Prediction of Quality of Water According to a Random Forest Classifier

Shahd Maadi Alomani[1], Najd Ibrahim Alhawiti[2], and A'aeshah Alhakamy[3]
Faculty of Computers and Information Technology, Master of Artificial Intelligence at University of Tabuk, Saudi Arabia[1,2,3]
Industrial Innovation & Robotics Center (IIRC), and Faculty of Computers and Information Technology,
Department of Computer Science at University of Tabuk, Saudi Arabia[3]

***Abstract*—Potable or drinking water is a daily life necessity for humans. The safety of this water is a concern in many regions around the world, since polluted waters are increasing and causing the spread of disease among populations. Continuous management and evaluation of the water which is meant for drinking is very essential and must be taken seriously. Often, the quality of water is evaluated through regular laboratory testing and analysis which can be tiresome and time consuming. On the other hand, advanced technologies using big data with the help of machine learning can have better results in terms of potability evaluation. For this reason, several studies have been conducted on predicting the quality of water and the several factors and classification that affect the prediction model. In this study, a random forest model was developed using PySpark classification to predict the potability of river water by relying on ten different features: pH, hardness, presence of solids, presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and finally potability. In addition, The developed model was able to predict water potability classification with a 1.0 accuracy, and 1.0 F1-score.**

***Keywords*—Big data; machine learning; classification; random forest; water quality; PySpark**

I. Introduction

When there is no water, there's no life. Freshwater is the most essential natural resource without which life, all forms of life, would not exist. Humans of all the other living organisms rely on the water not just for drinking but also for various aspects of their lives such as bathing, cooking, and watering their agricultural fields. In fact, there's even increased demand for water due to the increase in wide spreading urbanization, the development and expansion of the economic movement, and the general rapid increase in human population [1].

However, the water quality and its safety for use for different purposes is a complex issue. The overuse of the water both underground and on the surface in addition to other factors are causing the deterioration of water quality. One of the factors that are having a significant impact on water is the global climate change since it doesn't just affect the availability of water resources but also affects their future quality. Add to that the dangerous pollution resulting from the human activities where individual humans don't only dump their waste into rivers and wells, but also large factories could pollute rivers and underground water as a result of their chemical wastes [2]. As a matter of fact, the poor-quality water is the source of many water-bourne illnesses such as diarrhea. This means that using non-clean water especially for drinking raises health issues that can be avoided but choosing the appropriate water to drink [3].

The most common estimation of water quality has been the laboratory analysis which is time-consuming, expensive, and not very practical. The laboratory analysis of water requires the collection of water samples from different areas over a period of time, then transporting these samples in suitable conditions before they can be analyzed. Of course, this method is still being applied, but with the current development of technology, these processes can be made much more efficient by applying machine learning and big data tools [4]. Machine learning ML is a method of programming software in a way that allows them to learn from historical data and adapt accordingly such that they learn, assess their performance, and improve. Machine learning algorithms are often used to detect patterns in data and the non-visible behavior of data. There are several algorithms already in ML divided into classes: unsupervised, semi-supervised, supervised, and reinforced algorithms [5].

Random Forest RF is one of the machine learning algorithms through which several decision trees are merged together to achieve more accurate results. The term random forest also corresponds to the randomness of the method where the choice of samples is random. More specifically, a number of samples are randomly chosen from the training dataset in order to form what is called the "root node" samples. Furthermore, the choice of attributes in RF is also random, where the candidate attributes are selected at random, and after that the most suitable attribute is picked to be the "split node". The RF model starts with shuffles input sample data, creates many training sets that make up the decision trees, and finally chooses the output prediction results based on the majority of votes from the collection of decision trees [6].

In this paper, PySpark for the classification is utilized to evaluate water potability using a well-known Water Quality dataset. The Random Forest Classifier was used to build a model that asses various properties, including temperature, acidity, turbidity, and hardness, to arrive at an accurate decision. The developed model is evaluated to answer the following research questions for a better understanding of the presented work.

- **RQ1:** *What factors directly affect the potability of the water?*
- **RQ2:** *Can a random forest model effectively predict the quality of water based on these factors?*

The topic of water quality assessment was chosen due to its importance, and we have selected the Random Forest model to be our predictive model for the quality of water after reviewing the literature. Through literature, it was evident that

Random Forest provides the most effective and accurate results in evaluating the quality of water.

II. Related Work

A lot of studies discuss the quality of potable water and quality of river waters or near-shore water, and many of them also focus on finding the relationships between several factors affecting the quality of water and which of them have the greatest influence. Numerous studies applied machine learning algorithms to predict the influential factors affecting water quality, including the Random Forest algorithm.

The term water potability refers to the characteristic that the water is safe for human consumption, specifically drinking or cooking. For example, potable water must be free from micro-organisms or harmful chemicals [7]. Other factors that indicate the quality of water are shown in Fig. 1, such as the chemical pH, the clarity of the water, abundance of nutrients, presence or absence of pest animals and vegetation, etc. [8].

