Frontiers of Environmental Science & Engineering
2025, 19(1): 12 https://doi.org/10.1007/s11783-025-1932-8

RESEARCH ARTICLE

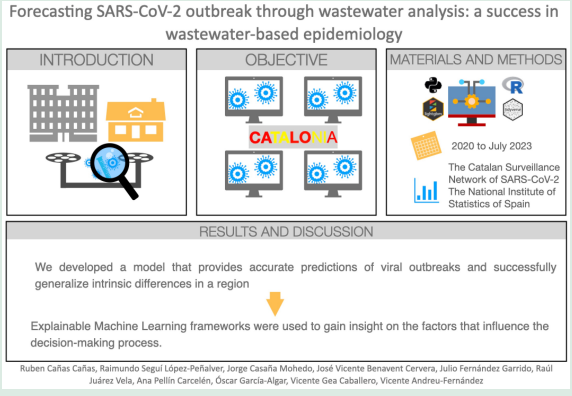
Forecasting SARS-CoV-2 outbreak through wastewater analysis: a success in wastewater-based epidemiology

Rubén Cañas Cañas[1,2,3,4], Raimundo Seguí López-Peñalver[1], Jorge Casaña Mohedo[1,5], José Vicente Benavent Cervera[1], Julio Fernández Garrido[6], Raúl Juárez Vela[7], Ana Pellín Carcelén[1], Óscar García-Algar[3,4,8], Vicente Gea Caballero[1,#], Vicente Andreu-Fernández ✉ [1,3,9,#]

1. Faculty of Health Sciences, Valencian International University (VIU), Valencia 46002, Spain
2. Global Omnium, Valencia 46005, Spain
3. Grup de Recerca Infancia i Entorn (GRIE), Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona 08036, Spain
4. Department de Cirurgia i Especialitats Mèdico-Quirúrgiques, Universidad de Barcelona, Barcelona 08036, Spain
5. Faculty of Health Sciences, Universidad Católica de Valencia San Vicente Mártir, Valencia 46001, Spain
6. Department of Nursing, University of Valencia, Valencia 46001, Spain
7. Faculty of Health Sciences, La Rioja University, Logroño 26006, Spain
8. Department of Neonatology, Instituto Clínic de Ginecología, Obstetricia y Neonatología (ICGON), Hospital Clínic-Maternitat, BCNatal, Barcelona 08028, Spain
9. Biosanitary Research Institute, Valencian International University (VIU), Valencia 46002, Spain

HIGHLIGHTS

- Virus detection in wastewater is a valuable tool for anticipating outbreaks.
- Feature engineering has proven valuable for developing predictive models.
- LightGBM models robustly generalize predictions across an entire region.
- Explainable ML frameworks are crucial for confidence in model predictions.
- WBE is a valuable tool that can help Public Health authorities in decision-making.



ABSTRACT: The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), triggered a global emergency that exposed the urgent need for surveillance approaches to monitor the dynamics of viral transmission. Several epidemiological tools that may help anticipate outbreaks have been developed. Wastewater-based epidemiology is a non-invasive and population-wide methodology for tracking the epidemiological evolution of the virus. However, thorough evaluation and understanding of the limitations, robustness, and intricacies of wastewater-based epidemiology are still pending to effectively use this strategy. The aim of this study was to train highly accurate predictive

✉ Corresponding author. E-mail: vandreu@universidadviu.com

# These authors contributed equally to this work.

Article history: Received 11 June 2024, Revised 4 November 2024, Accepted 5 November 2024, Available online 21 November 2024

models using SARS-CoV-2 virus concentrations in wastewater in a region consisting of several municipalities. The chosen region was Catalonia (Spain) given the availability of wastewater SARS-CoV-2 quantification from the Catalan surveillance network and healthcare data (clinical cases) from the regional government. By using various feature engineering and machine learning methods, we developed a model that can accurately predict and successfully generalize across the municipalities that make up Catalonia. Explainable Machine Learning frameworks were also used, which allowed us to understand the factors that influence decision-making. Our findings support wastewater-based epidemiology as a potential surveillance tool to assist public health authorities in anticipating and monitoring outbreaks.

KEYWORDS: SARS-CoV-2, Wastewater based epidemiology, Surveillance, Machine learning, Predictive models, Model explainability

1 Introduction

The WHO officially declared COVID-19 - caused by the SARS-CoV-2 virus - a pandemic on March 11, 2020, and announced its end on May 5, 2023 (Sarker et al., 2023). The evolution of the virus challenged health systems and government institutions worldwide, resulting in 765 million diagnosed cases and 6.9 million deaths. Despite the official end of the pandemic, COVID-19 waves persist, most times triggered by much more virulent variants of the virus, which become dominant by taking advantage of the mobility phenomena in the context of globalization (Islam et al., 2022).

Throughout the pandemic, different diagnostic strategies for virus detection have been developed and used to track the behavior and evolution of the virus during the different waves. Individual diagnostic tests, primarily PCR and antigen tests, have been available since nearly the beginning of this health crisis. Beyond diagnosis and population screening, these tests have also enabled effective contact tracing, suppression of transmission chains, and containment of outbreaks (Vandenberg et al., 2021; Daza-Torres et al., 2023). Moreover, the implementation of massive virus detection tools, based on the detection and quantification of the virus in wastewater collection systems, has gained relevance in Wastewater-Based Epidemiology (WBE). Implemented in numerous cities worldwide, these tools have proven effective for early detection, assessing trends in viral circulation and monitoring community health (López-Peñalver et al., 2023; Chen et al., 2024). This methodology can be applied in large urban centers, in smaller districts, or in specific facilities such as schools, nursing homes, social care centers, etc., as it provides valuable information that can be used by public administrations and government agencies to support decision-making.

However, despite the potential of monitoring wastewater collection systems, the interpretation of the resulting data is complex, as it depends on several factors such as population size, traffic and transport infrastructure, season, weather conditions, or wastewater treatment plant (WWTP) factors, among others (Shang et al., 2023; Silva, 2023; Liao et al., 2024).

Statistical models are needed to explain the involved covariates and make accurate predictions of epidemiological trends. The development of these models could be highly effective in providing a deep understanding of the epidemiological context, ultimately aiding public health authorities by offering a surveillance system to support the formulation of adaptation and prevention policies (Jeng et al., 2023).

In recent years, predictive models, using algorithms to identify patterns in data and forecast future trends, and in particular machine learning (ML) models, have become key in the epidemiological surveillance of COVID and other infectious diseases. These models have proven effective in predicting disease progression (Joseph-Duran et al., 2022). They can help governments prepare for future pandemics, identify the risk factors that contribute to their spread, and/or assess the effectiveness of the implemented control measures (Santangelo et al., 2023).

Statistical models using AI and ML to improve screening, diagnosis, predictive analysis, public health monitoring, and vaccine development have been widely used in the context of the COVID-19 pandemic. Statistical models have proven to be essential in multiple aspects in the management of pandemic, underscoring the critical role of data science in tackling unprecedented public health challenges (Lalmuanawma and Hussain, 2020; Booth et al., 2021; Chadaga et al., 2021).

