



OPEN Machine learning comparison for biomarker level estimation in wastewater dynamics monitoring

Francisco Javier Maldonado Carrascosa^{1✉}, Sebastián García Galán¹, Paweł Wityk², Aneta Łuczkiwicz³ & Małgorzata Szczerska⁴

Wastewater surveillance is an emerging strategy that enables monitoring of the presence and dynamic changes of targeted substances, facilitating improved allocation of preventive actions and public health interventions. This paper investigates the application of machine learning models for classifying wastewater samples based on varying concentrations of C-Reactive Protein (CRP), a critical biomarker for inflammation, whose levels may rise due to the presence of certain drugs. Using absorption spectroscopy spectra, classification tasks were conducted to distinguish between five concentration classes ranging from zero to 10^{-1} $\mu\text{g/ml}$. Rather than relying on a single model, this study evaluates and compares multiple machine learning algorithms to determine the most effective approach for this classification task. Additionally, performance metrics including accuracy, precision, recall, F1 score, and specificity were calculated for each model. The comparative analysis revealed accuracies ranging from 64.88% to 65.48% for the best model, Cubic Support Vector Machine (CSVM), using both full-spectrum and restricted-range spectral data. Confusion matrices and Receiver Operating Characteristic (ROC) curves are presented to visually interpret classification performance. The results highlight the potential of machine learning techniques to moderately classify CRP levels in wastewater, offering promising insights for future biosensor development and real-time environmental monitoring.

Keywords Wastewater surveillance, Machine learning, C-reactive protein, Biomarker detection, Optical spectroscopy

The world today faces various challenges, from viruses outbreaks to drug spreading^{1,2}, creating a need for fast response actions preventing or minimizing the potential dangerous consequences. The sensitive and quick detection of molecules, with real-time surveillance and automated data analysis, can deliver an extremely effective tool for public health condition monitoring and its protection with targeted actions at the point of need³. Traditional laboratory methods like enzyme-linked immunosorbent assay (ELISA) or nucleic acid sequence-based amplification (NASBA) are used for samples investigations which deliver reliable and highly-accurate information^{4,5}. However, this approach requires much time (from sample taking to the result), trained personnel and significant funds for devices and reagents⁶. As a result, scaling of such methodology is not possible, creating a need for a comprehensive system capable of monitoring larger areas and their populations. Recent research highlights that traditional diagnostic systems often suffer from latency and network congestion when processing large healthcare datasets through centralized cloud infrastructures. A novel regional computing approach has been proposed to address these challenges, allowing decentralized data handling and improving real-time response capabilities in public health surveillance⁷.

An interesting strategy to solve this problem is wastewater monitoring, especially in the light of the latest advancements in biosensing^{4,8–10}. Wastewater-based epidemiology (WBE) involves analyzing sewage to monitor the health of a population, focusing on detecting human health-related molecules. For example, viral particles can traverse to sewage pipelines through infected people body fluids and excretions, introducing the viral material in wastewater samples^{11–13}. In over 100 studies across the world, infectious diseases and pathogens were detected through wastewater surveillance¹⁴. Devianto and Sano have reviewed and identified protein markers

¹Telecommunication Engineering Department, University of Jaén, Linares, Jaén, Spain. ²Biotechnology and Microbiology Department, Faculty of Chemistry, Gdansk University of Technology, Gdansk, Poland. ³Technology in Environmental Engineering Department, Faculty of Civil and Environmental Engineering, Gdansk University of Technology, Gdansk, Poland. ⁴Metrology and Optoelectronics Department, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland. ✉email: fjmaldon@ujaen.es

that are found in relatively high concentrations within wastewater samples, that can be useful for WBE systems development¹⁵. Hence, this strategy allows to identify potential threats: by analyzing wastewater it is possible to detect and track the targeted molecules, like pathogens, proteins, pharmaceuticals, chemical contaminants, and other substances^{6,15–18}. A study analyzing crude wastewater from South Wales treatment plants demonstrated the robust capability of WBE to detect not only illicit drugs such as cocaine and cannabis but also medicinal compounds like buprenorphine and methadone. This supports the growing use of WBE in tracking a wide range of chemical and pharmaceutical contaminants, emphasizing its relevance for both public health and environmental monitoring¹⁹.

Installation of such system in municipalities, like schools, hospitals, city offices, sewage treatment plants etc. would provide a net able to indicate the area needing intervention. If a given substance is detected or its level changes, the system will automatically notify authorities about the situation, providing a basis for further actions. Therefore, such system can provide a continuous overview of the community health, follow the changes in tracked substance dynamics, and guide timely interventions in the place of need, saving time and resources of the services²⁰. Empirical evaluations from China show that implementation of smart city infrastructure, which often includes enhanced wastewater management technologies, significantly improves sewage treatment capabilities. The integration of technological and socio-economic tools through smart city frameworks enhances wastewater monitoring and treatment effectiveness across urban environments²¹.

To date, several research groups reported successful detection of tracked molecules in wastewater samples. Alvarez-Serna et al. presented a label-free and portable field-effect transistor (FET) sensor coupled with reverse transcription loop-mediated isothermal amplification (RT-LAMP) reaction for detecting SARS-CoV-2 genome in wastewater samples. The developed sensor enabled detection and qualification of nucleic-acids in wastewater samples within 30 min and has shown potential to other pathogens customization²². Sharma et al. developed a field-deployable system, based on Nanotrap microbiome particles and RNA Isothermal Co-Assisted and Coupled Amplification¹¹. The system allows targeted virus detection in wastewater, with the whole procedure taking around 60 min. Guerrero-Estaban et al. reported a carbon nanodot-amplified electrochemiluminescence immunosensor for Spike S1 protein from SARS-CoV-2 coronavirus detection in wastewater. The disposable platform has shown a low detection limit and broad linear response²³. Sen et al. proposed integration of an electrochemical aptamer-based sensor with an FPE (filtration, purification, extraction) system for in-field wastewater surveillance. The platform of 50x70 cm allows detection in less than 60 min²⁴. Sokołowski et al. demonstrated an optical method based on spectroscopy for tracking dynamics of C-Reactive Protein (CRP) levels in complex biological matrix²⁰. The primary benefit of the proposed method lies in its use of an optical system that enables near-real-time monitoring of CRP level fluctuations and permits its installation with minimal alterations to sewage infrastructure. With the use of machine learning (ML) algorithms, the system constitutes a promising platform for real-time CRP monitoring. Jagadeesan et al. presented a proof of concept for wastewater-based infectious diseases warning system²⁵. The mass spectrometry-based approach was presented for concurrent monitoring of SARS-CoV-2 and CRP, chosen as representatives for pathogen and protein occurring in wastewater samples.

In this study, a concept of a point-of-need system able to monitor the wastewater samples and classify them using ML model CSVM is proposed. For the purpose of this research, a well-known and tested biomarker—CRP—has been used and added its known concentrations to the wastewater samples, however, it is also possible to detect other substances using the presented approach. CRP serves as a critical biomarker for inflammation and infection within the human body²⁶. Traditionally utilized in clinical settings for diagnosing various ailments, recent research has extended its utility to unconventional domains, notably wastewater analysis. A recent study confirmed the feasibility of CRP detection in human urine using an optical interferometric biosensor integrated with ML classifiers, achieving up to 100% classification accuracy. This supports the continued exploration of CRP in WBE systems and highlights the potential of optical and data-driven methods in health monitoring through wastewater analysis^{27,28}.

CRP serves as a robust and practical biomarker for assessing the inflammatory pathways activated by environmental stressors. Its responsiveness to a broad spectrum of pollutants^{29,30}, medical conditions³¹ or psychosocial stress^{32,33}, coupled with the accessibility of its measurement, positions CRP as a critical tool for environmental health surveillance. By integrating CRP monitoring into public health strategies, researchers and officials can better understand the population-level impacts of environmental quality, identify vulnerable communities, and evaluate the effectiveness of interventions aimed at mitigating environmental health risks. Mounting evidence supports the use of CRP, a sensitive marker of systemic inflammation—bacterial, viral, cancerous and others^{34,35}, as a valuable indicator for monitoring the health impacts of environmental exposures. This readily measurable biomarker offers a means to assess the physiological stress induced by a wide range of pollutants, providing a crucial link between environmental quality and public health^{29,30}. Scientific research has increasingly demonstrated a strong association between elevated levels of CRP and exposure to various environmental contaminants. Initially recognized for its link to cardiovascular disease, studies have now established connections between heightened CRP concentrations and exposure to ambient air pollutants, including particulate matter (PM_{2.5}) and ozone^{36,37}. This association holds true for both short-term and long-term exposures across diverse populations, encompassing children, healthy adults, and the elderly. Beyond air quality, investigations have revealed a similar inflammatory response, indicated by increased CRP, following exposure to other common environmental toxins. Notably, studies have linked elevated CRP levels to chronic exposure to heavy metals such as lead, mercury, cadmium, and arsenic^{38,39}. Furthermore, emerging research suggests a correlation between pesticide exposure and systemic inflammation as measured by CRP⁴⁰. The utility of CRP as an environmental health indicator is underscored by several key advantages. Elevated CRP can result from various conditions, including infections, chronic diseases, and lifestyle factors such as smoking and obesity. The development of high-sensitivity wastewater CRP sensing allows for the precise detection of inflammation,

offering a more nuanced understanding of populations chronic or increased environmental exposures and health.

