HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

Quick Clean Water: IoT and Machine Learning-Based Water Contamination Detection System

Nehal Singh

Texas Academy of Mathematics and Science

ABSTRACT

Quick Clean Water can detect water contamination in private wells, piped water, surface water, and water used for agriculture, recreation, and other purposes in developed and developing countries. Current testing systems are slow, costly, low in availability, and give back minimal results of 10-16 contaminants using expensive strips. Quick Clean Water is reusable, easy-to-use, portable, affordable, and gives advanced results of presently 21 contaminants. First, the Quick Clean Water device calculates the pH, turbidity, temperature, total dissolved solids, conductivity, and salinity using sensors. Inputting the pH, turbidity, conductivity, and TDS, a machine learning model using the Random Ensemble algorithm predicts whether the water is safe at a 55% accuracy rate, which can be improved through data augmentation. Several algorithms were tested and evaluated by the precision, recall, f-score, specificity, negative predictive value, and accuracy rates. The hypothesis was that the K-Means Clustering would result in the best model, but Random Ensemble was the most efficient. If the water is classified as non-potable, users can enter the odor, color, and taste of their water into a 99% accurate ML model using the Random Ensemble algorithm to identify the exact contaminant in their water, which can be advanced by researching more contaminants.

Introduction

Common water contamination sources are nearby agricultural fertilizers causing high nitrogen concentrations, industrial or textile work causing chemical or heavy metal contaminants, metal pipes leaching impurities, sewage or wastewater, organic chemicals, and radioactive elements.[1] Water contamination commonly occurs in private wells, farms, surface water, and piped water.

- Private Wells

In 2015, about 43 million US households used private wells for water and were responsible for their groundwater and any possible contaminants. However, a study in 2009 indicated that 23% of samples of groundwater from over 2,000 domestic wells contained metallic ions or organic substances at a level that is a concern to human health.[2]

- Farms

At farms including plantations, market farms, and shifting cultivation, contaminated water used for irrigation, food processing, and hygiene of workers can affect the health of workers and crops and eventually consumers. In farming techniques like pastoral nomadism and mixed crop and livestock, low-quality water can harm livestock and surrounding communities through waste.[3]

- Surface Water

144 million people use surface water and 435 million use unprotected wells and springs that with increasing offshoring and industrialization in developing countries cause industrial waste to flow into water systems.[4]

- Piped Water

Water that travels through pipes can also be contaminated through the leaching and corrosion of copper, lead, and metal pipes and faucets. This causes heavy metals to be present in water and is the main cause of lead contamination in homes.[5]

Drinking water contaminated with nitrites/nitrates can cause methemoglobinemia and “blue baby syndrome”. Heavy metals including copper and lead put people at risk of liver, kidney, and intestinal damage, anemia, and cancer. Organic chemicals result in liver, kidney, circulatory, reproductive, and nervous system damage. Radionuclides can cause potential kidney damage and cancer. An excess amount of fluoride tooth discoloration and skeletal fluorosis. High amounts of sewage and bacterial contamination cause the spread of disease and fatal disease. Therefore, it is imperative to ensure clean water sources at farms, wells, and surface waters.

Water contamination is an increasingly pervasive issue caused by the expansion of industrial and agricultural industries. Beginning in the 1970s, mass consumption around the globe led to mass production and an increase in the number of factories; however, at that time, few government policies on environmental safety were enacted and the developing technology was prone to explosions and accidental leaks. The biggest water-related disaster was possibly the incident in Lanzhou, China, where over 20 times the legal and safe amount of benzene was detected in water sources due to explosions at a nearby plant. Environing communities facing prolonged exposure to the carcinogenic chemical experienced escalated rates of cancer and hematopoietic system damage. To worsen the issue, 34 tons of benzene seeped into the groundwater throughout the 27 years (1987 to 2014) the toxic buildup remained unnoticed. Similar devastating events include the accumulation of cyanide from a mine in Ghana in 2001. As stricter regulations are being made on the disposal of industrial wastewater and locations of storage facilities, oil spills and sewage are still running into rivers. Major lead contamination from corroding pipes was found in Flint, Michigan in 2014. Increased cases of illness and lead poisoning followed in the city of Flint but no legislation to prevent further catastrophes. One of the most detrimental oil spills to date occurred in Bligh Reef in 1989. The 38 gallons of crude oil can as of 2015 be found in the ocean and has proved life-threatening to aquatic animals especially several pods of orcas that died off completely. Furthermore, in recent decades the Yamuna River water supply for 70% of India, 500 million gallons of sewage enters the river every day making the river unrecognizable to older locals. Additionally from 2009 to the present, residents of Mutare, Zimbabwe are exposed to wastewater from local diamond mining operations containing high levels of chromium, nickel, and bacteria causing typhoid and cholera.

For these reasons and incidents of unexpected and unnoticed contamination, water testing is crucial to know whether or not the water is safe to be consumed or used for other purposes to prevent widespread communal illness. Current water testing systems do not meet the needs of consumers as they are slow, costly, low in availability, and give back minimal results. Inspections can take a couple of hours, and tests generally take 3 days to 2 weeks. Considering my experiences, when I send a water test for my drinking water, it would take days or weeks to receive results even if I need them quickly. Sending tests to laboratories costs 25-400 dollars. Quality testing kits are 100-200 dollars. Water testing centers are not available at all locations. Testing kits check for about 1-2 possible contaminants, and water testing strips that test for more impurities can be misread. Do-it-yourself tests simply provide information on whether or not pH or another metric are off. And neither inform about the reasons for contamination.

The purpose of Quick Clean Water is to provide a quick, inexpensive, reusable, portable, easy-to-use, and advanced water testing method for piped water, wells, surface water, and farms in developed and developing countries. The objective is to develop the most efficient machine learning models to predict potability and the contaminant. The engineering goal of Quick Clean Water is to create an IoT and machine learning-based water contamination detection system.

Methods

Hardware

The hardware system includes pH, turbidity, temperature, and total dissolved solid sensors, an Arduino microprocessor, and a 3D printed exterior. These sensors collect the pH, turbidity, temperature in Celsius and Fahrenheit, and total dissolved solids.

The total dissolved solids (ppm) found using the TDS meter and temperature sensor can be converted to conductivity and salinity.

$$Conductivity(\mu s/cm) = Total\ Dissolved\ Solids(ppm)\ /\ .67$$

$$Salinity(mg/L) = Conductivity(\mu s/cm)\ *\ .55$$

The schematics for the circuits are shown below (Figure 1).

