COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 2, Jul-Dec 2023, pp. 98-113
DOI: 10.17492/computology.v3i2.2305
https://www.journalpressindia.com/cjacsit

Water Quality Monitoring on Streaming Data

Bhawnesh Kumar[1], Tinku Singh[2], Anuj Kumar[1] and Naveen Kumar[3]*

ABSTRACT

The increasing contamination of natural water bodies due to diverse human activities necessitates a comprehensive approach to monitoring water quality, especially considering its widespread use in daily life. This study addresses the escalating contamination of natural water bodies, emphasizing the need for a robust real-time water quality monitoring system. Focused on evaluating Triveni Sangam, Prayagraj, where the Ganga and Yamuna rivers converge, the study recognizes the crucial role of continuous monitoring in safeguarding precious water resources. To achieve this, a sophisticated framework has been proposed, leveraging a Spark server to simulate streaming data. This dynamic approach ensures uninterrupted and real-time assessment of water quality, crucial for the effective management of water resources. The system categorizes training data using the Water Quality Index (WQI) and employs Naive Bayes classification for real-time data, achieving an impressive accuracy of 82.21%. The results underscore the effectiveness of learning from streaming data, emphasizing its utility for monitoring water quality in real-time. This study contributes significantly to ongoing water resource management initiatives but also highlights the pivotal role of machine learning in addressing pressing environmental challenges.

Keywords: *Streaming Data; WQI; Real-Time Monitoring; Classification; Incremental Learning.*

1.0 Introduction

Water, a paramount natural resource and a vital national asset constitutes the

[1]*Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, Uttarakhand, India*
[2]*School of Information and Communication Engineering, Chungbuk National University, Cheongju-si, Chungcheongbuk-do, Republic of Korea*
[3]*School of Computing, DIT University, Dehradun, Uttarakhand, India*
**Corresponding author e-mail: tinku.singh@cbnu.ac.kr*

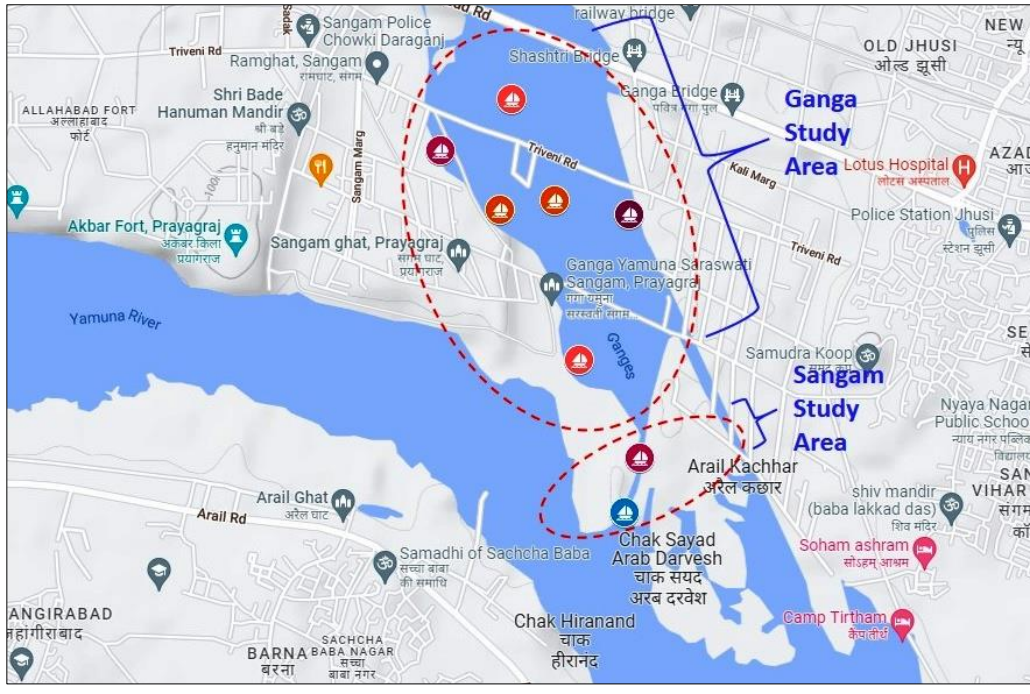
primary element of the ecosystem. The assessment of water quality in a specific area or source involves considering physical, chemical, and biological parameters. Utilizing a Water Quality Index (WQI) is one of the most effective approaches to characterize water quality [1]. The WQI condenses extensive information into a single value, offering a comprehensive representation of the water system's overall status by combining data from various sources. In the context of India, the Ganga and Yamuna rivers, revered as sacred waterways, play a crucial role in sustaining numerous communities in northern India. Unfortunately, both rivers face significant pollution challenges [2].