In addressing the significance of monitoring targeted substances in wastewater, this study undertakes the data analysis and classification stage, as the majority of WBE systems reports focuses on the sensors development. Two distinct classification tasks are presented in order to highlight the use of ML in the field. The first classification focuses on discerning CRP presence and concentration thresholds within wastewater samples using the complete UV–Vis spectrum data. This classification is made across five distinct CRP concentration levels ranging from 10^{-4} µg/ml to 10^{-1} µg/ml. In parallel, the paper extends its analysis by conducting an additional classification restricted to a specific spectral range—utilizing solely the spectrum from 400 nm to 700 nm, which not only streamlines computational resources but also holds potential cost-saving implications for the potential development of biosensors currently in the pipeline. This spectral-based approach aligns with recent findings where UV–Vis spectroscopy was successfully applied to assess the removal of complex dye compounds in textile wastewater using biosynthesized nanoparticles. This confirms its applicability as a reliable tool for detecting biomarkers and environmental contaminants⁴¹. Thus, the contributions of this paper are the following:

- The application of machine learning techniques to the monitoring of CRP concentrations in wastewater has been explored, providing valuable insights into environmental sample analysis.
- The dynamism of CRP concentration levels in wastewater has been illustrated through the use of multiple classes, enabling the observation of varying CRP concentrations over time within the same wastewater region. This multi-class approach allows for capturing temporal fluctuations and spatial variability, reflecting the dynamic nature of CRP presence in complex wastewater matrices.
- A 5-class classification scheme, distinguishing five different CRP concentration levels, has been implemented for wastewater samples, representing a novel categorization not previously reported in the literature.

In summary, the primary objective of this study is to be achieved through the development and evaluation of machine learning models for the classification of CRP concentrations in wastewater samples using UV–Vis spectra. Specifically, the CSVM algorithm is applied within a multi-class framework to distinguish between five concentration levels of CRP. Two classification tasks are conducted: one utilizing the broadband spectral range (220–750 nm) and another limited to a spectral range (400–750 nm), reflecting potential cost and computational efficiencies for future biosensor development. Model performance is assessed through repeated experiments to ensure robustness and reproducibility. These objectives, including the design, training, and evaluation of the CSVM models, are described in more detail later in the manuscript.

The structure of the paper is as follows: section “[Materials and methods](#)” describes the materials and methods used, including data collection, preprocessing, and the application ML. Section “[Results and discussion](#)” details the results of the classification tasks, comparing model performances across multiple metrics. Finally, section “[Conclusion](#)” concludes the study with a summary of key insights and discusses potential implications for future developments in wastewater monitoring and environmental biosensing.

Materials and methods

This section outlines the methodology adopted to investigate the classification of CRP concentrations in municipal wastewater samples using UV–Vis spectrometry measured signal and machine learning techniques. It begins with a description of the dataset, which consists of real influent samples spiked with known CRP concentrations. The spectral characteristics of these samples are then discussed, followed by a summary of the classification approach, including the selection of spectral markers. Finally, the machine learning models employed for multi-class classification are introduced, detailing their configurations and relevance to the study’s objectives. Together, these subsections provide a comprehensive overview of the experimental and computational framework that underpins the analysis.

Dataset

In this study, a dataset comprising 840 distinct wastewater samples, each exhibiting varying concentrations of CRP (including samples without CRP), serves as the foundational basis for analysis. These samples encapsulate a diverse spectrum of CRP levels, ranging from negligible concentrations to notable quantities, reflecting the dynamic composition inherent to wastewater matrices. The base for dataset construction were measurements carried out with the spectrophotometer (NanoDrop ND-1000, Thermo Fisher Scientific Inc., Waltham, MA, USA). The methodology, samples description and measurement process description can be found elsewhere²⁰.

Although the dataset comprises 840 individual samples, it is important to highlight that all were derived from real municipal wastewater. Specifically, 14 composite influent samples were collected between August and October 2023 from the Gdynia-Dębogórze Wastewater Treatment Plant (WWTP), located along the Baltic Sea coast in northern Poland. This WWTP is the second-largest facility of its kind in the region and serves both the city of Gdynia and neighboring municipalities. A 24-h composite, flow-proportional sampling procedure was employed to represent the daily composition of raw municipal wastewater. The influent stream is composed predominantly of domestic wastewater, with industrial and hospital sources contributing approximately 1% and 0.1%, respectively. During the sampling campaign, the plant operated at a hydraulic load of roughly 450,000 population equivalents (PE), with an average daily influent flow of $61,886.1 \pm 3760.2$ m³/day. The treatment process employed is mechanical-biological, including advanced nutrient removal and occasional chemical phosphorus precipitation.

The physicochemical characteristics of the collected influent samples reflected typical raw wastewater complexity. On average, samples showed a chemical oxygen demand (COD) of 1268.3 ± 203.6 mg O₂

/L, biochemical oxygen demand (BOD_5) of 614.2 ± 149.2 mg O_2 /L, and total suspended solids (TSS) of 561.7 ± 90.0 mg/L. Nitrogen and phosphorus levels were also characteristic of high-load influents, with total nitrogen (TN) at 97.2 ± 5.2 mg N/L, ammonium nitrogen ($N-NH_4^+$) at 68.8 ± 2.7 mg N/L, total phosphorus (TP) at 12.3 ± 2.3 mg P/L and orthophosphates ($P-PO_4^{3-}$) at 5.9 ± 0.1 mg P/L. Other measured parameters included pH of 8.0 ± 0.1 , and conductivity of 936.0 ± 73.7 $\mu S/cm$. These values confirm that the experimental conditions were grounded in the complex and variable matrix of real influent wastewater, as encountered in full-scale operational facilities.

To enable ML classification of CRP levels, controlled spiking of CRP into the real wastewater samples was performed. This approach preserved the authentic background variability and interferences of raw municipal wastewater, while allowing for reliable class labeling and model evaluation. As no synthetic matrices or laboratory-prepared waters were used, the dataset represents a controlled experimental design built upon genuine environmental samples. However, while the complexity of the matrix strengthens the ecological relevance of the study, it is acknowledged that the generalizability of the models could be further enhanced by validating them on wastewater from other geographical locations or treatment configurations. Such external validation would confirm the robustness of the models under different operational and environmental conditions.

All wastewater samples utilized in this study were obtained from municipal influent streams of municipal WWTP collected over multiple days, thereby inherently reflecting the natural temporal and compositional variability characteristic of real wastewater matrices. These samples comprised complex mixtures of organic, inorganic, and colloidal constituents, including variable concentrations of nitrogen and, phosphorus species, organic carbon compounds, and suspended solids, without artificial filtration or significant matrix alteration beyond minimal preprocessing for spectrophotometric analysis. Controlled additions of CRP were performed solely to establish known concentration classes for model training and evaluation. Consequently, the dataset preserves the authentic physicochemical heterogeneity and spectral interferences present in operational wastewater treatment environments. This approach ensures that the reported classification performance accounts for the challenges associated with real-world sample complexity, enhancing the ecological validity and practical relevance of the developed machine learning models.

The size and breadth of this dataset facilitate robust statistical analyses and ML model training across the classification tasks outlined in the research. By encompassing a wide array of CRP concentrations, the dataset affords a comprehensive understanding of CRP distribution patterns within wastewater, laying the groundwork for nuanced insights into the interplay between CRP levels and environmental dynamics.

Classification

For the classification task, 176 markers were considered, each coinciding with a specific point of the UV–Vis absorption spectrum. These markers correspond to 176 distinct points along the wavelength, providing a comprehensive dataset that captures the spectral characteristics necessary for distinguishing between different classes of wastewater samples. The wavelength range for each spectrum was 220–750 nm with accuracy of 1 nm.

Additionally, for the restricted range (400–720 nm) of spectrum classification, 116 markers were chosen. By leveraging this detailed spectral information, the classification model can effectively identify subtle variations in the absorption profiles, which are indicative of the presence and concentration of CRP in the wastewater. This approach ensures a robust analysis by utilizing the full breadth (or a restricted range) of the absorption spectrum, enhancing the accuracy and reliability of the classification results. In Fig. 1 a representation of the UV–Vis spectroscopy-based absorption spectrum can be observed.

