frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Original Research
PUBLISHED 25 November 2024
DOI 10.3389/fbioe.2024.1508964

Check for updates

OPEN ACCESS

EDITED BY
Nar Singh Chauhan,
Maharshi Dayanand University, India

REVIEWED BY
Sanjay Kumar Singh Patel,
Hemwati Nandan Bahuguna Garhwal University, India
Asiya Nazir,
Abu Dhabi University, United Arab Emirates

*CORRESPONDENCE
Pankaj Bharali,
pankajbharali98@gmail.com,
pbharali@neist.res.in
Hridoy Jyoti Mahanta,
hridoy69@gmail.com,
hridoy@neist.res.in

RECEIVED 10 October 2024
ACCEPTED 12 November 2024
PUBLISHED 25 November 2024

CITATION
Gogoi G, Singh SD, Koch D, Kalyan E, Boro RR, Devi A, Mahanta HJ and Bharali P (2024) Leveraging environmental microbial indicators in wastewater for data-driven disease diagnostics.
Front. Bioeng. Biotechnol. 12:1508964.
doi: 10.3389/fbioe.2024.1508964

COPYRIGHT
© 2024 Gogoi, Singh, Koch, Kalyan, Boro, Devi, Mahanta and Bharali. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Leveraging environmental microbial indicators in wastewater for data-driven disease diagnostics

Gayatri Gogoi[1,2], Sarangthem Dinamani Singh[1,2], Devpratim Koch[1,2], Emon Kalyan[1], Rashmi Rani Boro[1], Aradhana Devi[3], Hridoy Jyoti Mahanta[2,4]* and Pankaj Bharali[1,2]*

[1]Centre for Infectious Diseases, Biological Sciences and Technology Division, CSIR-North East Institute of Science and Technology, Jorhat, India, [2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India, [3]Materials Sciences and Technology Division, CSIR-North East Institute of Science and Technology, Jorhat, India, [4]Advanced Computation and Data Sciences Division, CSIR-North East Institute of Science and Technology, Jorhat, India

Introduction: Wastewater-based surveillance (WBS) is an emerging tool for monitoring the spread of infectious diseases, such as SARS-CoV-2, in community settings. Environmental factors, including water quality parameters and seasonal variations, may influence the prevalence of viral particles in wastewater. This study aims to explore the relationships between these factors and the incidence of SARS-CoV-2 across 28 monitoring sites, spanning different seasons and water strata.

Methods: Samples were collected from 28 sites, accounting for seasonal and spatial (surface and intermediate water layers) variations. Key physicochemical parameters, heavy metals, and minerals were measured, and viral presence was detected using RT-qPCR. After data preprocessing, correlation analyses identified 19 relevant environmental parameters. Unsupervised learning algorithms, including K-means and K-medoid clustering, were employed to categorize the data into four distinct clusters, revealing patterns of viral positivity and environmental conditions.

Results: Cluster analysis indicated that seasonal variations and water quality characteristics significantly influenced SARS-CoV-2 positivity rates. The four clusters demonstrated distinct associations between environmental factors and viral prevalence, with certain clusters correlating with higher viral loads in specific seasons. The clustering patterns varied across sample sites, reflecting the diverse environmental conditions and their influence on viral detection.

Discussion: The findings underscore the critical role of environmental factors, such as water quality and seasonality, in shaping the dynamics of SARS-CoV-2 prevalence in wastewater. These insights provide a deeper understanding of the complex interplay between environmental contexts and disease spread. By

utilizing WBS and advanced data analysis techniques, this study offers a robust framework for future research aimed at enhancing public health surveillance and interventions.

KEYWORDS

SARS-CoV-2, wastewater-based surveillance (WBS), environmental factors, machine learning (ML), public health

1 Introduction

Wastewater-based surveillance has emerged as a powerful tool for monitoring public health and environmental contamination, playing a pivotal role in the early detection and management of various waterborne diseases and pollutants. This approach, rooted in the analysis of physicochemical parameters within wastewater, has gained importance due to its ability to provide real-time, cost-effective, and community-wide insights into the presence of contaminants. In an era, characterized by the continuous generation of massive amounts of data, harnessing the potential of data-driven approaches has become imperative in wastewater surveillance to enhance its accuracy and efficiency (Wigginton et al., 2015; Mathew and Kanmani, 2020; Srikanth et al., 2019).

The World Health Organization (WHO) emphasizes the importance of effective wastewater management and surveillance in preventing waterborne diseases, such as cholera and typhoid, which continue to pose a significant threat to global public health (World Health Organization, 2019). Traditionally, wastewater surveillance relied on periodic sampling and laboratory testing of water samples, a time-consuming and resource-intensive process. However, recent advancements in sensor technologies and data analysis methods have revolutionized the field by enabling continuous monitoring and analysis of physicochemical parameters in real-time (Newhart et al., 2019).

Physicochemical parameters, including pH, dissolved oxygen, turbidity, and the concentration of specific chemicals, provide critical information about the quality of wastewater. These parameters serve as indicators of potential contamination and can help in the early detection of pollutants, pathogens, and emerging chemical constituents, such as pharmaceuticals and microplastics (Arora et al., 2021).By continuously collecting and analyzing data from various wastewater treatment plants, data-driven approaches can effectively detect abnormal patterns and deviations, thus alerting authorities to potential issues before they escalate into public health crises (Pan et al., 2022; Levin et al., 2024; Moretti et al., 2024). Heavy metal content in wastewater plays a pivotal role in wastewater surveillance by providing essential insights into the overall water quality and potential contamination risks. Monitoring heavy metal concentrations, including elements like mercury (Hg), cadmium (Cd), lead (Pb), selenium (Se), and arsenic (As), contributes to a comprehensive assessment of environmental health and safety of that particular region (Zeiner et al., 2007). These heavy metals are considered priority pollutants due to their toxicity and persistence in aquatic ecosystems (Tchounwou et al., 2012).

The integration of data-driven approaches in wastewater surveillance leverages the power of machine learning, artificial intelligence, and big data analytics. These technologies enable the development of predictive models that can forecast contamination events, optimize treatment processes, and guide effective policy decisions. Such models learn from historical data, adapt to changing conditions, and provide actionable insights that empower decision-makers to respond proactively to emerging challenges in wastewater management (Moretti et al., 2024; Sahu et al., 2023; Van der Werf et al., 2023). High mutation rates lead to analytical limitations, requiring frequent updates to primers and probes used in RT-PCR assays. Moreover, wastewater samples contain complex microbial communities that may hinder variant identification accuracy. Other challenges include the need for improved data resolution to differentiate among closely related variants, which is vital for effective public health responses and anticipating variant-driven case surges (Thakur et al., 2022; Gogoi et al., 2024).

Machine learning (ML) and Deep Learning (DL) models have been employed for time-series predictions and track COVID-19 outbreaks in multiple communities as well as pre-screening tool for the identification of differences among the variant composition of different wastewater samples (Ai et al., 2022; Férez et al., 2023). Utilizing unsupervised ML algorithms, a quantifiable model for characterizing peaks and gaps in multiple waves of COVID-19 across 120 countries not only reveals the complexity in predicting growth or decline rates within each wave, but also identifying common features among the clusters, offering potential insights into anticipating future developments (Mahanta and Narahari Sastry, 2022).

Extensive seroepidemiological and genomic investigations for SARS-CoV-2 have been conducted across India, encompassing smaller regions like Jorhat district of Assam in the north east India. These studies provided evidence of a significant number of positive cases in the area that are linked to common Omicron and Delta variations (Naushin et al., 2021; Wahengbam et al., 2023). These studies have delved into the dynamics of COVID-19 progression within the Indian population, employing transcriptomic data analysis. Further studies also reveal serological surveys in the north eastern region of India which involves ML approach to discern infection statuses among Covaxin recipients (Kshattry et al., 2022; Singh et al., 2022). These studies show the potential of data-driven ML approaches for deciphering complex questions in diverse epidemiological studies.

