E-tongue -A smart tool to predict safe consumption of ground water

A.M.P.B.Alahakoon
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
pandukalahakoon@gmail.com

M.M.Nibraz
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
nibraznibraz@gmail.com

P.M.S.S.B.Gunarathna
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
sanomikabandara@gmail.com

S.Thenuja
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
sthenuyah@gmail.com

K.A.D.C.P.Kahandawaarchci
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
chathurangika.k@sliit.lk

N.D.U.Gamage
Faculty of Computing
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
narmada.g@sliit.lk

***Abstract*— In Sri Lanka, reportedly 59.4% of population depends on water from natural sources which grabs the attention to make sure that these people are receiving and dealing with a safe water with a good quality for the usage. Although government is taking necessary action to provide a better quality of water, there has been always a need for a better educational session to educate people about the importance of maintaining water quality, importance of using a better-quality water and necessary precautionary to be taken to avoid Chronic Kidney Disease (CKD). It is mainly recognised issue that a smart solution must be implemented to solve the identification of water quality problem. E-Tongue: a smart device to predict safe consumption of ground water is an attempt to assist any kind of users to identify the water quality of a groundwater sample in real time by analysing the water quality parameters to predict the Water Quality Index (WQI). This task is achieved by designing a hardware device that embeds a set of sensors to read the value of water quality parameters and GPS to fetch location which will be then transferred to cloud environment for an easy access by the machine learning model to process and identify the WQI. It will then predict the water quality parameter levels that could be changed in future and check the possibility of the CKD. All the outputs will be finally displayed through mobile application. The solution is developed using supervised leaning techniques, optimization techniques and android, python as programming languages. The results of the study give the predictions with 73% of accuracy. The final product provides easily maintainable, real-time, high performance system to users.**

Keywords—water quality predicting, machine learning, predict risk of CKDu, Internet of Things, WQI

I. INTRODUCTION

Water plays a vital role in making up human body, regulating body temperature and performing many wonderous functions within the body. In that context, the water being consumed by human beings must be clean and contamination free to make sure the body is disease free and the functions are not being interrupted. One of the major sources of water for people in Sri Lanka is ground water. For 94% of water consumers, natural water sources directly or indirectly supply water whereas 6% of the remaining purchase water from the vendors [1][2].

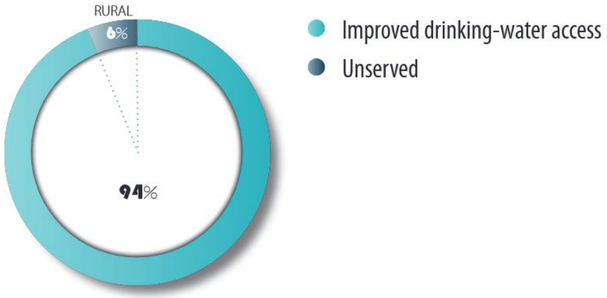

Figure 1: Population with access to improved drinking water sources

According to World Health Organization (WHO), nearly 1.8 million people worldwide dies annually due to water borne diseases including diarrhea and cholera and almost 90% are children under 5 years age [3]. Furthermore, up to 88% of water borne diseases arise from unsafe water supplies and inadequate sanitation and hygiene [4].

The National water Supply & Drainage Board of Sri Lanka as well as the water resources Authorities have been working on testing the quality of the water sources which has more human interactions and supply. However, the lack of advancement as well as the lack of implementation of technological predictions into the testing procedures still makes this process time consuming and heavily frantic.
To address this problem, this research proposes introduction of a single numeric index called Water Quality Index (WQI) which could be used to understand the entire purity of a water sample. The e-Tongue device is specifically fashioned to read the water quality data in real time and feed it to the Machine learning model to predict the WQI of the sample water in real time.

Although the device is designed to predict the quality of the water that the user is testing, there were implementation of Machine learning models to predict the possibility of a disease outbreak like Chronic Kidney Disease (CKDu) as well as future predictions of the behaviors of the water quality parameters over time.