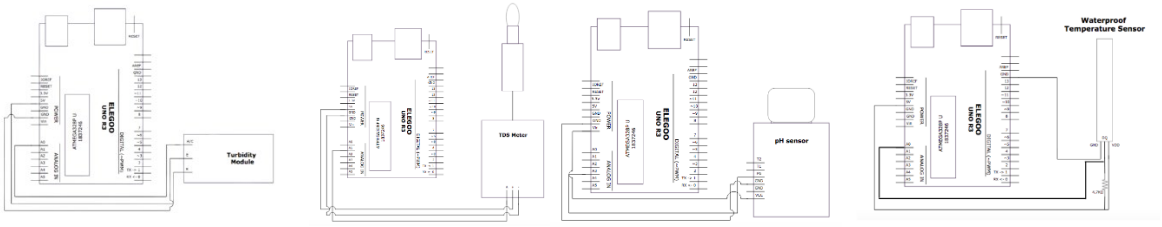

Figure 1 *Quick clean water hardware schematics*

Machine Learning Models: Predicting Potability

When multiple factors can affect the output, it is best to use Machine Learning to receive accurate predictions.[9] In this case, the output is dependent on the pH, conductivity, TDS, and turbidity; therefore, machine learning is used to make predictions. A model utilizing machine learning would consider all readings before making a prediction allowing for more precise results.

The data in this model is acquired from the water quality dataset available online on Kaggle.[10] The first steps in coding a machine learning model are data cleaning and feature engineering. To do this, the Chloramines, Sulfate, Organic carbon, Trihalomethanes, and Hardness columns are deleted, which are irrelevant to the experiment. Rows that contain null values are deleted, and the Solids column is divided by 100. Data visualization assists in identifying and deleting outliers to slowly make the features more correlated with one another and data more accurate. This can be executed by using pairwise bivariate distributions of data to identify outliers and using pandas to delete data points in each column that are above or below a certain value that defines the boundary between outliers and correlated points. In the pH column, outliers lie below 2 or above 12.5. For Solids, outliers lie above 500. For Conductivity, outliers lie above 700. For Turbidity, outliers lie above 6. The data is then split into X and y datasets, and the X dataset is standardized using StandardScaler. The scaled data can be arranged into two components using principal component analysis. To begin training, the data is split so that 30% is used for testing and 70% is used for training.

The correlation of the online data is not known; therefore, the algorithm that returns the most efficient model is tested for. There are several regression and classification algorithms including linear regression, logistic regression, decision tree, random forest, K nearest neighbors, K-means clustering, and support vector machines that can be applied to the data. The random_state is set to 101 to ensure that the data split is consistent throughout all of the algorithm tests. The X_train and y_train dataset are fit to each of the algorithms being tested as described above. Each model and class are evaluated using precision, recall, f-score, specificity, negative predictive value, and accuracy. After first calculating these five metrics for each feature, they can be averaged to find the macro-average metrics for the model as a whole. The dependent variables are the precision, recall, f-score, negative predictive value, and overall accuracy rates of the model. The independent variable is the algorithm used. The training and testing data, procedure to clean data, and all Python code except for the algorithm stays the same within each model. There is no control group.

Machine Learning Models: Predicting the Contaminant

This model is built by engineering a dataset using background research on the odor, color, and taste of water containing Algae; Ammonia; Arsenic; Bacteria; Clay, Silt, and Sand; Copper; Disinfectant; Fluoride; Gasoline/Petrol; Hydrogen Sulfide; Lead/Iron; Manganese; Nitrates/Nitrites; Potassium Permanganate; Radioactive elements (Radium, Radon, or Uranium); Sodium; and Zinc. To begin training, the data is split into X and y datasets. Since the data contains qualitative data, not quantitative data, the OneHotEncoder Transformer can be applied to convert all categories in the columns into numeric values presented in a matrix. Next, the transformed data (X and y datasets) is fit to the Random Forest algorithm. The model is then used to predict the testing data, and the predictions are used to evaluate the model and classes using the precision, recall, f-score, specificity, negative predictive value, and accuracy. The data is self-engineered; therefore, it is known that the Random Forest algorithm best fits this data.

Quick Clean Water Testing

This test evaluates the accuracy of the Quick Clean Water system against water testing strips that test for 16 different contaminants. First, 300mL water samples are collected from Kirkland Signature Purified Water with Added Minerals, tap water from Irving, Texas, Pensacola Beach, Big Lagoon State Park, and public restroom in Destin, Florida. These samples were picked due to being from different reservoirs and areas from the Southeast of the United States. The samples are versatile in what contaminant they could contain, which emulates the unpredictability of water testing in real life. One sample is then poured into a larger bucket of 29.4 cm by 19.8 cm by 13.9 cm dimensions, and a water strip from the 16 in 1 Drinking Water Test Kit is inserted into the sample for 2 seconds. Once the results are read in 30 seconds, each of the 16 metrics and their values is recorded in a table with metric names and a yes or no for whether or not the water is safe. To test the device, the Quick Clean Water device is inserted into the bucket so that all the sensors are touching the water sample. Like before, the pH, turbidity, temperature in both units, TDS, conductivity, salinity, and potability metrics are recorded in a table with metric names, values, and a yes or no for whether or not the water is safe.[38-40] If the water is not safe, the contaminant quiz or the contaminant-predicting model is used to find the specific contaminant. When taking the quiz, the sample is wafted to find the odor, and if the water has not been tasted or is not from the ocean, which has a salty taste, the taste can be entered as None. Similarly, the odor, taste, and color entered in the Machine Learning model and the output given is recorded in a table. If the water was predicted to be safe, the contaminant quiz does not need to be taken. According to UC San Diego, if the pH values are safe, and there are no oils or radioactive elements in the samples, the sample is safe to dispose of in the drain. Otherwise, the sample is buffered with tap water.[41]

Results and Discussion

Hardware

Improvements

The dimensions of the Quick Clean Water device (210 mm by 190mm by 75mm) (Figure 2) can be reduced significantly to 150 mm by 80 mm by 40 mm using a printed PCB board. This would make the device more portable and easier to hold and handle.