The Ganga River's water serves various purposes, including agriculture, domestic use, and industrial activities, necessitating a comprehensive assessment of its quality across different sectors [3, 4]. Our experimental focus was the Sangam region in Prayagraj, Uttar Pradesh, India, where the Ganga and Yamuna rivers converge. This specific location was chosen to enable a comparative analysis of the individual rivers. Moreover, cultural factors significantly contribute to the degradation of water quality. The site holds immense religious importance, with millions of people taking holy dips at the Sangam throughout the year.

The peak of tourist activity occurs during the Kumbh Mela, the world's largest gathering, which takes place every 12 years. In the 2013 Kumbh Mela, a staggering 120 million people took a dip at Sangam, with a single-day maximum count reaching 30 million [5]. Figure 1 illustrates the Google map of the Sangam region, with boat symbols denoting the survey points for the Ganga and Sangam during the investigations. WQI is a crucial tool for assessing water quality [7]. Particularly, the National Sanitation Foundation Water Quality Index (NSFWQI) stands out as a popular technique for categorizing surface water quality. Although definitions and parameters for computing WQI may vary, the primary purpose remains consistent: the quantification of water quality for objective analysis.

WQI essentially determines the optimal water usage for various purposes by considering multiple parameters such as temperature, total dissolved solids (TDS), pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), conductivity, fluoride, oxidation-reduction potential (ORP), mercury, cobalt, oil, and grease [8]. By synthesizing these diverse water quality parameters, the WQI generates a single value, offering a comprehensive assessment of the overall water quality at a specific location and time. WQI plays a crucial role in translating intricate water quality data into actionable information and gauging the suitability of water for different applications. The quality index typically ranges from 0 to 100, with lower scores indicating better water quality for various usages.

Figure 1: Study Area of Ganga River and Sangam Area [6]



Machine learning methods are widely embraced for their capacity to analyze data, uncover patterns, and predict outcomes, especially when dealing with extensive datasets collected from diverse scenarios. Water quality monitoring using machine learning involves the application of advanced algorithms and computational models to analyze and interpret data related to water characteristics. Machine learning algorithms excel in processing large datasets, making them valuable tools for assessing and predicting water quality. These models can analyze parameters such as pH, DO, turbidity, and pollutant concentrations, deriving meaningful insights and patterns that might be challenging for traditional methods. Machine learning enables the development of predictive models that can forecast changes in water quality over time, providing early warnings for potential issues. By leveraging machine learning in water quality monitoring, authorities and environmental agencies can make informed decisions, implement preventive measures, and contribute to the sustainable management of water resources.

This study aims to achieve real-time monitoring of water quality at Triveni Sangam, Prayagraj. Traditionally, water quality assessment relies on laboratory testing, which is a time-consuming process. To address this limitation, we propose a real-time approach that continuously updates information on water quality using reliable methods. The study focuses on monitoring five key parameters: pH, DO, conductivity, temperature, and ORP. IoT devices are employed to collect streaming data, which is then processed using a pretrained machine learning algorithm. The assessed water quality parameters are promptly displayed on a PC or mobile device in real time. The study extends our previous research on Quality Assessment and Monitoring of River Water Using IoT Infrastructure [9] by proposing a real-time water quality monitoring system.