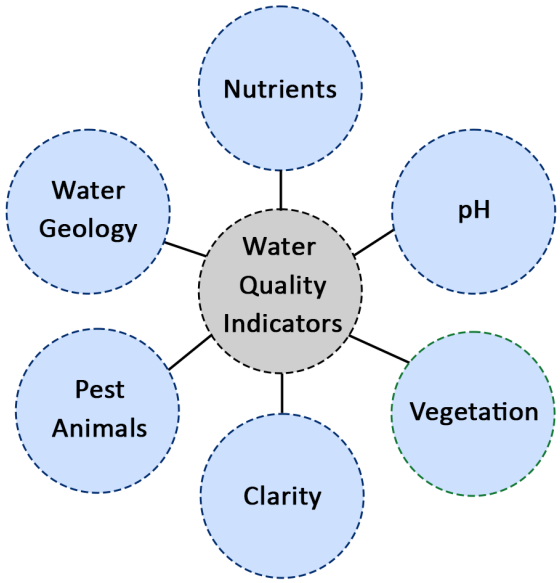

Fig. 1. Some of the Factors that Indicate Water Quality.

Back in 1960, a water quality index was created as a means to evaluate the safety of water [9]. After that period, many water assessment and quality management programs have been developed. In order to make appropriate decisions about drinkable water, one must be educated about the many factors that affect the safety of the water. Among other factors, potable water is affected by the source of origin, as well as whether it was treated before being delivered to houses, and the storage containers or water pumps and pipelines [10]. That's in addition to the factors within the water itself such as its temperature, its content of certain salts and minerals, its electrical conductivity, its pH, and other features [11]. The various water quality indicators can be divided into four classes: biological, chemical, physical, and radiological, as described in the Table I.

The aime of Xu, et al. research [12] is to design a framework for the prediction of the water quality in two

TABLE I. Classification of Water Indicators into Biological, Chemical, Physical and Radiological

Classes of Water Quality Indicators	Examples
Biological	Bacteria, parasites
Chemical	pH, dissolved oxygen, salts
Physical	Temperature, electric conductivity
Radiological	Radioactive elements like Uranium

regions, inland river water and nearshore water, based on different factors. The researchers were investigating the effect of several factors such as turbidity, temperature, dissolved gasses, ammonia concentrations, and dissolved solids on the total nitrogen level in the tested water. Inland water testing occurred at 2-hour intervals and total nitrogen levels were also collected at 4-hour intervals. The collected samples made a total of 1917 creating the dataset which was then subjected to normalization and correlation analysis. 90% of these data were used to train machine learning algorithms including Decision Tree, KNN, SVR, MLR, Random Forest, Ridge Regression, and GBRT. On the other hand, 10% were used to evaluate the models by comparing the predicted results of total nitrogen with the actual collected results.

This evaluation was based on correlation coefficient as well as the following metrics: Root Mean Square Error (RMSE), Mean Absolute Percentage Error, Nash–Sutcliffe efficiency coefficient, Mean Square Error, and Mean Absolute Error. The evaluation results showed that Random Forest achieves the best prediction results with 0.967 correlation coefficient and 0.509 RMSE, and that the ensemble models in general outperformed the non-ensemble models. Random Forest was also the focus when testing nearshore waters in comparison to the other ML models, where 147 new data were gathered and only three metrics were used since the acquired data differ from before (only temperature and salinity). The results of the second testing also came in favor of Random Forest compared to the other algorithms, achieving the lowest MAE and MAPE values.

In another study, Bachir Sakaa and his colleagues developed a Random Forest model as well as a Sequential Minimal Optimization-Support Vector Machine method for the determination of water quality in Saf-Saf river [13]. The researchers chose to collect the data from 35 areas in wet and dry seasons in order to gather 70 total samples that make up their dataset. The dataset was divided into training and testing subsets made up of 80% and 20% of data respectively. It was decided that the two models will be evaluated according to the root mean square error (RMSE), relative absolute error, mean absolute error, and root relative square error. In addition to these values, sensitivity analysis was carried out to assess how the independent variables (factors) are affecting the dependent variable (water quality). In order to determine the effector factors, a method called recursive feature elimination-linear "RFEL" was used where 15 different features were selected including suspended solids, ammonium, chemical, and biochemical oxygen demand, temperature, oxygen saturation, conductivity, and pH. Upon analyzing the results, it was evident that the results greatly differ in the upstream river area compared to the downstream river area. The same was noticed between data from wet seasons vs. dry seasons. RFEL was also used to determine a subset of combinations of some features

to notice their effects as a group. In this regard, for Random Forest the third created combination scored the best values (RMSE=5.17 and correlation R2=0.82) whereas for SMO-SVM the fourth input was better than the rest (RMSE=7.43 and R2=0.71). When comparing the overall performance of RF compare to SMO-SVM, it was concluded that they have similar results even though the error metrics show the superiority of RF.

RF was also employed in another study to predict the concentration of dissolved oxygen in the water of Potomac River [14]. More specifically, the purpose of the study was to determine the most important factors influencing the concentration of oxygen and to evaluate the efficiency of the Random Forest model in predicting the latter with respect to a varying combination of factors. Dissolved oxygen data were collected from a publicly available water quality database by the USGS. The variables chosen in this study are the gauge height, water temperature, turbidity, water pH, instantaneous discharge, and specific conductance. As a method of data preparation and pre-processing, a noise removal process was done by eliminating the individual predictor with its respective co-measured factors in order to avoid missing data. The Kolmogorov-Smirnov test was used to check the normality of each predictor variable, and also Box-Cox transform was used on the datasets. Furthermore, the water temperature was transformed into the Kelvin metric, and the data were standardized with the Z-score technique. After pre-processing, the data were divided 80-20 for training and testing respectively and the performance of two models: RF and MLR was evaluated through RMSE and R2 values. The correlation matrix revealed a significantly strong correlation with water temperature, and a significantly weak correlation with water salinity. There is also a strong multi-collinearity in the data due to correlation between the multiple variables. In conclusion, temperature, pH, and salinity were able to explain 98.7% of the data variance.