II. LITERATURE REVIEW

According to Shafi U in [5], classical Machine learning algorithms like Su*pport Vector Machines (SVM), Deep Neural Networks (Deep NN)* and *Neural Networks (NN)* were used to measure the water quality with a highest accuracy of 93%. This level of accuracy distinctly indicates the importance of training the selected mathematical model under both controlled and open field conditions. It can also be stated that NN and Deep NN models are highly fitting in training a model which includes complex functions. Furthermore, it is important to note here that out of 30 water quality variables which were defined by World Health Organization (WHO), 25 variables are used in order to achieve this highest accuracy. However, using 25 different types of sensors makes this system economically infeasible due to budgetary restraints. Sakizadeh, M [6] used 16 water quality parameters along with an *Artificial Neural Network (ANN)* with Bayesian regularizations. This study capitulated correlation coefficients between the observed and the predicted values of 0.94 and 0.77. Even though the reduction of number of used variables didn’t impact on a vast difference in the accuracy and mean error, using 16 sensors on respective variables puts the progress of the study into a tight spot.

In a comprehensive overview [7], the study suggests a different and more efficient and scalable approach when it comes to selecting the variables for the WQI calculation. Once a data set has been collected, it is initially passed through a *Principal Factor Analysis (PFA)* where all the variables present in the dataset will be preprocessed to select the best suiting variables while preserving the overall variance as much as possible. This step addresses a principal issue when it comes to designing the device with using suitable sensors that could be within the planned budget. This paper also states the importance of parallelly training multiple algorithms, selecting the best algorithm with the highest accuracy and a least mean error to proceed with an effective mathematical model for a reliable WQI output [7][8].

As it mentioned in [9] CKDu is one of the major health issues in Sri Lanka and main responsible for it has not yet been identified. Among the suggested number of risk factors, it relates with the certain drinking water quality parameters strongly. The research was carried on Ulagalla cascade in Anuradhapura. According to the study it suggested that cumulative levels of heavy metals may aggravating the CKDu. Study that carried on Giradurukotte, Sri Lanka [10] thirty-two water samples were selected representing CKDu prevalent and non-prevalent communities and experimented the water quality parameter such as PH, Electrical conductivity, TDS. Based on measured water quality parameters twenty five percent of ground water bodies were identified as doubtful, whereas all the natural surface water bodies were identified as suitable for drinking purposes.

Higher TDS content was recorded in CKDu non-prevalent areas compared to prevalent areas.[11] Also, a significant difference ($p<0.05$) was observed in groundwater and surface water for TDS and higher TDS values were recorded in groundwater. Average fluoride content of shallow wells and surface water bodies in both areas were varied from 0.29 mg/l to 1.36 mg/l and 0.17 mg/l to 0.88 mg/l respectively. Based on results of ANOVA, a significant difference ($p>0.05$) was not observed in CKDu prevalent and non-prevalent areas. But a significant difference ($p<0.05$) was observed in drinking water sources while, groundwater recorded the higher fluoride content and high calcium concentrations were recorded in groundwater bodies located in CKDu non-prevalent areas.

General spotlight is given on predicting a least number of parameters with solitary algorithm or predicting a solitary parameter with a least number of algorithms. In [12], Dissolved Oxygen (DO) is predicted by using two types of ANN (Multi Linear Perception (MLP), Radial Based Function (RBF)), Linear Genetic Programming (LGP) and Support Vector Regression (SVR) techniques, above the models were evaluated by Root Mean Square Error (RMSE), Means Absolute Relative Error (MARE) and correlation coefficient. Get best performance model and which is used to understand deeper relationship of the water quality parameters but only considered a single parameter that was predicted by different regression problem. In [13], authors use five different machine learning models such as

General Neural Network (GNN), Back-Propagation Neural Network (BP- NN), SVR and Multi Linear Regression algorithm for predict the concentration of DO in a hypoxic river in China. Besides, it was considered as a single parameter prediction.