Feature engineering in the present study was primarily conducted through explicit marker selection based on spectral resolution, encompassing 176 markers across the full UV–Vis wavelength range of 220–750 nm, and 116 markers within a restricted range of 400–720 nm. This comprehensive selection of spectral points provided

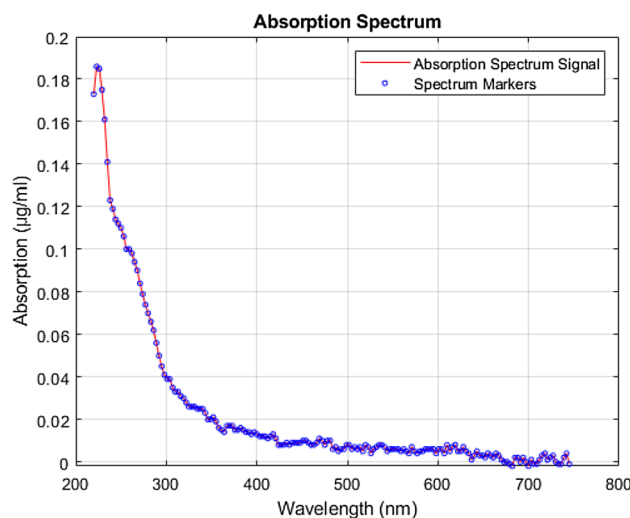


Fig. 1. Absorption spectrum signal and markers from UV–Vis spectrometry.

a detailed dataset capturing the absorption characteristics relevant for CRP classification in wastewater samples. The dataset used in this study includes five distinct CRP concentration classes, distributed as follows: wastewater with no detectable CRP (176 samples), CRP at 10^{-4} $\mu\text{g/ml}$ (166 samples), 10^{-3} $\mu\text{g/ml}$ (169 samples), 10^{-2} $\mu\text{g/ml}$ (165 samples), and 10^{-1} $\mu\text{g/ml}$ (164 samples). This relatively balanced sample distribution across all classes minimizes potential bias and supports robust and fair model training and evaluation. No manual spectral transformations, such as derivatives, integrals, or spectral band ratios, were applied. Instead, higher-order interactions among spectral features were implicitly modeled via the cubic polynomial kernel of the Cubic Support Vector Machine (CSVM) within the Error-Correcting Output Codes (ECOC) framework. This model-driven approach enabled the capture of complex, non-linear relationships between absorption spectra and CRP concentration classes without handcrafted feature engineering. Although detailed analysis of spectral region importance was not the primary focus, comparable classification accuracies exceeding 65% were obtained using both spectral range and the restricted 400–720 nm region, suggesting that the visible spectrum alone contains a substantial portion of the discriminative information. This finding also indicates the potential for hardware simplification in future sensor development. Further internal analyses revealed that wavelengths corresponding to amide and aromatic ring absorption zones (approximately 400–500 nm), as well as protein- and lipid-associated shoulders (approximately 600–700 nm), were significant contributors to classification confidence. These observations are consistent with established absorbance characteristics of CRP and its interactions within wastewater matrices. Future work will incorporate advanced explainability methods, such as SHAP and permutation feature importance, to more precisely quantify wavelength-specific contributions, thus facilitating the optimization of optical sensor design and improving interpretability of spectral signatures related to CRP presence.

Machine learning models

In the realm of classification, several ML models can be leveraged to enhance decision-making processes and support the proposed wastewater classification approach^{42,43}. Models such as Support Vector Machines (SVM), Neural Networks (NN), K-Nearest Neighbors (KNN), Ensemble models, Decision Trees, Discriminants, and Naive Bayes were explored for their potential in handling the complexities of wastewater data. Among these, the cubic Support Vector Machine (CSVM) demonstrated the highest performance in both classification tasks, as shown in the results. Consequently, this subsection focuses on CSVM, providing a detailed explanation of its methodologies and advantages in effectively addressing wastewater classification challenges^{44,45}.

SVMs are a powerful set of supervised learning methods used primarily for classification, though they can also be applied to regression and outlier detection tasks⁴⁶. The fundamental concept behind SVMs is to find the optimal hyperplane that separates data points of different classes with the maximum margin, acting as a decision boundary. In a two-dimensional space, this hyperplane is a line; in three dimensions, it is a plane; and in higher dimensions, it is a hyperplane. The goal is to ensure that data points from different classes are on opposite sides of this hyperplane, achieving the best possible separation.

A key feature of SVMs is the margin, the distance between the hyperplane and the nearest data points from each class, known as support vectors. The margin should be as wide as possible because a larger margin implies better generalization and a lower risk of misclassification on new, unseen data. The optimal hyperplane is the one that maximizes this margin, making SVMs highly effective at ensuring that the classifier is robust and performs well on unseen data. An example of a binary classification optimal hyperplane is presented in Fig. 2.

One of the strengths of SVMs is their effectiveness in high-dimensional spaces. They are particularly useful when the number of dimensions exceeds the number of samples, a scenario that can be challenging for many other algorithms. SVMs handle this by transforming the input data into a higher-dimensional space where it becomes easier to segregate classes that are not linearly separable in the original space. This transformation is done using kernel functions, which map the input data into a higher-dimensional feature space. Kernel functions are crucial to the power of SVMs. Commonly used kernels include the linear kernel, polynomial kernel, and

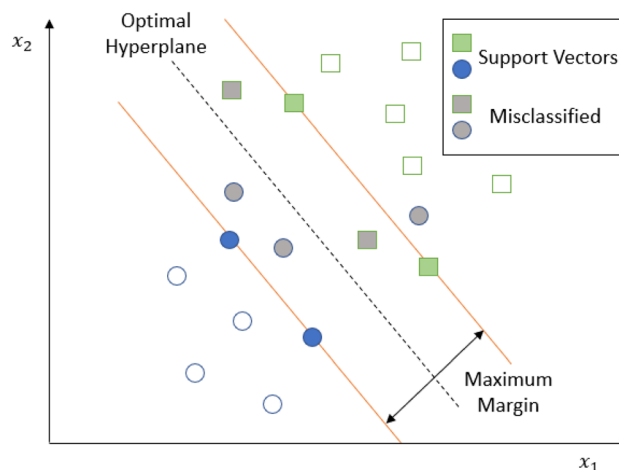


Fig. 2. Hyperplane separation for a binary SVM classifier with two classes represented by circles and squares.

radial basis function (RBF) kernel. These kernels allow SVMs to create complex decision boundaries that can handle a variety of classification problems, even when the data is not linearly separable in the original feature space. By applying these kernel functions, SVMs can effectively capture the underlying structure of the data, making them a versatile and powerful tool for a wide range of classification and regression tasks.

In this work, the proposed ML model designed to handle multi-class classification problems uses the Error-Correcting Output Codes (ECOC) method. ECOC decomposes a multi-class problem into multiple binary classification problems, with each solved by a binary classifier. The results from these binary classifiers are combined to make the final multi-class prediction. In this specific model, the response variable being predicted is the CRP concentration level, and the model aims to classify instances into one of five classes labeled 0, 1, 2, 3, and 4 (no CRP, 10^{-4} $\mu\text{g/ml}$, 10^{-3} $\mu\text{g/ml}$, 10^{-2} $\mu\text{g/ml}$ and 10^{-1} $\mu\text{g/ml}$ CRP concentration levels). There are 10 binary learners in this model, with a one-vs-one strategy, in which every binary classifier is trained for every pair of classes. With five classes, this results in ten binary classifiers, matching the number of binary learners specified.

Each binary learner in this model is a CSVM, which uses polynomial kernel functions of degree three. This means that the decision boundary is a cubic polynomial function of the input features, allowing the model to capture more complex relationships than a linear SVM. Importantly, since each CSVM is a separate model, each binary learner has its own bias. This bias is a characteristic of the specific classifier and influences how it separates the classes it is trained on. During the prediction phase, each of the ten binary classifiers makes a prediction. The final class prediction is determined by combining these results by considering which class has the highest aggregated score from the binary classifiers.

Table 1 summarizes the configurations of various ML models evaluated for the classification task. Each model type is listed alongside its key hyperparameters and learning settings, providing insight into the diversity of algorithms tested and the range of configurations applied. This breadth of modeling approaches helps ensure a robust assessment of which types of algorithms are most suitable for predicting CRP concentration categories in wastewater samples.

Ensemble models combine multiple base learners to produce a more robust and accurate classifier. By aggregating the predictions of several weaker models, they reduce the risk of overfitting and improve generalization. Techniques like bagging (e.g., Bagged Trees) reduce variance by averaging over models trained on different data subsets, while boosting (e.g., RUSBoosted Trees) focuses sequentially on correcting the errors of previous learners, improving performance particularly on harder-to-classify instances. Subspace ensembles further enhance diversity by training each learner on a random subset of features. These models are well-suited to complex, noisy datasets and often deliver strong performance in multi-class tasks.

Decision trees are hierarchical models that split the data based on feature values to form a tree-like structure of decision rules. Each node in the tree represents a feature, and each branch corresponds to a decision based on that feature's value. They are simple to interpret and fast to train. However, standalone trees can be sensitive to small data fluctuations, leading to overfitting, especially with deep trees (e.g., Fine Tree). Simpler trees (e.g., Coarse Tree) offer greater generalization but may underfit. Decision trees form the foundation of many ensemble methods and serve as a baseline for understanding feature importance and interactions.

KNN models are instance-based, non-parametric classifiers that predict the label of a data point based on the majority class among its k nearest neighbors in the training set. The choice of distance metric (e.g., Euclidean, cosine, Minkowski) and the number of neighbors significantly affects performance. Weighted versions further refine predictions by giving more influence to closer neighbors. KNN is simple and effective in low-dimensional problems but can become computationally expensive and less accurate in high-dimensional or imbalanced datasets. Despite these challenges, it can perform well when local structure in the data is informative.