III. Method

The main purpose of this study is to be able to assess the quality of rivers' water and whether it is drinkable or not based on a machine learning technique, namely Random Forest.

Quite literally, the random forest is a large collection of several decision trees that are used in unity, where the group of decisions can be collected to come up with one decision as an output. This happens after each decision tree dictates a specific class as its prediction result, and the class that collects the most votes from the tree ensemble is finally chosen as an output of the model. Fig. 2 shows a simple example of how a random forest chooses a prediction based on the collective results from each decision tree.

Machine learning is the automated data analysis process. Instead of being conventionally performed by a data scientist, nowadays, machines can replace manual analysis while using the same math and statistical techniques. The main difference though is that in machine learning, the techniques are integrated into algorithms that are capable of learning and improving themselves on their own. Machine learning has become the key to facilitating artificial intelligence (AI), where automated decisions can replace human decisions. And even though data science, machine learning, and artificial

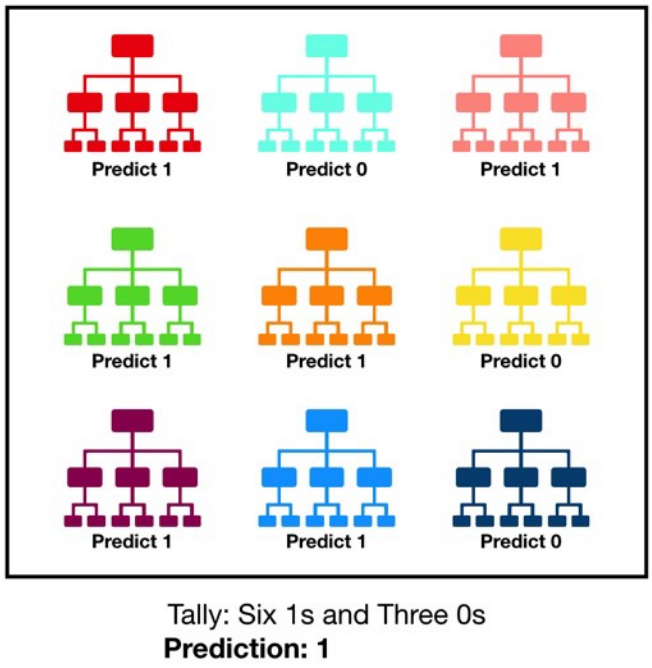

Fig. 2. Example of Decision Making using Random Forest Model by Tony [15].

intelligence are puzzle pieces in the same field, yet each has its own applications and its own meaning.

Machine learning approaches in general can be divided into supervised and unsupervised machine learning. In the unsupervised models such as Principle Component Analysis of K-mean clustering, the algorithm finds hidden patterns within the data without them being labeled. On the other hand, supervised machine learning like Decision Tree or Random Forest, operates on previously labeled data and it is their objective to perform classification predictions based on how they were trained with the respective labels [16].

In machine learning, often several steps are done in sequence starting with data pre-processing, extraction of features, fitting of the model, and finally the evaluation of the performance of the developed model. These steps require a lot of transformation for data, which can be easily done by the machine learning pipeline to keep everything in order. The role of a pipeline is to keep the data flowing properly and that the transformations are adequately done to make sure that the result reached is accurate and without error.

Machine learning is one of the very effective methods by which Big Data can be processed, visualized, and interpreted [17], [18]. In this study, for the execution of our model, we relied on the Spark framework since spark is capable of performing large processing tasks quickly and allows the distribution of tasks over several computers for processing [19]. More specifically, PySpark was utilized as it allows the use of Python as a programming language. In Apache Spark, machine learning algorithms can be employed through Spark MLlib which we also relied on.

The ML pipeline requires a chain of command where the stages are assigned and it can run smoothly on Spark.

The stages involved within a pipeline can be transformers or estimators that have different functions. The function of a transformer is to convert one type of data frame into another type of data frame which can be done through updating the categorical values within one column into numeric values, or through user-defined logic to map the data in the column to other values. On the other hand, an estimator is capable of developing a model based on a fit method. Here, for instance, the Random Forest classifier is an estimator.

There are several steps to be done for the completion of the study. These involve exploring the data after acquiring the dataset, preparing the data, and performing correlation analysis. After that, model design and testing are followed by model assessment or evaluation, see Fig. 3.

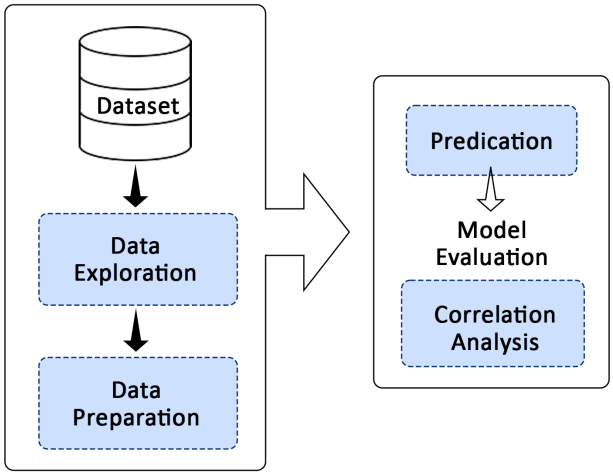

Fig. 3. Flowchart Explaining the Steps Followed in this Study: Data Exploration and Preparation, Correlation Analysis, Model Design and Evaluation.