2.0 Background and Context

2.1 Water Quality Index

A Water Quality Index (WQI) is a mathematical technique that provides briefs description of the overall water quality in a particular location. WQI is defined as a rating that reflects the composite influence of different water quality parameters [10].The purpose of a WQI is to simplify the complexity of water quality data and make it more understandable for the general public. The calculation of a Water Quality Index typically involves assessing multiple water quality parameters such as pH, DO, biochemical oxygen demand, nutrient levels, temperature, and the presence of pollutants. Each parameter is assigned a weight or importance factor based on its significance to water quality. The individual scores for each parameter are then aggregated, and the final WQI score is calculated. There are various methods [11] to calculate the water quality index based on the parameters are used shown in Table 1 :

Table 1: Water Quality Index [11]

a.	Weighted Average Water Quality Index (WAWQI)
b.	National Sanitation Foundation Water Quality Index (NSFWQI)
c.	Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI)
d.	Oregon Water Quality Index (OWQI)

2.2 Weighted Average Water Quality Index (WAWQI)

[11]:The Weighted Average Water Quality Index (WAWQI) is a specific type of Water Quality Index (WQI) that incorporates a weighted approach to reflect the importance of different water quality parameters such as pH,DO, conductivity,

temperature and ORP. The calculation of a WAWQI involves assigning weights to various water quality parameters based on their significance to overall water quality. These weights are then used to compute a weighted average, providing a single numerical value that represents the composite water quality at a particular location. The basic purpose of this method is to classified the water quality according to the water purity by using the most commonly measured water quality parameter. Here's a simplified formula for calculating a weighted average water quality index:

$$WAWQI = \frac{\sum_{i=1}^{n} W_i * Q_i}{\sum_{i=1}^{n} W_i} \quad \ldots(1)$$

where:

WAWQI is the Weighted Average Water Quality Index.

Wi is the weight assigned to the ith water quality parameter.

Qi is the quality rating scale of ith water quality parameter.

The value of *Qi* or each parameter is calculated using this expression:

$$Q_i = 100 * \left[\frac{V_i - V_0}{S_i - V_0}\right] \quad \ldots(2)$$

where,

Vi = estimated concentration of ith parameter in the analysed water.

V0 = ideal value of this parameter in pure water. *Si* = recommended standard value of ith parameter

For each water quality parameter, the value of unit Weight(Wi) is calculated by using the following formula:

$$W_i = \frac{K}{S_i} \quad \ldots(3)$$

where, K is proportionality constant and it can be calculated by this formula :

$$K = \frac{1}{\sum_{i=1}^{n}\left(\frac{1}{S_i}\right)} \quad \ldots(4)$$

The rating of water quality refers to the assessment and measurement of various parameters and characteristics in water to determine its suitability for specific purposes or to identify potential risks to human health and the environment.

Table 2: Water Quality Rating as per WAWQI[11]

WQI Value	Rating of Water Quality
0 − 25	Excellent water quality
26 − 50	Good water quality
21 − 75	Poor water quality
76 − 100	Very Poor water quality
Above 100	Unsuitable for drinking purpose

Water quality ratings are often expressed on a scale or in categories that reflect the level of contamination, pollutants, or other factors affecting the water. According to the Weighted Average Water Quality Index shown in Table 2, the Water quality rating may use different scales or categories, often ranging from excellent to poor. Water Quality Rating as per WAWQI is shown in the Table 2.

3.0 Related Work

Water quality monitoring is a critical aspect of environmental management, addressing the challenges of pollution and ensuring the sustainable use of water resources. Traditional methods of water quality assessment often involve extensive experimental requirements, making real-time monitoring a complex task. In recent years, researchers have explored innovative approaches, particularly leveraging machine learning techniques, to automate and enhance the accuracy of water quality monitoring. A framework employing a variety of sensors for real-time monitoring of water flow, conductivity, temperature, turbidity, pH, and more is proposed in [12]. The constant data feeds from the Internet of Things (IoT) devices contribute to effective flood predictions, allowing authorities to issue early warnings and minimize casualties during floods. This approach highlights the potential of machine learning in disaster management through continuous monitoring. [13] delves into various applications of AI algorithms for assessing water quality across different conditions, including surface water, groundwater, drinking water, sewage, and seawater. The study anticipates future implementations of AI approaches for water quality management, emphasizing the versatility of these techniques across diverse water environments.