Machine learning algorithms, including

Autoregressive Integrated Moving Average (ARIMA), Neural Networks, and Random Forest methods, have been successfully used to estimate COVID-19 trends (Jeng et al., 2023) and predict the effects of various vaccination strategies, such as those modeled by the recently developed SEIR models (Schneider et al., 2023). Furthermore, ML methods offer the advantage of generating accurate predictions without the challenges associated with models that require prior knowledge. The flexibility and versatility of ML models make them highly effective for prediction purposes (Weinan, 2020). Regarding WBE specifically, several research groups have made significant strides in WBE for tracking and predicting COVID-19 trends. Vallejo et al. (2022) employed both linear and non-parametric models to estimate the actual prevalence of COVID-19 infections using data from a WWTP. Similarly, Hill et al. (2023) demonstrated the ability to predict COVID-19 hospitalizations up to 10 days in advance through the use of Generalized Linear Mixed Models (GLMM). Expanding beyond single-site studies, other researchers, such as Ai et al. (2022) and Joseph-Duran et al. (2022), developed predictive models for multi-community or regional cohorts, extending the applicability of WBE to larger populations and broader geographic areas.

From a societal perspective, refining and improving these models with complementary data from wastewater monitoring can reduce healthcare costs and optimize resource allocation, helping to prevent further strain on the health system.

The aim of the present study was to analyze the use of ML models to predict future COVID-19 in past outbreaks, using clinical cases counts from various populations in Catalonia between 2020 and 2022. This research focused in developing wastewater-based ML models to predict clinical COVID-19 cases across an entire region, encompassing multiple municipalities. The goal was to create a generalized model that could be applied across diverse populations. To enhance its adaptability, the model incorporated contextual variables and feature-engineered variables, allowing it to account for population-specific factors and improve its predictive accuracy. The models were evaluated using the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). R^2 assesses how well the model explains the variability in the data, while MAE and RMSE indicate the size of the prediction errors. Furthermore, the effectiveness of machine-learning based methods creates new opportunities for their application in the surveillance of other infectious diseases.

2 Materials and methods

2.1 Software, tools, and code availability

Data transformation, aggregation, and feature engineering were performed using R 4.3.2. (R Core Team, 2024), while for model development and validation Python 3.10.12 (van Rossum, 1995) was used. The main packages used in this work were the tidyverse suite (Wickham et al., 2019), the scikit-learn suite, the LightGBM package (Ke et al., 2017) and the SHAP package (Lundberg and Lee, 2017).

2.2 Data acquisition

WWTP SARS-CoV-2 concentrations were obtained from the Catalan Surveillance Network of SARS-CoV-2 in sewage (Guerrero-Latorre et al., 2022). Viral concentrations were calculated by RT-qPCR amplifications of the N1, N2, IP4, and E gene targets. Variables for precipitation and measured flow of the WWTP at the time of sampling were also collected from the data set. Quantitative determination of SARS-CoV-2 in wastewater (by RT-qPCR and the N1, N2, IP4, and E gene targets) was the method adopted by the Ministry for the Ecological Transition and the Demographic Challenge of Spain (based on ISO 15216-1:2017) and was made the reference guideline since the beginning of the pandemic (Randazzo et al., 2020). Use of multiple genetic targets provides high sensitivity and specificity of the RT-qPCR results (Lu et al., 2020). N1 and N2 are two different regions within the nucleocapsid (N) gene (essential for viral replication and assembly, and highly conserved among SARS-CoV-2 strains); E stands for the Envelope (E) gene, which codes for a small protein involved in viral assembly; and IP4 refers to a region within the ORF1ab polyprotein translated from the viral genome, essential for viral replication. The results of the amplifications were only considered valid if at least two genetic targets were amplified from the samples (Kumar et al., 2021). This work compiled the information for every WWTP of Catalonia, collecting the information for every major population in the region starting July 2020.

The clinical data from Catalonia was retrieved from the open data provided by the transparency portal of the government of Catalonia (Generalitat de Catalunya, 2023). This data set dates from early 2020 to July 2022. Population data for the municipalities were downloaded from the National Institute of Statistics of Spain (Instituto Nacional de Estadística, 2023).

2.3 Data exploration and preprocessing

The three data sets mentioned earlier were merged by municipality and truncated to provide data up to July 2022. Any record missing relevant variable was excluded from the final data set. The variables for the four amplified genetic targets in the data set were combined into a single variable, with the higher concentration for each record chosen as the SARS-CoV-2 concentration (Joseph-Duran et al., 2022). SARS-CoV-2 concentrations were treated with locally estimated scatterplot smoothing (LOESS) using a span of 10% to mitigate different factors that add noise to the data, including temperature, precipitation, sampling site, different molecular targets, or water levels and flow in the WWTP (Arabzadeh et al., 2021).

Virus concentrations and case data were also processed with Box-Cox transformations and zero-mean, unit-variance normalization (to the transformed data) to optimize for normality and adapt to the properties of each variable individually, which will be favorable for the ML algorithms to fit better to the data (Blum et al., 2022; Marimuthu et al., 2022).

The entire data set was aggregated to represent weekly data. To introduce context for clinical cases to the model, a number of variables were introduced using 2-week and 4-week moving averages for cases and SARS-CoV-2 concentrations in wastewater for each municipality. This methodology was successfully explored by other research groups (Zhu et al., 2022).

Finally, since the detection of SARS-CoV-2 in wastewater has proven highly effective for forecasting of COVID (López-Peñalver et al., 2023), the target variable in the predictions for this work are the 1-week leaded cases. Moreover, the final data set excludes same-week cases, making the trained models agnostic to "current" cases. This inherent delay in data collection by healthcare authorities makes it unrealistic to include this variable into the model.

2.4 Model selection and training

A preliminary screening of several models was conducted using the LazyPredict package in Python (version 0.2.12), which offers a broad selection of models for this purpose and has been widely utilized in research (Pirzada et al., 2023). The screening involved training the models on 75% of the data and testing them on the remaining 25%. This train-test size was selected to ensure the metrics reflect the model's true adaptability to the data and not because of overfitting. The measures utilized for the scoring and evaluation of the models were the R^2 and the Root Mean Squared Error (RMSE). The R^2 metric quantifies the model's ability to explain variance, while the RMSE measures the magnitude of prediction errors.

This way we estimated which modeling strategy adjusted better to the data set.

The models chosen from the preliminary screening were evaluated using 5-fold cross-validation. This method divides the data sets into five 80/20 splits covering the whole data set as testing data to discard any model over-fitting the data or any relevant problem in the training process. The same scoring measures were used in the training process.

Finally, the selected model was fine-tuned by exploring hyperparameters through Grid and Random Search techniques, incorporating cross-validation. Fine-tuning was performed on 80% of the data set, on which a grid (hand-selected or randomized) of different parameter combinations were evaluated to find the optimal combination to fit the data. R^2 between the actual data and predictions was used as the evaluation metric to maximize during the hyperparameter tuning process, guiding the evaluation of different parameter settings. The evaluation of each model configuration was carried out using 5-fold cross-validation to ensure robustness and minimize the risk of overfitting, meaning the hyperparametrization process was performed with 64% and 16% of the whole data set for training and testing, respectively. Finally, the best performing model configuration was tested against the remaining 20% of the data set to evaluate the final metrics.

The best performing model was employed to generate the final predictions and assessments.

2.5 Model explainability

Model explainability is key for interpreting the predictions of ML models. To enhance interpretability of the final model used in this work, the Shapley Additive exPlanations (SHAP) framework was employed (Lundberg and Lee, 2017). SHAP is a model-agnostic method for explainable AI based on Shapley values, a concept grounded in cooperative game theory (Shapley, 1952). This methodology offers a unified measure of importance by attributing each feature's contribution to the model's output, providing both local and global interpretability of the predictions. This approach facilitates a thorough understanding of which features contribute most significantly to model predictions. As a result, it has become a leading method for model explainability (Clement et al., 2023).