This study aims to investigate how data-driven approaches can augment wastewater surveillance by correlating physicochemical parameters and heavy metals content with incidences of viral loads. This correlation study will indicate the early detection of the new pandemic for a particular region. It seeks to improve the early detection of contaminants and abnormal patterns, thereby strengthening public health and environmental safety.

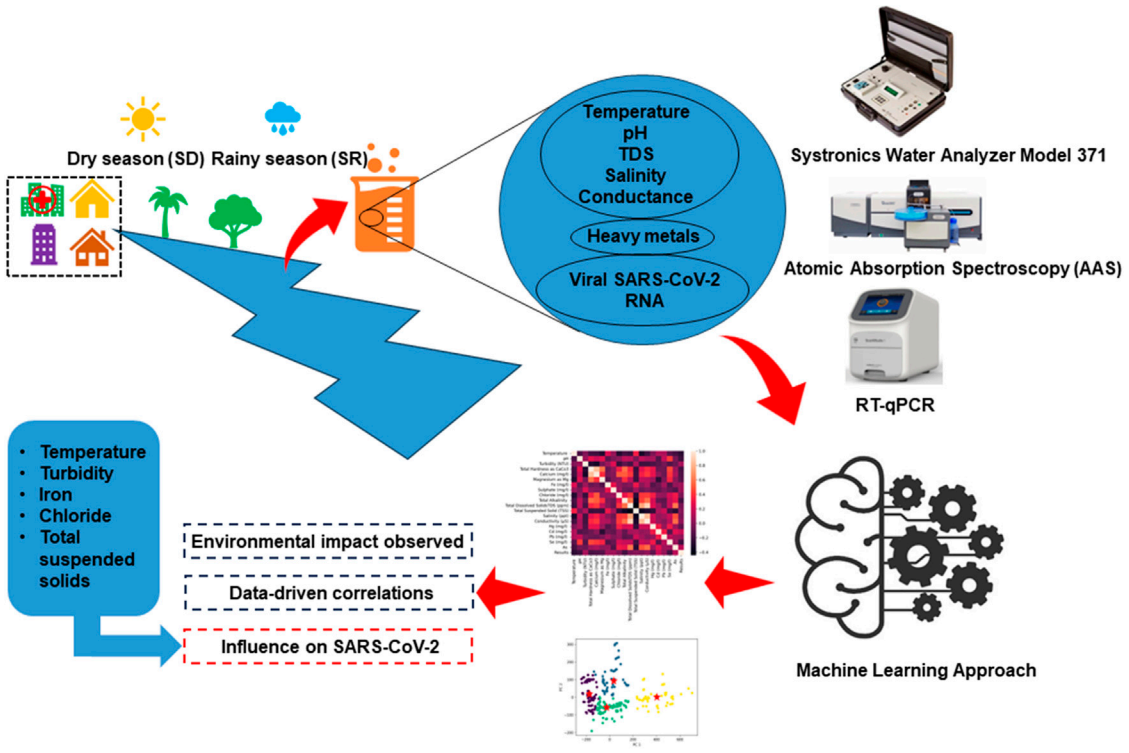

FIGURE 1
Overview of the study methodology, detailing the process from sample collection, through various stages of data processing, to final outcome representation. The diagram highlights key steps, including environmental sampling, laboratory analysis, data preprocessing, and clustering analysis for interpreting the results.

2 Materials and methods

2.1 Wastewater sample collection

The North-East region of India is a captivating region consisting of eight states, each brimming with unique cultural and geographical richness and shares international borders with Bhutan, China, Myanmar, and Bangladesh. Assam is the second largest state out of these eight states and share borders with all others. The current study was conducted in Jorhat district of Assam which has diverse indigenous communities, residing within low-resource and low-income settings. These communities heavily rely on medical facilities situated within their respective localities. These facilities serve as lifelines, providing essential healthcare support to those people. Wastewater samples have been collected from 28 different sites stretched across Jorhat district, representing three settings: healthcare, residential, and river bodies. Sampling was carried out during both rainy and dry seasons, spanning from September 2022 to March 2023. Two water layers were considered for sampling, the surface layer and the intermediate layer (30 cm depth) which resulted into four spatiotemporal conditions: surface layer during the rainy season (SR), intermediate layer during the rainy season (IR), surface layer during the dry season (SD), and intermediate layer during the dry season (ID). The samples were manually collected in Polypropylene bottles (PP). Sampling took place during morning hours to capture the viral peak load, and no rainfall was reported within the 24 h prior to collection. Additionally, we measured various physicochemical parameters, including temperature, pH, total dissolved solids (TDS), salinity, and conductance, using a Systronics Water Analyzer Model 371 during the sampling process (Férez et al., 2023; Yadav and Chauhan, 2023) and the analysis of heavy metals was carried out by using Atomic Absorption Spectroscopy (AAS) manufactured by Analytikjena model Zeenit 700p. Sample preparation involves careful collection, labelling, and homogenization of environmental samples, with subsequent digestion using appropriate acids. Certified reference standards are employed to create a calibration curve, establishing the relationship between absorbance and known heavy metal concentrations. Instrument setup encompasses optimizing AAS parameters, including lamp current, wavelength, and slit width, while regular alignment checks ensure instrument accuracy and standard limit of quantification for the selected element is upto 1,000 mg/L. In the measurement procedure, the sample is aspirated into the AAS instrument, and the absorbance is recorded at the specific wavelength for each heavy metal of interest. We have selected mercury (Hg), lead (Pb), cadmium (Cd), selenium (Se), and arsenic (As) for the present study (Zeiner et al., 2007). The complete methodology, from sample collection through processing to outcome representation, is illustrated in Figure 1.

2.2 Wastewater sample processing

The wastewater samples were processed in a BSL2 environment with strict adherence to personal protective equipment (PPE) protocols to ensure the workers safety (Spurbeck et al., 2021).

The sample processing steps consisted of Sample Homogenization and Pasteurization, Sample Filtration, and Sample Concentration using PEG-NaCl Method followed by viral extraction (Supplementary Figure S1). These meticulous steps ensured the proper processing and concentration of wastewater samples, while stringent safety measures were in place throughout the procedure (Smyth et al., 2022).

2.3 Viral RNA extraction and reverse transcription polymerase quantitative polymerase chain reaction (RT-qPCR)

The viral RNA extraction from wastewater samples was carried out using the QIAamp® Viral RNA Mini Kit (250) from Qiagen, Germany. The extraction process adhered strictly to the manufacturers' instructions, with a focus on obtaining SARS-CoV-2 viral nucleic acid from the PEG pellets post-virus concentration. In the 40 mL protocol, the RNA was eluted to a final volume of 40 µL and stored at −20°C when immediate processing was not possible, although every effort was made to process the samples on the same day.

Subsequently, the extracted RNA samples underwent analysis through reverse transcription-quantitative polymerase chain reaction (RT-qPCR) conducted on the QuantStudio™ 5 real-time PCR system by Applied Biosystems™ Inc., United States. The RT-qPCR assay targeted specific genes, namely, ORF-1ab and N genes, in a confirmatory test using the CoviPath™ COVID-19 RT-qPCR Kit from Applied Biosystems™ Inc., United States. Each 25 µL PCR reaction mixture consisted of 10 µL of the RNA extract, 6.25 µL of CoviPath™ COVID-19 Assay Multiplex, 1.25 µL CoviPath™ 1 Step Multiplex Master Mix (No ROX™), and the volume was adjusted to 25 µL using molecular grade water supplied by Sisco Research Laboratories Pvt. Ltd. Negative controls utilized ultrapure nuclease-free water, and for positive controls, CoviPath™ COVID-19 was employed, following the manufacturer's dilution guidelines.