A variety of classification techniques have been used for streaming data depending on the particular use case. In [14] ~~conducts~~ a thorough evaluation of artificial intelligence approaches—specifically, support vector machines (SVM), group method of data handling (GMDH), and artificial neural networks (ANN for forecasting water quality in the Tireh River, southwest Iran is discussed.. The study demonstrates the effectiveness of both ANN and SVM models in predicting various water quality aspects. A cost-effective water quality monitoring system leveraging cloud computing, machine learning, and the Internet of Things is proposed in [15]. This model not only offers an alternative to existing monitoring methods but also dynamically regulates water temperature based on ambient air temperature. Fitore's research [16] tackles time series challenges in water quality data using diverse models such as SVM, ANN, deep neural network (DNN), and more. The F-score metric is employed for performance evaluation,

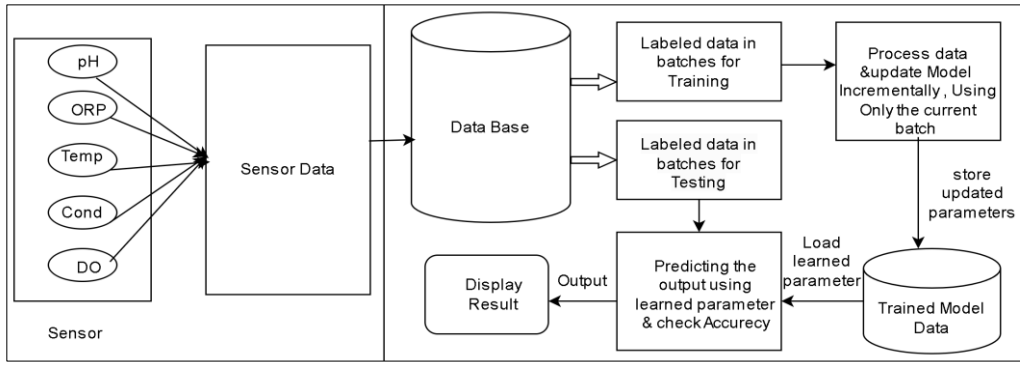
validated through a replication study. The other real-time forecasting and ML studies [17–19] show the importance and efficiency of ML methods in this domain.

Miller's work [20] introduces a novel method for predicting theWater Quality Index (WQI) using machine learning algorithms over a two-decade period in an urban lake. The study not only predicts WQI but also uncovers intricate relationships between various water-quality parameters. Machine learning, grounded in both statistical and computational principles, is showcased as a powerful tool for handling complex environmental datasets. The use of IoT devices can also be used for continuous water quality monitoring, as demonstrated in [9]. The study presents an IoT infrastructure-based river water quality monitoring and assessment system. The research uses sensor probes to monitor specific parameters like pH, DO, temperature, conductivity, and oxidation–reduction potential. The incorporation of machine learning techniques, such as principal component analysis and factor analysis, aids in feature selection and weight assignment for accurate water quality assessment. In conclusion, the integration of machine learning techniques into real-time water quality monitoring systems can offer promising results in terms of accuracy, continuous assessment, and early detection of environmental issues. This study focuses on the potential of these approaches in diverse water environments, paving the way for more effective and automated water quality management systems.

4.0 Methodology

Utilizing labeled data andWQI values, our model is trained on five parameters: pH, DO, conductivity, temperature, and ORP, as depicted in Figure 2.

Figure 2: Proposed Framework



This training enables the classification of water samples into specific classes. Subsequently, the model can classify samples from streaming data. Spark is employed for streaming-related processes, facilitating data processing, as discussed in the Spark streaming section. The initial ML model object is trained using the random parameter weights, while it is further updated incrementally on receiving the new data samples. Different ML models have been utilized to evaluate the performance of the proposed framework. A dedicated server simulates streaming data, continuously sending data at a predetermined interval. In our Spark application, we connect to the server using the socket API. The application listens continuously at the specified port, collecting data over a defined interval (e.g., 10 seconds) to update model parameters during the training phase and predict labels during testing. Each phase of the proposed framework is explained in detail in the subsequent sections.

4.1 Gaussian Naive Bayes (GNB)

Naive Bayes serves as a classification algorithm suitable for both two-class and multi-class classification problems. It simplifies the calculation of probabilities for each hypothesis, making computations manageable. The algorithm assumes independence among all attributes, leading to the calculation of their Bayes probability values through the product rule. For real-valued attributes following a Gaussian distribution, Naive Bayes can be applied. The Gaussian Naive Bayes (GNB) is a specific variant that adheres to the Gaussian Distribution Model. It requires the calculation of mean and standard deviation from the training data, providing an estimate of the distribution of the training data.