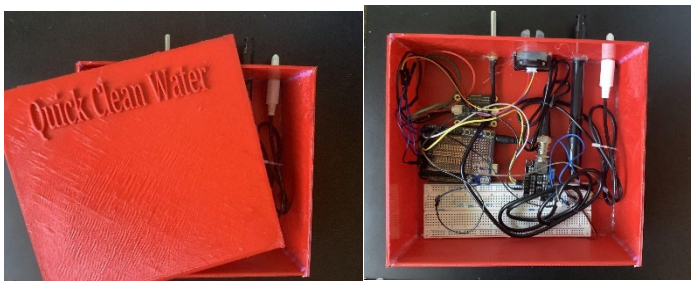

Figure 2 *Quick clean water hardware and interior.*

Machine Learning Models: Predicting Potability

Linear Regression

Table 1 *Class evaluation metrics of linear regression model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	0	236	0	361	0%	0%	0%	100%	40%	40%
0	236	0	361	0	40%	100%	57%	0%	0%	40%

Table 2 *Overall evaluation metrics of linear regression model.*

Precision	20%
Recall	50%
F-Score	28%
Specificity	50%
Negative Predictive Value	20%
Overall Accuracy	40%

This model classifies all test data as safe. Due to this trend in predictions, this network cannot be applied in the real world as evaluating all water samples as potable would not be accurate. The precision rate is 0% for Class 0 and 40% for Class 1. This indicates that the model and both classes need to be better trained to not predict false data as true. The recall rate is 0% for Class 0 and 100% for Class 1. The large difference shows that the model needs to be better trained to predict true points as true. This model predicts all points as Class 1. The specificity rate was high for Class 0, showing the model can successfully predict negative points; however, it must be taken into consideration that all inputs of this class were outputted as false. The negative predictive value indicates that 0% of negative values in Class 0 were predicted correctly, and 60% of negative values in Class 1 were classified accurately. For the same reason of the skew in outputs, the specificity of Class 0 was 0%. The accuracy rate is 40% for both classes. This demonstrates the model's ability to correctly predict both true and false data points (Table 1). The precision rate of this model is

20%, indicating that 210 of the points that were predicted as positive were actually true. 50% of the data that was actually true was evaluated by the model as positive. The f-score is another accuracy metric used when an emphasis on the incorrectly predicted values is needed as it is the harmonic mean of the precision and recall rates. Furthermore, 50% of the negative data samples were predicted to be false by the computer. The negative predictive value of this model is 30%, suggesting that 310 of the points that were predicted as negative were actually false. 60% of the testing data was classified correctly (Table 2).

Logistic Regression

Table 3 *Class evaluation metrics of logistic regression model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	361	0	236	0	60%	100%	75%	0%	0%	60%
0	0	361	0	236	0%	0%	0%	100%	60%	60%

Table 4 *Overall evaluation metrics of logistic regression model.*

Precision	30%
Recall	50%
F-Score	38%
Specificity	50%
Negative Predictive Value	30%
Overall Accuracy	60%

This model predicts all inputs as non-potable. This cannot be applied in deployment. The precision rate is 60% for Class 0 and 0% for Class 1. This indicates that the model and both classes need to be better trained to not predict false data as true. The recall rate is 100% for Class 0 and 0% for Class 1. The large difference between the two recall rates shows that the model needs to be strengthened in predicting true points as true. As shown by the matrix as well, the linear regression model predicts all points as Class 0, classifying all Class 0 data points correctly but never a Class 1 point. The specificity rate was high for Class 1, showing the model can successfully predict negative points; however, it must be taken into consideration that all inputs of this class were outputted as false. The negative predictive value indicates that 0% of negative values in Class 0 were predicted correctly, and 60% of negative values in Class 1 were classified accurately. The accuracy rate is also low for both classes. This demonstrates the model's ability to correctly predict both true and false data points (Table 3). 30% of the points that were predicted as positive were actually true. 50% of the data that was actually true was evaluated by the model as positive. The f-score is approximately 38%. Furthermore, 50% of the negative data samples were predicted to be false by the computer. 30% of the points that were predicted as negative were actually false. 60% of the testing data was classified correctly (Table 4).

Decision Tree

Table 5 *Class evaluation metrics of decision tree model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	227	79	157	134	59%	63%	61%	33%	37%	51%
0	79	227	134	157	37%	33%	35%	63%	59%	51%

Table 6 *Overall evaluation metrics of decision tree model.*

Precision	48%
Recall	48%
F-Score	48%
Specificity	48%
Negative Predictive Value	48%
Overall Accuracy	51%

The model did not accurately predict all inputs in each class. Both classes have high false positive and false negative values. The model is better at accurately predicting non-potable water samples, but compared to the previous models, this model has a clearer distinction between non-potable and potable water. The classes have a low precision and recall rate. The low precision rates show that a small amount of the data predicted as true is actually true. Class 0 has a precision rate above 50% indicating less false data was predicted to be in Class 0 than true data was predicted to be in Class 0, but Class 1 has a precision rate below 50% indicating more false data was predicted to be in Class 1 than true data was predicted to be in Class 1. The classes are weak in predicting true as positive and are not able to form a clear distinction between the two classes. The specificity scores are also relatively low, ranging from 63% and 33%, and both classes need to be better trained in predicting false as false according to the low negative predictive values. Both classes have accuracy rates of 51%. This is due to the high number of false positives in each class (Table 5). The model only precisely evaluates 48% of predicted positive samples and only 48% of true testing data is predicted true. The low precision and recall rates result in a low f-score of 48%. However, the model has proven that it is capable of predicting 48% of false data points as negative, and out of points evaluated as false, 48% were actually false. These metrics are almost 50%, implying a high amount of error in predictions. It can be concluded that the model accurately predicts 51% percent of the testing data, indicated by the overall accuracy (Table 6).

Random Ensemble

Table 7 *Class evaluation metrics of random ensemble model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	247	82	154	114	62%	68%	65%	35%	71%	55%
0	82	247	114	154	71%	35%	52%	68%	62%	55%

Table 8 *Overall evaluation metrics of random ensemble model.*

Precision	66%
Recall	52%
F-Score	59%
Specificity	52%
Negative Predictive Value	66%
Overall Accuracy	55%

The model did not accurately predict all inputs. The safe water class has a high false positive value, and the unsafe water class has a high false negative value. The model predicts more inputs as non-potable. The classes have a high precision rate compared to other model results. This implies a low number of false positives. There is a large difference of approximately 33% between the recall rates of both classes. Class 0 has a higher recall rate as more points were predicted as Class 0, allowing for a comparatively low number of false negatives. The same reason that more predictions being Class 0 applies to the large difference between specificity rates. Both classes have high negative predictive values because both classes have a low number of true negatives. Both classes have accuracy rates of 55% (Table 7). This model has an overall accuracy of 55%, suggesting that the model accurately predicts 55% of given data. As per the precision rate, 66% of the data predicted to be true was actually true. However, 52% of the water metrics that were actually true were predicted as true, which is concluded by the recall rate. As an accuracy metric that emphasizes false predictions, the f-score of this model was 59%. 52% of the negative points were predicted as false by the model, which is represented by the specificity score. Furthermore, out of data classified as false, 66% were true negatives (Table 8).