IV. Data

The chosen dataset comprises a total of ten features that will be used to predict the quality of water and whether it is good for drinking or not. Fig. 4 shows the ten values that describe the quality of the water, which are: pH, hardness, presence of solids presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and finally potability. The corresponding definitions and values of these features are described below.

pH. The pH metric is an evaluation of the concentration of hydrogen ions within a solution, and it allows the differentiation between acid media, basic media, and neutral media as water [20]. The pH recommended by the world health organization determines that the pH of drinkable water must range between 6.52 and 6.83.

Hardness. Hardness is the resultant of both magnesium and calcium salts that deposit from the geologic surrounding of running water [21]. The period of time in which the water is in contact with hardness-producing material determines how much hardness there is in raw water [22].

Total Dissolved Solids. Dissolved solids in water refer to the salts that can be present including potassium, magnesium, calcium, bicarbonates, sodium, chlorides, etc. [23]. The presence of these dissolved solids in water leads to changing its flavor in addition to affecting its safety [24]. The ideal concentration for TDS is 500 mg/l and should not go above 1000 mg/l for drinking water.

Chloramines. Chloramine alongside chlorine is often used for the treatment of water and disinfecting it from bacteria and other microorganisms [25], [26]. For safety, the amount of chloramine in drinkable water should not exceed 4 mg per liter.

Sulfate. Sulfates are natural elements present in the soil, minerals, food, groundwater, plants, and rocks. Yet they are heavily used in the chemical industry. The sulfate concentration in freshwater should be between 3 and 30 mg per liter [27].

Conductivity. Electric conductivity is a measure of conducting electricity through water. Pure water does not conduct electricity, rather it is considered an insulator [28]. However, ionic water has an increased electric conductivity as a result of the ionic compounds in it [29]. The safe electric conductivity level should be less than 400 $\mu S/cm$.

Organic Carbon. Total organic Carbon TOC resembles the total quantity of carbon from organic matter within the water [30]. This organic carbon can originate from either the decay of natural organic matter or from an unnatural synthetic source. The normal values of organic carbon should be less than 2 mg per liter for drinkable water, and less than 4 mg per liter for the water to be treated.

Trihalomethanes. Trihalomethanes are referred to as THMs in short, and these are molecules abundant in the case of chlorine treatment of water [31]. The factors that affect the amount of THMs are the temperature of treated water, the required chlorine concentration, and the level of organic matter within the water [32]. In order for water to be drinkable, the THM value must be below 80 ppm.

Turbidity. Turbidity is a description of the state of water and whether solids are suspended in it or not [33]. The turbidity of water can be calculated by the light emitting characteristics of water, which represents the quality of waste discharge in regard to the colloidal matter. The turbidity value recommended by the World Health Organization is turbidity=5.00 NTU.

Potability. Potability is a term given to describe whether the water is safe for human consumption or drinking or not [34]. In fact, it should also be considered if the same water is good for watering plants. If the given value=1 then the water is potable or drinkable, whereas value=0 means the water is not suitable for consumption.

A. Data Exploration and Preparation

As part of data pre-processing, the data were converted into float after being in a string. In addition, the data that were in repetition were deleted, so only the necessary data were kept, see Fig. 5.

Initially, we will check whether there are NULL values or not. This is important to ensure that the algorithm can run smoothly without any missing data since null values indicate

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
null	204.8904554713363	20791.318980747026	7.300211873184757	368.51644134980336	564.3086541722439	10.379783070804	86.9909704615088	2.9631353806316407	0
3.7160800753869	129.42292051494425	18630.057857970347	6.635245883862	null	592.8853591348523	15.180013116357259	56.32907628451764	4.500656274942408	0
8.099124189298397	224.23625939355776	19909.541732292393	9.275883602694089	null	418.6062130644815	16.868636929550973	66.42009251176368	3.0559337496641685	0
8.316765884214679	214.37339408562252	22018.417440775294	8.05933237743854	356.88613564305666	363.2665161642437	18.436524495493302	100.34167436508008	4.628770536837084	0
9.092223456290965	181.10150923612525	17978.98633892625	6.546599974207941	310.13573752420444	398.41081338184466	11.558279443446395	31.997992727424737	4.075075425430034	0
5.584086638456089	188.3133237696164	28748.68773904612	7.54486878877965	326.6783629116736	280.4679159334877	8.399734640152758	54.917861841994466	2.5597082275565217	0
10.223862164528773	248.07173527013992	28749.716543528233	7.513408465831302	393.66339551509645	283.6516335078445	13.789695317519886	84.60355617402357	2.672988736934779	0
8.635848718500734	203.36152258457054	13672.091763901635	4.563008685599703	303.3097711592812	474.60764494244853	12.363816698705	62.798308962925155	4.401424715445482	0
null	118.98857909025189	14285.583854224515	7.804173553073094	268.646940746221	389.3755658712614	12.706048968657911	53.928845767512236	3.5950171809576155	0
11.180284470721592	227.23146923797458	25484.50849098786	9.077200016914393	404.04163468408996	563.8854814810949	17.92780641128502	71.97660103221915	4.370561936655497	0
7.360640105838258	165.52079725952862	32452.614409143884	7.550700906704114	326.62435345560164	425.38341949538733	15.586810438033126	78.74001566430479	3.6622917828524573	0
7.974521648923869	218.69330048866644	18767.65668181348	8.110384501123875	null	364.09823046204866	14.525745697593209	76.48591117965157	4.011718108339787	0
7.119824384264552	156.70499334039215	18730.813653342713	3.606036090857203	282.3440504739606	347.71502726194376	15.929535908825699	79.5007783369744	3.445756223321899	0

Fig. 4. The Ten Feature for Assessing the Potability of Water.