4.2 Incremental learning

In a contemporary system with vast data volumes, the adoption of machine learning algorithms capable of incremental learning becomes crucial. Incremental learning allows for parameter updates whenever new data, often in the form of streams, becomes available. In scenarios involving online parameter updates, it is neither practical nor efficient to have the entire dataset in memory and reapply the entire algorithm. Gaussian Naive Bayes, a simple yet effective system in the machine learning and statistics literature, provides a solution for incremental learning in classification. The ease with which conditional probability estimates are derived in Naive Bayes makes it a preferred choice for handling streaming data. GNB stands out as an incremental, online, or one-time version of the naive Bayes algorithm. This characteristic allows GNB to fully leverage previously trained classifiers, adding value to past efforts. The continuous learning process merely necessitates the presence of the new training set in memory,

ensuring a swift and efficient updating process [21]. For Incremental learning, the updated mean (μ) and variance (σ^2) of the data in the case of Gaussian Naive Bayes can be computed easily as follows:

Online Mean updation

$$\mu_{\text{updated}} = (n_{\text{new}} * \mu_{\text{new}} + n_{\text{past}} * \mu_{\text{past}}) / n_{\text{total}}$$

where, n_{new} is the size of incoming stream of data, μ_{new} is the mean of the incoming stream of data, n_{past} is the size of data on which our model is currently trained, μ_{past} is previous mean, and n_{total} can be represented as:

$$n_{\text{total}} = n_{\text{new}} + n_{\text{past}}$$

Online Variance updation

$$ssd_{\text{old}} = \text{var}_{\text{past}} * n_{\text{past}}$$

$$ssd_{\text{new}} = \text{var}_{\text{new}} * n_{\text{new}}$$

$$ssd_{\text{total}} = ssd_{\text{old}} + ssd_{\text{new}} + (n_{\text{old}} * n_{\text{new}} / n_{\text{total}}) * (\mu_{\text{past}} - \mu_{\text{new}})^2$$

$$var_{\text{updated}} = ssd_{\text{total}} / n_{\text{total}}$$

where, ssd_{old}, ssd_{new}, ssd_{total} represents the previous, new and total sum of squared differences respectively. $\text{var}_{\text{updated}}$ is the updated variance.

4.3 Real time stream handling

In this paper, Spark Streaming was employed due to its inherent capabilities in managing both streaming and batch data, aligning with the necessity for processing continuous sensor readings and predicting water quality. Apache Spark Streaming is a component of the broader Spark API, providing a suite of tools for extensive data processing. Recognized for its fault tolerance and seamless integration with diverse data sources, Spark Streaming is a widely adopted tool in the field. Following the application of necessary transformations and extraction of pertinent information from the data stream, the processed data can be directed to live websites, updated at specified intervals, or stored in databases as dictated by the application. The continuous data stream received through DStream is further elaborated in the next subsection.

4.3.1 Discretized streams

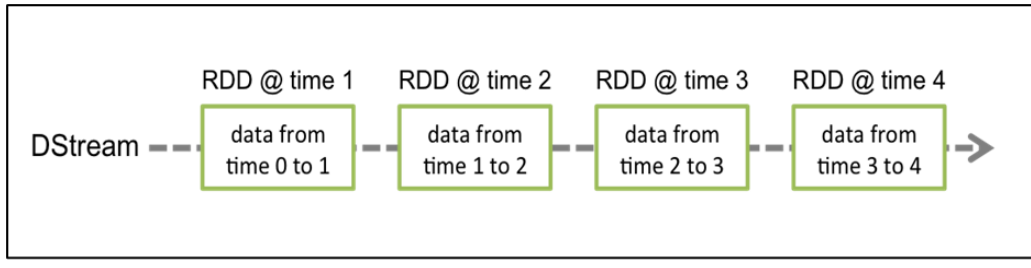
The foundational concept here is DStream, or Discretized Streams, depicted in Figure 3. This abstraction represents any continuous form of raw or processed data. It's crucial to understand that, despite its name, DStream is essentially an abstraction, encompassing a sequence of Resilient Distributed Datasets (RDDs), which are internal data structures in Spark. These RDDs, while immutable, can undergo various transformations to yield new RDDs. In the diagram, each RDD corresponds to data for a

specific time interval, such as time 0 to 1. Spark Streaming provides two categories of streaming sources:

i. Basic sources: These are default sources provided by the Spark Streaming context. Examples include socket connections, as utilized in our case.
ii. Advanced sources: Extra utility classes facilitate integration with sources like Kafka, Kinesis, etc.

