Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

# WATER QUALITY PREDICTION USINGMACHINE LEARNING

**Dr. M. Suchithra** Associate Professor, Dept. of Computing Technologies SRM University, Kattankulathur Chennai, India.
**Bala Vamsi Krishna Y, Harihara Mahidhara** B. Tech, Students, Dept. of Computing Technologies SRM University, Kattankulathur Chennai, India.

## ABSTRACT

It would be impossible for agriculture and industry to function without access to fresh water. An essential part of managing freshwater resources is conducting regular water quality tests. The annual study from the World Health Organization (WHO) shows that many people, particularly pregnant women, and children, are becoming ill or dying because of the lack of access to clean drinking water. Water quality testing is essential before using water for anything, including watering animals, spraying chemicals (pesticides, etc.), or drinking. Testing the water's quality is one method for locating safe supplies. Hence, effective water monitoring is fundamental to ensuring potable, clean, and hygienic water supplies. Doing water quality tests is essential for examining the genuine operation of water sources, ensuring the quality of water for consumption, detecting disease outbreaks, and authorising procedures and safety measures. Quantifying the degree to which physiological, chemical, and biological characteristics of water make it suitable for a certain use is what we mean when we talk about water quality.

***Keywords:*** *Machine learning; water monitoring; disease detection.*

## I. INTRODUCTION

Water not only controls all the body's functions but is also the primary medium via which energy is transported to all the cells. Eighty percent of the brain is made up of water. Severe dehydration may cause brain fog and impair one's ability to think clearly. Water is among the most crucial regular resources for the continuation of all species on Earth. The versatility of water as a resource explains why it is used for so many things beyond just hydration. Water is essential for the survival of all living things, including plants and animals. Simply put, water is essential to the survival of all known forms of life on Earth. For both agriculture and manufacturing, freshwater is a crucial resource. Important steps in the management of freshwater assets include collecting and analyzing data on water quality. According to the World Health Organization's annual report, many people, especially pregnant women, and children, are dying due to a lack of clean drinking water. Whether the water will be used for drinking, bathing, or watering animals, it is essential to verify its purity before use. Finding clean water requires a tool that measures water purity. Hence, proper water testing is crucial to preserving pure water supplies. The results of water tests are crucial in determining the reliability of water systems, gauging the safety of drinking water, identifying disease outbreaks, and green-lighting preventative and maintenance schedules.

The physical, chemical, and biological characteristics of water are all considered when determining water quality. Surface water and groundwater testing may help us answer questions about whether a body of water is fit for various uses, including those of drinking, washing, and operating a water system. Using the results of water quality tests, it is possible to examine water quality at a regional, state, or national scale, moving from one body of water to the next. Since pathogenic bacteria, infections, tapeworms, and so forth

constitute the most well enough and limitless health risk associated with drinking water, microbiological purity is usually the most urgent problem. A health danger is posed by an excess of certain synthetic compounds in water supplies. Fluoride, arsenic, & nitrate are all components of these man-made chemicals. The customer should be given clean drinking water for their own use in cooking, cleaning, and personal hygiene. Purity standards established at the point of delivery to consumers must be met. Breiman's random forests is one of the best machines (statistical) learning methods for practical uses (RF). Until recently, random forests were relatively unknown in water research, especially in hydrological applications, despite their obvious value. This means that the potential of "Breiman's" one-of-a-kind computation and its variants in water assets and applications is being wasted. While RF-based calculations are often used in relapsing and grouping concerns and the computation of significant measures, its use for decile expectation, durability inquiry, and off-the- cuff guessing seems to be less well recognized among water experts and researchers. In terms of water resources, random forests are generally considered to be a kind of data-driven model.