M Deqing, Z. Ying and C.Shangsong, in [14], utilized the Global System for Mobile Communications to detect the nature of water remotely. In their proposed framework, the basic water quality parameters, in particular, pH level, conductivity, dissolved oxygen, and turbidity are perused from the water through the individual sensors and it is then dissected by the controller and in the event that it is past the standard range, it is sent to relevant parties through an SMS, simultaneously. The information is also put away in a database and it is plotted to a graph for additional analysis. Be that as it may, this product is moderate for large water provider organizations or ventures since it comprises of costly parts.

As the main objectives of this paper is to design a smart device that could be used to identify the ground water quality by examining the water sample and introduce WQI value to Srilanka as a measurement of water quality. As additional functionalities following functions were implemented. (1) Predict risk of CKDu, (2) Predict future water quality parameters.

III. SYSTEM IMPLENTATION

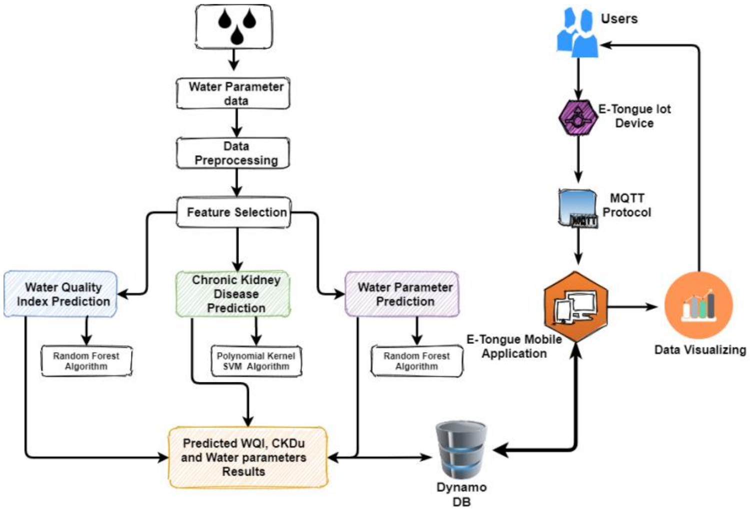
Users
Water Parameter data
E-Tongue Iot Device
Data Preprocessing
MQTT Protocol
MQTT
Feature Selection
Water Quality Index Prediction
Chronic Kidney Disease Prediction
Water Parameter Prediction
E-Tongue Mobile Application
Data Visualizing
Random Forest Algorithm
Polynomial Kernel SVM Algorithm
Random Forest Algorithm
Predicted WQI, CKDu and Water parameters Results
Dynamo DB

Figure 2: System overview diagram

As in Figure 2 shows system comprises of three principal segments. IoT based device that is utilized to gather the values of water quality parameters from the water samples is one of the fundamental parts. The subsequent part is the server backend which has services to decide water quality (WQI), foresee the risk of CKDu and water borne infections and predict future water quality parameters. Third component contains the web application and the mobile application where client can see the outcomes.

A. IoT Device Implementation

In this component, there is essentially two sections included, the first is equipment and second one is software. The equipment part has sensors which help to gauge the constant qualities, another is Arduino Nano, and Nodemcu esp8266 gives us inbuilt Wi-Fi capability which sets the device to communicate with the end user. An Arduino Nano microcontroller is utilized as the center controller of the framework. When the code is transferred to the microcontroller, no PC framework, console command is required to operations. The device capacities consequently and freely as indicated by the code transferred to the microcontroller. In this implementation, four sensors are utilized to quantify the fundamental water parameters. As it was concentrated from the past explores, the most basic water boundaries should have been observed by the normal clients are water pH level, water turbidity (darkness), water temperature, and the conductivity of water. Hence, five basic water boundaries which are temperature, pH level, turbidity, conductivity, and total dissolved solids can be estimated by this IoT device. From the implemented design we were able to measure sensor readings by submerging all the above-mentioned sensors inside the water container. All sensors read the water quality parameters and send the information to the microcontroller as electrical and analog signals.