In our specific application, we leverage Basic sources, creating a DStream that represents streaming data from a TCP source, with a specified hostname (e.g., localhost) and port (e.g., 9999). For each RDD in the datastream, we systematically process and extract data to incorporate the current batch into our model and update parameters accordingly.

Figure 3: Discretized Streams



5.0 Results and Discussion

This section outlines the outcomes of the streaming data classification experiments conducted using Google Colaboratory (GC), Apache Kafka, and an Apache Spark Cluster. The GC setup utilizes Python 3.7 and offers a single GPU cluster featuring an NVIDIA K80 GPU, 12 GB RAM, and a clock speed of 0.82 GHz. The deployed sensor network collected real-time data for pH, DO, conductivity, temperature, and ORP, storing the data in a database. Apache Kafka 2.13 was employed to stream the input data in real-time. The realtime streamed data frames were processed on the heterogeneous Apache Spark Cluster (ASC), consisting of one master and four worker nodes. All nodes ran on the Ubuntu 16.04 LTS operating system, Python 3.7.3, and Spark version 3.0.0. Among the worker nodes, two featured an Intel® Xeon(R) CPU E5-2630 v3 @ 2.40GHz processor with 32 cores and 32 GB RAM, while the other two had an Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz with 8 cores and 8 GB RAM. For end-to-end training, the GNB model underwent training with a batch size of 10 and testing

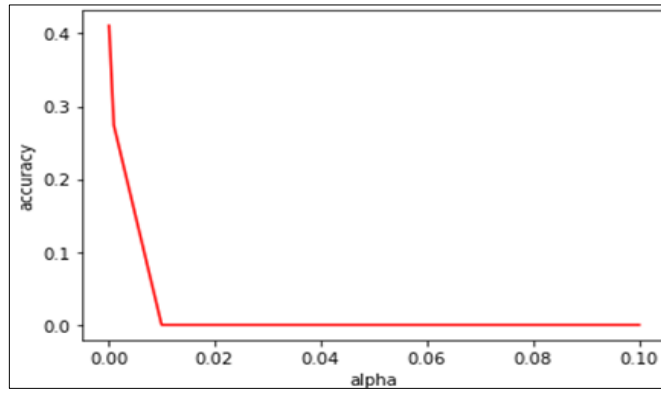
with a batch size of 7, utilizing the SGD classifier. The dataset was split into training and testing sets, containing 80% and 20% of the data, respectively.

Table 3: Alpha vs Accuracy Evaluation using SGD

Alpha	Accuracy
0.1	0
0.01	0
0.001	0.273
0.0001	0.305
0.00001	0.4107

In the case of SGD, the training model utilized two customized parameters: alpha, representing the weight of regularization, and the loss function, which assesses the model's performance. Additionally, several hyper parameters were set to their default values. The regularization parameter, denoted as : "alpha," governs the strength of regularization. A higher alpha value enforces stronger regularization, potentially resulting in a simpler model with smaller coefficients. The model underwent testing with varying alpha values, and the highest accuracy was achieved at alpha = 0.00001, as indicated in Table 3, presenting the correlation between Alpha and accuracy for the SGD experiments. Alpha serves as the weight determining the extent of regularization. Initially set at 0.1, it undergoes a tenfold reduction in each iteration. The accuracy plotted against alpha is illustrated in Figure 4.

Figure 4: Alpha vs Accuracy Graph using SGD



The second parameter utilized in the SGD classifier is the loss function, which evaluates the model's performance by quantifying the difference between the expected and actual outputs generated by the model. SGD enhances model performance by iteratively adjusting its parameters to minimize the loss function. Three loss functions—Hinge Loss, Log Loss, and Mean Squared Loss—were applied in this model. After computation, the Mean Squared Loss function yielded the highest accuracy of 0.4107 among the three functions, as depicted in Table 4.