## II. LITERATURE SURVEY

Intelligent machines Machine learning, an AI subfield that develops predictive models via past data, provides statistical methods with which to probe, interpret, and analyze information. First and foremost, it gives the machine the ability to think independently of human input. Data from the past, known as training data, helps it make predictions about the future. Analyzing and forecasting water quality metrics using machine learning and the K Means method. When applied to data, K means is an unsupervised learning model that uses training data sets to analyse the data and generates a hyperplane to partition the new inputs. Muddy water, flavored seltzer, salt, tap water, and drinkable water are just a few of the many types of water used in various training regimens. Predicted new data sets with comparable values for pH, conductivity, and turbidity may be grouped together. During the training phase, the network is taught to categorise all the quality values of the water it encounters into distinct clusters, each of which represents a certain kind of water, such as mud and water, flavored seltzer, salt water, city water, and drinking water. While testing a new dataset that use the K Means technique, the machine will choose the number of clusters that need to be generated, offer the centroid for each cluster, and categorise the new dataset such that it fits into the cluster with the closest name. As a result, the K Means method from the area of machine learning is used to make predictions about water quality parameters.

Analytics of large data sets for water quality administration: IoT devices use various sensors to continuously collect data on stream water's turbidity, ORP, temperature, pH, conductivity, etc. (Mohammed Salah Uddin Chowdurya, Talha Bin Emranb, Subhasish Ghosha, Abhijit Pathaka, Mohd. Manjur Alama, Nurul Absara, Karl Anderssonc, Mohammad Shahadat Hossaind, 2019). As a result, IoT gadgets may send the collected data in a continuous stream to a distant cloud server known as a Data Aggregator. Apart from this, the amount of unstructured data is growing at a rate at which only Big Data Analytics programmes can keep up with it in terms of efficiently storing and analysing it. As a result, the information official's layer may be sent to the Apache Hadoop distribution and put into use. Hadoop facilitates the distributed processing and preparation of massive data over a cluster of computers. Hadoop is deficiency tolerant since employments are redirected naturally to the operating hubs whenever hubs are fizzled.

IoT applications demand rapid of browsing of information and extremely available documents in the database. Apache HBase, a NoSQL database implemented on Hadoop, will be used to store massive amounts of data at this time. As a result, the Hadoop dispersed record system is used to circle the data

(HDFS). HBase can handle groups of records in parallel with consistent queries. Elevated data is supplied by the Data stores because it's taken care of in HDFS. Hadoop clusters are distributed over a cluster of computers, with Apache Zookeeper in charge of coordination. The IoT app will aid customers in continuously seeing the outcomes of water quality investigations generated by the data layer throughout a variety of time arrangements. The information representation programme operates on consumer devices, for example, Phones, PCs, & work settings. The root customers will have the ability to make day day by to month/yearly groundwater resources report from the data the boards layer and image in the customer devices [3].

**Internet of things**

The concept of an interconnected network of everyday objects, known as the Internet of Things (IoT), has been discussed for years. With the development of cutting-edge wireless technology, it is gaining ground in this case. A sensor, a Node microcontroller, and a network all work together to allow for device-to-device communication. With the help of IoT gadgets, we can automatically gather data in real time. An IoT system consists primarily of sensors, a sensor gateway, a CPU, and a software application. Cloud service providers gather sensor data, store it, analyse it, and make decisions based on what they find. During the last several years, WSN and IoT have become indispensable tools for environmental monitoring. With the aid of sensed data, cloud storage makes data retrieval simple. There is now a chance to process, analyse, and visualise the data. By using these factors, it can accurately predict water quality. Online analytical tools allow the expected outcome to be displayed in the shape of plots, graphs, and graphs.

## III. BACKGROUND WORKDIFFERENT ALGORITHMS

Studies in the past have employed several different machine learning algorithms. Technologies such as Neural Networks, Support Vector Machines (or its variants such as LS-SVM and Random Forest), and Genetic Programming (GP and LGP) are examples (ANN, BP-NN, and GR-NN). There was no usage of any gradient enhancing methods. This research uses the predictive power of Random Forest & extreme Gradient Boosting (XGBoost) to analyse and identify the nine distinct variables.

**FACTOR PREDICTION**

Most of the consulted articles [1–5] attempted to forecast separate aspects of water quality. potability [1, 2], faecal coliform [4, 5], chlorophyll [5, 6], and water temperature [2] were some of these. These variables were predicted from a pool of five - eighteen different water quality indicators based on the literature reviewed for this study [1–6]. In contrast, this research may be used to prediction of all nine components and the investigation of their interdependence. The results of feature prediction are investigated, and it is made clear how these results might be put to practical use.