The SHAP values obtained were utilized to interpret the final model predictions, validate the relevance of

each feature, and identify potential interactions among variables.

Figure 1 summarizes how the data were structured to build the final data set used to train the models (Fig. 1A) and the workflow followed to produce the final model (Fig. 1B).

3 Results

3.1 Data exploration results

Regarding SARS-CoV-2 molecular targets, the N1 gene was usually the most amplified, representing 72.5% of the total data set, consistently exhibiting the highest concentrations. The second most amplified target was N2 (23% of the time); IP4 and E genes exhibited the highest concentrations on 4.4% and 0.07% of the observations respectively.

These variables were also analyzed to determine the number of missing cases for each one. The most prevalent target in the data set was N1, missing only in 0.25% of the observations. Genes N2 and IP4 were missing in 35.63% and 66.87% of the observations, respectively. Again, the least represented variable was the E gene, missing in 96.95% of the observations. Nonetheless, the combined feature of all four genetic targets were present across the entire set of observations (Supplementary Fig. S1).

SARS-CoV-2 concentrations in wastewater and cases, which are the central features in the data set, exhibited a skewed distribution (Fig. 2A). When processed using Box-Cox transformation and the subsequent scaling, these variables approximated normal distributions (Fig. 2B).

After aggregating WWTP data sets and clinical data, a clear concordance between SARS-CoV-2 concentrations and clinical cases was observed (Fig. 3). Moreover, the variables showed a distinct distribution on population size when municipalities were categorized into bins according to their number on inhabitants.

3.2 Model training and selection

The preliminary analysis for model selection showed that models trained on the original distributions for SARS-CoV-2 concentration in wastewater and clinical cases achieved superior performance in the evaluations, i.e., higher R^2 between real data and predictions on the testing set (Supplementary Table S1). Nonetheless, during cross-validation, certain folds of the data sets exhibited inconsistencies and produced inaccurate evaluation metrics. For example, the extra-trees regressor, which initially achieved an R^2 of 0.84 without data transformation showed significant

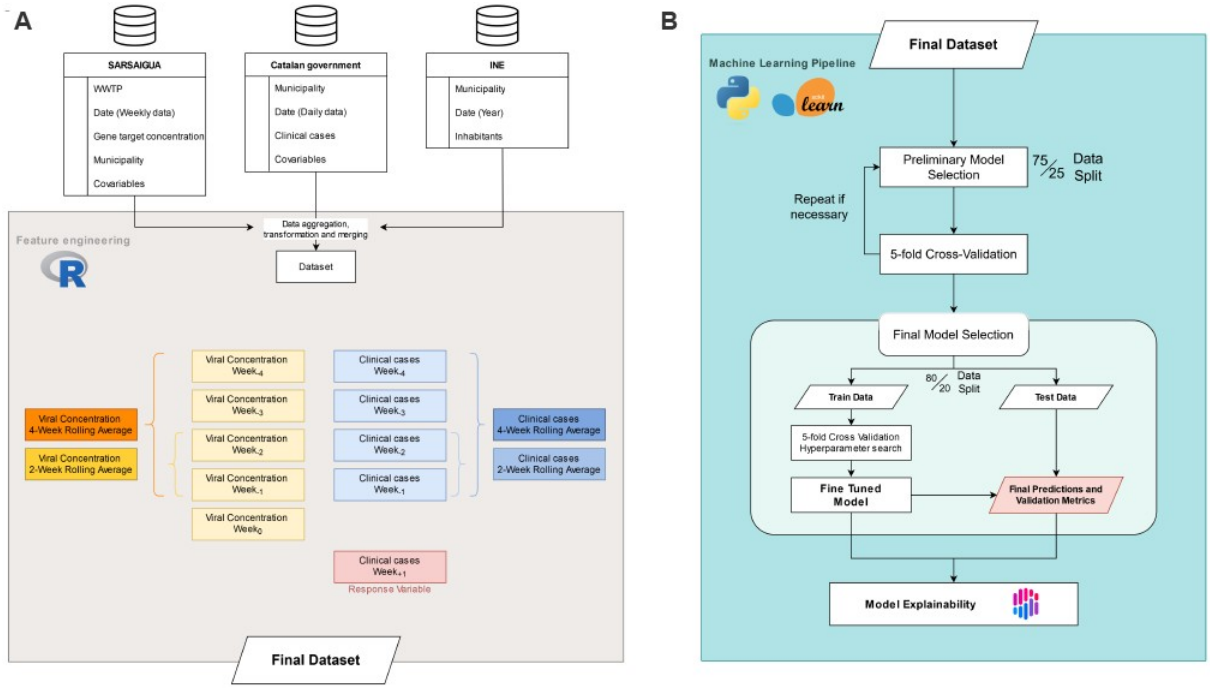


Fig. 1 Summary of Data Processing steps (A) and Machine Learning Workflow (B).

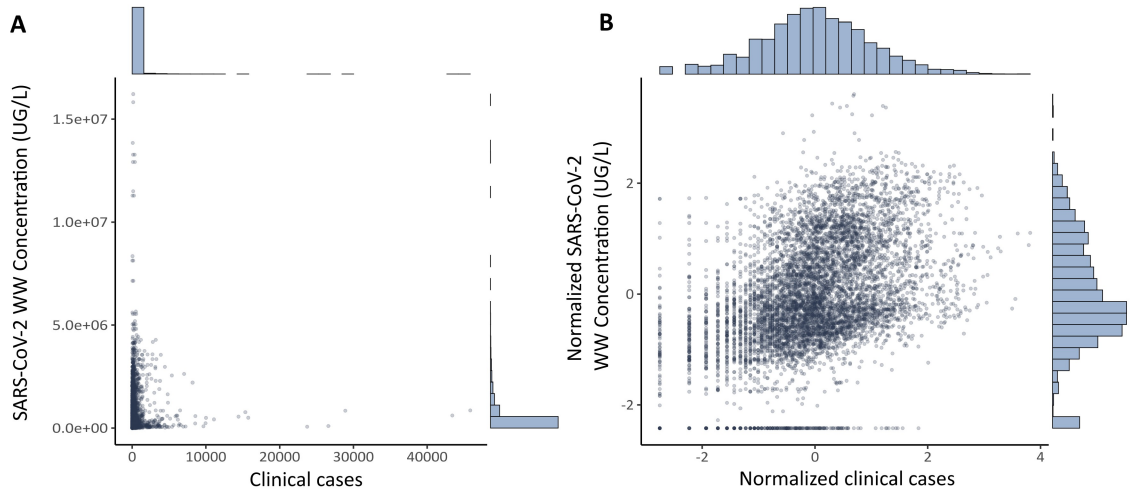


Fig. 2 Distribution of SARS-CoV-2 concentrations in wastewater and COVID-19 cases without transformation (A) and Box-Cox transformation (B). The data are represented in each situation as a scatterplot and a histogram distribution for the two variables.