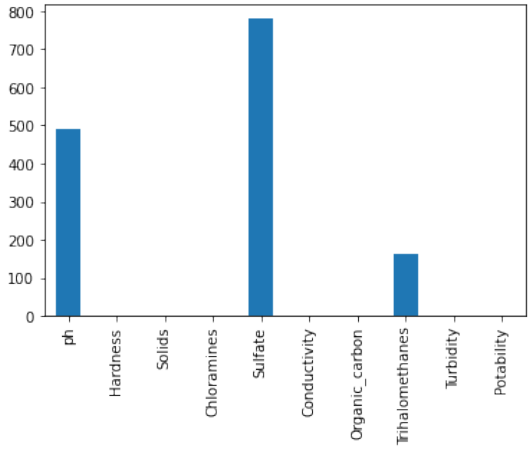

Fig. 5. The Value of Necessary Data Needed for our Model.

missing data. Furthermore, the algorithm can obtain more accurate results when the null values are replaced. As can be seen in the image below, pH, Sulfate, and Trihalomethanes have NULL values. As a solution, the null values are usually replaced by the average or mean of the specific category. After that the information in the dataset is checked again.

The mean value is calculated by measuring the sum of the available values divided by the total number of values in the categories. This mean calculation is used to handle the missing data relative to NULL values, see Fig. 6.

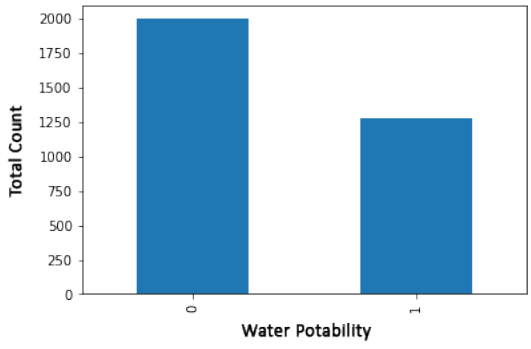

Fig. 6. The Sum of the Available Values for Water Potability in the Dataset: 0 Water is not Potable, 1 Water is Potable.

The value of feature probability density is also measured for each feature. This function describes the probability of a certain feature falling between a range of values. It is also described as density as it shows the mass distribution of said feature over the total scale. Continuous variables or features would produce a curvature shape, either described as a normal distribution or non-normal distribution. In Fig. 7, the distribution probability of all ten features is shown (as continuous bar graphs mixed into a curve).

V. Result

At the beginning of the study, we presented the first research question **RQ1:** *What factors directly affect the potability of the water?* To which the answer can be deduced from the correlation analysis. The correlation is done using the heat map function of seaborn, see Fig. 8. This function shows the degree that which two factors affect each other. The correlation matrix below shows that each feature is only strongly correlated with itself (scoring +1). On the other hand, when seeing the correlation between the features with potability, no strong correlations exist. Yet there exists a weak correlation between two of the factors: pH and hardness (0.08). This analysis means that the dimensions can't be reduced in this study due to the absence of correlations between the variables. Further, the data are divided into independent and dependent features. All are independent features except Potability because Potability is our dependent feature.

Answering the second research question **RQ2:** *Can a random forest model effectively predict the quality of water based on these factors?* The dataset was divided into 80% training, and 20% testing on a Random Forest Classifier model (Fig. 9). At first, the model was trained, then it was fed the testing dataset to observe and assess the predictions. The answer to our second research question can be found by evaluating the model through numbers. The most commonly used metrics for model evaluation are accuracy, precision, recall, and f1-score.

The accuracy is the measurement of how close the predicted value is to the actual value, whereas the precision is the measure of how much the model can produce a repeated prediction value. The recall shows how much the model is good at identifying true positives. Based on both precision and recall, the F1-score value is calculated, and the greater the value the better. The results obtained by our Random Forest model can be summed up in Table II. These results show good performance by our Random Forest model, and they are very satisfactory.