**DATA USED**

Several of the underlying research used limited data sets for both training and testing algorithms. These articles [1, 2, 5, 6] employed datasets with anything from 132 through 2063 rows of information. This research used the Random Forest and XGBoost algorithms, with a cumulative value of 3277 rows utilised for training and testing purposes. Several methods were used to divide the data for training and testing. Typically, test and training sets were divided 80/20 [1][3] or 75/25 [2][4]. Instead, this research used a 90%-10% split between training and test data.

| Parameters | Quality Range | Units |
|---|---|---|

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

| | | |
|---|---|---|
| pH | 6.5-8.5 | - |
| Hardness | 60-160 | Mg/L or PPM |
| Solids | 50-150 | PPM |
| Chloramines | 0-4 | Mg/L or PPM |
| Sulphates | 0-500 | Mg/L or PPM |
| Conductivity | 200-800 | S/cm |
| Organic carbon | 0-25 | PPM |
| Trihalomethanes | 0-100 | PPB |
| Turbidity | 0-5 | NTU |
| Potability | 0 or 1 | - |

*Fig.1- Parameter Ranges*

**DATA PREPROCESSING**

Data Pre-processing is very much essential for extracting the most accurate result from the machine learning algorithms. For our project we had to deal with dummy or null values to get accurate results. After that, data balancing is also done to make sure the even split up of predicting variable for training.

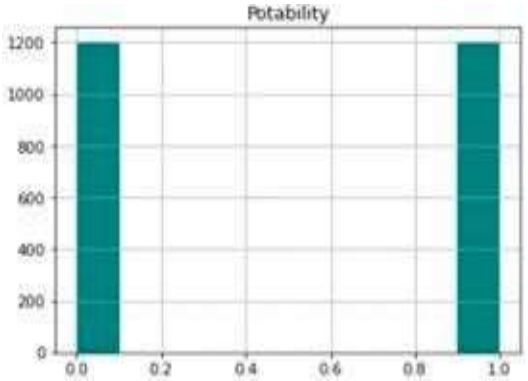


*Fig.2- Potability Distribution*

**DATA VISUALIZATION**

Data visualization is the graphical representation of the features in the dataset which allows us to analyse and make observations of each parameter value ranges.

