https://doi.org/10.48048/tis.2024.8528

Critically Reckoning Spectrophotometric Detection of Asymptomatic Cyanotoxins and Faecal Contamination in Periurban Agrarian Ecosystems via Convolutional Neural Networks

Soumyajit Koley

Department of Civil Engineering, Hooghly Engineering & Technology College, West Bengal, India

(Corresponding author's e-mail: samkoley5@gmail.com)

Received: 3 June 2024, Revised: 16 August 2024, Accepted: 23 August 2024, Published: 20 October 2024

Abstract

Based on a systematic review of convolutional neural networks (CNN), this study explores the efficacy of small imaging sensors in monitoring the real-time presence of cyanotoxins and hazardous contaminants in urban ecosystems. To develop a machine learning-based CNN, this study first investigated the relationships between the prevalence of hazardous algal blooms (HABs) and faecal indicator bacteria (FIB) in waterways and aquifers of certain semi-arid zones of Sri Lanka, Sweden and New York (United States). By incorporating a popularly known AbspectroscoPY framework to effectively process the spectrophotometric data of the obtained samples, the formulation subsequently reveals strong positive correlations between FIB coliforms and nutrient loads (particularly nitrate and phosphate). A corroborative association with the incidence of chronic kidney disease of uncertain aetiology (CKDu) among the residents of the studied regions further affirms the reliability of the methodology. These findings underline the need for policymakers to consider the geographical and land-use traits of urban habitats in strategies aimed at reducing water-borne health hazards.

Keywords: Cyanobacteria, Machine learning, *Escherichia coli*, Quantitative phase imaging, Pathogens, Chronic kidney disease, *Listeria monocytogenes*, Asymptomatic carriers

Introduction

Freshwater systems worldwide are experiencing the proliferation of hazardous algal blooms (HABs) caused by the rampant anthropogenic incursion of cyanobacteria [1]. Factors such as increased nutrient absorption owing to urban and agricultural runoff, higher water temperatures and longer residence times contribute to this trend [2]. Contemporary water treatment plants seldom restrict the obtrusion of cyanobacterial toxins [3,4]. Acute consumption of cyanotoxins leads to serious ailments, such as gastritis, liver damage, neurotoxicity and respiratory and allergic responses [5]. Chronic exposure to cyanotoxins can be fatal and amplifies tumour growth [6]. The risk of adverse health effects escalates when exposed to drinking water with cyanobacterial cell densities of 100,000 cells/mL or higher [7]. Moreover, invasive

offsets of the HABs can significantly impact the well-being and stability of aquatic life as well [8]. The growth of cyanobacteria can reduce the amount of energy that other species in the environment can obtain by decreasing the incoming intensity of solar radiation at water depths. Cyanobacteria assimilate dissolved CO_2 in aquatic environments, which restricts their accessibility to other plant species. This in turn dampens the oxygen production process, leading to rapid hypoxia and a subsequent decline in the mortality rate of aquatic organisms [9]. Harmful Algal Blooms (HABs) also have a detrimental impact on local economies because they render water containing cyanotoxins unsuitable for farming. Additionally, unpleasant odours and flavours associated with cyanotoxins have detrimental effects on local tourism prospects [10]. Climate change is expected to exacerbate the economic impact of HABs, because rising water temperatures can affect nutrient transport, which is influenced by the disturbed hydrology of a catchment region. In 2020, the Land and Water Resources Research and Development Corporation (Australia) commissioned an assessment that estimated the annual remediation cost of HABs in Australia to be between 180 and 240 million AUD [11]. To minimise mitigation expenses, monitoring the transport and fate of pathogens and cyanobacteria on a regular basis is essential for the early identification of HABs and targeted risk assessments [12].