Neural networks (NNs) are composed of interconnected layers of artificial neurons that transform inputs through weighted connections and non-linear activation functions. They are capable of modeling complex, non-linear relationships in data. In this study, both shallow (single- or two-layer) and deep architectures were explored, with varying layer sizes and activation functions like ReLU and Softmax. While neural networks require more data and computational power, they can outperform traditional models when sufficient training data is available and proper regularization is applied. Their flexibility makes them ideal for capturing nuanced patterns in complex datasets like spectrometry data.

Naive Bayes classifiers are probabilistic models based on Bayes' Theorem, with the simplifying assumption that features are conditionally independent given the class label. Despite this assumption often being violated in practice, Naive Bayes models tend to perform surprisingly well, especially on high-dimensional or noisy datasets. Gaussian Naive Bayes assumes normally distributed features, while Kernel Naive Bayes uses non-parametric density estimation to handle more flexible distributions. These models are fast, require little training data, and are useful as a strong baseline or in ensemble combinations.

Discriminant analysis models classify data by modeling the probability distributions of each class and using Bayes' Rule to assign labels. Linear Discriminant Analysis (LDA) assumes that classes share a common covariance structure, resulting in linear decision boundaries. Quadratic Discriminant Analysis (QDA) relaxes this assumption and allows each class to have its own covariance, yielding more flexible (quadratic) boundaries. These models are efficient and interpretable, especially when the data is approximately Gaussian and the class structure is well-separated.

Kernel models implicitly map input data into higher-dimensional spaces using a kernel function, allowing linear algorithms to learn non-linear relationships. While this is typically associated with SVMs, some other models (like kernel-based Naive Bayes or custom SVM implementations) also use this approach. Kernel learning is especially powerful for classification tasks with complex decision boundaries, offering a balance between model complexity and interpretability.

Model	Configuration		
FRBS system	Membership function	Number of rules	Number of inputs
	Triangular	5	3
SVM models	Kernel function	Function order	Classification
Cubic SVM	Polynomial (Scale 1.7)	3	ECOC 1vs1
Quadratic SVM	Polynomial (Scale 1.7)	2	ECOC 1vs1
Medium Gaussian SVM	Gaussian (Scale 1.7)	–	ECOC 1vs1
Fine Gaussian SVM	Gaussian (Scale 0.43)	–	ECOC 1vs1
Linear SVM	Linear (Scale auto)	–	ECOC 1vs1
Coarse Gaussian SVM	Gaussian (Scale 6.9)	–	ECOC 1vs1
Ensemble models	Number of splits	Learning cycles	Classification
Bagged Trees	20	30	Adaptive Boosting
RUSBoosted Trees	20	30	RUSBoost
Subspace Discriminant	–	30	Subspace
Subspace KNN	–	30	Subspace
Tree models	Number of splits	Split criterion	
Fine Tree	100	Gini's diversity index	
Medium Tree	20	Gini's diversity index	
Coarse Tree	4	Gini's diversity index	
K-Nearest neighbor models	Number of neighbors	Distance	Weight
Weighted KNN	10	Euclidean	Squared inverse
Cosine KNN	10	Cosine	Equal
Fine KNN	1	Euclidean	Equal
Medium KNN	10	Euclidean	Equal
Cubic KNN	10	Minkowski (exponent 3)	Equal
Neural network models	Layer size	Activation	Iteration
Wide Neural Network	100	ReLU	1000
Narrow Neural Network	10	ReLU	1000
Bilayered Neural Network	[10 10]	ReLU	1000
Medium Neural Network	25	ReLU	1000
Trilayered Neural Network	[10 10 10]	ReLU	1000
Deep Neural Network	[10 10 10 10 10 10 10 10 10 10]	ReLU & Softmax	1000
Naive Bayes models	Distribution	Support	
Kernel Naive Bayes	Kernel	Unbounded	
Gaussian Naive Bayes	Gaussian	Unbounded	
Discriminant models	Discriminant type		
Quadratic Discriminant	Quadratic		
Linear Discriminant	Linear		
SVM Kernel	Learner	Iteration	Classification
	SVM	1000	ECOC 1vs1

Table 1. Model configurations with parameters.

Results and discussion

In the classification endeavor, following the previous section, several ML models are taken into account, leveraging their unique capabilities to unravel the complexities of CRP analysis within wastewater samples. These models undergo rigorous training and evaluation using a five-fold cross-validation approach, ensuring robustness and reliability in the analyses. By partitioning the dataset into five subsets, the models are iteratively trained on four folds while validating its performance on the remaining fold. This iterative process allows to gauge the models' efficacy across different subsets of data, mitigating bias and variance concerns while enhancing generalizability. Through this meticulous methodology, the aim is to harness the predictive power of ML to unravel the intricate relationships between CRP levels and environmental dynamics.

In the first 5-class classification, the assignment of classes 0 through 4 delineates varying levels of CRP concentrations within wastewater samples considering all UV–Vis absorption spectrum. Specifically, class 0 denotes samples without CRP, class 1 corresponds to those with CRP concentrations at 10^{-4} µg/ml, class 2 encompasses concentrations at 10^{-3} µg/ml, class 3 signifies concentrations at 10^{-2} µg/ml and class 4 represents the highest concentrations at 10^{-1} µg/ml. Additionally, a range restricted 5-class classification is considered with a +400 nm UV–Vis absorption spectrum. Through these classifications, distinct thresholds are established

to categorize wastewater samples based on their CRP content, aiding in the characterization and analysis of CRP levels in the context of wastewater management and monitoring.

In assessing the performance, a set of standard metrics has been employed for highlighting the model. These metrics, namely Accuracy, Precision, Recall, F1 Score, and Specificity, offer comprehensive insights into the model's predictive capabilities. Accuracy quantifies the ratio of correct predictions to the total predictions made. Precision measures the proportion of true positives among all positive predictions, while Recall calculates the proportion of true positives among all actual positives. The F1 Score, a harmonic mean of Precision and Recall, provides a balanced evaluation of the model's performance. Lastly, Specificity gauges the ratio of true negatives to the sum of true negatives and false positives. By leveraging these metrics, a robust framework is established to assess the efficacy of the classifier model.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (5)$$

In the comprehensive analysis, the ML model underwent rigorous assessment to discern its efficacy in classifying CRP levels within wastewater samples. Through meticulous evaluation and validation procedures, the model is identified for each classification task regarding the accuracy of the model in the average result of 30 repetitions. To showcase the model's performance on classification tasks, a stratified 5-fold cross-validation was implemented, with each fold utilizing 80% of the data for training and 20% for validation. As no hyperparameter tuning was performed, the cross-validation procedure was repeated 30 times to obtain robust and reliable performance estimates. This approach makes efficient use of the entire dataset and mitigates the risk of overfitting, eliminating the need for a separate test set. However, to streamline presentation and focus on the most promising results, the performance of the top repetition of model for each classification is selectively showcased. The model represents the pinnacle of predictive accuracy and robustness within their respective categories, offering insights into the optimal methodologies for the analysis of the CRP concentration analysis in wastewater samples.

To assess whether the performance differences among models were statistically meaningful, a non-parametric Wilcoxon signed-rank test was conducted across 30 repetitions for the five best-performing classifiers. The resulting p-values, summarized in Tables 4 and 5, indicate that, particularly under the restricted-spectrum condition, the Cubic Support Vector Machine (CSVM) significantly outperformed the other models across all evaluated metrics (accuracy, precision, recall, F1 score, and specificity), with p-values far below the conventional 0.05 threshold-reaching values as low as 10^{-23} . Even in the full-spectrum scenario, CSVM demonstrated statistically significant superiority over models such as QSVM (e.g., $p = 0.0109$ for accuracy). These findings provide strong statistical evidence that CSVM offers a more robust classification performance in this context. While the absolute improvements may appear modest, the consistency and statistical reliability of CSVM's performance underscore its suitability for moderately classifying CRP concentrations in complex real wastewater matrices.

The Tables 2 and 3 presented below outline the performance metrics for each classification task, providing insights into the accuracy, precision, recall, F1 score, and specificity of the ML models employed. Accompanying these metrics, figures depict the confusion matrix and Receiver Operating Characteristic (ROC) curve, illustrating the classification performance of the most effective ML model repetition within each task, particularly emphasizing accuracy as a primary evaluation criterion. These visual aids offer a comprehensive understanding of the classification outcomes and highlight the robustness of the chosen model in effectively discerning between different concentrations of CRP in wastewater samples.

Table 2 presents the performance of the ML models in classifying wastewater samples into five different categories based on their CRP concentration levels. The categories are: wastewater without CRP, and wastewater with CRP concentrations of 10^{-4} µg/ml, 10^{-3} µg/ml, 10^{-2} µg/ml and 10^{-1} µg/ml. The table provides several key performance metrics for both the general performance of the CSVM model and its best repetition.