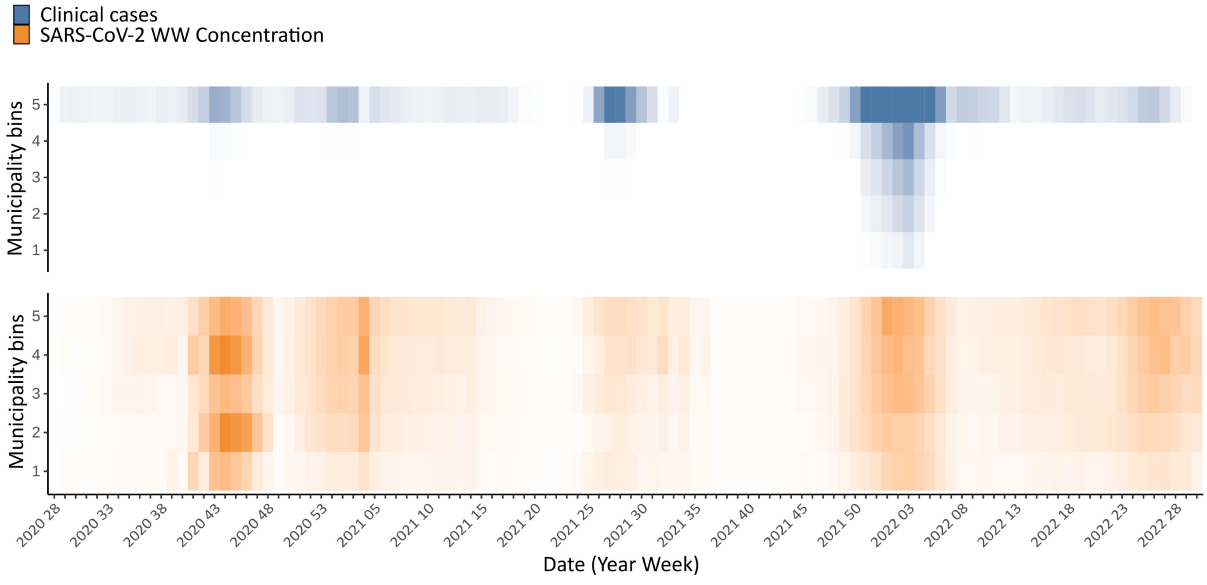


Fig. 3 Heatmap depicting the qualitative progression of clinical cases (top) and SARS-CoV-2 wastewater treatment plant concentrations over time (weekly resolution). Municipalities were grouped into five bins and ordered from top to bottom based on population size for visualization. Darker colors indicate higher values of the represented variable.

variation in R^2 values when evaluated using 5-fold cross-validation. The R^2 values ranged from −2.91 to 0.87 (Supplementary Table S2), indicating poor generalization and an incorrect fit for the entire data set.

With the Box-Cox transformation, the best performing model was the Light Gradient Boosting Model (LGBM) regressor (R^2 of 0.79). When subjected to 5-fold cross-validation, the R^2 was 0.788 ± 0.01, suggesting consistent performance across the whole data set.

Moreover, the models that performed best during the preliminary evaluation were those based on gradient boosting, random forest, and neural network types. From these three different types of ML algorithms, the best performing from each group was also evaluated with cross-validation. The selected models were the aforementioned LGBM, the Extra-Trees Regressor Model (ETRM), and the Multi-layered Perceptron Neural Network (MLPNN). The results of the 5-fold cross-validation are presented in Table 1. It important

to note that the use of the MLPNN required additional data transformation, specifically maximum-minimum normalization, to scale all variables within the 0 to 1 range.

The MLPNN exhibited the lowest correlation between real data and predictions among the chosen models; however, the RMSE is the lowest of the selected models. Furthermore, while the ETRM provides similar predictions to the LGBM, the execution time of the models is much shorter with the latter.

Thus, the MLPNN and the ETRM were excluded from further fine-tuning.

3.3 Model fine-tuning and final evaluation

Among the various adjusted hyperparameters, those that had the most significant impact on the algorithm's performance were the boosting learning rate, the maximum tree depth for base learners, the total number of estimators, the maximum tree leaves for base learners, the subsample ratio of columns when constructing each tree, and the L1 and L2 regularization terms on weights. These parameters were manually selected for the grid search. The top-performing parameters identified through the grid search were further refined using Random Search for optimal results. Results for all hyperparameters in Grid and Random Search are in Supplementary Table S3. The results of the predictions for the validation data set were an R^2 of 0.803 and an RMSE of 0.454.

Table 1 Performance comparison of the top-performing models using 5-fold cross-validation

Model	Fit time (s)	Score time (s)	R^2	RMSE
LGBM	0.04	0.00	0.78	0.480
	0.03	0.00	0.79	0.458
	0.03	0.00	0.81	0.447
	0.03	0.00	0.78	0.469
	0.03	0.00	0.78	0.447
ETRM	0.56	0.02	0.78	0.480
	0.56	0.02	0.78	0.458
	0.57	0.02	0.80	0.458
	0.56	0.02	0.77	0.480
	0.57	0.02	0.78	0.458
MLPNN	0.14	0.00	0.71	0.100
	0.14	0.00	0.72	0.100
	0.14	0.00	0.71	0.100
	0.14	0.00	0.68	0.100
	0.14	0.00	0.70	0.100

Notes: RMSE: Root Mean Squared Error; LGBM: Light Gradient Boosting Model; ETRM: Extra-Trees Regressor Model; MLPNN: Multi-Layered Perceptron Neural Network.

Using the above-mentioned parameters, the LGBM was retested using 5-fold cross-validation. Results were highly consistent, with an R of 0.806 ± 0.01 and an RMSE of 0.440 ± 0.01. Model fine-tuning slightly improved prediction accuracy of compared to the previous analysis.

The learning curve for the fine-tuned model (Supplementary Fig. S2) illustrates the model's adaptability, as it reaches a point of diminishing returns relatively quickly, approximately after 300 observations. This observation suggests that the model has successfully generalized from the data and has effectively captured the underlying patterns in the data set.

The model's final predictions seem very consistent across the entire range of input values (Fig. 4(A)). However, the residual plot (Fig. 4(B)) reveals that the model's predictions exhibit slightly more variance in the lower range of clinical case values.

For the final validation test to assess the model's performance, data from three randomly selected municipalities (Montcada i Reixac, Vallirana, and Vilanova i La Geltrú) were removed from the training data set, specifically the data from July 2021 onwards. Next, predictions were made on that time window in an attempt to recreate the epidemiological curve for these municipalities (Fig. 5). The fine-tuned model accomplished a MAE of 0.23, 0.45, and 0.36 for each municipality respectively, which corresponds to 26, 34, and 83 clinical cases of MAE when applying the inverse transformation to return the variables to its original scale.

3.4 Model explainability

The results from the SHAP framework (Fig. 6) showed that the most important features were the 2-week moving average for the clinical cases, the population size of each municipality, and the 2-week moving average for SARS-CoV-2. These three variables all have a direct relationship with the predicted variable: an increase of any of these features leads to an increase in the model's output, though the effects vary in distributions.