At the point when the device starts dc current of 12V given to the unit and Arduino and WIFI jumps on. The parameters of water tested are transferred to the dashboard of the app over Wi-Fi with an interval of 12 seconds. The information bundle is sent as a JavaScript object notation (JSON). A Message Queuing Telemetry Transport (MQTT) association is built up with the server by the inbuilt Wi-Fi module in Nodemcu esp8266. An Arduino Nano board is utilized to actualize the information handling code. This Arduino code helps the sensors to get the sensor inputs by analog or digital forms, process them and setting up the Wi-Fi connection to send these values to the real-time database.

The temperature sensor passes data to the microcontroller through a data pin. While turbidity and conductivity sensors send an analog signal, which needs to converted to a digital signal. The output analog value ranges from 0 to 1023 and its converted using the (1).

$$ecVolt = ecSensorVal * (5.0 / 1024.0) \qquad (1)$$

After analog conversion we are able to differentiate each value for its respective dimensions. Conductivity is ability to pass electricity. The conductivity is dependent with temperature. The following equation is built to capture the amount of voltage pass between two electrodes.

$$EC25 = EC / (1 + tempCoef * (temp - 25.0)) \qquad (2)$$

The logic level convertor used as a bridge to transmit and receive signals hence the Nodemcu esp8266 and Arduino Nano works with two different voltages. For a smooth run all the sensors are connected to the Arduino Nano board and nodemcu esp8266 used only for data transmission purpose. We can observe real-time data on our smart phone at any place over Wi-Fi.

B. Server Back-End

Server backend consists the most important functionalities of the system. The output generated from the IoT device are consists the combination of structured, semi-structured and unstructured data. These data will be preprocessed and transformed to a proper defined structure. Then the data will be stored in the database. DynamoDB is the database of the system. Since the major functions are deployed on the cloud platform, serverless functions has developed to trigger the relevant functionality. Then the obtained results will be stored in the database. REST APIs also has developed to access results from the mobile application and web application.

C. Predict Water Quality

The heart of the research depends on the prediction of quality of a water sample that has been used in the device. As abundant as the algorithms are present in order to achieve such goal, it is mandatory to select the best and the most efficient algorithm to finalize the prediction value. Based on the data obtained from National Water Board and Drainage Department, different regression algorithms are used to train models and their results were compared and a best algorithm is selected based on RMSE and r2 values.

As per the 4 input variables (temperature, dissolved oxygen, pH and turbidity), Radom Forest Regressor seem to give a highest r^2 of 0.9782 and a lowest RMSE value of 1.2836. Random Forest precisely is a type of an additive model where it sums up the decisions from a sequence of base models to make a final prediction as the trees in random forest runs in parallel and there is no interaction between trees while building the trees.

$$g(x)=f0(x)+f1(x)+f2(x)+... \qquad (3)$$

g = sum of simple base models fi
The average prediction is saved to the central database for customer references via mobile app.

D. Predict Risk of CKDu

According to previously conducted studies, number of water quality parameters associate with CKDu. Hardness, fluorides, total dissolved solids are the most impact factors of water quality parameters that are associated with CKDu. From the water quality parameters that associate with CKDu pH, turbidity, temperature, dissolved oxygen, TDS and hardness is selected as the input parameters for the prediction. The location of the water source and the history of the CKDu positive patients also taken account for the prediction.

Four different machine learning classification algorithms were trained to get the best accuracy comparing the accuracy of the of the results. According to the average accuracy and average error, Polynomial Kernel SVM algorithm which has 0.76543 value if accuracy is selected to be implemented in system for prediction of CKDu risk in given water source.

$$K(x,y) = (x^Ty + c)^d \qquad (4)$$

Polynomial kernel is defined as above for the degree-d polynomials where x and y are inputs. Predicted results are stored in the database for user reference.

E. Predict future values of water quality parameters

Forecasting water quality parameter involves with a numerous dataset according to the different location or site. Multivariate Random Forest Regression (RFR) is the popular supervised machine learning algorithm for multi parameter regression problem. It has arbitrary number of decision tree that predict the multiple independent water quality parameters (temperature, pH, turbidity, dissolved oxygen) that consists the previous measured value of the water quality parameters for certain location or site. Thus, number of decision trees lead to predict the accuracy of the model.

