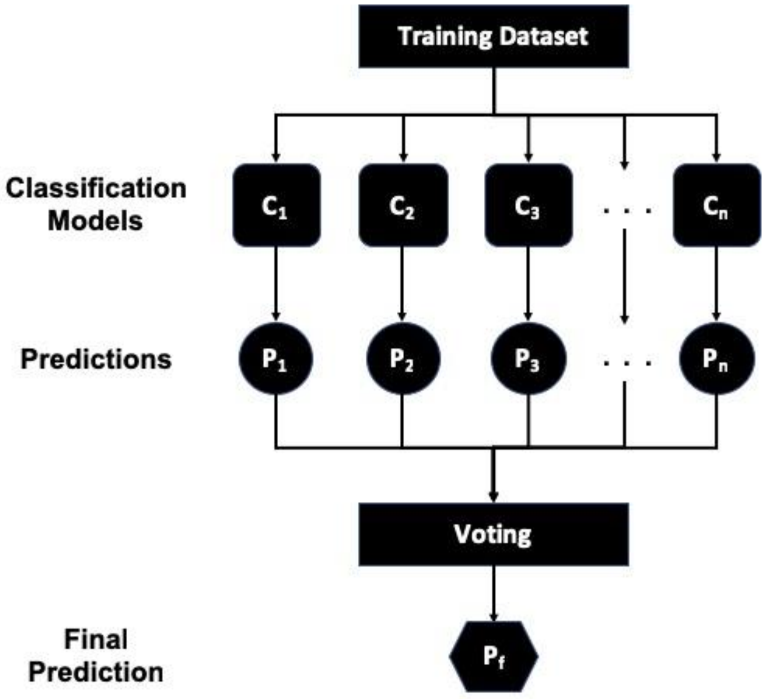


Figure 5. Ensemble learning algorithms and voting method

In hard voting method, we predict the final class label as the class label that has been predicted most frequently by the classification models [13]. Hard voting is the simplest case of majority voting. Here, we predict the class label $\hat{y}$ via majority (plurality) voting of each classifier Cj as shown in Equation 1.

$$\hat{y} = mode[C_1(x), C_2(x), \cdots, C_n(x)] \qquad (1)$$

Assuming that we combine three classifiers that classify a training sample as follows: classifier 1 = class 0; classifier 2 = class 0; classifier 3 = class 1. Via majority vote:

$$\hat{y} = mode[0, 0, 1] = 0 \tag{2}$$

In addition to the simple majority vote (hard voting), we can compute a weighted majority vote by associating a weight *wj* with classifier *Cj* shown in Equation 3.

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{n} w_j \chi_A\big(C_j(x) = i\big) \tag{3}$$

where χ_A is the characteristic function [Cj(x) = 1 ∈ A], and A is the set of unique class labels. Using the previous case with assigned weights 0.2, 0.2, 0.6 would yield a prediction $\hat{y} = 1$:

$$\arg[0.2(i_0) + 0.2(i_0) + 0.6(i_0)] = 1 \tag{4}$$

Although ensemble method can help improve machine learning competitions by devising sophisticated algorithms and producing results with high accuracy, it is often not preferred in the industries where interpretability is more important [14]. Nonetheless, the effectiveness of these methods is undeniable, and their benefits in appropriate applications can be tremendous especially in water potability classification.

Data and Results

We combined three machine learning algorithms and developed the voting classifier model from the sensor training dataset using Python-based Scikit Learn [16] software tool. The process of building the model is shown in Figure 6. To evaluate the model, we used the following classification metrics: accuracy, kappa statistics, precision, recall and precision [17].