Table 4: Loss Function vs Accuracy using SGD

Function	Accuracy
Hinge Loss	0.273
Log Loss	0.19
Mean Squared Loss	0.4107

Moreover, the GNB classifier is a pre-trained model employed for incremental learning. The dataset is split into two sets: 80% for the training phase and 20% for the testing phase. During the training phase, the model is trained for 3 epochs with a batch size of 10. The sample training data, which includes information on five parameters, is illustrated in Table 5. Each row in the table represents one set of sensor readings. The obtained batch is utilized to update the model using online mean and variance updating for GNB, as discussed earlier. Finally, when the streaming process is completed, the model is saved as a pickle file, named gnb model.pkl.

In the subsequent testing phase, the first step involves loading the saved model (gnb model.pkl). For each set of labeled data obtained, we predict the class using our model with a batch size of 7, as illustrated in Table 6.

Table 5: Sample Training Data for GNB

S. No.	DO	pH	ORP	Cond	TEMP	Label
1.	12.43	13.21	0.15	54.68	18.19	very Poor
2.	10.83	13.28	0.15	54.65	18.34	very Poor
3.	9.24	13.29	0.14	54.44	18.24	very Poor
4.	8.78	13.27	0.14	54.57	18.45	very Poor
5.	8.61	13.33	0.14	54.10	18.23	very Poor

Table 6: Sample Testing Data for GNB

S. No.	DO	pH	ORP	Cond	TEMP	Label	Predicted
1.	9.43	8.19	0.11	406.45	18.16	Excellent	very Poor
2.	9.43	8.18	0.11	406.34	18.16	Excellent	very Poor
3.	9.43	8.16	0.11	406.32	18.16	Excellent	very Poor
4.	9.43	8.19	0.11	406.43	18.16	Excellent	very Poor
5.	9.43	8.17	0.11	406.33	18.16	Excellent	very Poor

Upon completion of the streaming process, the accuracy is computed by dividing the number of correct predictions by the total number of data points, resulting in an accuracy of 82.21%.

Table 7: Comparison of Classifier's Accuracy

Classifier	Hyperparameters	Accuracy
SGD	Alpha: 0.00001 Loss function: squared loss	41.07%
Gaussian Naïve Bayes	None	82.21%

In the discussion of both SGD and GNB classifiers within the context of this research, the main issue centered around accuracy, as analyzed in subsection 6.2 above. Upon evaluating the results of both classifiers, it becomes evident that the GNB classifier outperforms SGD, as detailed in Table 7. The accuracy of SGD is merely 41.07%, which is significantly lower compared to the GNB classifier with an accuracy of 82.21%. Therefore, the GNB classifier stands out as the more suitable model for the classification of water quality parameters such as pH, DO, Conductivity, ORP, and Temp.

6.0 Conclusion and Future work

This study affirms the superiority of online learning algorithms, particularly when dealing with the escalating volume of data. Leveraging Apache Spark Streaming, we efficiently handled the continuous influx of data, demonstrating its efficacy in real-time applications. Two incremental learning methods, GNB and SGD, were implemented for water quality monitoring. The results revealed a substantial difference in accuracy, with Naive Bayes showcasing superior performance at 82.21%, compared to

SGD's 41.07%. The research emphasizes the critical role of continuous water quality monitoring, showcasing its applicability for real-time scenarios, such as the monitoring of the Ganga River in the Sangam area. The prompt identification of deteriorating water quality enables timely interventions by authorities, safeguarding aquatic life from prolonged exposure to poor water conditions. The findings underscore the potential of machine learning models, advanced streaming technologies, and continuous monitoring methodologies, advocating for their widespread adoption in comprehensive water quality monitoring initiatives. Future work should focus on enhancing the accuracy of online learning models, exploring additional machine learning algorithms, and extending the study to diverse geographical regions for a comprehensive water quality monitoring framework. Additionally, integrating more advanced sensor technologies and expanding the dataset would contribute to refining the predictive capabilities of the models.

References

[1] Verma, R., Ahuja, L. & Khatri, S. K. (2018). Water quality index using iot. I*n 2018 International Conference on Inventive Research in Computing Applications*, pp. 149–153. Retrieved from https://doi.org/10.1109/ICIRCA.2018.8597357

[2] Chaudhary, M. & Walker, T. R. (2019). River ganga pollution: Causes and failed management plans (correspondence on dwivedi et al. 2018. ganga water pollution: A potential health threat to inhabitants of ganga basin. Environment International, 117, 327–338). *Environment International*, *126*, 202–206.