As the top-performing model across all classification tasks, the CSVM achieved a mean accuracy of 65.19% over 30 repetitions, with the best repetition reaching a slightly higher accuracy of 65.48%. This means that the model correctly predicts the CRP concentration category of the wastewater sample about 65% of the time. While this is significantly better than random guessing, which would be 20% for five classes, it indicates that there is still considerable room for improvement. The precision for the CSVM model is 65.44%, and for the best repetition, it is 65.81%. High precision means that when the model predicts a certain CRP concentration category, it is correct about 65% of the time. This suggests that the model is relatively good at avoiding false positives. The recall for the CSVM model is 65.11%, and for the best repetition, it is 65.39%. This implies that the model correctly identifies about 65% of the samples belonging to each CRP concentration category, indicating moderate sensitivity. The F1 score for the CSVM model is 65.27%, and for the best repetition, it is 65.60%. An F1 score in this range suggests a balanced performance in terms of precision and recall, but still indicates that the model's performance could be

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)
Cubic SVM	65.1865	65.4368	65.1106	65.2733	91.3037
Quadratic SVM	63.0873	63.5658	63.0150	63.2890	90.7774
Wide NN	61.3413	61.5153	61.2501	61.3823	90.3401
Fine KNN	61.1151	61.2204	61.0810	61.1505	90.2827
Mediumk NN	60.0000	60.1178	59.9057	60.0114	90.0049
Weighted KNN	56.7778	56.8529	56.6942	56.7734	89.1972
Narrow NN	56.5397	56.5711	56.4303	56.5003	89.1397
Medium Gaussian SVM	54.8452	57.5217	54.6942	56.0716	88.7165
Bilayered NN	54.7222	54.6426	54.6102	54.6260	88.6843
Linear Discriminant	54.4524	54.3174	54.2785	54.2978	88.6194
Bagged Trees (Ensemble)	54.2619	54.2287	54.1116	54.1699	88.5621
Cosine KNN	53.4960	53.9166	53.3587	53.6359	88.3780
Trilayered NN	53.1310	53.1528	53.0250	53.0884	88.2883
Cubic KNN	52.9325	53.3018	52.8167	53.0579	88.2380
Medium KNN	51.8254	51.9055	51.6961	51.8005	87.9619
Coarse Gaussian SVM	51.0675	52.0138	50.9129	51.4569	87.7635
Linear SVM	49.1270	50.1322	48.9277	49.5224	87.2866
Fine Tree	40.1429	40.2596	40.0796	40.1692	85.0381
RUSBoosted Trees (Ensemble)	37.8492	39.3179	37.7663	38.5218	84.4690
Medium Tree	34.5595	36.2584	34.4763	35.3367	83.6398
Subspace KNN (Ensemble)	32.3730	35.2327	31.9179	33.4847	82.9937
Kernel Naive Bayes	32.0833	36.2916	32.0859	34.0566	83.0314
Subspace Discriminant (Ensemble)	25.4722	29.6809	24.8633	26.9619	81.2313
Gaussian Naive Bayes	23.9206	32.4346	24.1401	27.6410	81.0288
Best result (CSVM)	65.4762	65.8134	65.3923	65.6022	91.3755

Table 2. 5-class classification: wastewater without CRP, 10^{-4} $\mu\text{g/ml}$, 10^{-3} $\mu\text{g/ml}$, 10^{-2} $\mu\text{g/ml}$, and 10^{-1} $\mu\text{g/ml}$ CRP (all spectrum). The reported values represent the average performance across 30 repetitions. The best value for each metric, highlighted in bold, corresponds to the single best-performing repetition among all trials.

enhanced. The specificity for the CSVM model is 91.30%, and for the best repetition, it is 91.38%. High specificity means the model is very good at identifying samples that do not belong to a certain CRP concentration category, thus effectively avoiding false negatives.

On the other hand, Table 3 shows the performance of the ML models when using only the range restricted spectrum for classification. The metrics for the CSVM model are slightly different from the full spectrum results, indicating how its performance can vary when evaluated on a subset of the data. Accuracy for this setup is 63.35% for the general model and 64.88% for the best repetition, which is slightly lower than the full spectrum results. This suggests that the range restricted spectrum provides almost the same performance results compared to the full spectrum using a lower amount of markers for classification. Precision, Recall, F1 Score, and Specificity are also slightly lower compared to the full spectrum results. This indicates a generally lower performance when using the range restricted spectrum alone, but the similar results imply that the range restricted spectrum provides almost the same comprehensive information for classifying the wastewater samples than all spectrum. In order to emphasize the results of the CSVM, other models such as Quadratic SVM, Ensemble Subspace KNN, Wide NN and Fine Gaussian SVM have been considered for comparison (Tables 4 and 5).

In Figs. 3 and 4, the confusion matrices and ROC curves of the best repetition of all spectrum and range restricted spectrum classification are showcased, respectively. Specifically, the confusion matrix provides a detailed breakdown of the model's predictions versus the actual labels. It shows how many samples from each actual category were predicted correctly or incorrectly by the model. This helps in understanding specific areas where the model may be making more errors and computing the performance metrics. In addition, the ROC (Receiver Operating Characteristic) curve plots the true positive rate (Recall) against the false positive rate (1-Specificity) for different threshold values. The area under the ROC curve (AUC) provides a single measure of the model's ability to distinguish between classes. A higher AUC indicates better overall performance. In the figures, the ROC curve is plotted for all classes, and AUC indicates values ranging from 79.89% to 92.19%, which indicates good overall performance. Overall, the performance metrics indicate that while the model is fairly good at making correct predictions, there is still significant room for improvement, especially in terms of increasing precision, recall, and overall accuracy. In general, CSVM shows strong class-wise accuracy with high diagonal values, yet a more granular error analysis reveals specific areas of confusion and interpretability insights. In the 400 nm range, the model performs best for Classes 1 and 5, while the most frequent misclassifications occur between Classes 2 and 3, and between Classes 4 and 5, suggesting overlapping spectral features within this limited range. Expanding to spectral range improves classification accuracy across most classes, particularly

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)
Cubic SVM	63.3532	63.3168	63.2059	63.2612	90.8444
Fine Gaussian SVM	59.3492	59.5536	59.2146	59.3835	89.8426
Wide NN	57.4206	57.4147	57.2792	57.3468	89.3640
Mediumk NN	55.7738	55.7569	55.6309	55.6937	88.9518
Fine KNN	55.9127	56.1932	55.8686	56.0303	88.9809
Weighted KNN	54.3294	54.5295	54.2231	54.3757	88.5824
Cosine KNN	53.7897	53.5448	53.6330	53.5888	88.4496
Medium Gaussian SVM	53.5556	56.2935	53.4102	54.8125	88.3999
Narrow NN	51.8254	51.8435	51.6785	51.7607	87.9654
Bilayered NN	51.4127	51.3504	51.2731	51.3112	87.8609
Linear Discriminant	51.3968	50.8151	51.1057	50.9599	87.8571
Cubic KNN	51.1230	51.4212	50.9866	51.2027	87.7823
Bagged Trees (Ensemble)	50.7619	50.6356	50.6067	50.6211	87.6863
Trilayered NN	50.3611	50.2789	50.2067	50.2425	87.5984
Medium KNN	50.1825	50.3821	50.0457	50.2132	87.5479
Coarse Gaussian SVM	49.0159	51.6679	48.7692	50.1755	87.2443
Linear SVM	48.0595	48.5859	47.8147	48.1967	87.0181
Fine Tree	38.0357	38.1712	37.9857	38.0779	84.5114
RUSBoosted Trees (Ensemble)	36.9524	38.8702	36.8793	37.8441	84.2478
Medium Tree	34.9762	37.4039	34.8757	36.0884	83.7449
Kernel Naive Bayes	33.2143	35.6113	33.1137	34.3160	83.2970
Subspace KNN (Ensemble)	28.4405	32.4879	27.8565	29.9719	81.9716
Subspace Discriminant (Ensemble)	27.7659	31.7558	27.1160	29.1964	81.8153
Gaussian Naive Bayes	24.0635	32.1147	24.3561	27.6863	81.0744
Best result (CSVM)	64.8810	64.8072	64.7473	64.7773	91.2273

Table 3. 5-class classification: wastewater without CRP, 10^{-4} $\mu\text{g/ml}$, 10^{-3} $\mu\text{g/ml}$, 10^{-2} $\mu\text{g/ml}$, and 10^{-1} $\mu\text{g/ml}$ CRP (+400 nm spectrum). The reported values represent the average performance across 30 repetitions. The best value for each metric, highlighted in bold, corresponds to the single best-performing repetition among all trials.

Model	Accuracy	Precision	Recall	F1	Specificity
QSVM	0.010882	0.025942	0.011672	0.017009	0.010016
WNN	3.50×10^{-14}	2.04×10^{-15}	3.46×10^{-14}	6.48×10^{-15}	3.17×10^{-14}
FKNN	6.55×10^{-13}	1.22×10^{-13}	9.27×10^{-13}	3.07×10^{-13}	6.22×10^{-13}
MNN	2.94×10^{-14}	4.84×10^{-15}	2.72×10^{-14}	1.05×10^{-14}	2.84×10^{-14}

Table 4. P-values comparing CSVM to other models (all spectrum).