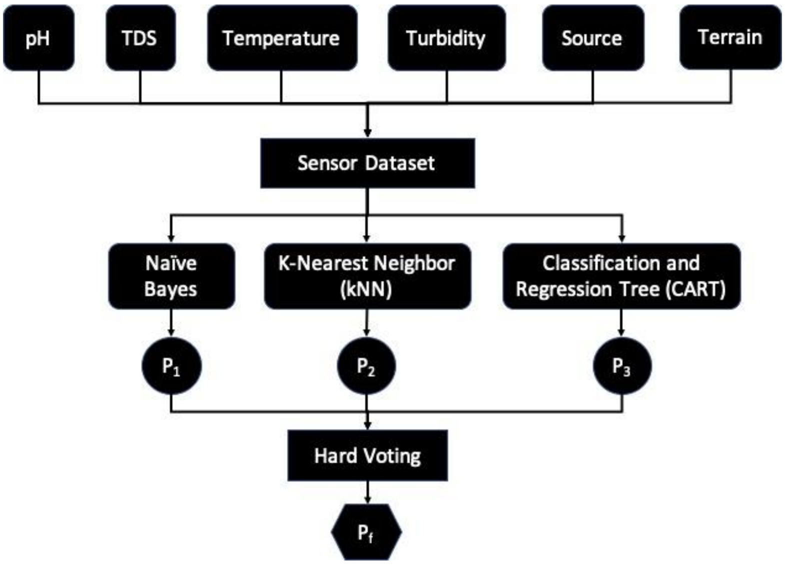
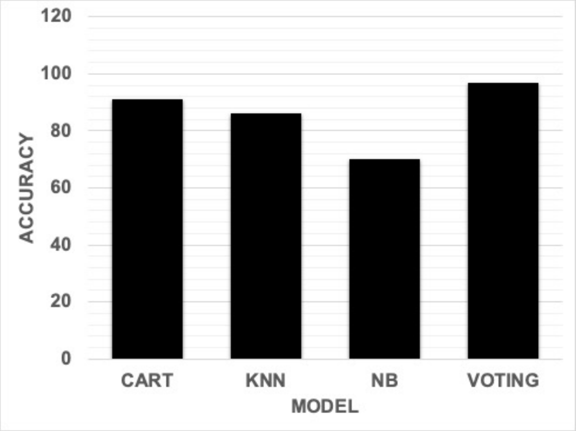
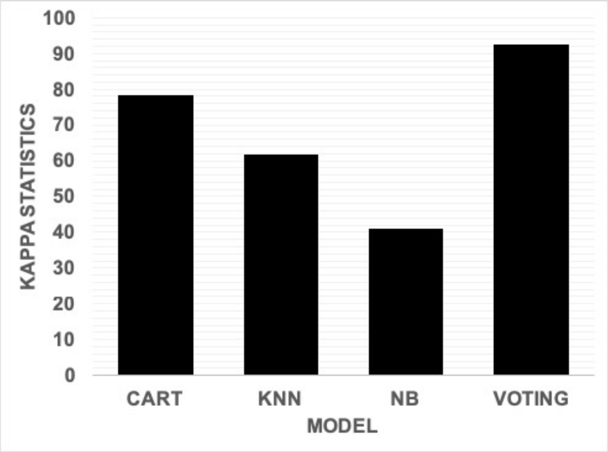


Figure 6. Classification model using hard voting method

Figures 7a and b show the comparison of accuracy and kappa statistic between the voting method and other stand-alone machine learning algorithms. As seen from the result, voting classifier obtained the highest values of 97% and 92.48%, for accuracy and kappa statistics respectively. A very high kappa-statistic indicates strong level of agreement between two parties, or simply an excellent interrater reliability. CART, KNN, and Naive-Bayes exhibit below average scores for these metrics. Naive Bayes is not suitable in producing such classification model. On the other hand, KNN is very dependent on the number of neighbors (k) where the testing sample is near. Changing the value of k can highly affect the accuracy of the model. It also requires scaling for large values of a specific parameter since it dominates the distance calculation of nearest neighbors. Scaling involves altering the data which is very critical since it can also affect the performance of the model. Moreover, CART tends to overfit when the algorithm captures noise in the dataset. The prediction model produced by this algorithm might get unstable with a very small variance in data. For highly complicated decision trees, they tend to have a low bias which makes it difficult for the model to work with new data.