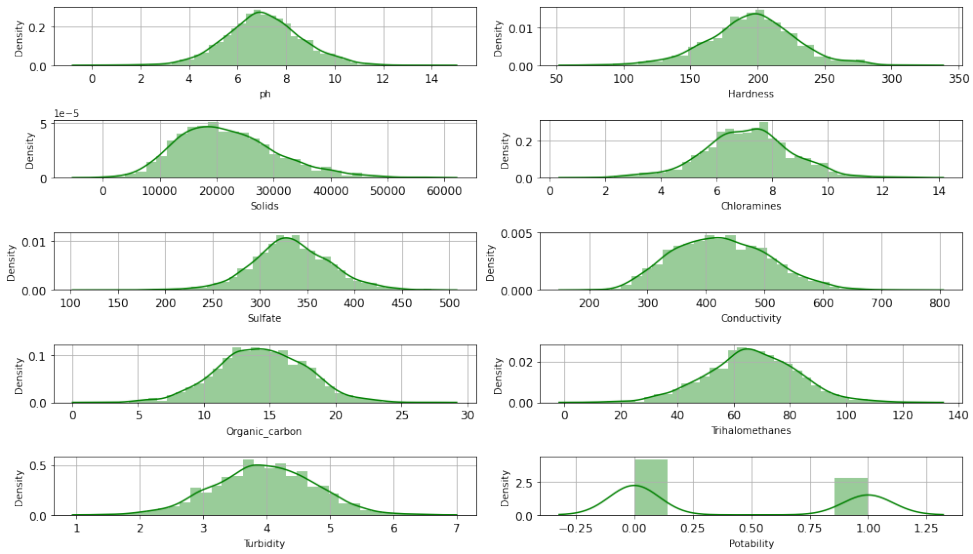

Fig. 7. Feature Probability Distribution as Calculated for each Feature.

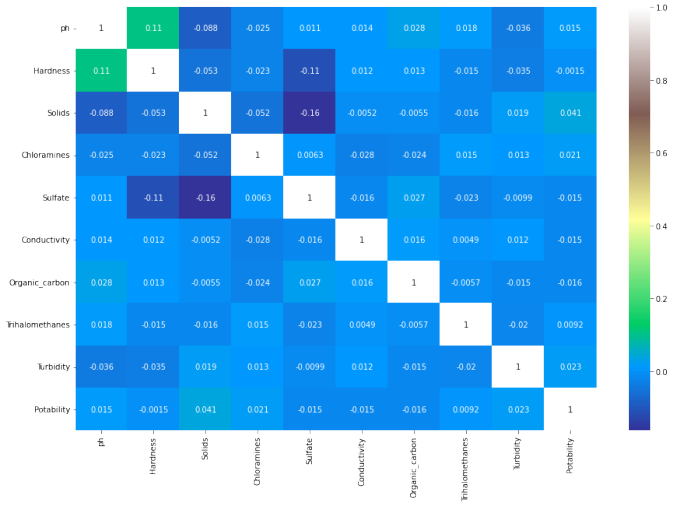

Fig. 8. Correlation Matrix Showing no Correlations between the Different Variables.

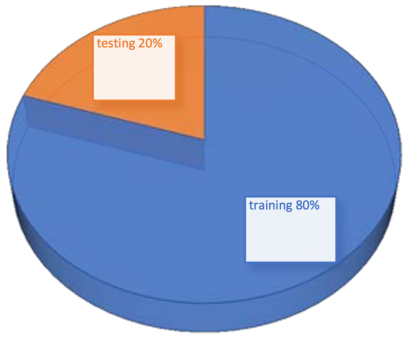

Fig. 9. Distribution of Dataset into Testing and Training.

TABLE II. MODEL PERFORMANCE EVALUATION BASED ON ACCURACY, RECALL, PRECISION, AND F1-SCORE

Test accuracy	1.0
Test recall	1.0
Test F1-score	1.0
Test Precision	1.0

VI. DISCUSSION

In this study, water quality prediction model was designed using machine learning classification in form of random forest classifier in predicting the water potability. Various variables are considered to perform a variety of calculations related to the water quality. These included pH, hardness, presence of solids, presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity. The results of the study revealed that different machine learning models performed differently when it came to predicting water potability. In order to compare our model with previous work presented by Xu, et al. [12] and Devi [35], we used RMSE (Root Mean Square Error) as metrics to evaluate the accuracy of random forest model in each study with ours. We can visually compare the accuracy of these models using Fig. 10.

One of the main advantages of ensemble learning is its ability to improve the system's overall performance. In addition, random forest methods can also reduce the likelihood of overfitting due to their random nature.. After successfully completing the water random forest prediction task, we then used remote sensing bands to perform water quality prediction. The previous variables were used as targets, while the independent variables were used as the distribution of the data. Through the use of the Jupyter platform, we were able to perform an inverted analysis of the data to fit the water quality prediction model. As it can be seen, our model provide a better performance score.

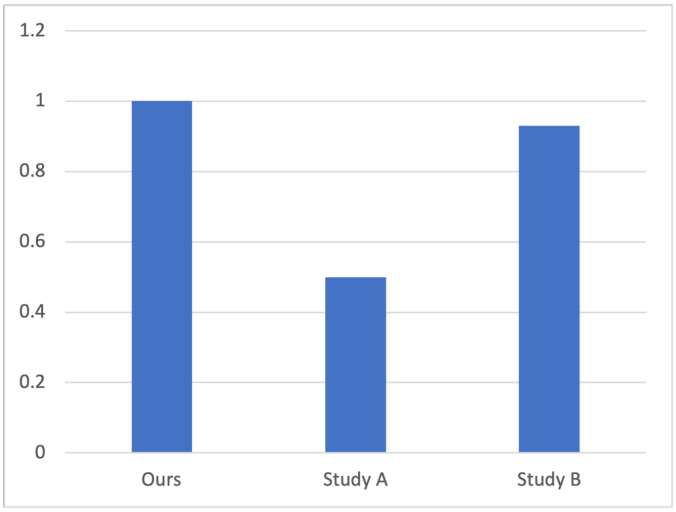

Fig. 10. Accuracy Histograms of Different Models: Ours, Study A [12] and Study B [35].