Model	Accuracy	Precision	Recall	F1	Specificity
FGSVM	7.01×10^{-11}	3.09×10^{-11}	6.24×10^{-11}	4.11×10^{-11}	7.02×10^{-11}
WNN	1.09×10^{-16}	7.29×10^{-17}	1.04×10^{-16}	8.11×10^{-17}	1.08×10^{-16}
MNN	2.22×10^{-23}	1.85×10^{-23}	2.39×10^{-23}	1.90×10^{-23}	2.08×10^{-23}
FKNN	7.17×10^{-25}	2.40×10^{-24}	8.96×10^{-25}	1.35×10^{-24}	6.59×10^{-25}

Table 5. P-values comparing CSVM to other models (restricted spectrum).

reducing confusion between Classes 2 and 3, indicating that critical discriminative spectral bands likely lie outside the 400 nm region, potentially in the near-infrared (NIR) or shortwave infrared (SWIR) regions.

Additionally, the confusion matrices for both the whole and restricted spectral ranges provide insights into the model's classification performance across the five CRP concentration classes. For the whole spectral range, classes representing no CRP and the lowest concentration (10^{-4} $\mu\text{g/ml}$) show relatively high correct classification rates, with 128 and 107 samples correctly identified, respectively. However, misclassifications

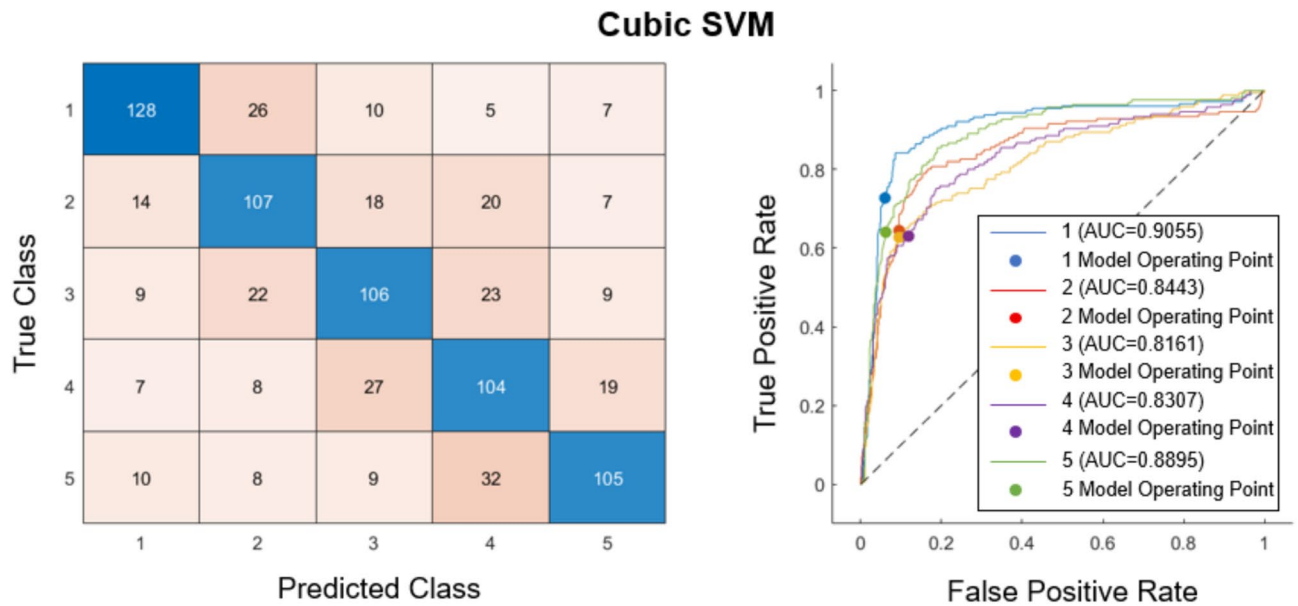


Fig. 3. Confusion matrix and ROC curve of best repetition of CSVM of Table 1.

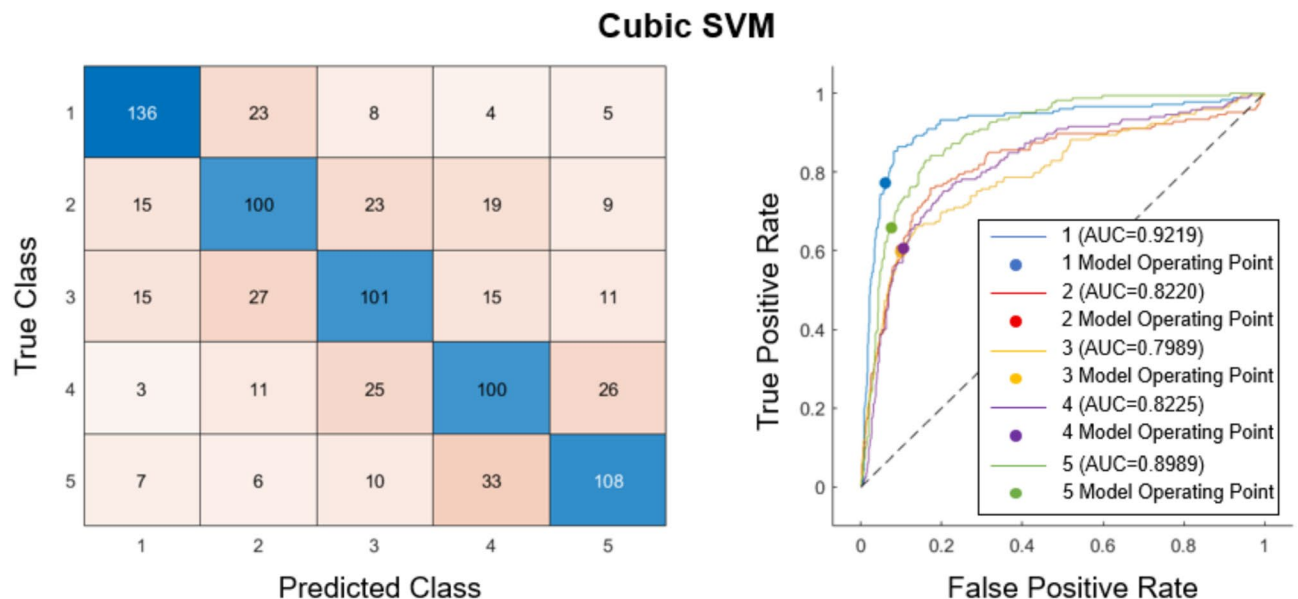


Fig. 4. Confusion matrix and ROC curve of best repetition of CSVM of Table 1.

frequently occur between adjacent classes, such as between 10^{-4} and 10^{-3} $\mu\text{g/ml}$, indicating challenges in distinguishing closely spaced CRP levels. A similar trend is observed in the restricted spectral range, where overall accuracy improves slightly for certain classes. Notably, the correct classification of samples with no CRP and the highest concentration (10^{-1} $\mu\text{g/ml}$) increases to 136 and 108, respectively. The number of misclassifications between neighboring classes decreases modestly, suggesting that focusing on a targeted spectral range can enhance the model's ability to differentiate between close concentration levels. These findings highlight that the most frequent errors occur between adjacent CRP classes, reflecting inherent spectral similarities in these concentration ranges. This error pattern underscores the need for further model refinement and targeted sensor development to improve discrimination in these critical boundaries. Incorporating such improvements could enhance predictive accuracy and robustness in real-world monitoring applications.

While multiple ML models were evaluated for comparative purposes, the performance differences among them were generally modest. Although the CSVM consistently outperformed the other models across most metrics, statistical testing across the 30 repetitions revealed that the observed differences were not large enough to be considered statistically significant. This suggests that, despite minor variations in mean accuracy, precision,

recall, F1 score, and specificity, the models perform comparably when accounting for variability across runs. Therefore, no strong conclusion can be drawn about the superiority of one model over another solely based on these performance metrics, and further investigation using larger datasets or more robust feature representations may be necessary to establish clearer distinctions in model performance.

The findings indicate that utilizing a limited spectral range is a viable option for applications such as sensor-based tools designed to classify CRP concentrations in wastewater, facilitating deeper understanding of its dynamics. Although there is a slight decrease in accuracy compared to using spectral range, the performance metrics remain close, suggesting that the reduced spectrum still contains much of the essential information required for reliable classification. This comparable effectiveness, achieved with fewer spectral features, is especially beneficial in settings with limited resources, like portable or embedded sensor devices, where reducing data collection and processing is critical. Furthermore, the slight decreases in precision, recall, and F1 score represent only a minor compromise in predictive accuracy, which could be acceptable in practical, rapid evaluation contexts. The consistently high specificity ensures that false positive rates stay low, which is crucial for preserving confidence in automated monitoring solutions. As a result, employing the limited spectral range offers a practical and efficient choice for implementation in field-deployable diagnostic systems or mobile platforms, striking a balance between classification performance, operational efficiency, and cost.