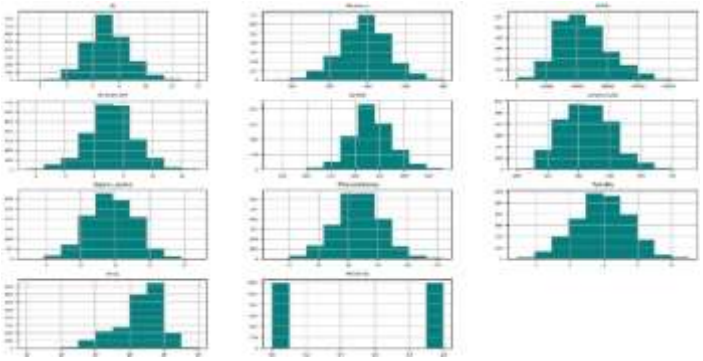

*Fig.3- Display of the count of features*

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

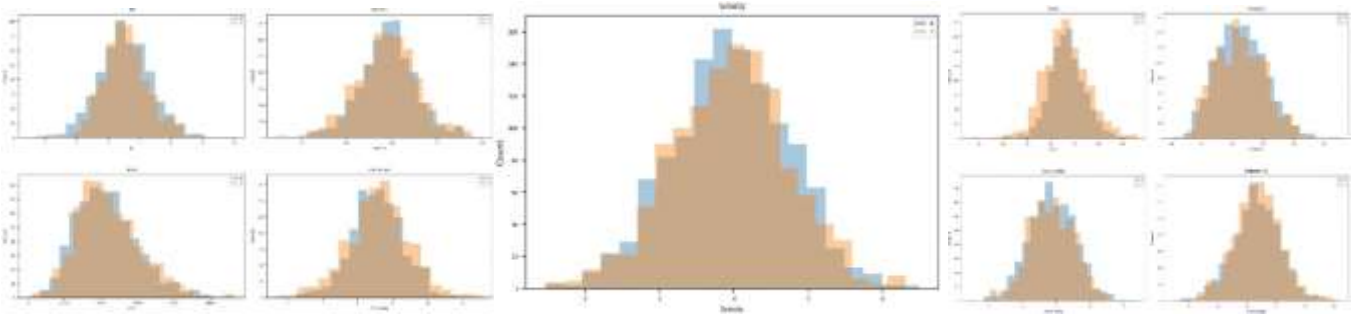

*Fig.4* *Fig.5* *Fig.6*

*Fig.4,5,6- 0's and 1's associated with specificparameters.*

## IV. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is used to show the distribution of data points through the specific measure. It is usually done as box-plot analysis in which the major part lies inside the box.

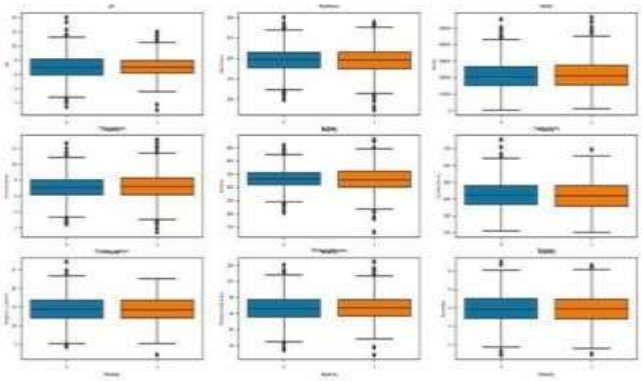

*Fig.7- Box-plot analysis of features*

## CORRELATION MATRIX

Correlation Matrix is used to depict the internal relations between each feature in the data and henceforth any unwanted data which are notrequired for the processing can be removed.

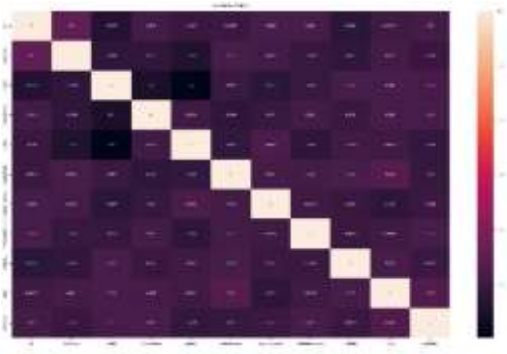

*Fig.8- Correlation Matrix*

## METHODS

Random Forest and XGBoost, a gradient boosting technique, are used to evaluate water quality variables (eXtreme Gradient Boosting). Python was used in Pycharm to develop the algorithms, along with libraries like SciKit-Learn, Pandas, NumPy, and Seaborn. Each algorithm was trained and evaluated using dataset collected from Kaggle. There were 3277 rows and 9 columns, which each algorthim is implemented successfully.

Among the dataset's total of 10 parameters, all ten were included in the model's training and validation processes. Consideration of the time frame during which data was gathered helped choose which variables to use. The quality and consistency of dataset sometimes varied from report to report. There are no eliminated variables because every variable has influence with the final prediction. The data set was obtained by way of a download from the Kaggle website. The success rates of the various factors varied, in both Random Forest & XGBoost. Both models achieved a best prediction of 77.5 in Random Forest, proving to be effective.

Rest of the components have a prediction accuracy between 53% and 75%. Very little fluctuation occurred outside of a one- or two-degree range for the majority. It took the Random Forest algorithm and XGB around two minutes to make a forecast. In most cases, notably for potability, its predictions were more accurate than those of the rest, which predicted component in around 10 to 15 seconds.