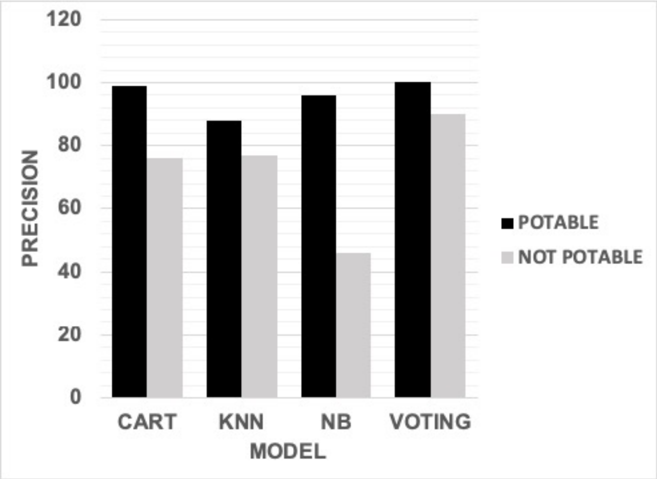
(a) Accuracy



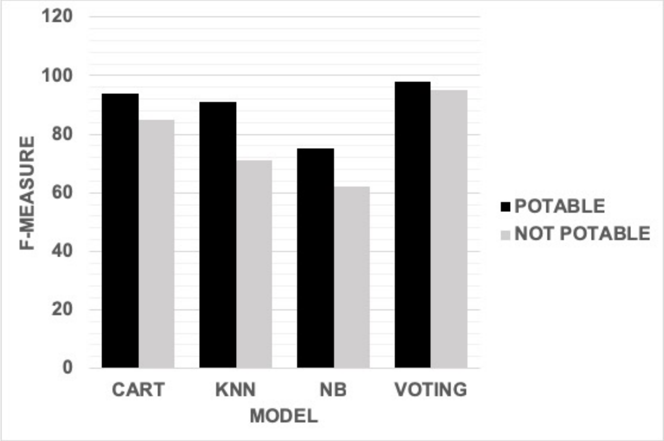
(b) Kappa-statistics

Figure 7. Accuracy test on classification model

Figure 8a shows the comparison of precision and recall among the machine learning algorithms between classifying potable and non-potable samples. It can be deduced that the voting classifier achieved the highest precision with 100% and 90% for potable and not potable, respectively. On the other hand, it achieved recall values of 96% and 100% for potable and not potable, respectively. These values indicate that the voting classifier has a high rate of expressing the proportion if the data points are actually relevant based on the precision and the ability to find all the relevant instances in a dataset based on the recall parameter. Moreover, Figure 8b shows the comparison of F-measure scores of all four algorithms. It can be observed that the voting classifier still obtained the highest F-measure with 98% and 95% for potable and not potable, respectively. In line with this, the voting classifier indicates an even class distribution on the harmonic mean of recall and precision.



(a) Precision



(b) F-measure

Figure 8. Classification metric score

A comparison of the water potability test between our system and the ideal laboratory test was also conducted by the study. 30 water samples were randomly taken from different rural areas and were testbed in both industrial laboratory set-up and using our portable device. Results shows that 27 out of 30 samples are match with conventional laboratory grade test process which gave the system an accuracy of 90%.

Conclusions

In this work, a data-driven system for monitoring water quality and classifying water potability particularly in Philippine rural residential areas was developed. A sensor node capable of monitoring physico-chemical properties of water such as potential hydrogen (pH), turbidity, total dissolved solids (TDS), and temperature was built. The data collected from the sensors nodes is sent to local base station which performs prediction of water potability based on a classification model. The classification model was developed based on a 500-sample dataset using voting method in ensemble learning. The result of water potability classification is sent to residential households via 2G/3G communication system for real-time dissemination. Our results show that the portable system is 90% matched with the conventional industrial laboratory test for water samples. In addition, the classification model achieved 97% accuracy.

In the future, other water parameters can be considered such as dissolved oxygen, electrical conductivity, and color to improve the accuracy of classification. Since the study focused on drinking water, other physico-chemical parameters should be also considered such as the amount of Escherichia coli (E-coli). The higher accuracy of potability prediction will be obtained. Moreover, aside from the type of terrain and water source, it would also be useful to add as an instance the weather condition of the area where the water sample is being acquired. Finally, it is recommended to use of advance machine and deep learning algorithms to increase the viability of the system.