The month used as a seasonality variable proved to be especially relevant. Examining the predictions for each month, we can observe that months with a higher number of clinical cases such as December and January, typically have a greater impact on the model's values (Fig. 7). Similarly, most of the lowest-impact

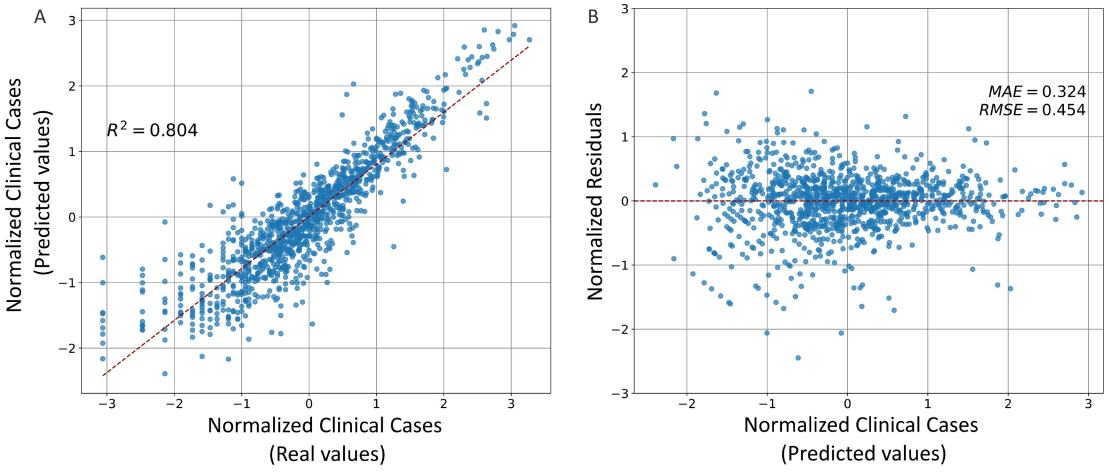


Fig. 4 Analysis of correlation and residuals of the final model predictions on the validation data set. (A) Correlation between real and predicted values; the linear fit is shown in the figure as a red dashed line and R^2 is annotated in the figure. (B) Residuals (predicted values-real values) plotted against the predicted values for clinical cases; MAE and RMSE are annotated in the figure.

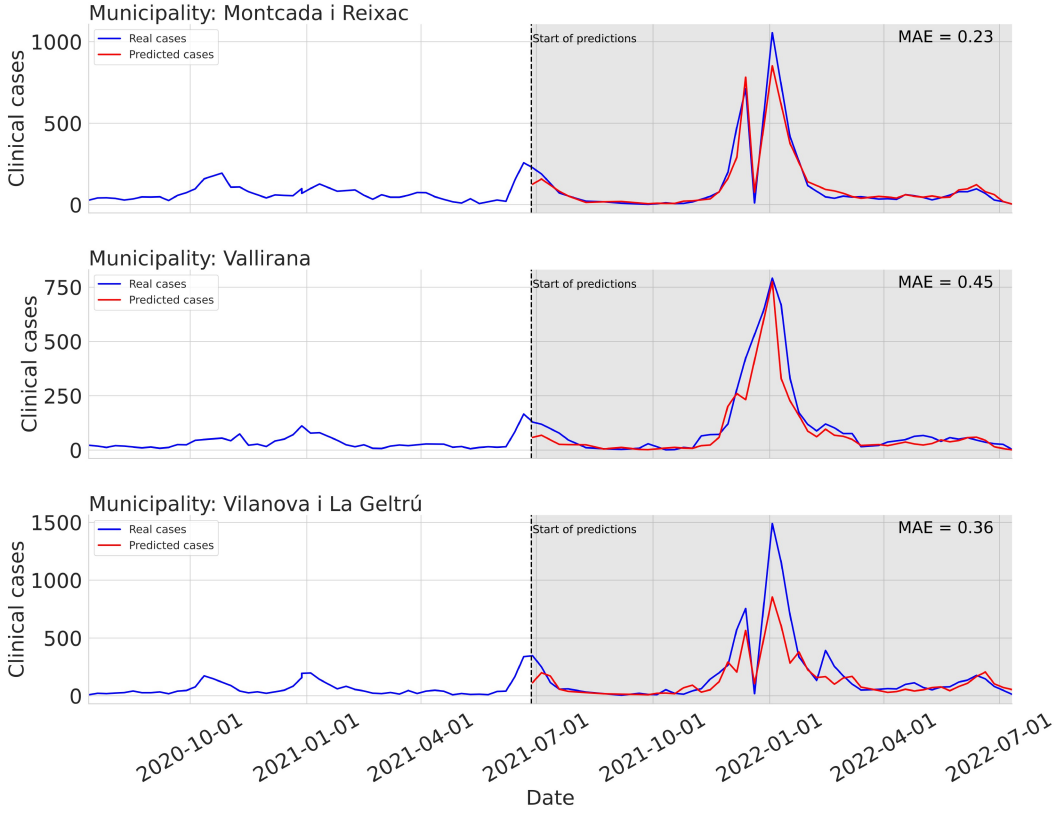


Fig. 5 Recreation of epidemiologic curves for three municipalities of Catalonia: Montcada i Reixac, Vallirana, and Vilanova i La Geltrú. The complete time series of data for the municipalities are represented by a blue line and the predictions by a red line, starting July 2021. Values for MAE for each municipality are annotated in the figure.

values on the model occur in months with fewer clinical cases, such as October.

Interestingly, the current concentration of SARS-CoV-2 in wastewater is the fourth least relevant variable for the model, with higher values of the variable influencing predictions where case numbers

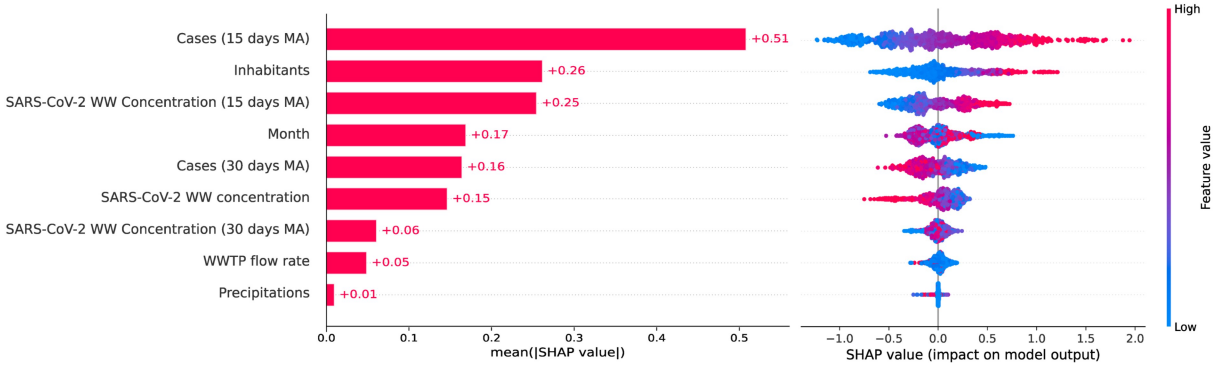


Fig. 6 Model explainability using the SHAP framework. (Left) Feature importance of each variable introduced into the model is represented by the mean absolute SHAP value of each variable across all local predictions. (Right) Beeswarm plot showing the impact of each variable on the model output across all local predictions; the color gradient represents the value for each feature, red for higher values and blue for lower values.

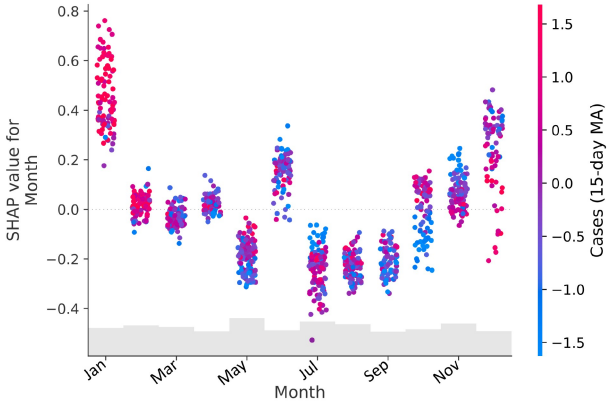


Fig. 7 SHAP dependence plot between seasonality (Month) and clinical cases (15d-MA). The SHAP values for the variable "Month" in each prediction are represented by a gradient color, with red indicating higher values and blue indicating lower values according to the 2-week moving average of clinical cases. This illustrates how the impact of this contextual feature varies.

are lower.