K-Nearest Neighbors

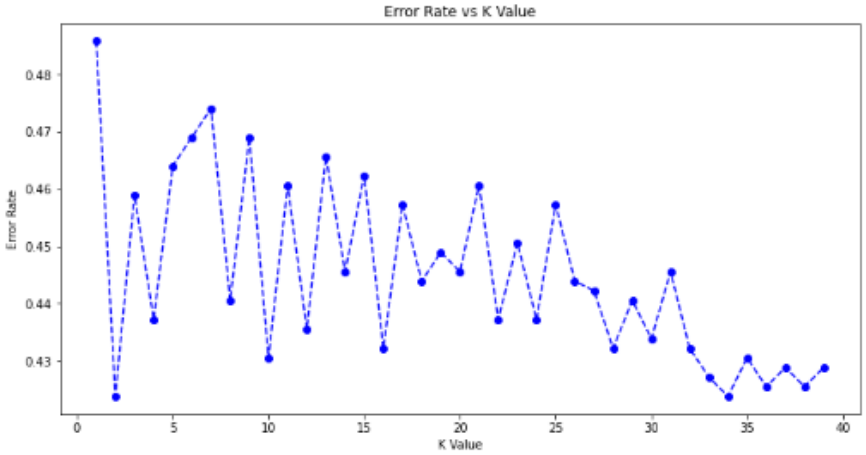

Figure 3. *Error Rate vs K Value*

Overall K values of 12, 16-18, and 20-21 inclusive performed the best, and earlier values out of which 5 resulted in the highest error rate. The K value of 16 resulted in the lowest error rate by a small margin.

Table 9 *Class evaluation metrics of k-nearest neighbors model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	307	37	199	54	61%	85%	71%	16%	41%	57%
0	37	307	54	199	41%	16%	23%	85%	61%	57%

Table 10 *Overall evaluation metrics of k-nearest neighbors model.*

Precision	51%
Recall	50%
F-Score	47%
Specificity	50%
Negative Predictive Value	51%
Overall Accuracy	57%

The model did not correctly predict all test data. Similar to the last model, the safe water class has a high false positive value, and the unsafe water class has a high false negative value. Compared to the models using Random ensemble and Decision Tree, this model predicts more inputs as non-potable. This indicates that this model has a more blurred distinction between the classes. The classes have a low precision rate in the range of approximately 40-60%. This indicates that the number of false positives for both non-potable and potable is almost equal to the number of true positives. The model needs to be better trained in identifying false data as false and not true. On the other hand,

the recall rates, as in the prior model, have a difference of 69% between them. This model tends to classify more points as of Class 0, causing Class 1 to have more false negatives. The specificity score presents the conclusion that the algorithm does not have a clear line between the two classes and results in more water samples to be predicted as Class 0. Similar to the precision rates, the negative predicted values are in the range of approximately 40-60%. The number of false negatives is almost equal to the number of true negatives. The overall accuracy rate is 57%. The number of true positives is higher compared to most models, excluding the Linear and Logistic models, even though the model predicts more Class 0, indicating that Class 1 needs to be better distinguished (Table 9). The K-Nearest Neighbor model precisely classifies 51% of predicted positive points, and 50% of true testing data is predicted true. Consequently, the f-score is 47%, by a small margin below 50%. With current testing, the model is capable of predicting 50% of false data samples as negative. Additionally, the negative predictive values 51%, indicating that the model needs to be better trained. Both the recall and specificity rates are low because of the skew of predictions to be Class 0. Considering the overall metrics, it can be observed that the model accurately predicts 57% percent of water samples (Table 10).

K-Means Clustering

Table 11 *Class evaluation metrics of k-means clustering model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	180	120	116	181	61%	61%	50%	55%	40%	50%
0	120	180	181	116	41%	40%	51%	45%	61%	50%

Table 12 *Overall evaluation metrics of k-means clustering model.*

Precision	50%
Recall	50%
F-Score	50%
Specificity	50%
Negative Predictive Value	50%
Overall Accuracy	50%

The model did not correctly predict all data. Both classes have a high false negative and high false positive value. This number of inputs predicted as potable and non-potable are approximately equal. The overall precision rate is in the range of 40-60%, suggesting that approximately between 25 and 35 of the data points that were predicted as true were actually true. The recall rates and specificity rates, unlike in the prior models, do not have a significant difference between them. This is because the model tends to classify an equal number of points as Class 1 and Class 0, but considering the unbalanced data, the accuracy rate is low. The negative predictive value is also in the range of the precision rate and shows that between 25 and 35 of the data points that were predicted as false were actually false. The overall accuracy rate is 50%, meaning that the model can predict approximately 50% of the testing data correctly (Table 11). The K-Means Clustering model precisely evaluates 50% of predicted true points, and 50% of positive data is predicted true. The f-score is 50%. The model evaluates 50% of false points as negative. The negative predictive

value is 50%. The precision, recall, f-score, specificity, and negative predictive scores are all about 50%, showing that there are a significant number of false positives and false negatives. But all predictions in this model are not concentrated in Class 0; instead, both classes have almost equal predictions. Considering the accuracy, the model accurately predicts 50% percent of points (Table 12).

Support Vector Machine

Table 13 *Class evaluation metrics of support vector machine model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
0	361	0	236	0	60%	100%	75%	0%	0%	60%
0	0	361	181	236	0%	0%	0%	100%	60%	60%

Table 14 *Overall evaluation metrics of support vector machine model.*

Precision	30%
Recall	50%
F-Score	38%
Specificity	50%
Negative Predictive Value	30%
Overall Accuracy	60%

This model classifies all test data as unsafe which is not appropriate for production. The Support Vector Machine algorithm classified the data in the same way as the Logistic Regression algorithm, and the same conclusions can be made (Table 13 and Table 14).

Machine Learning Models: Predicting the Contaminant

Table 15 *Class evaluation metrics of the predicting the contaminant model.*

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
Algae	4	30	0	0	100%	100%	100%	100%	100%	100%
Ammonia	1	33	0	0	100%	100%	100%	100%	100%	100%
Arsenic	1	33	0	0	100%	100%	100%	100%	100%	100%
Bacteria	2	32	0	0	100%	100%	100%	100%	100%	100%
Clay, Silt, Sand	2	32	0	0	100%	100%	100%	100%	100%	100%

Class	TP	TN	FP	FN	Precision	Recall	F-Score	Specificity	Negative Predictive Value	Accuracy
Copper	2	32	0	0	100%	100%	100%	100%	100%	100%
Disinfectant	2	32	0	0	100%	100%	100%	100%	100%	100%
Fluoride	0	33	0	1	0%	0%	0%	100%	97%	97%
Gasoline/ Petrol	2	32	0	0	100%	100%	100%	100%	100%	100%
Hydrogen Sulfide	6	28	0	0	100%	100%	100%	100%	100%	100%
Lead/Iron	4	30	0	0	100%	100%	100%	100%	100%	100%
Manganese	2	32	0	0	100%	100%	100%	100%	100%	100%
Nitrates/ Nitrites	0	33	0	1	0%	0%	0%	100%	97%	97%
Potassium Permanganate	1	33	0	0	100%	100%	100%	100%	100%	100%
Radioactive Elements (Radium, Radon, Uranium)	1	31	2	0	33%	100%	65%	94%	100%	94%
Sodium	1	33	0	0	100%	100%	100%	100%	100%	100%
Zinc	1	33	0	0	100%	100%	100%	100%	100%	100%