The RT-qPCR reactions were carried out with an initial step at 53°C for 10 min, followed by 95°C for 2 min, and then cycled 40 times at 95°C for 3 s and 60°C for 30 s in the QuantStudio™ 5 real-time PCR system. Notably, the interpretation of SARS-CoV-2 positivity in this study underwent a revision, aligning with the CoviPath protocol. A cycle threshold (Ct) value of 35 was adopted as the criterion for positivity, ensuring a standardized and reliable approach for the detection and characterization of SARS-CoV-2 variants. This modification facilitates downstream Whole Genome Sequencing (WGS) analysis, making the study a more consistent and robust method for SARS-CoV-2 variant identification (Spurbeck et al., 2021; Ravi et al., 2024).

2.4 Quantitative analysis through machine learning

2.4.1 Correlation analysis of features

The water samples analysis and RT-qPCR tests generated 27 parameters out of which 24 parameters were distinct features as independent parameters and 3 features as dependent variables (ORF 1 ab, N gene, and RNaseP) in building the dataset for training the ML models. During initial preprocessing it was found that 5 out of the 24 features suffered from missing value and noisy value problem. Due to the complexity of the problem as well as size of data, imputation methods were refrained from being used and all these 5 features were not considered further. The remaining 19 features were then subjected to find their correlation with the three dependent variables and the final RT-qPCR outcome using Sparman's rank correlation analysis. This correlation will help to find the monotonic relationship between each feature with the dependent variables as well as with one another. This is a prominent approach of feature engineering which helps to identify the most relevant features for modelling. The value ranges from −1 to 1, where −1 and 1 resembles a perfectly negative and perfectly positive correlation respectively with 0 indicating no linear correlation.

2.4.2 Partitional clustering of sample sites

Clustering is a fundamental unsupervised machine learning technique used to discover hidden structures or patterns within a dataset. It involves grouping similar data points into clusters, where the members within a cluster are more similar to each other than to those outside the clusters. This is an iterative process which continues to form a new set of clusters till the optimum result is obtained. In the current study, K-means has been employed for interpretation of the underlying data. The K-means clustering method groups data points into clusters based on their similarity. It partitions data into K clusters, aiming to minimize within-cluster sum of squares. The algorithm starts with random cluster centroids and iteratively assigns data points to the nearest centroid, then updates the centroids based on the assigned points, repeating until convergence. During the assignment, each point is assigned to the nearest cluster mean with the least Euclidean distance such that,

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \forall_j, 1 \le j \le k \right\} \tag{1}$$

Where the point *xp* is assigned to exactly one *S(t)*.

The centroids are then recalculated and the observations are reassigned to new clusters. This process is repeated till the algorithm converges and reaches an optimal state.

$$m_i^{(t)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2}$$

K-means is widely used for data segmentation, pattern recognition, and feature engineering in various fields (Ahmed et al., 2021; Pearson, 1895).

2.4.3 Elbow method for optimal number of clusters

The Elbow Method is a widely used heuristic method for determining the optimal number of clusters (K) in a K-means clustering analysis. It involves performing the K-means clustering algorithm on the dataset for a range of K values and evaluating within-cluster sum of squares (WCSS) for each K. The WCSS is a measure of the total variance within the clusters. The point at which the reduction in WCSS starts to slow down, forming an "elbow" in the plot, is considered the optimal K value (Arthur and Vassilvitskii, 2007). WCSS is computed by,

$$WCSS = \sum_{i=1}^{K} \sum_{x \epsilon C_i} \|x - c_i\|^2 \quad (3)$$

Where K is the number of clusters, C_i represent the i^{th} with *ci* as the centroid of the i^{th} cluster and *x* is any data point. As the number of clusters increases, the WCSS will generally decrease, as each data point will be closer to its cluster centroid. However, at a certain point, the WCSS will start to plateau, as increasing the number of clusters will no longer significantly reduce the distance between data points and their cluster centroids. The elbow in the WCSS plot indicates this point, and the optimal number of clusters is the value of K at the elbow (Rdusseeun and Kaufman, 1987).

2.4.4 Performance metrics

A number of cluster validation metrics have been used in this study to evaluate the performance and quality of the clusters generated by the two clustering algorithms, K-means and K-medoid clustering.

2.4.4.1 Silhouette score

This is the measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Its value ranges from −1 to 1, where higher values indicate better-defined clusters.

$$S(i) = \frac{(b(i) - a(i))}{max(b(i), a(i))} \quad (4)$$

Where, *S(i)* is the silhouette score for data point *i*, *a(i)* is the average distance between i^{th} data point and other data points in the same cluster, and *b(i)* is the minimum average distance, minimised across clusters, between *ith* data point and the data points in a different cluster.

2.4.4.2 Davies-Bouldin index (DBI)

DBI estimates the average similarity between each cluster and the most comparable one. Lower numbers imply better grouping.

$$DBI(C) = \frac{1}{K} \sum_{i=1}^{K} \underset{i \neq j}{max} \frac{\Delta(C_i) + \Delta(C_j)}{\delta C_i, C_j} \quad (5)$$

Where the intra-cluster distance is represented by $\Delta(Ci)$ and the inter-cluster distance by δ (*Ci*, *Cj*).

2.4.4.3 Dunn index (DI)

DI is the measurement of the compactness of clusters and the separation between them. Higher values are better, indicating better-defined clusters.

$$DI = \frac{min(\delta(Ci, Cj))}{max(\Delta(Ci))} \quad (6)$$

Where, δ (*Ci, Cj*) represents the minimum intra-cluster distance between clusters *Ci* and *Cj*, while $\Delta(Ci)$ is the diameter of cluster *Ci*. Better clustering is indicated by a higher DI, which denotes tightly packed, well-separated clusters.

2.4.4.4 Inertia

Inertia measures the overall compactness of clusters through total squared distance between each data point and the cluster centre. Tighter and better-defined clusters are indicated by lower inertia levels and *vice versa*. Inertia is one of the key ideas in evaluating the quality of clustering solutions which is frequently used in conjunction with techniques like the Elbow Method to establish the ideal number of clusters (Hastie et al., 2009; Xu and Tian, 2015).

$$inertia = argmin \sum_{i=1}^{k} \sum_{x \epsilon S_i} \|x - \mu_i\|^2 \quad (7)$$

Where *S* is a set of observations with *x* as a data point and μ as the mean. Any distance metric, including the cosine, Manhattan, and Euclidean distances, may be used to compute the inter-cluster distance. Usually, the greatest distance between any two locations in the cluster is used to compute the intra-cluster distance.

These metrics assess the compactness, separation, and general cohesiveness of the clustering findings and allow to statistically assess the suitability and efficacy of the clustering techniques while making it easier to choose the best algorithm for the particular dataset. As the ground truth was not known, these metrics would help in converging that the clusters formed were correctly formed.

3 Results

3.1 Sample collection and analysis

For this study, 448 wastewater samples were collected from 28 different sites within Jorhat district of Assam, India (Supplementary Figure S2) during different season (Supplementary Table S1). Among these samples, 144 were reported to be positive for SARS-CoV-2 with an illustrative portrayal of the distribution of SARS-CoV-2 ORF 1ab, N, and RNAseP genes, alongside the corresponding results indicating positivity and negativity in wastewater samples (Supplementary Figure S3). The analysis covered a range of physicochemical parameters, including temperature, pH, turbidity (NTU), total hardness as $CaCO_3$, calcium (mg/L), magnesium as Mg, iron (Fe) content (mg/L), sulphate (mg/L), chloride (mg/L), total alkalinity, total dissolved solids (TDS) in ppm, total suspended solids (TSS), salinity (ppt), and conductivity (μS). Additionally, the study examined the presence of heavy metals such as Hg (mercury), Cd (cadmium), Pb (lead), Se (selenium), and As (arsenic) in the samples, providing a comprehensive overview of the environmental conditions and the presence of SARS-CoV-2 in the region.