The classification performance, particularly the observed accuracy of approximately 65% in the five-class scenario, can be attributed to several factors. First, the intrinsic complexity and variability of the spectral data derived from real wastewater matrices contribute to a non-negligible level of class overlap. This is compounded by the presence of multiple interfering substances and natural fluctuations in wastewater composition, which can affect spectral signatures even in the absence of variation in the target analyte (CRP). Additionally, the limited interclass separability between adjacent CRP concentration ranges may have further constrained the classifier's discriminative power. Although higher classification accuracy was observed in simplified binary classification tasks, taking in to account CRP and no CRP samples, a five-class framework was adopted to better reflect the continuous and dynamic nature of CRP presence in wastewater environments. The objective of the study was not solely to achieve maximum classification accuracy, but rather to provide a more nuanced view of CRP level variation. The five-class model, despite its moderate accuracy, enables the identification of concentration trends and transitions that may be masked in dichotomous scenarios. It should also be noted that the moderate performance reflects a balance between model complexity and ecological relevance. The use of real wastewater, rather than synthetic or idealized matrices, inherently introduces variability that is both a challenge and a strength. This variability ensures that the model evaluation is more representative of real-world deployment scenarios, though at the cost of reduced classification performance when using conventional ML classifiers.

CRP is primarily recognized as a systemic biomarker of inflammation in clinical diagnostics. Its detection in municipal and hospital wastewater can serve as an indicator of human biological waste and, more broadly, as a proxy for anthropogenic biochemical load. Elevated CRP levels in wastewater have been associated with increased discharge from healthcare facilities or densely populated buildings, thereby providing valuable data for public health surveillance and environmental exposure assessments. While this manuscript does not describe the development of an actual sensor, it outlines a conceptual framework for future deployment of sensitive photonic sensor technologies for in situ CRP monitoring in complex matrices such as raw or partially treated wastewater. Such sensor systems could be integrated into decentralized monitoring points within sewage infrastructure – including influent pipelines, hospital discharge outlets, and community wastewater collection nodes – enabling early warning systems for sanitary contamination and near real-time epidemiological surveillance. Furthermore, the envisioned monitoring system is compatible with existing smart water infrastructure. By coupling sensor outputs with remote telemetry and cloud-based data platforms, the approach could support next-generation digital water networks that combine chemical sensing with health-relevant biomarker tracking. This integration would provide actionable insights for wastewater-based epidemiology (WBE), enhancing environmental health strategies and preparedness efforts, particularly in the context of urban health management and post-pandemic monitoring.

Conclusion

In this study, two classification tasks were conducted to discern various concentrations of CRP in wastewater samples based on the UV–Vis absorption spectroscopy spectra. Across these tasks, CSVM model, alongside additional ML comparison models, underwent rigorous performance assessment, with the top-performing model repetition selected for evaluation based on accuracy. The classification tasks encompassed distinguishing wastewater samples categorizing samples into five distinct CRP concentration classes ranging from zero to 10^{-1} µg/ml. Each classification task was evaluated across 30 repetitions using both the entire broadband UV–Vis absorption spectrum and a limited spectral range from 400 nm onwards. The model showcased varying degrees of performance across tasks, with the top-performing model repetitions achieving accuracies ranging from 64.88% to 65.48%. Specifically, ML model such as SVMs were found to be the best models for these classification tasks due to their robustness and effectiveness, working with a limited amount of data. Overall, the results underscore the efficacy of ML algorithms in moderately classifying wastewater samples based on CRP concentrations, with potential applications in biosensor development and water quality dynamics monitoring.

This manuscript outlines our concept of an expert system using point-of-need optical sensors and machine learning for wastewater surveillance. As a next step for development a practical application, the following steps are crucial for a more comprehensive investigation and future work. From metrological point of view one of the most crucial steps is to evaluate the sensing performance under diverse wastewater conditions, such as varying pH, temperature, and the presence of interfering substances, to ensure real-world robustness. On the other hand, a suitable machine learning model which will be trained on those data will be developed. In

the next step, this prototype will be implemented in a controlled, small-scale wastewater treatment plant or a simulated environment to test its functionality, data acquisition capabilities, and initial performance in a semi-realistic setting. We'll gather continuous data over an extended period to assess the system's stability. Finally, the investigation with the system which will cover larger geographical areas and be integrated with existing public health infrastructure will be completed.

In the final step, implementing our system in collaboration with public health authorities and wastewater treatment plants enables the identification of anomalies in wastewater data and the prediction of disease trends as part of a population-level approach that supports real-time monitoring for wastewater-based epidemiology (WBE). Integrating such information enables public health decision-making to be faster, more accurate, and more efficient—both in terms of epidemiological response and economic resource allocation, such as more targeted deployment of funding and public health services.

This leads to a concept of an expert system incorporating point-of-need optical sensors for wastewater surveillance including ML algorithms to deliver reliable information. Such system can process significant volumes of data, identify trends and anomalies, and provide actionable insights. For instance, monitoring CRP levels in wastewater can offer evidence of population-wide inflammatory disease dynamics, supporting public health officials in making informed decisions through quick detection and automated classification of the samples from a given area. The system can be tailored to the specific needs by incorporating dedicated optical biosensors and adapting the algorithm to the specific data acquired. Additionally, while the current study appropriately relied on default hyperparameters to mitigate overfitting risks due to dataset limitations, incorporating even modest hyperparameter optimization could enhance model accuracy and generalization. Such optimization becomes especially valuable when scaling the system for real-world deployment.

Data availability

The data supporting this study can be achieved from the corresponding author upon a reasonable request.

Received: 16 May 2025; Accepted: 13 August 2025

Published online: 24 September 2025

References

1. Bagre, A., Patel, P. R., Naqvi, S. & Jain, K. Emerging concerns of infectious diseases and drug delivery challenges. In *Nanotheranostics for Treatment and Diagnosis of Infectious Disease* 1–23 (2022). <https://doi.org/10.1016/B978-0-323-91201-3.00013-X>
2. Baker, R. E. et al. Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205. <https://doi.org/10.1038/s41579-021-00639-z> (2022).
3. Wityk, P. et al. Celebrating biomimicry: Bioinspired layers in optical biosensors. *Opt. Quantum Electron.* **56**, 1303. <https://doi.org/10.1007/s11082-024-06798-6> (2024).
4. Hosseini, S., Vázquez-Villegas, P., Rito-Palomares, M. & Martínez-Chapa, S. O. Advantages, disadvantages and modifications of conventional ELISA. In *Enzyme-Linked Immunosorbent Assay (ELISA): From A to Z* (eds Hosseini, S. et al.) 67–115 (Springer, 2018). https://doi.org/10.1007/978-981-10-6766-2_5.
5. Yan, Y., Chang, L. & Wang, L. Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): Current status, challenges, and countermeasures. *Rev. Med. Virol.* **30**, e2106. <https://doi.org/10.1002/rmv.2106> (2020).
6. Krishnan, S. & ul Quasim Syed, Z. Colorimetric visual sensors for point-of-needs testing. *Sens. Actuators Rep.* **4**, 100078. <https://doi.org/10.1016/j.snr.2022.100078> (2022).
7. Alsaifi, T., Badshah, A., Aboulola, O. I. & Daud, A. Optimizing healthcare big data performance through regional computing. *Sci. Rep.* **15**(1), 3129 (2025).
8. Srivastava, P., Mittal, Y., Gupta, S., Abbassi, R. & Garaniy, V. Recent progress in biosensors for wastewater monitoring and surveillance. In *Artificial Intelligence and Data Science in Environmental Sensing* (eds Asadnia, M. et al.) 245–267 (Academic Press, 2022). <https://doi.org/10.1016/B978-0-323-90508-4.00010-1>.
9. Chadha, U. et al. Recent progress and growth in biosensors technology: A critical review. *J. Ind. Eng. Chem.* **109**, 21–51. <https://doi.org/10.1016/j.jiec.2022.02.010> (2022).
10. Ejeian, F. et al. Biosensors for wastewater monitoring: A review. *Biosens. Bioelectron.* **118**, 66–79. <https://doi.org/10.1016/j.bios.2018.07.019> (2018).
11. Sharma, V., Takamura, H., Biyani, M. & Honda, R. Real-time on-site monitoring of viruses in wastewater using Nanotrap® particles and RICCA technologies. *Biosensors* **14**, 115. <https://doi.org/10.3390/bios14030115> (2024).
12. Kumblathan, T., Liu, Y., Uppal, G. K., Hruđey, S. E. & Li, X.-F. Wastewater-based epidemiology for community monitoring of SARS-CoV-2: Progress and challenges. *ACS Environ. Au* **1**, 18–31. <https://doi.org/10.1021/acsenvironau.1c00015> (2021).
13. Maryam, S. et al. COVID-19 surveillance in wastewater: An epidemiological tool for the monitoring of SARS-CoV-2. *Front. Cell Infect. Microbiol.* **12**, 978643. <https://doi.org/10.3389/fcimb.2022.978643> (2023).
14. Kilaru, P. et al. Wastewater surveillance for infectious disease: A systematic review. *Am. J. Epidemiol.* **192**, 305–322. <https://doi.org/10.1093/aje/kwac175> (2023).
15. Devianto, L. A. & Sano, D. Systematic review and meta-analysis of human health-related protein markers for realizing real-time wastewater-based epidemiology. *Sci. Total Environ.* **897**, 165304. <https://doi.org/10.1016/j.scitotenv.2023.165304> (2023).
16. Andrés-Costa, M. J., Andreu, V. & Picó, Y. Liquid chromatography–mass spectrometry as a tool for wastewater-based epidemiology: Assessing new psychoactive substances and other human biomarkers. *TrAC Trends Anal. Chem.* **94**, 21–38. <https://doi.org/10.1016/j.trac.2017.06.012> (2017).
17. Gracia-Lor, E. et al. Measuring biomarkers in wastewater as a new source of epidemiological information: Current state and future perspectives. *Environ. Int.* **99**, 131–150. <https://doi.org/10.1016/j.envint.2016.12.016> (2017).
18. Petrie, B., Youdan, J., Barden, R. & Kasprzyk-Hordern, B. Multi-residue analysis of 90 emerging contaminants in liquid and solid environmental matrices by ultra-high-performance liquid chromatography tandem mass spectrometry. *J. Chromatogr. A* **1431**, 64–78. <https://doi.org/10.1016/j.chroma.2015.12.036> (2016).
19. Davies, B., Paul, R., Osselson, D. & Woolley, T. Analysis of crude wastewater from two treatment plants in South Wales for 35 new psychoactive substances and cocaine, and cannabis. *Sci. Rep.* **14**(1), 20129 (2024).
20. Sokolowski, P. et al. Optical method supported by machine learning for dynamics of C-reactive protein concentrations changes detection in biological matrix samples. *J. Biophotonics* <https://doi.org/10.1002/jbio.202300523> (2023).
21. Zhang, X., Yu, X. & Hu, X. Evaluating the impact of smart city construction on sewage treatment in China from a synergistic perspective. *Sci. Rep.* **15**(1), 260 (2025).