Table 16 *Overall evaluation metrics of the predicting the contaminant model.*

Precision	84%
Recall	88%
F-Score	86%
Specificity	100%
Negative Predictive Value	100%
Overall Accuracy	99%

The model predicts data not in the Fluoride or Nitrates/Nitrites classes accurately with values of zero for the false positive and false negative. The Fluoride and Nitrates/Nitrites classes were trained to have no distinct smell, odor, or color. The Radioactive Elements class is trained with the same characteristics, implying that the two incorrectly predicted data points were correctly evaluated. Therefore, for the data and contaminants inputted, the model can be deployed successfully. The classes with the lowest precision rates that are not 100% include the Fluoride, Nitrates/Nitrites, and Radioactive Elements (Radium, Radon, and Uranium) classes. The low rate shows that a small

amount of the data points predicted as that class is actually true; however, due to the small amount of testing data, the Fluoride and Nitrates/Nitrites classes have true positive and false positive values of 0. Similarly, the Fluoride and Nitrates/Nitrites all have recall rates of 0%. These classes are weak in predicting true as positive. This does not imply that the Fluoride will never have a true positive value above 0 since very little data is provided that no clear conclusion can be drawn. The negative predictive value is always above 97% due to a large number of true negatives compared to the false negatives. Overall, this model has high overall accuracy and specificity rates. For most classes, the precision, recall, f-score, specificity, negative predictive value, and accuracy rates are all 100%. The Fluoride and Nitrates/Nitrites classes that do not have 100% accuracy both were trained to have no distinct smell, odor, or color. The Radioactive Elements class is trained with the same characteristics, implying that the two incorrectly predicted data points were correct. The distinction between fluoride, nitrates/nitrites, and radioactive elements will be made through the app and not the model. Considering real-life situations, the accuracy is approximately 100% but not quite because of the self-engineered data and lack of variability in data. The model predicts all data accurately for the contaminants that it was trained for.

Sources of Error

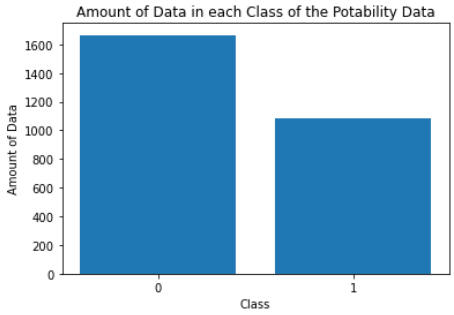

Figure 19 *Amount of data in each class of the predicting potability model data set.*

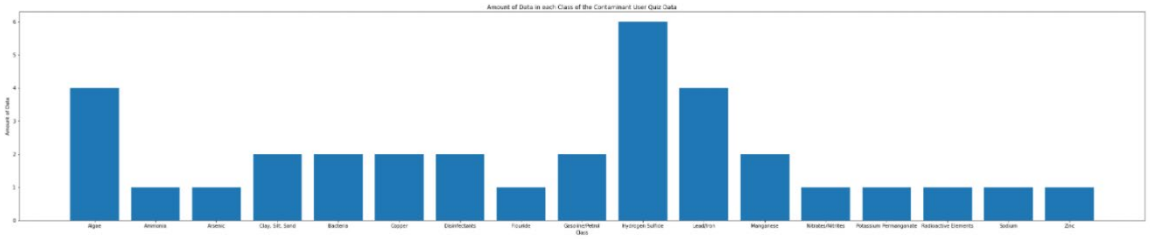

Figure 20 *Amount of data in each class of the predicting the contaminant model data set.*

The bar graphs above show the differences between data available for classes within each dataset. In both models, specifically, the contaminate user quiz model, there was a low amount of data available. This small amount of data does not accurately represent the potential of the model, for example, in the aluminum and ammonia classes. This results in inaccurate precision, recall, f-score, specificity, negative predictive value, and accuracy as a low number of data points were tested. Another source of systematic error includes that the dataset is unbalanced or does not have about the same amount of data between classes, which also returns imprecise evaluation metrics for each class. For example, in the water quality dataset, Class 0 has 1,666 points while Class 1 has 1,083.

Improvements

Any future advancement in this innovation would require more data. Low amounts of data do not train the model to its full potential using as many scenarios as possible. With more data samples, models have the capability to predict more varied data and be more efficient. In addition to this problem and area of possible improvement, the current data is unbalanced when regarding data points in each class, which results in imprecise precision, recall, f-score, specificity, negative predictive value, and accuracy scores. This can be improved through data augmentation. This is when data is duplicated and altered in some way to provide more variability. For example, changing the pH of 7 to 7.2 but keeping the rest the same provides another data point for training. By performing data augmentation, there will be more data that is also more balanced, and this method will prevent overfitting the network as well. The Quick Clean Water Django Python application allows users to input their metrics into Quick Clean Water anonymously to improve ML datasets if the user has completed a lab test on their water. Regarding the model for the contaminant test, the odor, color, and taste of more contaminants such as Aluminum and dyes/paints need to be researched, to identify taste, or tested for, to record odor and color in water. Validating this model only through current data would be insufficient and inaccurate since it was generated through research. As a more promising alternative to reduce bias, Quick Clean Water can partner with a local water treatment lab to conduct anonymous surveys with consumers about their water odor, color, taste, and contaminant. Data is the base of a good machine learning model; therefore, the model should train using relevant data.

Quick Clean Water Testing

Kirkland Signature Purified Water with Added Minerals

Water is sourced from Niagara Falls and is filtered using Advanced Filtration, Ozone, and Reverse Osmosis technologies. Ingredients include purified water, potassium bicarbonate, sodium bicarbonate, calcium citrate, sodium chloride, and magnesium oxide.