3.2 Correlation analysis

The independent features considered in this study were subjected to Spearman's rank correlation with the dependent parameters, i.e., N gene, ORF 1 ab, RNaseP and the RT-qPCR result (positive and negative). In correlation with N gene, features such total hardness, Magnesium content and total alkalinity had high positive correlation (≈0.22) compared to temperature, pH, calcium, and Cadmium (≈0.15). Whereas, turbidity, Iron and lead showed a negative correlation (≈− 0.14) with N gene. For the ORF 1ab gene, temperature displayed higher positive correlation

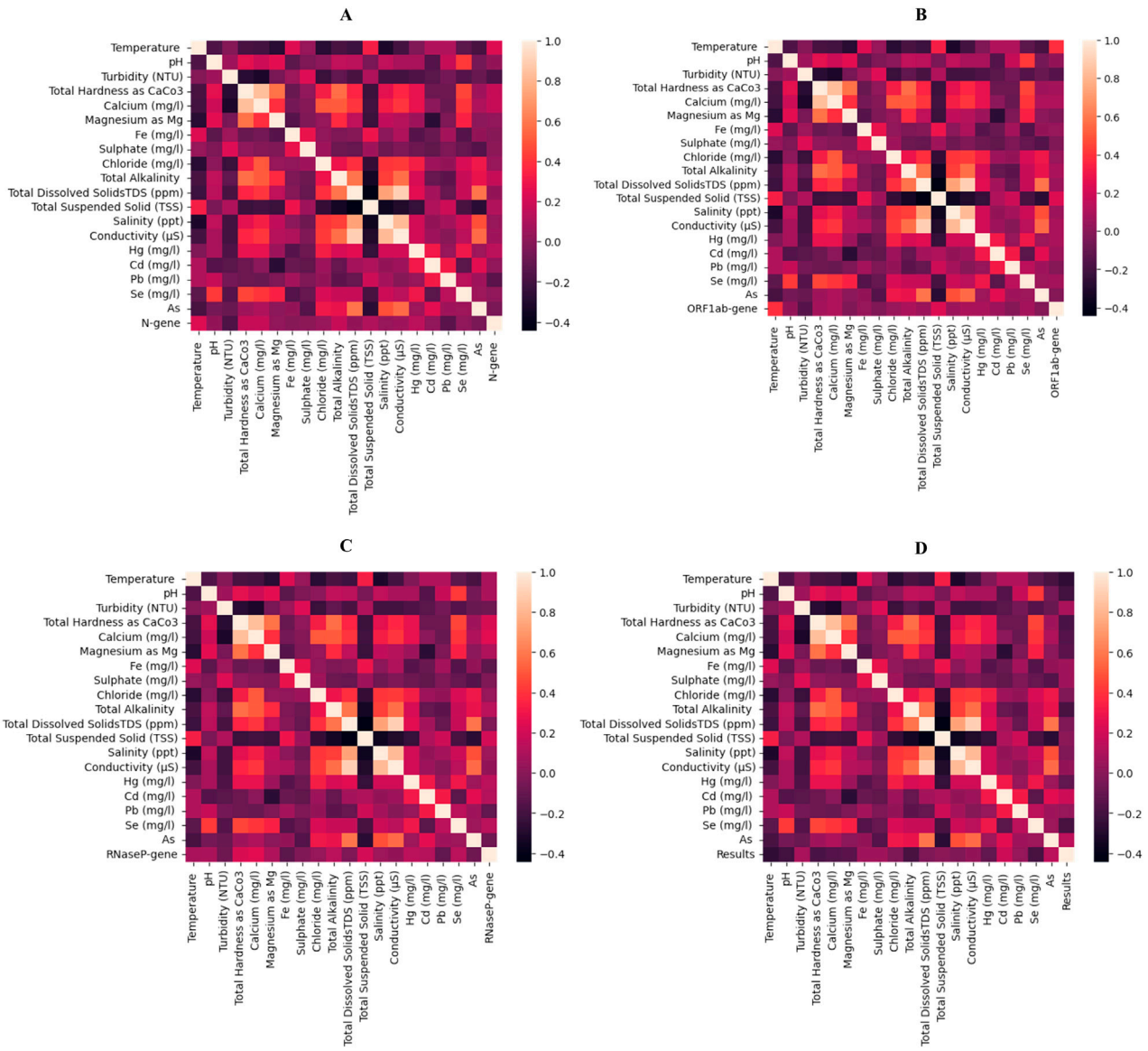

FIGURE 2
Heatmap represented Spearman's rank correlation of all the physicochemical properties considered as independent features in this study. The correlation plots are depicted for **(A)** N gene, **(B)** ORF1ab gene, **(C)** RNaseP gene and **(D)** for RT-qPCR results. The plot showing the presence of both + ve and -ve correlation with the selected features.

(≈− 0.27) compared to other features like pH, total hardness, Calcium, Magnesium, total alkalinity, TSS and Selenium (≈− 0.14). Whereas, Arsenic had a negative correlation of ≈ − 0.17. Similarly, in case of RNaseP gene it was seen that total hardness and calcium had high positive correlations (≈0.27 and ≈0.25 respectively) as compared to other features like pH, Magnesium, alkalinity and TSS. Whereas, features like Iron, Lead and Arsenic contents had a negative correlation (≈− 0.12). Considering the above correlation analysis, it can be seen that most of the features are correlated with the dependent features and will provide insights and contribute in dividing sample locations into relevant clusters. Figure 2 depicts all these findings in the form of heat maps with each dependent features as well as the final outcome, i.e., positive and negative results of RT-qPCR.

3.3 Clustering of sample sites with k-means

The elbow curve (Equation 3) which plots the variance or WCSS against the number of clusters is used to determine the optimal cluster size for a given dataset. The idea is to find the "elbow" point in the curve, which represents the point where increasing the number of clusters ceases to significantly reduce the variance. In the current study, the choice of four clusters was driven by the point on the curve where further increasing the number of clusters resulted in diminishing returns in terms of reducing variance (Supplementary Figure S4) making them adequate to capture the underlying patterns in the present dataset.

K-means clustering (Equations 1, 2) was then to the dataset of this study which resulted in the formation of four distinct clusters

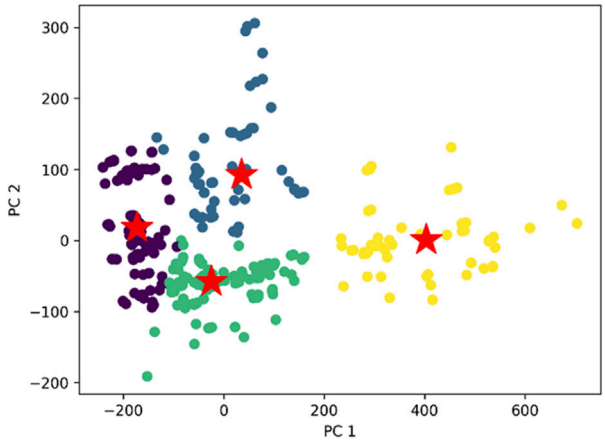

FIGURE 3
The comprehensive plot of K- means clustering with distinct four clusters and their centroids. The cluster we obtained with respect to the two principal component PC1 and PC2 derived using principal component analysis (PCA).

TABLE 1 Performance analysis of clustering algorithms. These analyses were done to verify if the clusters were formed correctly. Silhouette coefficient, Davies-Bouldin indicator.

Silhouette coefficient	Davies-bouldin indicator	Inertia	Dunn indicator
0.38	0.78	3.3×10^6	1.30

(Figure 3). Within this visualization, four unique clusters are delineated by different colours, each serving as a visual indicator of the distinct groupings within our dataset. To signify the centroid, or the central point, of each cluster, we have marked it with a red and green star (Rdusseeun and Kaufman, 1987; Hastie et al., 2009).