Lastly, the features related to the WWTP, such as water flow and precipitation, had minimal impact on the model and were among the least relevant variables. The impact of water flow appears somewhat random, showing inconsistent SHAP values when the variable's values are high. Nonetheless, with high precipitation, the model's output would shift toward lower values for clinical cases.

4 Discussion

A strong correlation has been shown between SARS-CoV-2 concentrations in wastewater and clinical cases and hospitalizations (López-Peñalver et al., 2023), with the correlation becoming even more pronounced when a lead time of 1–2 weeks is considered. Based on previous research, this study used ML to build predictive models of future outbreaks. These models can usually fit to complex data sets and generate advanced and sophisticated predictions without a complete understanding of the underlying variables, compared to classical statistical models (Gregovic et al., 2023).

Data processing and transformation have enabled better control of model behavior across all data ranges, leading to improved prediction accuracy. Smoothing techniques, such as LOESS, require careful application, due to the potential risk of reducing or even eliminating significant patterns or signals, which could compromise the robustness of the final predictions. Nevertheless, smoothing of viral concentrations has been an explored topic in WBE predictive models to reduce the inherently noisy nature of wastewater data (Arabzadeh et al., 2021). The implementation of LOESS with a narrow span (10%) has proven beneficial in controlling the trade-off between smoothing and signal retention. Additionally, the Box-Cox transformation has proven useful in ML modeling, as it is tailored to each specific variable rather than relying on a generic transformation function like logarithmic transformation (Atkinson et al., 2021; Blum et al., 2022).

The results in this study show the success of various ML models in accurately fitting epidemiological data and virus concentrations in wastewater. Notably, Light Gradient Boosting models achieved the highest accuracy and lowest deviations from the actual data (RMSE = 0.44). In recent years, Gradient Boosting models have gained significant popularity, particularly in the medical and diagnostic fields, due to their remarkable ability to adapt to various data sets

(Karabayir et al., 2020; Adamidi et al., 2021; Li et al., 2022). This finding is further supported when training is performed on large data sets, as the success of these models lies in their ability to effectively generalize across large data sets.

Since one of the objectives of this study was to develop a model that could be broadly applicable to other populations, the data set used included the historical records for the entire population of Catalonia from 2020 to 2022. This data set enabled us to showcase the generalization capability of ML models while ensuring the reliability and accuracy of the predictions. For example, Spearman's rank correlation (ρ) between the normalized SARS-CoV-2 concentration in wastewater and the number of clinical cases is 0.37. Although this may be interpreted as a low correlation, it is influenced by the aggregation of data from different municipalities and the variation across distinct epidemiological waves. The correlation increased between 0.27 and 0.86 when each municipality is assessed separately.

The heterogeneity in the data set contributes to the observed decrease in correlation. Nonetheless, the model has proven its ability to adapt to different populations by incorporating population size as an additional variable and effectively generalizing across various epidemiological waves. This adaptability may be linked to contextual variables associated with clinical cases.

Numerous researchers have employed various ML and statistical models in wastewater-based epidemiology (Ai et al., 2022; Joseph-Duran et al., 2022). However, the inference of clinical cases across different populations within a large region remains relatively unexplored. Some researchers have addressed this issue by making separate predictions for each WWTP. Using the same data set from Catalonia, Joseph-Duran et al. (2022) achieved an $R^2 > 0.9$ between actual and predicted data in most cases. Nonetheless, incorporating a larger number of WWTP and municipalities would require recalibrating the model to accommodate the new data. In this study, we achieved a slightly lower level of accuracy, yet with the advantage of a robust implementation across a wide range of municipalities, without the need for recalibration when incorporating new populations. Ai et al. (2022) used a deep learning model to predict at a multi-community level using data from nine sewersheds, achieving an $R^2 = 0.81$ with a single model.

The aim of this study was to develop a model that provides personalized predictions and risk assessments for each individual municipality. This approach allows for detailed observations of epidemiological progression and its specific characteristics, ultimately enabling public health authorities to make more informed decisions and take appropriate measures to protect the public. Hence, ongoing wastewater-based epidemiology demonstrates its intrinsic value. This approach complements traditional epidemiological surveillance methods by offering non-invasive, cost-effective means to monitor epidemiological patterns during outbreaks within communities. It also provides accurate estimates of COVID-19 incidence relative to population size. This tool has the potential to assist in the early detection of emerging COVID-19 outbreaks, such as the recent surge in cases in June 2024, thereby facilitating the prompt implementation of necessary control measures.

The wastewater-based epidemiological approach has also proven effective in detecting other biomarkers of infectious diseases, such as the genetic material of influenza viruses (Zheng et al., 2023), enteroviruses (Randazzo et al., 2019) or the monkeypox virus (Tiwari et al., 2023). For other pathologies of infectious nature, integrating ML and predictive models into this approach would significantly enhance its effectiveness in proactively averting potential outbreaks or adverse situations. The feature engineering and model building guidance presented in this study could be adapted to expand the range of pathogens detected by wastewater-based epidemiological surveillance.

Additionally, it is important to recognize that model explainability is crucial in the development of predictive models. Not only does this allow deeper insight into the underlying processes and a better understanding of the covariates associated with the epidemiological evolution in that particular scenario, but it also eliminates the concept of predictive algorithms as opaque entities. Instead, it provides transparent, interoperable, and explainable solutions.

A major limitation in conducting this research was the availability of open data. Specifically, data on virus concentration in WWTP are still reported weekly in the Catalan Surveillance Network of SARS-CoV-2 in Sewage. However, clinical data for SARS-CoV-2 infections ceased in July 2022. This is crucial for building predictive models, as larger data sets enable algorithms that are more robust and more accurate predictions. The relevance of open data cannot be overstated, as it empowers researchers to develop effective models and fosters collaborative efforts to address these challenging scenarios.

Another limitation of this work is the reliance on pre-existing data to train the predictive models and the difficulty in extrapolating beyond the reference data,

which poses a challenge in the early stages of a crisis, such as a pandemic or an outbreak of a novel pathogen. However, in these initial stages, model explainability can help quickly identify key variables influencing the epidemic's progression and facilitate the implementation of preventive measures to mitigate its impact.

This study has revealed the crucial role of context variables through model explainability, underscoring their importance in providing accurate predictions. However, it is crucial to recognize that mishandling context variables can lead to data leakage in predictive models. This issue was thoroughly addressed in the study through a rigorous selection and training process. Among the context variables, previous infections emerged as one of the most important, though it was not the sole variable considered for predictions. Interestingly, current SARS-CoV-2 concentration in wastewater was found to be less relevant for the model compared to other features. This is somewhat surprising but could be explained by lagging effects between viral concentration in wastewater and clinical cases. The population size also significantly influenced predictions, indicating the model's ability to generalize across different population sizes. Additionally, seasonality was a key factor, with the Month variable being the fourth most relevant to the model. Higher infection rates were observed in December and January, which corresponds with the epidemiological data of the region, a fact that is also reflected in the predictions through the SHAP values. Colder months are known to promote virus transmission, influenced by various factors including environmental conditions (high temperatures shorten the half-life of the virus) or human behavior (Christmas Holidays increase travel between regions and contacts between individuals) (Zoran et al., 2022).