Table 17 *Kirkland signature purified water with added minerals water testing strips results.*

Metric	Reading	Safe for Drinking
Total Hardness (mg/L)	0	Yes
Free Chlorine (mg/L)	0.5	Yes
Iron (mg/L)	5	Yes
Copper (mg/L)	0.5	Yes
Lead (mg/L)	0	Yes
Nitrate (mg/L)	0	Yes
Nitrite (mg/L)	0	Yes
MPS (mg/L)	0	Yes
Total Chlorine (mg/L)	0.5	Yes

Metric	Reading	Safe for Drinking
Fluoride (mg/L)	0	Yes
Cyanuric Acid (mg/L)	0	Yes
Ammonia Chloride (mg/L)	0	Yes
QUAT/QAC (mg/L)	5	No
Total Alkalinity (mg/L)	20	Yes
Carbonate (mg/L)	20	Yes
pH	6.8	Yes

Table 18 *Kirkland signature purified water with added minerals quick clean water device results*

Metric	Reading	Safe for Drinking
pH	6.97	Yes
Turbidity (NTU)	0.4	Yes
Temperature (Celsius)	24.29	Yes
Temperature (Farenheit)	75.72	Yes
Total Dissolved Solids (ppm)	42.15	Yes
Conductivity (s/cm)	62.19	Yes
Salinity (mg/L)	34.6	Yes
ML Safe for Drinking	NOT SAFE	No

Table 19 *Kirkland signature purified water with added minerals quick clean water contaminant quiz results.*

Metric	Output
Odor	None
Color	None
Taste	None
Contaminant Prediction	Fluoride, Nitrites/Nitrates, or Radioactive Elements

Out of the sixteen metrics that the water testing strips tested for, only QUATs, which are disinfectant chemicals, were present in the water at an unsafe level. The Quick Clean Water software did alert the user that the water has a low amount of total dissolved solids and is non-potable. Since the water was evaluated to be unsafe, the user proceeded to take the contaminant quiz and entered None for all three categories. Quick Clean Water's ML predicted there to be fluoride, nitrites/nitrates, or radioactive elements in the water. This indicates that minimally distinct odors cannot be identified by the user to claim that the water has chemical-like or disinfectant properties. These could possibly be combated by training the model to recognize inputs of None for odor, color, and taste as disinfectants along with fluoride, nitrites/nitrates, or radioactive elements. Further discriminating between the contaminants could utilize sensor readings from additional sensors.

Tap Water from Irving, Texas

Water is sourced from Lake Ray Hubbard, Lake Tawakoni, Elm Fork of the Trinity River, Lake Chapman, Lake Grapevine, Lewisville Lake, and Lake Ray Roberts.

Table 20 *Tap water from irving, texas water testing strips results*

Metric	Reading	Safe for Drinking
Total Hardness (mg/L)	50	Yes
Free Chlorine (mg/L)	0.5	Yes
Iron (mg/L)	0	Yes
Copper (mg/L)	0	Yes
Lead (mg/L)	0	Yes
Nitrate (mg/L)	0	Yes
Nitrite (mg/L)	0	Yes
MPS (mg/L)	0	Yes
Total Chlorine (mg/L)	0.5	Yes
Fluoride (mg/L)	0	Yes
Cyanuric Acid (mg/L)	0	Yes
Ammonia Chloride (mg/L)	0	Yes
QUAT/QAC (mg/L)	0	Yes
Total Alkalinity (mg/L)	20	Yes
Carbonate (mg/L)	80	Yes
pH	7.2	Yes

Table 21 *Tap water from irving, texas quick clean water device results.*

Metric	Reading	Safe for Drinking
pH	7.17	Yes
Turbidity (NTU)	0.7	Yes
Temperature (Celsius)	23.84	Yes
Temperature (Farenheit)	74.92	Yes
Total Dissolved Solids (ppm)	221.35	Yes
Conductivity (s/cm)	330.37	Yes
Salinity (mg/L)	181.7	Yes
ML Safe for Drinking	NOT SAFE	No

Table 22 *Tap water from irving, texas quick clean water contaminant quiz results.*

Metric	Output
Odor	None
Color	None
Taste	None
Contaminant Prediction	Fluoride, Nitrites/Nitrates, or Radioactive Elements

The water testing strips claim that all metrics tested for are within safe levels; however, the Quick Clean Water device classified that sample as non-potable and that it contained fluoride, nitrites/nitrates, or radioactive elements. This can be concluded as an error by the Machine Learning model that predicts potability as the accuracy is 55%.

Pensacola Beach

Pensacola Beach is located in Pensacola, Florida, and provides a view of the Gulf of Mexico.

Table 26 *Pensacola beach water testing strips results.*

Metric	Reading	Safe for Drinking
Total Hardness (mg/L)	425	Yes
Free Chlorine (mg/L)	0.5	Yes
Iron (mg/L)	0	Yes

Metric	Reading	Safe for Drinking
Copper (mg/L)	0	Yes
Lead (mg/L)	0	Yes
Nitrate (mg/L)	0	Yes
Nitrite (mg/L)	0	Yes
MPS (mg/L)	0	Yes
Total Chlorine (mg/L)	0.5	Yes
Fluoride (mg/L)	0	Yes
Cyanuric Acid (mg/L)	0	Yes
Ammonia Chloride (mg/L)	0	Yes
QUAT/QAC (mg/L)	5	No
Total Alkalinity (mg/L)	0	Yes
Carbonate (mg/L)	0	Yes
pH	6.6	Yes

Table 27 *Pensacola beach water quick clean water device results.*

Metric	Reading	Safe for Drinking
pH	6.89	Yes
Turbidity (NTU)	0.3	Yes
Temperature (Celsius)	23.74	Yes
Temperature (Farenheit)	74.74	Yes
Total Dissolved Solids (ppm)	778.45	No
Conductivity (s/cm)	1161.87	Yes
Salinity (mg/L)	639.03	Yes
ML Safe for Drinking	NOT SAFE	No

Table 28 *Pensacola beach water quick clean water contaminant quiz results.*

Metric	Output
Odor	None
Color	None
Taste	Salty
Contaminant Prediction	Sodium

According to the water testing strips, Pensacola Beach has an unsafe amount of QUATs in the water, but similar to other salt-water samples, the strips lack the ability to detect salinity. The Quick Clean Water device and quiz both are able to evaluate the high salinity but fail to distinguish disinfectants. The user is often not able to identify chemical-like odors, and the Arduino is unable to notify the user that the disinfectant levels should be unsafe. More and higher-quality sensor readings could be used to combat this problem.

Big Lagoon State Park

Big Lagoon State Park features a saltwater lagoon and natural environment.