3.4 Performance evaluation

The performance of the clustering algorithms used in this study were measured with four metrics Silhouette coefficient, Davies-Bouldin indicator, Inertia, and Dunn Indicator (Equations 4–7). These metrics serve as valuable tools to evaluate the cohesion, separation, and overall effectiveness of the clusters generated by each algorithm, aiding us in making informed and data-driven conclusions about the quality of our clustering solutions.

The Silhouette coefficient, which assesses cluster quality, produced values of 0.38 for K-means indicating comparable cluster cohesion and separation (Table 1). Likewise, the Davies-Bouldin indicator, a measure of cluster compactness and separation, returned values of 0.78 demonstrating closely aligned results. In our study, we have chosen to focus on samples that overlap in both algorithms, specifically selecting those data points for further analysis. The contributions of the principal components in the model's prediction and identifying the patterns in the dataset has been depicted through SHAP dependence plot in Figure 4.

3.5 Analysis of the clusters

With k-means clustering, the 448 samples were categorized into four distinct clusters: cluster 0, cluster 1, cluster 2, and cluster 3. Notably, our analysis revealed that the highest rate of viral positivity was observed within cluster 2, whereas the lowest rate was found in cluster 0 (Table 2). These findings offer valuable insights into the distribution of SARS-CoV-2 within our sample population, shedding light on potential patterns or associations that may be of significance in the context of the study's objectives.

Cluster 2, which exhibited the highest SARS-CoV-2 positivity rate (47.14%), with a remarkable 84.84% during the rainy season. Interestingly, we also noted that the intermediate layer of water displayed a substantial positivity rate of 57.57%, whereas the surface layer showed a slightly lower rate of 42.42%. Furthermore, this cluster consisted of the samples which were collected from sources at comparatively the higher average temperature (24.98°C) than other clusters. Notably, we observed that turbidity, sulphate, total alkalinity, and total dissolved solid (TDS) content were relatively lower in this cluster when compared to the other clusters. Conversely, factors such as iron, chloride, total suspended solid (TSS), and conductivity were found to be higher in this cluster. The combination of environmental factors within this cluster could potentially contribute to the elevated viral positivity observed, which may show relationship between environmental conditions and leading to the SARS-CoV-2 prevalence in this cluster in our study (Férez et al., 2023; Xu and Tian, 2015; Bishop, 2006; Kisand et al., 2023; Vasickova et al., 2010; Osborne et al., 2022; Weller, 2020).

In contrast, cluster 0, which exhibited the lowest positivity rate (26.25%), a higher positivity rate of 78.57% during the dry season which contrasts with cluster 2. Both cluster 2 and cluster 0 displayed a similar pattern where the intermediate water layer had a higher positivity rate at 61.90%, compared to 38.09% in the surface layer. Cluster 0 notably had the highest turbidity content at 17.69 NTU among all clusters, and it also showed higher levels of magnesium, sulphate, total alkalinity, and TDS content in comparison to the other clusters. However, iron, chloride, and TSS were comparatively lower in cluster 0. This distinctive combination of water quality parameters may contribute to the observed lower positivity rate within this particular cluster (Férez et al., 2023; Bishop, 2006; Kisand et al., 2023; Vasickova et al., 2010; Osborne et al., 2022; Weller, 2020).

In the remaining two clusters, i.e., cluster 1 and cluster 3, the SARS-CoV-2 positivity rates exhibited a relatively consistent range, with values falling between 30.15% and 31.61%, respectively. Notably, both clusters shared similar environmental conditions, with samples temperatures ranging from 22.55°C to 23.77°C and pH levels hovering between 7.09 and 6.97. Moreover, key water quality parameters such as total hardness as $CaCO_3$, magnesium, iron, chloride, total alkalinity, TDS, salinity, and conductivity demonstrated comparable values in both clusters. However, a significant distinction emerged in the levels of sulphate and TSS, with cluster 3 exhibiting higher concentrations of these elements in the water when compared to cluster 1. These nuanced differences in water quality factors may contribute to the slight variations observed in SARS-CoV-2 positivity rates between cluster 1 and cluster 3.

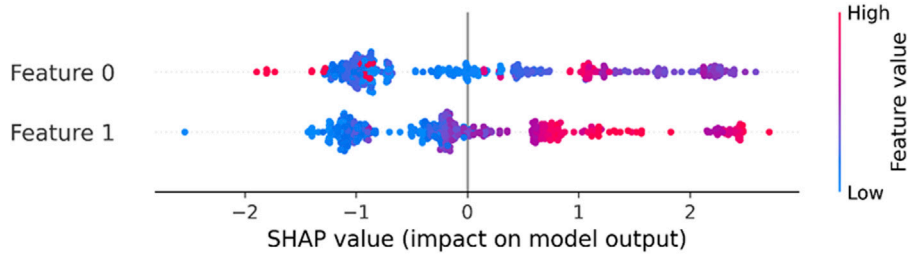

FIGURE 4
SHAP dependence plot of the features (in terms of PC1 and PC2) of the k-means clustering model.

TABLE 2 Cluster details with viral positivity from K- means algorithms.

Cluster	Total samples	Positive samples	Rainy (%)	Dry (%)	Surface (%)	Intermediate (%)
Cluster 0	160	42 (26.25%)	9 (21.43%)	33 (78.57%)	16 (38.09%)	26 (61.90%)
Cluster 1	63	19 (30.16%)	3 (15.79%)	16 (84.21%)	11 (57.89%)	8 (42.11%)
Cluster 2	70	33 (47.14%)	28 (84.84%)	5 (15.15%)	14 (42.86%)	19 (57.14%)
Cluster 3	136	43 (31.62%)	31 (72.09%)	12 (27.91%)	23 (53.49%)	20 (46.51%)

*Rainy, and dry are two different season and two water layers, the surface and the intermediate layer (30 cm depth).

(Férez et al., 2023; Bishop, 2006; Kisand et al., 2023; Vasickova et al., 2010; Osborne et al., 2022; Weller, 2020).

While analysing the physicochemical properties among the clusters, it was seen that there were distinct variations across each cluster (Supplementary Figures S5–S7). This observation may suggest that these specific parameters could potentially exert an influence on the pattern of SARS CoV-2 positivity. The diverse clusters exhibited in the figures underscore the significance of these parameters in understanding and potentially predicting the occurrence and spread of the virus.

3.6 Introspection of overlapping sample sites

In our analysis, we have observed that a few sample sites were associated with multiple clusters, such as 0/3, 1/3, 0/2, 1/2, 2/3, 0/1/2, and 0/2/3 for different samples of same site. After minute examination, it was found that the combination of clusters 0 and 3 appeared most frequently (8 different sites) making it the most prevalent two-cluster combination. Additionally, the combination of clusters 0, 1, and 2 occurred 4 times, which is the most common three-cluster combination in our observations (Table 3).

Further introspection of sites in the 0/3 cluster combination revealed that out of the 8 sites, 6 of them (75%) are situated near flowing water sources such as rivers and streams, while the remaining 2 (25%) are near stagnant water sources (Table 4). Similarly, for the sites associated with the 0/1/2 cluster combination, it was found that all 4 of them (100%) are in close proximity to medical colleges and hospitals (Supplementary Table S2). On a positive note, this shows that the clusters formed out of the sample sites also had similar patterns of data on similar geographical positionings.