**TRAIL AND ERROR**

To determine the effects of each variable, many methods were used. Prediction of potabilty was the primary focus of most of the tests. Initially, the Random Forest method was only applied to a short dataset, consisting of no and over 139 rows. On the first Random Forest attempt, the prediction percentage was negative . During the second try, we narrowed down the number of parameters to just four. As a result, more information was included into the predictive model, leading to greater precision. When the data was examined manually, however, it was discovered that several of the numbers in each column were missing. Each missing data point was refilled using the column means to improve the Random Forest prediction accuracy. Each blank was then filled in, and the Random Forest method was re-executed to provide predictions for the potability levels. After dealing with data gaps and importing massive quantities of data from the website, the Random Forest method and XGB was found to have an accuracy rate for potability predictions in the range of 77-79%. To improve accuracy, more data collection was conducted after this promising first run. When a procedure was shown to be effective again for Random Forest algorithm, it was simply transposed to the XGBoost method using the same data.

Data normalization was also tried to increase the predictive performance of Random Forest and XG Boost. Both algorithms tried to normalize their data using three different methods. For starters, we tried SciKit-normalization Learn's feature. In this case, the data was normalized from -1 to 1, a range supplied by using the. Norm() function. Unfortunately, this gave a negative prediction %. The second effort at data normalization included applying a simple algorithm (where X represents all nine columns and 3277 rows, and N is the normalized outcome).

$$N = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

*Normalisation Formula*

The normalisation range was narrowed from 0 to 1 thanks to the easy formula. With this method of normalisation, we were able to provide potability prediction ratings that were, on average, 92.60 percent accurate. The Random Forest method was substantially speedier over the normalised data, even though its prediction accuracy was lower than that of the untreated data.

**ALGOITHMS DESCRIPTION**

The algorithms each have their own advantages. Fast in its operation, XGBoost also performed well but not spectacularly on several measures of water quality. Random Forest's prediction accuracy was

higher than XGBoost's for only two of the nine parameters, but it was substantially slower overall. Results for both methods were satisfactory across the board. To determine which of the elements can be predicted more accurately, it would simply be a question of altering parameters inside both systems.

## V. RESULTS

Predicting potability levels within a safe margin of error was the first objective. The simplicity of testing meant that the other criteria could be anticipated as well. The outcomes ranged widely depending on both the method and the variable in question.

The accuracies of the implemented models are as follows:

LGBMClassifier: 74.72%
XGBClassifier: 77.22%
AdaboostClassifier: 53.05%
RandomForest - 77.22%
DecisionTree - 65.27%
KNeighborsClassifier - 66.94%
Bernoulli - 51.11%
GaussianNaiveBayes - 59.44%
SVC - 65.83%
LogisticRegression - 48.05%

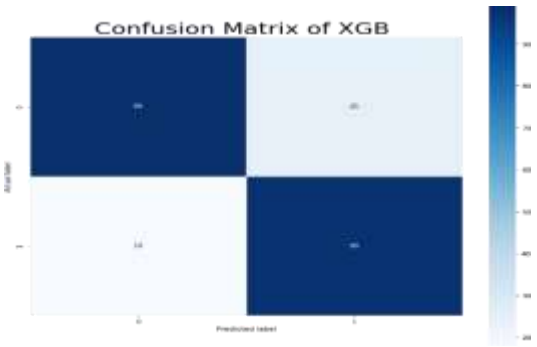


*Fig.9- Confusion Matrix of XGB*

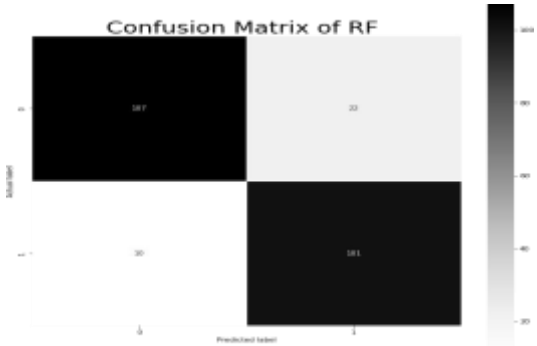


*Fig.10- Confusion Matrix of RF*

The Random Forest algorithm and XGB achieved prediction accuracies of 77.22% or higher, althoughthe time spent on each prediction was a cost. Potability of the water prediction had a roll of roughly two and a half minutes for every time the programme was run. But some factor estimations took almost three minutes to provide findings. The chart below reveals the predication accuracy rate for each model's forecast as well as the best and lowest prediction accuracy rate in the Random Forest algorithm.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