VII. Conclusion

The safety of our drinking water is a very essential matter, which should be monitored and managed effectively due to its importance. The quality of water that we used for drinking or cooking has a direct effect on our own health, which is why having perfectly safe water is not only a right for humans but also extremely critical. Several protocols and assessment criteria were developed to keep an eye on the safety and potability of water of different origins (underground, surface, inshore waters, etc.).

Using machine learning and PySpark classification for collection, storage, and analysis of water samples is a much more effective and efficient method for water quality evaluation than regular laboratory tests. This motivated us to create a machine learning model based on the Random Forest algorithm to evaluate the quality of river water based on 10 distinctive features: pH, hardness, presence of solids, presence of chloramines, presence of sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and finally potability. The obtained results show that the developed RF model is capable of predicting whether the collected water sample is potable or not with a 100% accuracy and 1.0 F1-score.

The water quality prediction model that we designed can be used in the field of water quality monitoring. In the future, it can be used to provide an online tool that allows users to monitor the water quality of their local waterways. This method can be used to collect the necessary data using sensors to perform the prediction. The next step in the development of the water quality prediction model will be to collect the stream data necessary to perform the prediction. This method will be carried out through a dynamic update of the model.

References

[1] M. Kachroud, F. Trolard, M. Kefi, S. Jebari, and G. Bourrié, "Water quality indices: Challenges and application limits in the literature," *Water*, vol. 11, no. 2, 2019. [Online]. Available: https://www.mdpi.com/2073-4441/11/2/361

[2] S. H. Ewaid, S. A. Abed, N. Al-Ansari, and R. M. Salih, "Development and evaluation of a water quality index for the iraqi rivers," *Hydrology*, vol. 7, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2306-5338/7/3/67

[3] O. T. Opafola, K. T. Oladepo, F. O. Ajibade, and A. O. David, "Potability assessment of packaged sachet water sold within a tertiary institution in southwestern nigeria," *Journal of King Saud University - Science*, vol. 32, no. 3, pp. 1999–2004, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1018364720300537

[4] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2073-4441/11/11/2210

[5] D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire, "Comparison of machine learning algorithms in statistically imputed water potability dataset," *preprint*, 2022.

[6] J. H. Lee, J. Y. Lee, M. H. Lee, M. Y. Lee, Y. W. Kim, J. S. Hyung, K. B. Kim, Y. K. Cha, and J. Y. Koo, "Development of a short-term water quality prediction model for urban rivers using real-time water quality data," *Water Supply*, vol. 22, no. 4, pp. 4082–4097, 02 2022. [Online]. Available: https://doi.org/10.2166/ws.2022.038

[7] M. Kejariwal, C. Patil, A. B. Tiwari, and S. K. Sahani, "Water potability testing case study of mumbai region," *Journal of Harmonized Research in Applied Science*, 2018.

[8] D. T. Burns, E. L. Johnston, and M. J. Walker, "Authenticity and the Potability of Coconut Water - a Critical Review," *Journal of AOAC INTERNATIONAL*, vol. 103, no. 3, pp. 800–806, 03 2020. [Online]. Available: https://doi.org/10.1093/jaocint/qsz008

[9] E. Ochungo, G. Ouma, J. Obiero, and N. Odero, "Water quality index for assessment of potability of groundwater resource in langata sub county, nairobi-kenya," *American Journal of Water Resources*, vol. 7, no. 2, pp. 62–75, 2019.

[10] Z. Jamshidzadeh and M. T. Barzi, "Groundwater quality assessment using the potability water quality index (pwqi): a case in the kashan plain, central iran," *Environmental earth sciences*, vol. 77, no. 3, pp. 1–13, 2018. [Online]. Available: https://doi.org/10.1007/s12665-018-7237-5

[11] P. Li and J. Wu, "Drinking water quality and public health. exposure and health, 11 (2), 73–79," 2019.

[12] J. Xu, Z. Xu, J. Kuang, C. Lin, L. Xiao, X. Huang, and Y. Zhang, "An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies," *Water*, vol. 13, no. 22, 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/22/3262

[13] B. Sakaa, A. Elbeltagi, S. Boudibi, H. Chaffaï, A. R. M. Islam, L. C. Kulimushi, P. Choudhari, A. Hani, Y. Brouziyne, Y. J. Wong *et al.*, "Water quality index modeling using random forest and improved smo algorithm for support vector machine in saf-saf river basin," *Environmental Science and Pollution Research*, pp. 1–18, 2022. [Online]. Available: https://doi.org/10.1007/s11356-022-18644-x

[14] M. H. Ahmed, "Prediction of the concentration of dissolved oxygen in running water by employing a random forest machine learning technique," *Preprints*, 2020.

[15] T. Yiu, "Understanding random forest: How the algorithm works and why it is so effective," *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[16] J. McArthur, N. Shahbazi, R. Fok, C. Raghubar, B. Bortoluzzi, and A. An, "Machine learning and bim visualization for maintenance issue classification and enhanced data collection," *Advanced Engineering Informatics*, vol. 38, pp. 101–112, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474034617305049

[17] N. Deshai, S. Venkataramana, B. V. D. S. Sekhar, K. Srinivas, and G. P. Saradhi Varma, "A study on big data processing frameworks: Spark and storm," in *Smart Intelligent Computing and Applications*, S. C. Satapathy, V. Bhateja, J. R. Mohanty, and S. K. Udgata, Eds. Singapore: Springer Singapore, 2020, pp. 415–424.