Further, we have observed that some of the sample sites (17 out of 28) were associated with multiple clusters, such as 0/3, 1/3, 0/2, and 1/2 with seasonal variation. The sites which were in cluster 0 and 1 during dry season shifted to cluster 2 and 3 during the rainy season. In this regard it was seen that the physicochemical parameters of the sites showed variations with the season. The sites which changed the cluster from 1 to 2 during dry to rainy season had decline values in total hardness, TDS, conductivity, Calcium, Magnesium whereas elevated values in turbidity, TSS, Iron and Sulphate. Similarly for sites which changed cluster from 0 to 2 had decline in values of hardness, TDS, calcium, magnesium and elevated values in iron, chloride, and TSS. For the sites which changes the clusters from 0 or 1 to 3 during dry to rainy season had decline values in most of the parameters like total hardness, TDS, conductivity, calcium, magnesium, sulphate, chloride but elevated values in iron and TSS (Figure 5).

4 Discussion

The present study utilized unsupervised machine learning algorithms to analyse a dataset comprising physicochemical parameters, heavy metal content, and SARS-CoV-2 positivity data from wastewater samples that has collected from 28 different sites in Jorhat district, Assam, India. To our present understanding, this study stands as a pioneering endeavour within the North Eastern region of India, marking the inaugural exploration into the influence of environmental variables on SARS-CoV-2 positivity through the application of a machine learning framework. This methodological approach represents an innovative step in comprehending the intricate interplay between environmental dynamics and viral prevalence, offering novel

TABLE 3 Details of the sites and number of samples comprising 0/1/2 cluster combination.

Locations	Cluster 0		Cluster 1		Cluster 2	
	Rainy	Dry	Rainy	Dry	Rainy	Dry
Athuvoga bridge (H S)	0	2	0	6	8	0
JMCH Hospital Outlet (H S)	1	1	4	4	3	3
Tarajan Kakoty gaon (H S)	3	0	0	8	5	0
Teok Tea Estate (H S)	1	8	1	0	6	0

*HS, hospital site.

TABLE 4 Details of the sites and number of samples comprising 0/e cluster combination.

Locations	Cluster 0		Cluster 3	
	Rainy	Dry	Rainy	Dry
Kuhum stream (Rv S)	4	6	4	2
Kamarbandha (Rv S)	4	8	4	0
CID, CSIR-NEIST (R S)	0	6	8	2
Bhogdoi river (Rv S)	0	8	8	0
FRU, Teok (R S)	2	8	6	0
Baghmora (Rv S)	5	8	3	0
Jhanji, Jorhat-Sibsagar border (Rv S)	1	8	7	0
Nimatighat (Rv S)	0	8	8	0

* RvS, river site; RS, residential site.

insights in this unexplored geographical domain (Cabral, 2010; Rimoldi et al., 2020).

The clustering results indicated K-means algorithm has effectively partitioned the data into four clusters, demonstrating comparable cluster quality as assessed by the Silhouette coefficient and Davies-Bouldin indicator. The clusters' differential characteristics elucidated potential associations between environmental conditions and viral prevalence. Notably, Cluster 2 displayed the highest SARS-CoV-2 positivity rates, especially during the rainy season and in the intermediate water layer. This cluster demonstrated higher temperatures, lower turbidity, sulphate, alkalinity, and total dissolved solids but higher concentrations of iron, chloride, and total suspended solids, hinting at specific environmental conditions favouring increased viral presence. These observations resonate with prior research highlighting the impact of temperature and water quality on viral persistence (Mahanta and Narahari Sastry, 2022; Rimoldi et al., 2020; Ahmed et al., 2020). Conversely, Cluster 0 exhibited the lowest positivity rates and distinct water quality parameters, notably higher turbidity and concentrations of certain minerals and ions. The distinctiveness of these clusters underscores the potential influence of environmental factors in shaping viral prevalence patterns within wastewater.

We further focused our analysis on samples that overlapped in during clustering, which resulted in the selection of 429 data points for further study. Among these samples, 31.93% tested positive for SARS-CoV-2, while the remaining 68.06% tested negative. Notably, our analysis revealed varying rates of viral positivity across the four clusters, with Cluster 2 exhibiting the highest rate of 47.14% during the rainy season, and Cluster 0 displaying the lowest rate of 26.25%, particularly during the dry season. Environmental factors, such as temperature, turbidity, water layer, and specific water quality parameters, were found to vary across the clusters, potentially influencing the observed SARS-CoV-2 positivity rates. Our findings suggest a potential link between environmental conditions and the prevalence of SARS-CoV-2 in different clusters. Additionally, we observed that certain sample sites were consistently associated with specific cluster combinations, indicating potential spatial patterns and associations. For instance, sample sites near flowing water sources, medical colleges, and hospitals exhibited distinct cluster combinations. But variations were also seen for the sites based on the samples collected in two different seasons. The spatial distribution analysis revealed intriguing associations between cluster combinations and geographic positioning.

In our analysis, we've observed that certain locations are associated with multiple clusters, such as 0/3, 1/3, 0/2, 1/2, 2/3, 0/1/2, and 0/2/3. After minute examination, we've identified that the combination of clusters 0 and 3 appears most frequently, occurring in 8 different locations, making it the most prevalent two-cluster combination. Additionally, the combination of clusters 0, 1, and 2 occurs 4 times, which is the most common three-cluster combination in our observations. The locations with the 0/3 cluster combination, we found that out of 8 different geographic locations, 6 (75%) are situated near flowing water sources such as rivers and streams, while the remaining 2 (25%) are near stagnant water sources. Similarly, for the geographic locations associated with the 0/1/2 cluster combination, we found that all four (100%) of these locations are in close proximity to medical colleges and hospitals. Different areas may yield varied viral loads due to population density and access to sanitation, potentially skewing viral presence estimates if not uniformly distributed. Viral load in wastewater can fluctuate seasonally or with rainfall, potentially affecting concentration consistency and detection sensitivity. Temperature, wastewater treatment processes, and chemical pollutants impact viral RNA stability, potentially leading to data inaccuracies in viral load measurement (Medema et al., 2020; Albastaki et al., 2021).

This study's findings align with existing literature emphasizing the role of environmental conditions in modulating viral persistence and transmission dynamics (Smyth et al., 2022; Rimoldi et al., 2020; Wurtzer et al., 2020; Kitajima et al., 2020; Gogoi et al., 2024).