Recently, various biochemical studies have analysed cyanobacterial species taxonomy and potential hazard levels at different cell densities [8-13]. Cell viability tests, which reveal the residual risk of treated water, are useful for evaluating potential threats to drinking water quality [14]. The effectiveness of different treatment methods in reducing microbial activity can be evaluated through cell viability assays, which measure cell membrane integrity using a cocktail of DNA-binding labelling agents [15]. To achieve the desired effects, cell-impermanent dyes are used for dead cells, and cell-permanent dyes are used for live cells. To ascertain the level of vitality without employing labelling techniques, one can utilise nondestructive and user-friendly imaging microscopy. Drawing on a comprehensive review of convolutional neural networks (CNN), this article examines the usefulness of image recognition and small imaging sensors for the real-time detection of cyanotoxins and harmful pollutants in urban ecosystems. To build a machine learning-based CNN, this study initially investigated the connections between the occurrence of hazardous algal blooms (HABs) and faecal indicator bacteria (FIB) in waterways and aquifers of a few semi-arid regions in Sri Lanka, Sweden and New York (United States). By incorporating the well-known AbspectroscoPY framework [16] for the effective processing of the spectrophotometric data of the obtained samples, the methodology subsequently uncovered strong positive correlations between FIB coliforms and nutrient levels, particularly nitrate and phosphate. Furthermore, the association between the incidence of chronic kidney disease of unknown cause (CKDu) among the inhabitants of the studied regions and the reliability of the method was confirmed. These results emphasise the importance of policymakers considering the geographic and land-use characteristics of urban areas in strategies aimed at reducing water-borne health risks.

Materials and methods

Background and objectives

The performance of neural networks in detecting microscopic cells depends on high-quality inputs and spatial resolution. FlowCam technology is a modern tool that helps quantify and classify cyanobacteria species at both the early and advanced stages of HABs in various freshwater systems. In recent case studies, researchers have used trained neural networks to identify FlowCam images of larger species such as *Oscillatoria sp.*, *Anabaena sp.* and whole colonies of *Microcystis sp*. However, the images of smaller species tend to be less detailed than those of larger species. One potential solution is Quantitative Phase Imaging (QPI) technology, which can increase the availability of information and accuracy of image identification by producing images with observable biophysical characteristics [17]. Flow cytometry is another method that involves the examination of cell cultures by passing them through a solution (sheath fluid) and evaluating their performance using image sensors that detect fluorescence and/or optical signals. This technology has several advantages over conventional methods that rely on bright-field microscopy and heterotrophic plate counts. First, it significantly reduces the amount of time, effort and accuracy required to measure water quality, as changes in these variables can lead to errors in techniques that require incubation or preservation of samples [18]. Second, it can detect live microorganisms in water samples, which may not be possible with traditional culture methods that use nutrient agars to selectively suppress certain microbes and may suffer from substrate inhibition [19,20]. Finally, flow cytometry can provide real-time, on-site analysis of live samples, which can reduce errors and the need for expert expertise and hands-on examination, thereby reducing turnaround times for bacterial samples from several days to a matter of hours and for viral samples from weeks to a matter of days [21,22]. These advantages make flow cytometry a valuable tool for water treatment, particularly in situations where the quick provision of clean water to customers is critical to avoid health concerns. The traditional flow cytometry analysis of moving particles in streams has been conducted without imaging. Particles, such as bacteria, viruses, protozoa, cell remnants and inorganic waste, can be identified. This technique relies on the emission of fluorescence or scattered light upon exposure to a laser beam. The side-scatter signal reveals information about the complexity and granularity of the sample, whereas the forward-scatter signal (FSC) is inversely proportional to the particle size. However, the non-imaging method used for analysis compromises the accuracy of cell population and morphological characterisation. However, the use of relative light scattering is an inherent drawback. The amount of scattered light is affected by factors, such as the surface roughness of the sample, its refractive index and the type of sheath fluid used. The FSC value of a cell is not necessarily twice that of an adjacent cell, even though it appears to be larger. This makes it difficult to determine cell taxonomy, particularly for irregular particles, whose orientation during examination affects the scatter signals. Clumping the cell clusters together can also lead to larger particles. The scattered light is more uniform and slightly scattered by larger cells, whereas smaller cells scatter it more strongly. These features can affect cell recognition and population analyses. Appendix A lists all notations defined in this study. This study explored machine learning techniques to determine the role of convolutional neural networks in enhancing the spectroscopic detection of asymptomatic cyanobacteria and faecal contamination in food chains. The objective of imaging cytometry is to enhance the visibility of target cells using 1 or more dyes. By combining the fluorescence