[3] Haritash, A., Gaur, S., Garg, S. (2016). Assessment of water quality and suitability analysis of river ganga in rishikesh, india. *Applied Water Science*, *6*(4), 383–392.

[4] Seeboonruang, U. (2012). A statistical assessment of the impact of land uses on surface water quality indexes. *Journal of Environmental Management, 101*, 134–142.

[5] Pandey, A., Pandey, M., Singh, N. & Trivedi, A. (2020). Kumbh mela: A case study for dense crowd counting and modeling. *Multimedia Tools and Applications, 79*, 17837-17858.

[6] Kumar, M., Singh, T., Maurya, M. K., Shivhare, A., Raut, A. & Singh, P. K. (2023). Quality assessment and monitoring of river water using iot infrastructure. *IEEE Internet of Things Journal*, *10*(12), 10280–10290. Retrieved from https://doi.org/10.1109/JIOT.2023.3238123

[7] Hoseinzadeh, E., Khorsandi, H., Wei, C. & Alipour, M. (2015). Evaluation of aydughmush river water quality using the national sanitation foundation water

quality index (nsfwqi), river pollution index (rpi), and forestry water quality index (fwqi). *Desalination and Water Treatment, 54*(11), 2994–3002.

[8] Sener, S., Sener, E. & Davraz, A. (2017). Evaluation of water quality using water quality index (wqi) method and gis in aksu river (sw-turkey). *Science of the Total Environment, 584*, 131–144.

[9] Kumar, M., Singh, T., Maurya, M. K., Shivhare, A., Raut, A. & Singh, P. K. (2023). Quality assessment and monitoring of river water using iot infrastructure. *IEEE Internet of Things Journal, 24*(2), 494.

[10] Sehnaz, S., Erhan, S. & Aysen, D. (2017). Evaluation of water quality using water quality index (wqi) method and gis in aksu river (sw-turkey). *Science of the Total Environment*, 131-144. Retrieved from https://doi.org/10.1016/j.scitotenv.2017.01.102

[11] Shweta, T., Bhavtosh, S., Prashant, S. & Rajendra, D. (2013). Water quality assessment in terms of water quality index. *American Journal of Water Resources, 1*(3), 34–38. Retrieved from https://doi.org/10.12691/ajwr-1-3-3

[12] Khan, M.A., Hoque, M.A., & Ahmed, S. (2021). Iot-based system for real-time water pollution monitoring of rivers. *In 2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp. 1–5. Retrieved from https://doi.org/10.1109/ICECIT54077.2021.9641483

[13] Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment Health, 1*(2), 107–116. Retrieved from https://doi.org/10.1016/j.eehl.2022.06.001

[14] Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal, 53*(1), 3–13.

[15] Koditala, N. K. & Pandey, P. S. (2018). Water quality monitoring system using iot and machine learning. *In 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pp. 1–5. Retrieved from https://doi.org/10.1109/RICE.2018.8509050

[16] Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set*. *Journal of Information and Telecommunication, 3*(3), 294–307. Retrieved from https://doi.org/10.1080/24751839.2019.1565653

[17] Singh, T., Kalra, R., Mishra, S., & Kumar, S. M. (2023). An efficient realtime stock prediction exploiting incremental learning and deep learning. *Evolving Systems, 14*(6), 919–937.

[18] Singh, T., Sharma, N., & Kumar, S. M. (2023). Analysis and forecasting of air quality index based on satellite data. *Inhalation Toxicology, 35*(1-2), 24–39.

[19] Singh, T., Rajput, V., & Prasad, S. U., & Kumar, M. (2023). Real-time traffic light violations using distributed streaming. *The Journal of Supercomputing, 79*(7), 7533–7559.

[20] Miller, T., Durlik, I., Adrianna, K., Kisiel, A., Cembrowska-Lech, D., Spychalski, I. & Tunski, T. (2023). Predictive modeling of urban lake water quality using machine learning: A 20-year study. *Applied Sciences, 13*(20). Retrieved from https://doi.org/10.3390/app132011217

[21] Losing, V., Hammer, B. & Wersing, H. (2016). Choosing the best algorithm for an incremental on-line learning task. *European Symposium on Artificial Neural Networks*. Retrieved from https://api.semanticscholar.org/CorpusID:15821623