In RFR each decision tree responses based on the predictor values that are selected independently and the predictor values of an original given dataset is the subset of the forest. Equation for optimal size of predictor input variables below.

$$\log_2 M + 1 \qquad (5)$$

M = number of input parameters (location or site, year, and month)

Prediction of the RFR is taken the average of each prediction of each tree is obtained. Following formula is used for RFR prediction.

$$p = \frac{1}{k}\sum_{k=1}^{k} k^{\text{th}} \text{ decision tree response} \qquad (6)$$

k = runs over the single decision tree in the forest.

According to the given result of the RFR model is to be implemented system for forecasting water quality parameters.

F. Mobile App

"E-tongue" is an android application that is developed using java as programming language and retrofit libraries that used to connect with the backend services. App is connected with the respected IoT device and display the results of the system to the users. Once IoT device takes a reading from a given water sample android application will display the predicted WQI value. Application also displays the risk of

CKDu of the given water sample and future water quality parameter values for given date range. As an additional feature that is useful for users, "E-tongue" provides the precautionary steps that can be used to improve the quality of the water, to reduce the risk of the CKDu and water borne diseases.

IV. TEST RESULTS

This section discusses the observations and the results of the solution. Initially IoT device detect the information from sensors that are installed. The detected information will be consequently sent to the web server database with an interval of 12 seconds, when an appropriate connection is set up with end device. Table no.1 shows the average device sensor readings that are stored into database for clean water sample and contaminated water sample.

Table 1: Results of clean and contaminated water samples

	Ph	EC(mS/cm)	Temperature(Celsius)	Turbidity (ntu)	TDS (ppm)
Clean water	7.12	0.23	25.44	0.02	161
Salty water	7.04	5.02	27.75	0.02	3550

The main objective of the research is to predict the quality of the given ground water sample. To acquire the expected outcome, selection of the machine learning algorithm is a crucial step. When predicting water quality, algorithm needs to predict a single numeric value using multiple inputs. The algorithm that uses for the testing and the accuracy of them has shown in the table 2. Among the algorithms Random forest algorithm which gives the highest r^2 value is selected for the prediction. Same as the water quality prediction, risk of CKDu prediction is tested with the different algorithms and selected best algorithm from them for the prediction of one Boolean value. For the prediction of CKDu Polynomial Support Vector Machine algorithm is used. The results of the algorithms are shown on table 3.

Table 2: Results of model training predicting water quality

Model	**RMSE value**	**R Squared value**
Linear Regression	1.9165	0.9249
k-Nearest Neighbors	4.1590	0.6424
Random Forest Regression	1.2836	0.9782
ANN	9.4993	0.2381
Ridge Regression	1.9167	0.9248
Lasso Regression	1.9217	0.9245
ENR	1.9358	0.9234

Table 3: Results of predicting ckdu.

Model	**Avg accuracy**	**Avg Error**
Polynomial Kernel SVM	0.7699	0.4201
Random Forest Regression	0.4969	0.5031
Gaussian Kernel SVM	0.5683	0.4317
Sigmoid Kernel SVM	0.5432	0.4568

Predicting water quality parameters for given geometrical location based on analyzing previous water parameter values needs an algorithm that is capable of multi value prediction. As same as the scenarios that discussed above, the best algorithm was selected via testing multiple algorithms. The test results can be found on table 4.

Table 4: Results of predicting water quality parameters

Model	**RMSE**
VAR	5.9604
RFR	2.9280
LSTM	9.1696
KNN	7.6127
SVR	15.0221

To check the performance of the solution, research group has used different water samples from different geometrical areas and water samples that specially prepared for testing. The test results for clean water sample and a laboratory salted water sample are mentioned on the following table 5.