and bright-field images of tagged cells, it is possible to determine the total number of cells, including both dead and active cells. To assess cell viability, it is crucial to analyse the proportion of living cells relative to the overall cell count. Imaging flow cytometry is a novel method for examining the cell shape and population dynamics. This technique involves capturing images of cells in motion through a liquid and sending these images to a computational database for counting and identification (**Figures 1(a) - 1(c)**).

Flow cytometry, which captures cell populations, allows for a faster analysis with less manual preparation. Two popular image flow cytometry systems are FlowCam (8000 Series, Yokogawa Fluid Imaging Technologies) and Amnis ImageStream (MkII, Luminex) [23-28]. Including microbial taxonomy, such as the presence of potentially harmful cyanobacteria, in water sample risk evaluations enhances the accuracy of risk assessment methodologies for humans, agriculture and environmental systems. This is easier and faster with automated imaging flow cytometry devices than with older, more labour-intensive methods such as manual counting of cells with a haemocytometer. The automated dye injection capabilities of 2 commercially available imaging flow cytometers are summarised in **Figure 1(d)**. The value of cytometers in determining cell viability is emphasised in this study. The aim of flow cytometry imaging is to achieve rapid and high-resolution imaging of cells. When evaluating the viability of a technique, it is essential to consider the important tradeoffs associated with using contemporary camera technology for imaging flow cytometry. As the imaging speed and fluid velocity increased, the sensitivity of the camera decreased because it could capture fewer photons from each specimen. Imaging flow cells with multiple parallel channels allows for a higher sample throughput with the same or lower imaging rates and fluid velocities. McCuskey *et al.* [14] performed high-throughput image flow cytometry using flow cells with 384 parallel channels. However, this method requires a large field of view, which in turn requires a lower-numerical-aperture imaging lens. As a result, the acquired images lose some spatial resolution. Therefore, there may be better alternatives to flow cytometry, which can process samples more quickly or produce better images. The practicality of such alternatives depends on the specific use cases. Previous studies have explored the potential of these methods for assessing cell viability. A sample of the microorganisms was affixed to a slide and observed under a brightfield microscope, which is typically used in conventional water quality microscopy, along with cultivating microorganisms from a water sample. However, because of the extensive preparation required for samples and the potential for errors in experiments, this approach is less efficient than the more advanced methodologies that are the focus of this research. The advantages and disadvantages of these methods, as outlined in scientific publications, provide insights into their constraints and practical applications in areas such as neural network recognition and water quality analysis. The presence of cyanobacteria in the samples was determined by fluorescence microscopy, which allowed for the estimation of the cell count and density. This method relies on the measurement of light intensity scattered by cellular pigments such as chlorophyll-a and phycocyanin. However, it is important to note that this technique can only detect cyanobacteria in a sample and does not provide information on the species or the number of species present. Additional microscopic studies are necessary for more comprehensive taxonomic identification and counting. Although conventional sample preparation and brightfield microscopy can be used to identify and quantify imaging cells in cyanobacterial blooms, this procedure is time-consuming and requires technical and morphological knowledge. Alternatively, AI image recognition