WWTP flow rate and precipitation both contribute to the dilution of viral RNA in wastewater. Nevertheless, neither of these variables seem to have a substantial impact on the predictions, only certain high precipitation values that shift the model to lower clinical cases predictions.

Future studies could benefit from including variables that affect the stability and concentration of the virus in model evaluations, such as daily temperature. Additionally, parameters related to case prevalence, such as the age distribution within each municipality, should be considered. Finally, the developed model will be tested in other infectious pathologies with airborne transmission patterns similar to SARS-CoV-2, such as the influenza virus. This testing will serve as a foundation for creating new models for other epidemic outbreaks that share some of the variables used in the current predictive model.

5 Conclusions and future perspectives

Wastewater-based epidemiology is becoming an essential tool for monitoring emerging outbreaks and supporting public health authorities. The success of the wastewater-based epidemiology approach presented in this study could enhance the surveillance of emerging pathogens and novel outbreaks. Additionally, this work shows how integrating ML or predictive algorithms into routine practice can enhance the tool's accuracy and validity in predicting clinical cases. This capability can aid health authorities anticipate challenging situations and make informed decisions to address potential issues.

This research also explores the relevance of model explainability, which is essential for building confidence in predictive modeling. By providing valuable insights into how the model makes predictions, explainability enhances the reliability and transparency of the model. A significant breakthrough in this study is the identification of the crucial impact of contextual variables, such as the moving averages of past cases and seasonality, on the model's output and accuracy.

Further research is needed to assess the effectiveness of this methodology in predicting new COVID-19 outbreaks and its applicability to other infectious diseases to fully validate its potential. Moreover, enhanced accuracy and general improvements in model performance could be attained through the integration of additional contextual variables pertaining to environmental conditions (e.g., ambient temperature), demographic data (e.g., age distribution, mobility patterns), and social factors (e.g., vaccine coverage data). Another prospect would be to apply the model to a wider range of geographical regions, encompassing more diverse environmental conditions and population densities, with a view to enhance the robustness and generalizability of the predictions.

Abbreviations

COVID-19: Coronavirus Disease 2019

ETRM: Extra-trees regressor model

INE: National Institute of Statistics (Spain)

LGBM: Light gradient boosting model

LOESS: Locally estimated scatterplot smoothing

MAE: Mean absolute error

MLPNN: Multi-layered perceptron neural network

PCR: Polymerase chain reaction

RMSE: Root mean squared error

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

SHAP: Shapley additive explanations

WBE: Wastewater-based epidemiology

WHO: World Health Organization.

WWTP: Wastewater treatment plant

CRediT Authorship Contribution Statement

Mr. Ruben Cañas Cañas: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - original draft, Writing - review and editing. **Dr. Seguí López-Peñalver:** Methodology, Formal analysis, Data Curation, Resources, Writing - original draft, Writing - review and editing. **Dr. Casaña Mohedo:** Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review and editing. **Dr. Benavent Cervera**: Conceptualization, Investigation, Resources, Formal analysis, Writing - review and editing. **Dr. Fernández Garrido:** Conceptualization, Investigation, Resources, Formal analysis, Writing - review and editing. **Dra. Pellín Carcelén:** Methodology, Investigation, Resources, Writing - review and editing. **Dr. García-Algar:** Methodology, Investigation, Resources, Writing - review and editing. **Dr. Juárez Vela:** Methodology, Investigation, Resources, Writing - review and editing. **Dr. Gea Caballero:** Conceptualization, Investigation, Visualization, Validation, Supervision, Writing - review and editing, Formal analysis. **Dr. Andreu-Fernández:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Funding acquisition, Supervision, Project administration, Writing - original draft, Writing - review and editing. All authors critically reviewed the manuscript and approved the final manuscript as submitted.

Acknowledgements This research was funded by the Valencian International University and Generalitat Valenciana (GVA) through the Grants to emerging research groups 2023 (CE2023) from the Regional Ministry of Innovation, Universities, Science and Digital Society (CIGE/2022/58). We would like to thank the Catalan Surveillance Network of SARS-CoV-2 in Sewage and the government of Catalonia for generating high-quality data and making it publicly available for research. We also would like to thank the Biosanitary Research Institute of Valencian International University (VIU) and Global Omnium for their support in the development of this project.

Conflict of Interests 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.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11783-025-1932-8 and is accessible for authorized users.

References

Adamidi E S, Mitsis K, Nikita K S (2021). Artificial intelligence in clinical care amidst COVID-19 pandemic: a systematic review. Computational and Structural Biotechnology Journal, 19: 2833–2850

Ai Y, He F, Lancaster E, Lee J (2022). Application of machine learning for multi-community COVID-19 outbreak predictions with wastewater surveillance. PLoS One, 17(11): e0277154

Arabzadeh R, Grünbacher D M, Insam H, Kreuzinger N, Markt R, Rauch W (2021). Data filtering methods for SARS-CoV-2 wastewater surveillance. Water Science and Technology, 84(6): 1324–1339

Atkinson A C, Riani M, Corbellini A (2021). The Box–Cox transformation: review and extensions. Statistical Science, 36(2): 239–255

Blum L, Elgendi M, Menon C (2022). Impact of Box-Cox transformation on machine-learning algorithms. Frontiers in Artificial Intelligence, 5: 877569

Booth A L, Abels E, Mccaffrey P (2021). Development of a prognostic model for mortality in COVID-19 infection using machine learning. Modern Pathology, 34(3): 522–531

Chadaga K, Prabhu S, Vivekananda B K, Niranjana S, Umakanth S (2021). Battling COVID-19 using machine learning: a review. Cogent Engineering, 8(1): 1958666

Chen H, Chen Z, Hu L, Tang F, Kuang D, Han J, Wang Y, Zhang X, Cheng Y, Meng J, et al. (2024). Application of wastewater-based epidemiological monitoring of COVID-19 for disease surveillance in the city. Frontiers of Environmental Science & Engineering, 18(8): 98

Clement T, Kemmerzell N, Abdelaal M, Amberg M (2023). XAIR: A systematic metareview of explainable AI (XAI) Aligned to the software development process. Machine Learning and Knowledge Extraction, 5(1): 78–108

Daza-Torres M L, Montesinos-López J C, Kim M, Olson R, Bess C W, Rueda L, Susa M, Tucker L, García Y E, Schmidt A J, et al. (2023). Model training periods impact estimation of COVID-19 incidence from wastewater viral loads. Science of the Total Environment, 858(Pt 1): 159680

Generalitat de Catalunya (2023). Register of COVID-19 tests performed in Catalonia. Catalunya: Generalitat de Catalunya

Gregovic M, Filipovic L, Katnic I, Vukotic M, Popovic T (2023). Machine learning models for statistical analysis. The International Arab Journal of Information Technology, 20

(Special Issue 3A): 505–514

Guerrero-Latorre L, Collado N, Abasolo N, Anzaldi G, Bofill-Mas S, Bosch A, Bosch L, Busquets S, Caimari A, Canela N, et al. (2022). The Catalan surveillance network of SARS-CoV-2 in sewage: design, implementation, and performance. Scientific Reports, 12(1): 16704