[18] S. Chen, "Research on big data computing model based on spark and big data application," *Journal of Physics: Conference Series*, vol. 2082, no. 1, p. 012017, nov 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/2082/1/012017

[19] L. Chen and L. A. Coulibaly, "Data science and big data practice using apache spark and python," in *Intelligent Analytics With Advanced Multi-Industry Applications*. IGI Global, 2021, pp. 67–95. [Online]. Available: https://doi.org/10.4018/978-1-7998-4963-6.ch004

[20] K. Myeong-ryul and L. Wontae, "Effects of water temperature and ph on water quality improvement by a mixture of beneficial microorganisms," *J Korean Soc Environ Eng*, vol. 40, no. 1, pp. 1–6, 2018. [Online]. Available: http://www.jksee.or.kr/journal/view.php?number=4088

[21] D. Shvachko, "Hardness of water and its impact on the human body," Ph.D. dissertation, Sumy State University, 2017.

[22] Z. K. Jabbar-Lopez, C. Y. Ung, H. Alexander, N. Gurung, J. Chalmers, S. Danby, M. J. Cork, J. L. Peacock, and C. Flohr, "The effect of water hardness on atopic eczema, skin barrier function: A systematic review, meta-analysis," *Clinical & Experimental Allergy*, vol. 51, no. 3, pp. 430–451, 2021. [Online]. Available: https://doi.org/10.1111/cea.13797

[23] J. Warrack and M. Kang, "Challenges to the use of a base of fresh water in groundwater management: Total dissolved solids vs. depth across california," *Frontiers in Water*, vol. 3, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frwa.2021.730942

[24] M. Jamei, I. Ahmadianfar, X. Chu, and Z. M. Yaseen, "Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: New approach," *Journal of Hydrology*, vol. 589, p. 125335, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169420307952

[25] J.-L. Boudenne, F. Robert-Peillard, and B. Coulomb, "Chapter two - inorganic chloramines analysis in water," in *Analysis and Formation of Disinfection Byproducts in Drinking Water*, ser. Comprehensive Analytical Chemistry, T. Manasfi and J.-L. Boudenne, Eds. Elsevier, 2021, vol. 92, pp. 31–49. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166526X21000027

[26] Z. T. How, I. Kristiana, F. Busetti, K. L. Linge, and C. A. Joll, "Organic chloramines in chlorine-based disinfected water systems: A critical review," *Journal of Environmental Sciences*, vol. 58, pp. 2–18, 2017, water treatment and disinfection by-products. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1001074216311822

[27] H. Tahraoui, A.-E. Belhadj, A.-E. Hamitouche, M. Bouhedda, and A. Amrane, "Predicting the concentration of sulfate (so4 2–) in drinking water using artificial neural networks: A case study: Médéa-algeria," *Desalination and Water Treatment*, vol. 217, pp. 181–194, 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03225482

[28] Y. Xianhong, L. Shijun, H. Jian, and X. Jie, "Application analysis of conductivity in drinking water quality analysis," *IOP Conference Series: Earth and Environmental Science*, vol. 784, no. 1, p. 012028, may 2021. [Online]. Available: https://doi.org/10.1088/1755-1315/784/1/012028

[29] T. Manzoor, S.Iqbal, F.Somorro, S.Imran, H. Alvi, and S. S. Haider, "Analysis of fluoride ion concentration in drinking water of karachi in relation with conductivity and ph," *Pakistan Journal of Chemistry*, vol. 9, no. 1-4, pp. 1–5, may 2019. [Online]. Available: https://doi.org/10.15228/2019.v09.i01-4.p01

[30] K. I.A., S. D.A., K. E.A., P. E.G., and B. L.A., "Approaches to regulating organic carbon and the necessity of its obligatory monitoring in drinking water," *Public Health and Life Environment – PH&LE.*, vol. 9, pp. 61–66, may 2020. [Online]. Available: https://doi.org/10.35627/2219-5238/2020-330-9-61-66

[31] S. Sriboonnak, P. Induvesa, S. Wattanachira, P. Rakruam, A. Siyasukh, C. Pumas, A. Wongrueng, and E. Khan, "Trihalomethanes in water supply system and water distribution networks," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/17/9066

[32] Y. Wang, G. Zhu, and B. Engel, "Health risk assessment of trihalomethanes in water treatment plants in jiangsu province, china," *Ecotoxicology and Environmental Safety*, vol. 170, pp. 346–354, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0147651318312910

[33] D. Wang, X. Chang, K. Ma, Z. Li, and L. Deng, "Estimating effluent turbidity in the drinking water flocculation process with an improved random forest model," *Water Supply*, vol. 22, no. 1, pp. 1107–1119, 07 2021. [Online]. Available: https://doi.org/10.2166/ws.2021.213

[34] V. Pradhan, "Assessment of potability of water with respect to physical parameters," *International Journal of Life Sciences*, vol. 2, no. 4, pp. 382–388, 2014.

[35] S. Devi, "Random forest advice for water quality prediction in the regions of kadapa district," *Int. J. Innov. Technol. Explor. Eng*, vol. 8, pp. 1–3, 2019.