using ML-trained neural networks has been demonstrated to be a practical substitute for identifying specimens using microscopic imaging data, as indicated by a literature review of AI applications in this field. The challenge of accurately and rapidly classifying cyanobacteria based on photographs continues to elude machine learning algorithms. One reason for this is the absence of reproductive characteristics that can provide taxonomic information. Additionally, the cells are small, vary in size, and exhibit polymorphism. Algorithms for picture recognition analyse the data and geographical features of a photograph in order to identify the subject. To enhance the interpretation of higher-dimensional data, it is beneficial to increase the complexity of the neural network by expanding the range of the recorded cellular characteristics. The final outcomes of this study were improved identification ability and more reliable statistical data. Imaging methods that can absorb more data have greater potential to enhance AI recognition. Therefore, methods that feed more information into neural networks are more valuable, and the current research is reviewed to provide recommendations. The investigation identified 3 quantitative phase imaging (QPI) methods: Method 1, which utilised a portable device; Method 2, which used Michelson interferometry; and Method 3, which used a commercially available *in-situ* probe. The 3 methods for accurate cell identification are compared in **Figure 1(e)**, with method 2 considered the most practical solution. **Figure 1(f)** depicts a possible setup for cyanobacteria monitoring using method 2 and sample collection buoys. Furthermore, this section provides a comprehensive review of current and future commercially available microscopic imaging methods as well as their application to water quality microscopy. **Figure 1(d)** offers a brief summary of the advantages and disadvantages of these options, which will be discussed in greater detail later. Sections M1 - M5 (of the Supporting Information document enclosed with the article) elucidate the key descriptions pertaining to the corroboration and relevancy of this study's novel findings with recently published case studies dealing with similar topics.

Historical discourse

A commercially available automated device, OC-300, has been developed by OnCyt Microbiology [23]. By connecting flow cytometry equipment to raw water samples gathered from various bodies of water, the OC-300-unit functions as an intermediary device. The OC-300 apparatus was programmed to automatically or manually collect the water samples. Twelve parallel sample streams were within the OC-300's capabilities. This allowed for the dependable acquisition of cell counts from several locations using single-flow cytometry equipment. The samples were prepared in a volume of 100 - 500 µL using autoloaders. Furthermore, a dye can be automatically added to a sample for ease of examination [24]. Researchers in Alberta, Canada, who keep tabs on cyanobacteria populations to detect both the early and late stages of hazardous algal blooms (HABs), have noticed problems with bacterial cells obstructing smaller imaging flow cells [25]. By dispersing the cells more uniformly across the flow cell, Lugol's solution improved the picture capture efficiency by decreasing the cell adhesion [26]. The autofluorescence of chlorophyll-a was not detectable when Lugol's solution was used [27]. However, a system that can identify cyanobacteria by using machine learning and image recognition has few limitations [28]. If the cell count exceeds a certain point, which might lead to blockage inside the imaging cells, OC-300 can further dilute the fluid to enhance sample processing, and the meticulously produced samples can either be utilised

to prepare samples for analysis by QPI or inserted into a nearby imaging flow cytometer, as described in **Figures 2(a) - 2(c)**. Precise sample preparation, cell counts, viability evaluation and taxonomic identification of the species present were made possible by combining these technologies. The proposed incorporation of OC-300 into a HAB monitoring station is discussed in later sections.