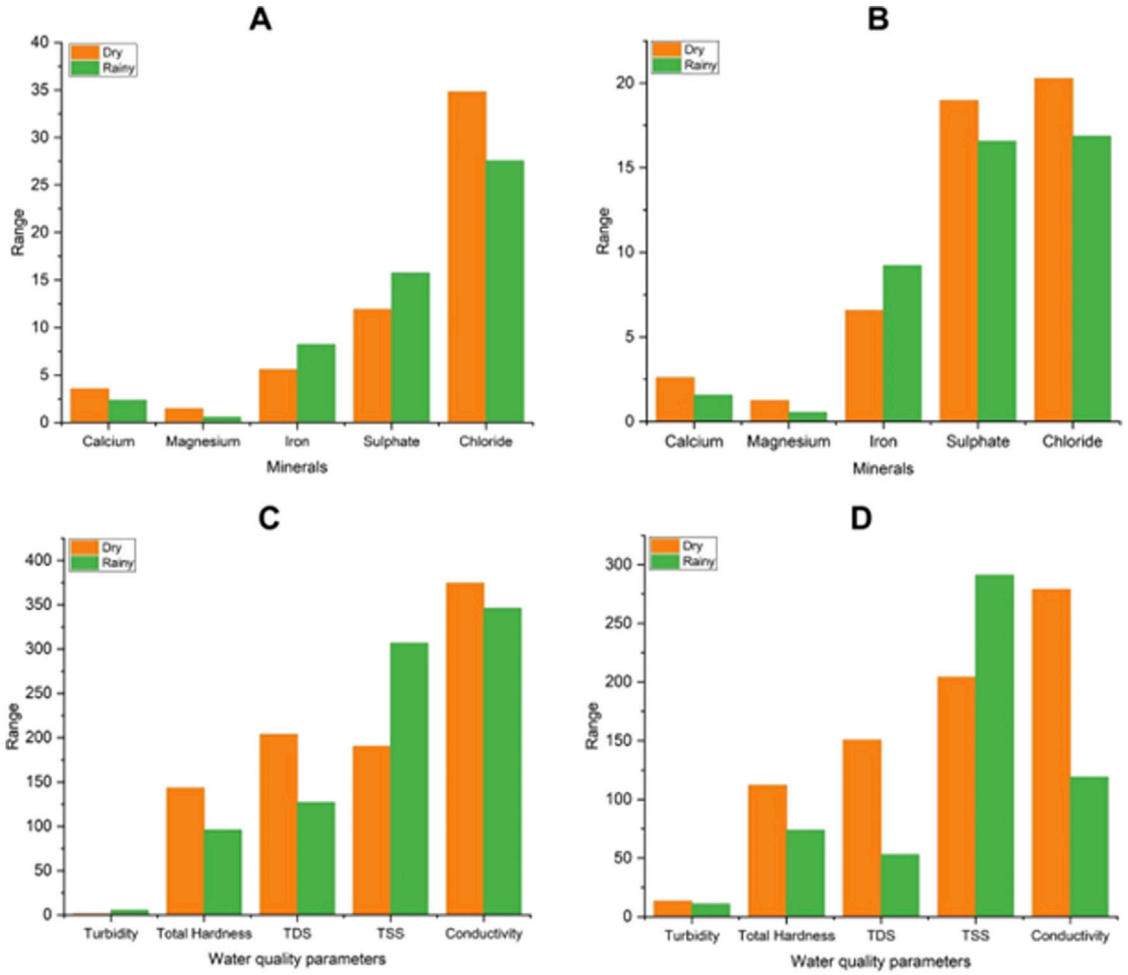

FIGURE 5
Variation in minerals and water quality parameters with season for changes in clusters. **(A)** variation in mineral content of sites for cluster change from 0/1 to 2 **(B)** variation in mineral content of sites for cluster change from 0/1 to 3 **(C)** variation in water quality parameters of sites for cluster change from 0/1 to 2 and **(D)** variation in water quality parameters of sites for cluster change from 0/1 to 3.

Understanding these associations aids in devising targeted intervention strategies, especially in locations with higher viral prevalence, and underscores the importance of considering environmental factors in public health management strategies. Further research is warranted to explore the complex relationships between environmental conditions and viral prevalence.

5 Conclusion

In summary, this study leveraged unsupervised machine learning algorithms to analyze physicochemical parameters, heavy metal content, and SARS-CoV-2 positivity data from wastewater samples collected from 28 sites in Jorhat district, Assam, India. This innovative approach marked the first exploration of the influence of environmental variables on SARS-CoV-2 positivity in the North Eastern region of India. The K-means clustering algorithm effectively partitioned the data into four clusters, revealing significant associations between environmental conditions and viral prevalence. Notably, Cluster 2 exhibited the highest SARS-CoV-2 positivity rates, particularly during the rainy season, and was characterized by higher temperatures, lower turbidity, and increased levels of iron and chloride. Conversely, Cluster 0, with higher turbidity and certain mineral concentrations, showed the lowest positivity rates, especially during the dry season.

The spatial distribution analysis underscored the potential impact of geographic and seasonal variations on viral prevalence, with sample sites near flowing water sources and medical institutions consistently aligning with specific cluster combinations. These findings align with existing literature on the role of environmental conditions in viral persistence and transmission dynamics.

In conclusion, our findings highlight the critical role of environmental factors in shaping SARS-CoV-2 prevalence patterns in wastewater. This study underscores the importance of integrating environmental considerations into public health surveillance and intervention strategies. By demonstrating the utility of machine learning frameworks in epidemiological studies, this research provides valuable insights for targeted

public health interventions in areas with higher viral prevalence. Further research is warranted to explore these relationships in different geographic and environmental contexts.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

GG: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review and editing. SS: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing - original draft. DK: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing - original draft. EK: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing - original draft. RB: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing - review and editing. AD: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing - review and editing. HM: Data curation, Formal Analysis, Investigation, Software, Supervision, Validation, Visualization, Writing - review and editing. PB: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is funded by DST-SERB (Ref. No. CRG/2022/000095).

Acknowledgments

PB thanks DST-SERB (Ref. No. CRG/2022/000095), GG thanks CSIR-NEIST for the support in form of OLP-2088, AD thanks CSIR-NEIST for the support in form of OLP 2078, HM thanks DBT, Govt of India, for Centre of Excellence in Advanced Computation and Data Sciences (Ref. No: BT/PR40188/BTIS/137/27/2021).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2024.1508964/full#supplementary-material

References

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., et al. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci. Total Environ.* 728, 138764. doi:10.1016/j.scitotenv.2020.138764

Ahmed, W., Tscharke, B., Bertsch, P. M., Bibby, K., Bivins, A., Choi, P., et al. (2021). SARS-CoV-2 RNA monitoring in wastewater as a potential early warning system for COVID-19 transmission in the community: a temporal case study. *Sci. Total Environ.* 761, 144216. doi:10.1016/j.scitotenv.2020.144216

Ai, Y., He, F., Lancaster, E., and Lee, J. (2022). Application of machine learning for multi-community COVID-19 outbreak predictions with wastewater surveillance. *Plos one* 17 (11), e0277154. doi:10.1371/journal.pone.0277154

Albastaki, A., Naji, M., Lootah, R., Almeheiri, R., Almulla, H., Almarri, I., et al. (2021). First confirmed detection of SARS-COV-2 in untreated municipal and aircraft wastewater in Dubai, UAE: the use of wastewater based epidemiology as an early warning tool to monitor the prevalence of COVID-19. *Sci. Total Environ.* 760, 143350. doi:10.1016/j.scitotenv.2020.143350

Arora, K., Kaur, P., Kumar, P., Singh, A., Patel, S. K. S., Li, X., et al. (2021). Valorization of wastewater resources into biofuel and value-added products using microalgal system. *Front. Energy Res.* 9, 646571. doi:10.3389/fenrg.2021.646571

Arthur, D., and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. *Soda* 7, 1027–1035. doi:10.1145/1283383.1283494

Bishop, C. M. (2006). *Pattern recognition and machine learning*, 2. Springer google schola, 1122–1128.

Cabral, J. P. (2010). Water microbiology. Bacterial pathogens and water. *Int. J. Environ. Res. public health* 7 (10), 3657–3703. doi:10.3390/ijerph7103657

Férez, J. A., Cuevas-Ferrando, E., Ayala-San Nicolás, M., Simón Andreu, P. J., López, R., Truchado, P., et al. (2023). Wastewater-based epidemiology to describe the evolution of SARS-CoV-2 in the south-east of Spain, and application of phylogenetic analysis and a machine learning approach. *Viruses* 15 (7), 1499. doi:10.3390/v15071499

Gogoi, G., Singh, S. D., Kalyan, E., Koch, D., Gogoi, P., Kshattry, S., et al. (2024). An interpretative review of the wastewater-based surveillance of the SARS-CoV-2: where do we stand on its presence and concern? *Front. Microbiol.* 15, 1338100. doi:10.3389/fmicb.2024.1338100

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, 2. New York: Springer, 1–758.