22. Alvarez-Serna, B. E. et al. Label-free and portable field-effect sensor for monitoring RT-LAMP products to detect SARS-CoV-2 in wastewater. *Talanta* **253**, 124060. <https://doi.org/10.1016/j.talanta.2022.124060> (2023).
23. Guerrero-Esteban, T. et al. Sensitive SARS-CoV-2 detection in wastewaters using a carbon nanodot-amplified electrochemiluminescence immunosensor. *Talanta* **247**, 123543. <https://doi.org/10.1016/j.talanta.2022.123543> (2022).
24. Sen, P. et al. Integrating water purification with electrochemical aptamer sensing for detecting SARS-CoV-2 in wastewater. *ACS Sens.* **8**, 1558–1567. <https://doi.org/10.1021/acssensors.2c02655> (2023).
25. Jagadeesan, K. K. et al. Wastewater-based proteomics: A proof-of-concept for advancing early warning system for infectious diseases and immune response monitoring. *J. Hazard. Mater. Lett.* **5**, 100108. <https://doi.org/10.1016/j.hazl.2024.100108> (2024).
26. Sohrabi, H. et al. Recent advances of electrochemical and optical biosensors for detection of C-reactive protein as a major inflammatory biomarker. *Microchem. J.* **158**, 105287. <https://doi.org/10.1016/j.microc.2020.105287> (2020).
27. Cierpiak, K. et al. C-reactive protein (CRP) evaluation in human urine using optical sensor supported by machine learning. *Sci. Rep.* **14**(1), 18854 (2024).
28. Szczerska, M. et al. Remote real-time wastewater surveillance by the optical sensor supported by machine learning. *Sens. Actuators Rep.* **10**, 100346 (2025).
29. Azzouz, M. et al. Air pollution and biomarkers of cardiovascular disease and inflammation in the Malmö Diet and Cancer cohort. *Environ. Health* **21**(1), 39 (2022).
30. Xu, Z. et al. Association between gaseous air pollutants and biomarkers of systemic inflammation: A systematic review and meta-analysis. *Environ. Pollut.* **292**, 118336 (2022).
31. Pritzker, K. P. H. Blood-based biomarkers of chronic inflammation. *Expert Rev. Mol. Diagn.* **23**(6), 495–504 (2023).
32. Johnson, T. V., Abbasi, A. & Master, V. A. Systematic review of the evidence of a relationship between chronic psychosocial stress and C-reactive protein. *Mol. Diagn. Ther.* **17**(3), 147–164 (2013).
33. Kennedy, E. & Niedzwiedz, C. L. The association of anxiety and stress-related disorders with C-reactive protein (CRP) within UK Biobank. *Brain Behav. Immunity-Health* **19**, 100410 (2022).
34. Ali, S. et al. Role of C-reactive protein in disease progression, diagnosis and management. *Discoveries* **11**(4), e179 (2023).
35. Rao, L. et al. The ability of inflammatory markers to recognize infection in cancer patients with fever at admission. *Immunol. Res.* **70**(5), 667–677 (2022).
36. Gogna, P. et al. A cohort study of the multipollutant effects of PM_{2.5}, NO₂, and O₃ on C-reactive protein levels during pregnancy. *Environ. Epidemiol.* **8**(3), e308 (2024).
37. Kim, J. H. et al. Long-term effects of ambient particulate and gaseous pollutants on serum high-sensitivity C-reactive protein levels: A cross-sectional study using KoGES-HEXA data. *Int. J. Environ. Res. Public Health* **19**(18), 11585 (2022).
38. Rahman, H. H., Niemann, D. & Bugajski, A. Association of environmental toxic metals with high sensitivity C-reactive protein: A cross-sectional study. *Occup. Dis. Environ. Med.* **9**(4), 173–184 (2021).
39. Wang, Y. et al. Low-grade systemic inflammation links heavy metal exposures to mortality: A multi-metal inflammatory index approach. *Sci. Total Environ.* **947**, 174537 (2024).
40. Fierro-Barrientos, G. N. et al. Effect of pesticide exposure on systemic inflammatory biomarkers: A meta-analysis, and trial sequential analysis. *J. Environ. Health Sci. Eng.* **23**(2), 1–15 (2025).
41. Ali, S. et al. Photocatalytic removal of textile wastewater-originated methylene blue and malachite green dyes using spent black tea extract-coated silver nanoparticles. *Sci. Rep.* **15**(1), 1851 (2025).
42. Yang, X. et al. A sensor array based on a nanozyme with polyphenol oxidase activity for the identification of tea polyphenols and Chinese green tea. *Biosens. Bioelectron.* **250**, 116056. <https://doi.org/10.1016/j.bios.2024.116056> (2024).
43. Kim, S. et al. Distinct plasma phosphorylated-tau proteins profiling for the differential diagnosis of mild cognitive impairment and Alzheimer's disease by plasmonic asymmetric nanobridge-based biosensor. *Biosens. Bioelectron.* **250**, 116085. <https://doi.org/10.1016/j.bios.2024.116085> (2024).
44. Ribeiro, D., Sanfins, A. & Belo, O. Wastewater treatment plant performance prediction with support vector machines. In *Advances in Data Mining, Applications and Theoretical Aspects* (ed. Perner, P.) 99–111 (Springer, 2013). https://doi.org/10.1007/978-3-642-39736-3_8.
45. Fairouz, T., McNamee, S. E., Finlay, D., Ng, K. Y. & McLaughlin, J. A novel patches-selection method for the classification of point-of-care biosensing lateral flow assays with cardiac biomarkers. *Biosens. Bioelectron.* **223**, 115016. <https://doi.org/10.1016/j.bios.2022.115016> (2023).
46. Xue, H., Yang, Q. & Chen, S. SVM: support vector machines. In *The Top Ten Algorithms in Data Mining* (ed. Vani, M.) (Chapman and Hall/CRC, 2009).

Acknowledgements

This work has been supported by the Research Project PID2023-146520OB-C21, funded by the Spanish Government. Additionally, this study was performed within the framework of COST Action CA21159 (Understanding interaction light—biological surfaces possibility for new electronic materials and devices—PhoBioS), supported by COST (European Cooperation in Science and Technology; <http://www.cost.eu>).

Author contributions

Conceptualization: F.J. Maldonado Carrascosa, S. García Galán, M. Szczerska. Methodology: F.J. Maldonado Carrascosa, S. García Galán. Software: F.J. Maldonado Carrascosa, S. García Galán. Validation: F.J. Maldonado Carrascosa, S. García Galán. Formal analysis: F.J. Maldonado Carrascosa, S. García Galán, M. Szczerska. Investigation: F.J. Maldonado Carrascosa, S. García Galán, M. Szczerska. Resources: A. Łuczkiwicz, P. Wityk, M. Szczerska. Data Curation: A. Łuczkiwicz, P. Wityk, M. Szczerska. Writing—original draft preparation: F.J. Maldonado Carrascosa, S. García Galán. Writing—review and editing: F.J. Maldonado Carrascosa, S. García Galán, A. Łuczkiwicz, P. Wityk, M. Szczerska. Visualization: S. García Galán, F.J. Maldonado Carrascosa. Supervision: S. García Galán, M. Szczerska. Project administration: S. García Galán, M. Szczerska. Funding acquisition: S. García Galán, M. Szczerska.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.J.M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025