Table 5: Results obtain from water quality prediction and ckdu prediction

	Ph	**EC**	**Temp**	**Turbidity (ntu)**	**TDS (ppm)**	**WQI**	**Risk of CKDu**
Clean water	7.12	0.23	25.44	0.02	161	80.4	false
Contaminated water	6.04	5.02	27.75	0.02	3550	63.2	false

For same clean water sample which has tested above and obtained on 2020.08.25, the test results of predicted water quality parameters for 2021.08.25 are shown on table 6**.**

Table 6: Results of predicting water quality parameters

Date	**Ph**	**EC**	**Temp**	**Turbidity (ntu)**	**TDS (ppm)**
2020.08.25	7.12	0.23	25.44	0.02	161
2021.08.25	7.08	0.22	26.64	0.013	159

The prototype model was tested under distinct circumstances and different water solutions. The results of the device were effective and as per the exploration goals. As referenced, the sensor readings are appeared on the screen of the E-Tongue mobile app.

V. CONCLUSION

This paper has presented a comprehensive approach to detect the quality of a given ground water sample, predict the risk of CKDu and future values of water quality parameters. The solution has used supervised learning techniques to obtain the expected research goals.

Results of the study shows that using machine learning approach, it is possible to build a real time device that detect the quality of the ground water and deliver users a better way to purify the water in order to consume.

As future work, “E-tongue” can be improved by adding the extra sensors such as flow rate and Calcium and increasing the accuracy of the prediction models. Turbidity sensor calibration can be improved with standard adjustments, if more monitory sources are to be invested on the turbidity sensor, a probe of high caliber and accuracy can be applied.

ACKNOWLEDGMENT

Sincere gratitude to the National Water Supply and Drainage Board and National Building Research Organisation of Sri lanka, National Hospital of Anuradhapura for providing required dataset and guidance for the research project.

REFERENCES

[1] UN-Water Global Analysis and Assessment of Sanitation and Drinking water - Sanitation, drinking-water and hygiene status overview*

[2] Handbook for water consumers – Ministry of water Supply & Drainage, national water Supply and Drainage Board.

[3] World Health Organization - Water sanitation hygiene, available at:https://www.who.int/water_sanitation_health/diseases-risks/diseases/diarrhoea/en/

[4] World Health Organization World Water Day Report - Water for Health-Taking Charge, available at: https://www.who.int/water_sanitation_health/takingcharge.html

[5] Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.

[6] Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. Model. EartSyst. Environ. 2016, 2, 8.

[7] Nabeel M. Gazzaz, Mohd Kamil Yosoff, Ahmad Zaharin Aris, Hafizan Juhair, Mohammad Firzul Ramil - Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors, Marine Pollution Bulletin 64 (2012) 2409–2420

[8]Umair Ahmed, Rafia Mumtaz , Hirra Anwar , Asad A. Shah, Rabia Irfan and José García-Nieto, Efficient Water Quality Prediction Using Supervised Machine Learning, Water 2019, 11, 2210; doi:10.3390/w11112210

[9]Wanasinghe, W., Gunarathna, M., Herath, H. and Jayasinghe, G., 2018. Drinking Water Quality on Chronic Kidney Disease of Unknown Aetiology (CKDu) in Ulagalla Cascade, Sri Lanka. *Sabaragamuwa University Journal*, 16(1)

10]International Journal of Advances in Agricultural and Environmental Engineering, 2016. Drinking Water Quality in Chronic Kidney Disease of Unknown Aetiology (CKDu) Prevalent and Non-prevalent Areas in Giradurukotte, Sri Lanka.

[11]Wasana, H., Aluthpatabendi, D., Kularatne, W., Wijekoon, P., Weerasooriya, R. and Bandara, J., 2015. Drinking water quality and chronic kidney disease of unknown etiology (CKDu): synergic effects of fluoride, cadmium and hardness of water. *Environmental Geochemistry and Health*, 38(1), pp.157-168.

[12] E. Olyaie, H. Z. Abyaneh and A. D. Mehr, "A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in delaware river", Geoscience Frontiers, vol. 8, no. 3, pp. 517-527, 2017.

[13] X. Ji, X. Shang, R. A. Dahlgren and M. Zhang, "Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of wen-rui tang river china", Environmental Science and Pollution Research, vol. 24, no. 19, pp. 16 062-16 076, 2017.

[14] M. Deqing, Z. Ying, C. Shangsong. (2012). Automatic Measurement and Reporting System of Water Quality Based on GSM. International Conference on Intelligent System Design and