Table 38 *Big lagoon state park water testing strips results.*

Metric	Reading	Safe for Drinking
Total Hardness (mg/L)	425	Yes
Free Chlorine (mg/L)	0.5	Yes
Iron (mg/L)	0	Yes
Copper (mg/L)	0	Yes
Lead (mg/L)	0	Yes
Nitrate (mg/L)	0	Yes
Nitrite (mg/L)	0	Yes
MPS (mg/L)	0	Yes
Total Chlorine (mg/L)	0.5	Yes
Fluoride (mg/L)	0	Yes
Cyanuric Acid (mg/L)	0	Yes
Ammonia Chloride (mg/L)	0	Yes
QUAT/QAC (mg/L)	10	No

Metric	Reading	Safe for Drinking
Total Alkalinity (mg/L)	10	Yes
Carbonate (mg/L)	0	Yes
pH	7.2	Yes

Table 39 *Big lagoon state park water quick clean water device results.*

Metric	Reading	Safe for Drinking
pH	7.5	Yes
Turbidity (NTU)	0.2	Yes
Temperature (Celsius)	23.97	Yes
Temperature (Farenheit)	75.15	Yes
Total Dissolved Solids (ppm)	772.11	No
Conductivity (s/cm)	487.66	Yes
Salinity (mg/L)	268.21	Yes
ML Safe for Drinking	NOT SAFE	No

Table 40 *Big lagoon state park quick clean water contaminant quiz results.*

Metric	Output
Odor	None
Color	Yellow
Taste	None
Contaminant Prediction	Arsenic

The lagoon at Big Lagoon State Park has one of the highest measured QUATs rates measured in this experiment, so Quick Clean Water efficiently and accurately evaluates the water to the unsafe for consumption. The water strips portray a high QUATs amount while the Quick Clean Water contaminate quiz predicted arsenic. It is true that the strips detect a limited number of contaminants, but in this case, it is more likely that the disinfectants showed no obvious sign of presence. This contaminant could be detected using additional sensors, and the contaminant ML model should be trained to classify disinfectants if the water has no odor, color, or taste. With arsenic and disinfectants, the sample is saltwater but was not detected to be so due to entering taste as None. To address multiple contaminants, the model can output the top predictions based on percent.

Public Restroom in Destin, Florida

Water is sourced from the upper Floridan Aquifer, which is made of limestone.

Table 41 *Public restroom in destin, florida water testing strips results.*

Metric	Reading	Safe for Drinking
Total Hardness (mg/L)	50	Yes
Free Chlorine (mg/L)	0.5	Yes
Iron (mg/L)	0	Yes
Copper (mg/L)	0	Yes
Lead (mg/L)	0	Yes
Nitrate (mg/L)	0	Yes
Nitrite (mg/L)	0	Yes
MPS (mg/L)	1	Yes
Total Chlorine (mg/L)	0.5	Yes
Fluoride (mg/L)	0	Yes
Cyanuric Acid (mg/L)	0	Yes
Ammonia Chloride (mg/L)	0	Yes
QUAT/QAC (mg/L)	5	No
Total Alkalinity (mg/L)	20	Yes
Carbonate (mg/L)	80	Yes
pH	7.6	Yes

Table 42 *Public restroom in destin, florida water quick clean water device results.*

Metric	Reading	Safe for Drinking
pH	7.3	Yes
Turbidity (NTU)	0.1	Yes
Temperature (Celsius)	23.97	Yes

Metric	Reading	Safe for Drinking
Temperature (Farenheit)	75.19	Yes
Total Dissolved Solids (ppm)	218.91	Yes
Conductivity (s/cm)	326.73	Yes
Salinity (mg/L)	179.7	Yes
ML Safe for Drinking	NOT SAFE	No

Table 43 *Public restroom in destin, florida quick clean water contaminant quiz results.*

Metric	Output
Odor	Chemical-Like
Color	None
Taste	None
Contaminant Prediction	Disinfectants

Water in a public restroom in Destin, Florida is found to contain high amounts of disinfectants. The non-potability of the water was predicted accurately by the Quick Clean Water device and Potability ML model. The water had a clear chemical-like smell and was not tasted, but the model was still able to classify the impurity in the sample accurately.

Conclusion

Machine Learning Models: Predicting Potability

The hypothesis was at first that the K-Means Clustering would allow the model to perform the best compared to other classification and regression algorithms tested. The K-Means Clustering algorithm proved to be limiting as the model using the Random Ensemble algorithm performed the best and better satisfied the objective of the research. The models Linear Regression, Logistic Regression, and Support Vector Machine all classified all test data as one class. These models cannot be applied in the real world as evaluating all water samples as non-potable or potable would not be accurate. The K-Nearest Neighbors model has a large difference between recall rates between the two classes. This model tends to classify more points as of Class 1, causing Class 0 to have more false positives and Class 1 to have more false negatives. Unlike this model, the K-Means Clustering model classifies approximately an equal number of data points into both of the classes, but they are evaluated mostly incorrectly indicated by the accuracy rate of 50%. Similar to the Linear, Logistic, and SVC models discussed above, applying the Random Ensemble and Decision Tree models in the real world would provide results slightly skewed towards Class 0 compared to the previously referred to models. Between the Decision Tree model and Random Ensemble model, the values of all metrics for the Random Ensemble model are greater, presenting a more efficient and accurate algorithm to utilize for this task.

The only contaminant present in harmful amounts in the samples is QUATs/QACs. According to the testing strips, the chemicals are present in four out of the five samples: Kirkland Signature purified water with added minerals, Pensacola beach, Big Lagoon State Park, and the public restroom in Destin, Florida. Out of four, only one was predicted to contain disinfectants by the Quick Clean Water device. This low accuracy is due to high concentrations of sodium present in the samples as well, which is a contaminant not detected by the strips but is detected by the Quick Clean Water device. Since the Quick Clean Water software does test for more contaminants, the system did evaluate the sample from Pensacola Beach better than the strips. Identifying more than one impurity can be done by utilizing sensor readings from additional sensors or by outputting the top predictions for each input based on percent. Another reason for the low accuracy rate in detecting disinfectants is that users are unable to identify the properties of these contaminants. As stated before, further distinctions could involve more sensors and training the model to recognize disinfectants as possibly having no odor, color, or taste. Samples from the Gulf of Mexico beaches were not evaluated as having unsafe amounts of salinity by the Arduino. This can be improved using higher-quality sensors. These investigations primarily focused on the classification of disinfectants, and to improve the device and software, varied water samples containing other contaminants must be tested.

Acknowledgements

I would like to recognize the invaluable assistance of Dr. Rajeev Dwivedi in introducing me to Arduino, his guidance, and critiques on my research work. I wish to thank my parents for their support throughout my studies.