Kisand, V., Laas, P., Palmik-Das, K., Panksep, K., Tammert, H., Albreht, L., et al. (2023). Prediction of COVID-19 positive cases, a nation-wide SARS-CoV-2 wastewater-based epidemiology study. *Water Res.* 231, 119617. doi:10.1016/j.watres.2023.119617

Kitajima, M., Ahmed, W., Bibby, K., Carducci, A., Gerba, C. P., Hamilton, K. A., et al. (2020). SARS-CoV-2 in wastewater: state of the knowledge and research needs. *Sci. Total Environ.* 739, 139076. doi:10.1016/j.scitotenv.2020.139076

Kshattry, M., Singh, S. D., Bharali, P., and Sastry, G. N. (2022). Study of progression of COVID-19 in Indian population based on transcriptomic approach. doi:10.20944/preprints202212.0431.v1

Levin, J. C., Curtis, C. J., and Woodford, D. J. (2024). A multi-spatial scale assessment of land-use stress on water quality in headwater streams in the Platinum Belt, South Africa. *Sci. Total Environ.* 927, 172180. doi:10.1016/j.scitotenv.2024.172180

Mahanta, H. J., and Narahari Sastry, G. (2022). COVID-19 impact on socio-economic and health interventions: a gaps and peaks analysis using clustering approach. *J. Statistics Manag. Syst.* 25 (8), 2123–2153. doi:10.1080/09720510.2022.2117335

Mathew, R. A., and Kanmani, S. (2020). A review on emerging contaminants in indian waters and their treatment technologies. *Nat. Environ. Pollut. Tech.* 19 (2), 549–562. doi:10.46488/NEPT.2020.v19i02.010

Medema, G., Been, F., Heijnen, L., and Petterson, S. (2020). Implementation of environmental surveillance for SARS-CoV-2 virus to support public health decisions: opportunities and challenges. *Curr. Opin. Environ. Sci. and health* 17, 49–71. doi:10.1016/j.coesh.2020.09.006

Moretti, A., Ivan, H. L., and Skvaril, J. (2024). A review of the state-of-the-art wastewater quality characterization and measurement technologies. Is the shift to real-time monitoring nowadays feasible? *J. Water Process Eng.* 60, 105061. doi:10.1016/j.jwpe.2024.105061

Naushin, S., Sardana, V., Ujjainiya, R., Bhatheja, N., Kutum, R., Bhaskar, A. K., et al. (2021). Insights from a Pan India sero-epidemiological survey (phenome-India cohort) for SARS-CoV2. *Elife* 10, e66537. doi:10.7554/eLife.66537

Newhart, K. B., Holloway, R. W., Hering, A. S., and Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: a review. *Water Res.* 157, 498–513. doi:10.1016/j.watres.2019.03.030

Osborne, E., Haddix, M., and Garner, E. (2022). Impact of hydraulic and physicochemical factors on spatiotemporal variations of particle-associated bacteria in a drinking water distribution system. *Front. Water* 4, 959618. doi:10.3389/frwa.2022.959618

Pan, B., Han, X., Chen, Y., Wang, L., and Zheng, X. (2022). Determination of key parameters in water quality monitoring of the most sediment-laden Yellow River based on water quality index. *Process Saf. Environ. Prot.* 164, 249–259. doi:10.1016/j.psep.2022.05.067

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58 (347-352), 240–242. doi:10.1098/rspl.1895.0041

Ravi, V., Shamim, U., Khan, M. A., Swaminathan, A., Mishra, P., Singh, R., et al. (2024). Unraveling the genetic evolution of SARS-CoV-2 Recombinants using mutational dynamics across the different lineages. *Front. Med.* 10, 1294699. doi:10.3389/fmed.2023.1294699

Rdusseeun, L. K. P. J., and Kaufman, P. (1987). "Clustering by means of medoids," in *Proceedings of the statistical data analysis based on the L1 norm conference*, 31. neuchatel, switzerland.

Rimoldi, S. G., Stefani, F., Gigantiello, A., Polesello, S., Comandatore, F., Mileto, D., et al. (2020). Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. *Sci. Total Environ.* 744, 140911. doi:10.1016/j.scitotenv.2020.140911

Sahu, S., Kaur, A., Singh, G., and Arya, S. K. (2023). Harnessing the potential of microalgae-bacteria interaction for eco-friendly wastewater treatment: a review on new strategies involving machine learning and artificial intelligence. *J. Environ. Manag.* 346, 119004. doi:10.1016/j.jenvman.2023.119004

Singh, P., Ujjainiya, R., Prakash, S., Naushin, S., Sardana, V., Bhatheja, N., et al. (2022). A machine learning-based approach to determine infection status in recipients of BBV152 (Covaxin) whole-virion inactivated SARS-CoV-2 vaccine for serological surveys. *Comput. Biol. Med.* 146, 105419. doi:10.1016/j.compbiomed.2022.105419

Smyth, D. S., Trujillo, M., Gregory, D. A., Cheung, K., Gao, A., Graham, M., et al. (2022). Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* 13 (1), 635. doi:10.1038/s41467-022-28246-3

Spurbeck, R. R., Minard-Smith, A., and Catlin, L. (2021). Feasibility of neighborhood and building scale wastewater-based genomic epidemiology for pathogen surveillance. *Sci. Total Environ.* 789, 147829. doi:10.1016/j.scitotenv.2021.147829

Srikanth, K., Sukesh, K., Rao, A. R., Pavan, G., and Ravishankar, G. A. (2019). Emerging contaminants effect on aquatic ecosystem: human health risks. *Agric. Res. Technol.* 19, 556104. doi:10.19080/ARTOAJ.2019.19.556104

Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., and Sutton, D. J. (2012). Heavy metal toxicity and the environment. Molecular, clinical and environmental toxicology. *Environ. Toxicol.* 3, 133–164. doi:10.1007/978-3-7643-8340-4_6

Thakur, V., Bhola, S., Thakur, P., Patel, S. K. S., Kulshrestha, S., Ratho, R. K., et al. (2022). Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. *Infection* 50, 309–325. doi:10.1007/s15010-021-01734-2

Van der Werf, J. A., Kapelan, Z., and Langeveld, J. (2023). Real-time control of combined sewer systems: risks associated with uncertainties. *J. Hydrology* 617, 128900. doi:10.1016/j.jhydrol.2022.128900

Vasickova, P., Pavlik, I., Verani, M. A. R. C. O., and Carducci, A. N. N. A. L. A. U. R. A. (2010). Issues concerning survival of viruses on surfaces. *Food Environ. Virology* 2, 24–34. doi:10.1007/s12560-010-9025-6

Wahengbam, R., Bharali, P., Manna, P., Phukan, T., Singh, M. G., Gogoi, G., et al. (2023). Seroepidemiological and genomic investigation of SARS-CoV-2 spread in North East region of India. *Indian J. Med. Microbiol.* 43, 58–65. doi:10.1016/j.ijmmb.2022.10.011

Weller, D. (2020). Cancer diagnosis and treatment in the COVID-19 era. *Eur. J. Cancer Care* 29 (3), e13265. doi:10.1111/ecc.13265

Wigginton, K. R., Ye, Y., and Ellenberg, R. M. (2015). Emerging investigators series: the source and fate of pandemic viruses in the urban water cycle. *Environ. Sci. Water Res. and Technol.* 1 (6), 735–746. doi:10.1039/C5EW00125K

World Health Organization (2019). *Water, sanitation, hygiene and health: a primer for health professionals. No. WHO/CED/PHE/WSH/19.149*. Geneva, Switzerland: World Health Organization.

Wurtzer, S., Marechal, V., Mouchel, J., Maday, Y., Teyssou, R., Richard, E., et al. (2020). Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. doi:10.1101/2020.04.12.20062679

Xu, D., and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Ann. data Sci.* 2, 165–193. doi:10.1007/s40745-015-0040-1

Yadav, M., and Chauhan, N. S. (2023). "Wastewater surveillance: a quick guide to check community health," in *Genomic surveillance and pandemic preparedness* (Academic Press), 187–224. doi:10.1016/B978-0-443-18769-8.00012-X

Zeiner, M., Rezic, I., and Steffan, I. (2007). Analytical methods for the determination of heavy metals in the textile industry. *Kem. Ind.* 56 (11), 587–595.