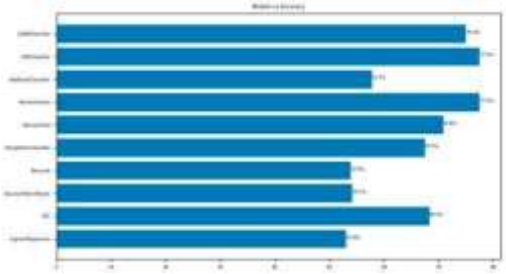

*Fig.11- Comparative Analysis of accuracies of all models*

## VI. CONCLUSION

One may find applications for supervised machine learning in a wide variety of settings and industries. The fields of environmental science and prediction have benefited greatly from the use of anomaly detection and prediction. Portland relies heavily on the Willamette River, therefore maintaining its high-water quality is crucial. To maintain a functional aquatic environment, we must act now. The dataset collection from website is very crucial. As it is, data analysis takes a lot of time and effort since it must be done manually. Machine learning techniques like Random Forest and XGBoost may be used to forecast these water quality parameters with accuracy rates more than 77%. It has been proven that water quality variables may be predicted with an accuracy of 80% or higher. Now that we know this, we can keep working to lessen the need for costly human inspection of data.

## VII. FUTURE WORK

Potability prediction "pseudo forecasting" was put through its paces in preliminary trials. The time frame within which a given element may be anticipated with any degree of certainty will be investigated. If a factor can be anticipated with enough data, then the sensors that gather this information may use it as a warning system. At now, every information is checked manually. If there are gaps in the data or the data doesn't really conform to typical patterns, this indicates a problem with the sensors, which must be addressed. For irregular readings, we either replace them with readings from other, more reliable sensors in the area, or we just ignore them. It is possible that human data inspection may be avoided if the algorithms evaluated in this research can be used to forecast future values for water quality factors. Alternatively, an alarm may be triggered to inspect the sensor for faults or issues in its surroundings if the algorithm can properly anticipate a data point, but the real data point is very different from the projected value. The Surveys and data collectors and analysers could save a tonne of money and time with this.

## REFERENCES

1. S. Malek, M. Mosleh, and S. M. Syed, “potability Prediction Using Support Vector Machine,” vol. 8, no. 1, p. 5, 2014.
2. E. Olyaie, H. Zare Abyaneh, and A. Danandeh Mehr, “A comparative analysis among computational intelligence techniques for potability prediction in Delaware River,” Geoscience Frontiers, vol. 8, no. 3, pp. 517–527, May 2017.
3. X. Ji, X. Shang, R. A. Dahlgren, and M. Zhang, “Prediction of potability concentration in hypoxic river systems using support vector machine: a case study of WenRui Tang River, China,” Environ Sci Pollut Res, vol. 24, no. 19, pp. 16062–16076, Jul. 2017.
4. M. S. Jadhav, K. C. Khare, and A. S. Warke, “Water Quality Prediction of Gangapur Reservoir (India) Using LS-SVM and Genetic Programming,” Lakes & Reservoirs: Science, Policy and Management for Sustainable Use, vol. 20, no. 4, pp. 275–284.
5. Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, “Development of early-warning

protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea," Science of The Total Environment, vol. 502, pp. 31–41, Jan. 2015.

6. S. Liu, H. Tai, Q. Ding, D. Li, L. Xu, and Y. Wei, "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction," Mathematical and Computer Modelling, vol. 58, no.3, pp. 458–465, Aug. 2013.

7. A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the bio-chemical oxygen demand (BOD) of Surma River," Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237–243, 2017.

8. Y. Khan and C. S. See, "Predicting and analysing water quality using Machine Learning: a comprehensive model," in 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1–6, Farmingdale, NY, USA, April 2016.

**9.** J. Yan, Z. Xu, Y. Yu, H. Xu, and K. Gao, "Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing," Applied Sciences, vol. 9, no. 9, p. 1863, 2019.