Hill D T, Alazawi M A, Moran E J, Bennett L J, Bradley I, Collins M B, Gobler C J, Green H, Insaf T Z, Kmush B, et al. (2023). Wastewater surveillance provides 10-days forecasting of COVID-19 hospitalizations superior to cases and test positivity: a prediction study. Infectious Disease Modelling, 8(4): 1138–1150

Instituto Nacional de Estadística (2023). Population by Municipality. Madrid: Instituto Nacional de Estadística

Islam S, Islam T, Islam M R (2022). New coronavirus variants are creating more challenges to global healthcare system: a brief report on the current knowledge. Clinical Pathology, 15: 2632010X221075584

Jeng H A, Singh R, Diawara N, Curtis K, Gonzalez R, Welch N, Jackson C, Jurgens D, Adikari S (2023). Application of wastewater-based surveillance and copula time-series model for COVID-19 forecasts. Science of the Total Environment, 885: 163655

Joseph-Duran B, Serra-Compte A, Sàrrias M, Gonzalez S, López D, Prats C, Català M, Alvarez-Lacalle E, Alonso S, Arnaldos M (2022). Assessing wastewater-based epidemiology for the prediction of SARS-CoV-2 incidence in Catalonia. Scientific Reports, 12(1): 15073

Karabayir I, Goldman S, Pappu S, Akbilgic O (2020). Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Medical Informatics and Decision Making, 20(1): 228

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T Y (2017). LightGBM: a Highly Efficient Gradient Boosting Decision Tree. Long Beach: Curran Associates Inc.

Kumar M, Joshi M, Patel A K, Joshi C G (2021). Unravelling the early warning capability of wastewater surveillance for COVID-19: a temporal study on SARS-CoV-2 RNA detection and need for the escalation. Environmental Research, 196: 110946

Lalmuanawma S, Hussain J L C (2020). Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review. Chaos, Solitons & Fractals, 139: 110059

Li K, Yao S, Zhang Z, Cao B, Wilson C M, Kalos D, Kuan P F, Zhu R, Wang X (2022). Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics, 38(6): 1631–1638

Liao X, Liu X, He Y, Tang X, Xia R, Huang Y, Li W, Zou J, Zhou Z, Zhuang M (2024). Alternate disinfection approaches or raise disinfectant dosages for sewage treatment plants to address the COVID-19 pandemic? From disinfection efficiency, DBP formation, and toxicity perspectives. Frontiers of Environmental Science & Engineering, 18(9): 115

López-Peñalver R S, Cañas-Cañas R, Casaña-Mohedo J, Benavent-Cervera J V, Fernández-Garrido J, Juárez-Vela R, Pellín-Carcelén A, Gea-Caballero V, Andreu-Fernández V (2023). Predictive potential of SARS-CoV-2 RNA concentration in wastewater to assess the dynamics of COVID-19 clinical outcomes and infections. Science of the Total Environment, 886: 163935

Lu X, Wang L, Sakthivel S K, Whitaker B, Murray J, Kamili S, Lynch B, Malapati L, Burke S A, Harcourt J, Tamin A, Thornburg N J, Villanueva J M, Lindstrom S (2020). US CDC real-time reverse transcription PCR panel for detection of severe acute respiratory syndrome Coronavirus 2. Emerging Infectious Diseases, 26(8): 1654–1665

Lundberg S M, Lee S I (2017). A unified approach to interpreting model predictions. Long Beach: Curran Associates Inc., 4768–4777

Marimuthu S, Mani T, Sudarsanam T D, George S, Jeyaseelan L (2022). Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal in medical research. Clinical Epidemiology and Global Health, 15: 101043

Pirzada R H, Ahmad B, Qayyum N, Choi S (2023). Modeling structure–activity relationships with machine learning to identify GSK3-targeted small molecules as potential COVID-19 therapeutics. Frontiers in Endocrinology, 14: 1084327

R Core Team (2024). R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing

Randazzo W, Cuevas-Ferrando E, Sanjuán R, Domingo-Calap P, Sánchez G (2020). Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. International Journal of Hygiene and Environmental Health, 230: 113621

Randazzo W, Piqueras J, Evtoski Z, Sastre G, Sancho R, Gonzalez C, Sánchez G (2019). Interlaboratory comparative study to detect potentially infectious human enteric viruses in influent and effluent waters. Food and Environmental Virology, 11(4): 350–363

Santangelo O E, Gentile V, Pizzo S, Giordano D, Cedrone F (2023). Machine learning and prediction of infectious diseases: a systematic review. Machine Learning and Knowledge Extraction, 5(1): 175–198

Sarker R, Roknuzzaman A S M, Nazmunnahar, Shahriar M, Hossain M J, Islam M R (2023). The WHO has declared the end of pandemic phase of COVID‐19: way to come back in the normal life. Health Science Reports, 6(9): e1544

Schneider K A, Tsoungui Obama H C J, Adil Mahmoud Yousif N (2023). A flexible age-dependent, spatially-stratified predictive model for the spread of COVID-19, accounting for multiple viral variants and vaccines. PLoS One, 18(1): e0277505

Shang M, Kong Y, Yang Z, Cheng R, Zheng X, Liu Y, Chen T (2023). Removal of virus aerosols by the combination of filtration and UV-C irradiation. Frontiers of Environmental Science & Engineering, 17(3): 27

Shapley L S (1952). A Value for n-Persons Games. Santa Monica: The Rand Corporation

Silva J A (2023). Wastewater treatment and reuse for sustainable water resources management: a systematic literature review. Sustainability, 15(14): 10940

Tiwari A, Adhikari S, Kaya D, Islam M A, Malla B, Sherchan S P, Al-Mustapha A I, Kumar M, Aggarwal S, Bhattacharya P, et al. (2023). Monkeypox outbreak: wastewater and environmental surveillance perspective. Science of the Total Environment, 856: 159166

Vallejo J A, Trigo-Tasende N, Rumbo-Feal S, Conde-Pérez K, López-Oriona A, Barbeito I, Vaamonde M, Tarrío-Saavedra J, Reif R, Ladra S, et al. (2022). Modeling the number of people infected with SARS-CoV-2 from wastewater viral load in Northwest Spain. Science of the Total Environment, 811: 152334

van Rossum G (1995). Python reference manual. Amsterdam: Centrum voor Wiskunde en Informatica

Vandenberg O, Martiny D, Rochas O, Van Belkum A, Kozlakidis Z (2021). Considerations for diagnostic COVID-19 tests. Nature Reviews. Microbiology, 19(3): 171–183

Weinan E (2020). Machine learning and computational mathematics. Communications in Computational Physics, 28(5): 1639–1670

Wickham H, Averick M, Bryan J, Chang W, Mcgowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43): 1686

Zheng X, Zhao K, Xu X, Deng Y, Leung K, Wu J T, Leung G M, Peiris M, Poon L L M, Zhang T (2023). Development and application of influenza virus wastewater surveillance in Hong Kong. Water Research, 245: 120594

Zhu Y, Oishi W, Maruo C, Bandara S, Lin M, Saito M, Kitajima M, Sano D (2022). COVID-19 case prediction via wastewater surveillance in a low-prevalence urban community: a modeling approach. Journal of Water and Health, 20(2): 459–470

Zoran M A, Savastru R S, Savastru D M, Tautan M N, Baschir L A, Tenciu D (2022). Assessing the impact of air pollution and climate seasonality on COVID-19 multiwaves in Madrid, Spain. Environmental Research, 203: 111849