Acknowledgement

The author would like to acknowledge the support of the Electronics and Communications Engineering Department of the De La Salle University in Manila, Philippines.

References

[1] "*The Water Crisis*" (n.d.) [Online]. Available: https://water.org/our-impact/water-crisis/ [Accessed: July 2019]

[2] "*Keeping Drinking-water Safe in the Philippines*" (n.d.) [Online]. Available: https://www.who.int/philippines [Accessed: July 2017]

[3] Q. Chen, G. Cheng, Y. Fang, Y. Liu, Z. Zhang, Y. Gao, and B.K.P. Horn, "Real-time learning-based monitoring system for water contamination," In: *2018 4th International Conference on Universal Village (UV)*, IEEE, Boston, Massachusetts, United States, pp. 1-5, 2018.

[4] N. Kitpo, Y. Kugai, M. Inoue, T. Yokemura, and S. Satomura, "Internet of things for greenhouse monitoring system using deep learning and bot notification services," In: *2019 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, Las Vegas, Nevada, United States, pp. 1-4, 2019.

[5] S. Imen, N. Chang, Y.J. Yang, and A. Golchubian, "Developing a model-based drinking water decision support system featuring remote sensing and fast learning techniques," *IEEE Systems Journal*, Vol. 12, No. 2, pp. 1358-1368, 2018. doi: 10.1109/JSYST.2016.2538082

[6] V.A. Usachev, L.I. Voronova, V.I. Voronov, I.A. Zharov, and V.G. Strelnikov, "Neural network using to analyze the results of environmental monitoring of water," In: *2019 Systems of Signals Generating and Processing in the Field of on Board Communications*, IEEE, Moscow, Russia, pp. 1-6, 2019.

[7] G.S. Menon, M.V. Ramesh, and P. Divya, "A low cost wireless sensor network for water quality monitoring in natural water bodies," In: *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, IEEE, San Jose, California, United States, pp. 1-8, 2017.

[8] P. Salunke, and J. Kate, "Advanced smart sensor interface in internet of things for water quality monitoring," In: *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, IEEE, Pune, India, pp. 298-302, 2017.

[9] A.N. Prasad, K.A. Mamun, F.R. Islam, and H. Haqva, "Smart water quality monitoring system," In: *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE),* IEEE, Nadi, India, pp. 1-6, 2015.

[10] "*Arduino*" (n.d.) [Online]. Available: https://www.arduino.cc/ [Accessed: March 2019].

[11] "*Xbee/xbee-pro s2c Zigbee*" (n.d.) [Online]. Available: https://www.digi.com/ [Accessed: January 2019]

[12] F. Moreno-Seco, J.M. Iñesta, P.J.P. de Leon, and L. Mico´, “Comparison of classifier fusion methods for classification in pattern recognition tasks,” *Structural, Syntactic, and Statistical Pattern Recognition*, Vol. 4109, pp. 705-713, 2006. doi: 10.1007/11815921_77
[13] Y. Zhang, D. Miao, J. Wang, and Z. Zhang, “A cost-sensitive three-way combination technique for ensemble learning in sentiment classification,” *International Journal of Approximate Reasoning*, Vol. 105, pp. 85-97, 2019. doi: 10.1016/j.ijar.2018.10.019
[14] J. Cao, S. Kwong, R. Wang, X. Li, K. Li, and X. Kong, “Class-specific soft voting based multiple extreme learning machines ensemble,” *Neurocomputing*, Vol. 149, pp. 275-284, 2015. doi: 10.1016/j.neucom.2014.02.072
[15] “*ThinkSpeak*” (n.d.) [Online]. Available: https://thingspeak.com/ [Accessed: July 2017]
[16] “*Scikit learn: Machine learning in python*” (n.d.) [Online]. Available: https://scikit-learn.org/stable/ [Accessed: October 2018]
[17] A. Tharwat, “Classification assessment methods,” *Applied Computing and Informatics*, in press, 2018. doi: doi.org/10.1016/j.aci.2018.08.003