References

1. Potential Well Water Contaminants and Their Impacts. https://www.epa.gov/privatewells/potential-well-water-contaminants-and-their-impacts (accessed Oct 5, 2021).
2. Contamination in U.S. Private Wells. https://www.usgs.gov/special-topic/water-science-school/science/contamination-us-private-wells (accessed Oct 5, 2021).
3. Water contamination. https://www.cdc.gov/healthywater/other/agricultural/contamination.html (accessed Oct 5, 2022).
4. Drinking-water. https://www.who.int/news-room/fact-sheets/detail/drinking-water (accessed Oct 5, 2021).
5. Drinking-Water. "Drinking Water Contaminant – Corrosive Water." *Drinking Water and Human Health*, 23 Aug. 2019, https://drinking-water.extension.org/drinking-water-contaminant-corrosive-water/.
6. 13 of the biggest water contamination disasters in the world. all-about-water-filters.com/biggest-water-contamination-disasters/ (accessed Oct 5, 2021).
7. Clean Water Testing. www.cleanwatertesting.com/resources/water-testing-faqs/ (accessed Oct 5, 2021).
8. Measuring Salinity - Department of Environment, Water and Natural Resources. www.landscape.sa.gov.au/mr/publications/measuring-salinity (accessed Oct 8, 2022).
9. ALPAYDIN, ETHEM. "Machine Learning." *Amazon*, MIT PRESS, 2021, https://docs.aws.amazon.com/machine-learning/latest/dg/when-to-use-machine-learning.html.
10. Kadiwal, A. Water quality. www.kaggle.com/adityakadiwal/water-potability (accessed Oct 20, 2022).
11. Ammonia in water. www.multipure.com/purely-social/science/ammonia-in-water/ (accessed Oct 6, 2021).
12. Arsenic in Groundwater, www.idph.state.il.us/envhealth/factsheets/arsenicwater.htm (accessed Oct 6, 2021).
13. Basic Information About Lead Drinking Water. www.epa.gov/ground-water-and-drinking-water/basic-information-about-lead-drinking-water (accessed Oct 7, 2021).

14. Copper in drinking water - Washington State Department of Health. doh.wa.gov/portals/1/Documents/pubs/331-178.pdf (accessed Oct 7, 2021).
15. Department of Health. www.health.ny.gov/environmental/water/drinking/oxygenates_in_drinking_water.htm (accessed Oct 7, 2021).
16. Facts on Drinking Water CSO G Uranium. www2.gnb.ca/content/dam/gnb/Departments/h-s/pdf/en/HealthyEnvironments/water/Uraniume.pdf (accessed Oct 8, 2021).
17. Fluoride in your drinking water: Good or bad? www.water-rightgroup.com/resources/fluoride-in-drinking-water/ (accessed Oct 7, 2021).
18. The Bronx River Ammonia Test. www.thirteen.org/edonline/studentstake/water/BronxRiver/Ammonia/ammonia.htm (accessed Oct 7, 2021).
19. How to Remove Iron Manganese and Odor from Well Water. www.cleanwaterstore.com/blog/remove-iron-manganese-odor-well-water-step-2/ (accessed Oct 7, 2021).
20. Hydrogen Sulfide and Sulfur Bacteria in Well Water. www.health.state.mn.us/communities/environment/water/wells/waterquality/hydrosulfide.html (accessed Oct 7, 2021).
21. Jannis Wenk Lecturer in Water Science & Engineering. Why a Canadian town's water supply turned pink. theconversation.2021/why-a-canadian-towns-water-supply-turned-pink-74315 (accessed Oct 8, 2022).
22. Learn about water. www.wqa.org/learn-about-water/water-q-a/manganese (accessed Oct 7, 2021).
23. Nitrates and Nitrites in Drinking Water. Vermont Department of Health. www.healthvermont.gov/health-environment/drinking-water/nitrates-and-nitrites (accessed Oct 7, 2021).
24. Orr, T. How to quickly get rid of algae from well water. waterpurificationguide.com/how-to-quickly-get-rid-of-algae-from-well-water/ (accessed Oct 6, 2021).
25. Potassium Permanganate Hazard Summary Workplace Exposure, nj.gov/health/eoh/rtkweb/documents/fs/1578.pdf (accessed Oct 8, 2021).
26. Radium in Drinking Water Fact Sheet, www.idph.state.il.us/cancer/factsheets/radium.htm (accessed Oct 6, 2021).
27. Radon in Drinking Water Wells. University of Rhode Island Water Quality Program. web.uri.edu/safewater/files/TipSheetC13-Radon.pdf (accessed Oct 8, 2021).
28. Salt (sodium chloride) in drinking water. ww2.health.wa.gov.au/Articles/S_T/Sodium-in-drinking-water (accessed Oct 8, 2021).
29. Sulfur, Hydrogen Sulfide, Sulfate and Sulfate-Reducing Bacteria. www.knowyourh2o.com/indoor-6/sulfur-hydrogen-sulfide-sulfate-and-sulfate-reducing-bacteria (accessed Oct 7, 2021).
30. How to tell if there is bacteria in your water: Contaminated water. www.lvecowater.com/water-purification/tell-tale-signs-bacteria-water/ (accessed Oct 6, 2022).
31. 331-286 revised February 2018 - home :: Washington State ... doh.wa.gov/Portals/1/Documents/Pubs/331-286.pdf (accessed Oct 6, 2022).
32. Water Treatment Solutions. Lenntech Water Treatment & Purification. www.lenntech.com/periodic/water/arsenic/arsenic-and-water.htm (accessed Oct 6, 2022).
33. *Wellcare® Information for You about Sediment & Well Water.* www.watersystemscouncil.org/download/wellcare_information_sheets/potential_groundwater_contaminant_information_sheets/Sediment-Well-Water-FINAL.pdf (accessed Oct 6, 2022).
34. Water Treatment Solutions. Lenntech Water Treatment & Purification. www.lenntech.com/turbidity.htm (accessed Oct 6, 2022).
35. What Are the Harmful Effects of Iron in Drinking Water? www.peninsulawater.com/is-iron-in-drinking-water-harmful/ (accessed Oct 7, 2022).
36. What Makes Gasoline on the Ground so Colorful? www.palmenfiat.com/blog/what-makes-gasoline-on-the-ground-so-colorful/ (accessed Oct 7, 2022).

37. *Zinc - Sask H2O*. www.saskh2o.ca/pdf-watercommittee/zinc.pdf (accessed Oct 8, 2022).
38. How to Check TDS Level, Ideal TDS Level of Drinking Water. www.bestrowaterpurifier.in/blog/how-to-check-tds-level-of-water/ (accessed Oct 8, 2022).
39. *Water Quality Salinity Standards*. mrccc.org.au/wp-content/uploads/2013/10/Water-Quality-Salinity-Standards.pdf (accessed Oct 8, 2022).
40. Sa health. www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/public+health/water+quality/salinity+and+drinking+water (accessed Oct 8, 2020).
41. Sewer Disposal: What Can Go down the Drain?, blink.ucsd.edu/safety/research-lab/hazardous-waste/disposal-guidance/sewer.html (accessed Nov 10, 2021).
42. Water supply. www.cityofirving.org/731/Water-Supply (accessed Nov 10, 2021).