One non-invasive way to analyse cell morphology is via Quantitative Phase Imaging (QPI), which is sometimes called holographic microscopy [29]. This is a phase contrast microscopy approach. To identify cells in a sample, QPI takes advantage of the scattering of specific wavelengths of light, which creates a sharp contrast with the background light, which is unhindered and non-scattering. The thickness and degree of change in the internal refractive index of the examined samples are directly related to the contrast. The total scattered light can be measured because of this contrast, which provides valuable information on the internal and exterior biophysical features of the cell. The density, form, mass, volume, membrane and cellular structures (including organelles) as well as their exterior, can be precisely investigated in this manner, as shown in **Figure 2(c)**. The use of QPI units in conjunction with flow cytometry allows for more precise analysis compared to conventional imaging flow cytometry methods. Flow cytometry systems that use quantitative phase imaging (QPI) can potentially analyse samples at a significantly faster rate than human analysts. This resulted in larger datasets for training the neural networks and expediting specimen identification. Not only does this advanced imaging method take high-resolution pictures but it also provides access to other biophysical characteristics, such as those mentioned previously, that are not available with standard imaging flow cytometry. Taxonomic identification attempts may make good use of these parameters, and machine learning can greatly benefit from the quantitative data acquired by QPI. After introducing the concept of QPI data, this study delves into the issue of identifying synergistic cells using trained neural networks. The refractive index of a specimen can be observed in 3 dimensions (3D) using tomographic quantitative phase imaging (QPI) [30]. This tomogram provides additional spatial details of the internal structure of the specimen [31]. For taxonomic identification, tomographic images can be created by merging many 2D quantitative phase imaging (QPI) images with a reconstructed neural network [32]. Significant processing power is required to create 3D tomograms for every cell being investigated [33]. This is due to the fact that capturing the cell's attributes digitally across its volume requires the tomographic reconstruction method to be applied at many focal depths [34]. Examples of QPI methods utilised in research include in-line holography, phase-shifting digital holography, image-plane holography, Gabor holography and off-axis Fresnel holography [35]. As they tend to produce twin pictures, several of these methods are not useful in biological research. Therefore, the pixel array efficiency suffers and the amount of data captured in each picture is reduced. Because in-line and phase-shifting holography eliminates twin pictures, they make full use of the hologram pixel count, which is beneficial for biological studies. A common method for conducting in-depth microscopic material examination is phase-shifting quantitative phase imaging (QPI). It creates an interference pattern on the image surface by using mirrors, beam splitters and lenses. The 'Recommendations' section delves more into Methods 1 and 2, 2 examples of this method. In-line QPI has become a popular real-time method because it requires few optical components for its construction. The fibre-optic cable termination point is an example of a focal point source that can illuminate in-line quantitative phase imaging (QPI) samples. Light from this source travels

through the solution under study and bounces off the nearby objects. The interference pattern created when the dispersed waves interact with the main wave is captured by a charge-coupled device (CCD). Subsequently, a picture of the objects being studied was created using a wave reconstruction method. Method 3 is only 1 example of a rapidly developing field of portable probes that employs in-line quantitative phase imaging (QPI) to conduct *in situ* studies.

The ability of QPI to offer high-resolution 3-dimensional (3D) imaging and collect more specimen information has led to its adoption in many areas of science and medicine. Because it enables the development of incredibly precise, cell-specific 3D pictures, quantitative phase imaging (QPI) of biological materials has grown in favour of medical settings. However, the diagnostic capability of this method was exceptional. The ability to detect and identify leukaemic RBCs and track T-cells that destroy cancer cells in real time has been made possible by the QPI analysis of RBCs. Blood samples analysed using QPI data have made it possible to diagnose diabetic blood cells by evaluating red blood cell morphologies, particularly cell membrane flexibility. Scientists have made 3-dimensional tomographic images of red blood cells infected with the malaria parasite Plasmodium falciparum. These images provide a wealth of data for both the internal structures of parasites and host cells. The structural and chemical characteristics of the parasite throughout its life cycle were investigated. An approach comparable to that employed by Hou *et al.* [36] may be employed in water quality assessments to identify different cyanobacteria species, including potentially dangerous strains, in a given water system. By comparing the cell integrity before and after therapy, Popescu's cell viability monitoring approach determined the efficacy of the treatment methods. In a study of unlabelled live cells, Bej *et al.* [37] photographed and identified Bacillus anthracis simultaneously using a portable quantitative phase imaging unit (QPIU). The following sections provide further details of the neural network employed and trained by Shivaram *et al.* [38]. Method 1 in the recommendations section elaborates on the advantages and disadvantages of this approach. Specifically, *B. anthraxis*, *B. cereus*, and *B. subtilis*, which were 1 μm in size, were imaged quantitatively using this approach. Using the QPI data as the input, the 'HoloConvNet' CNN was trained and was able to obtain a recognition accuracy of up to 96.3 %. The South Korean Agency for Defense Development's BSL-3 laboratory was the intended destination of this portable QPIU; hence, its design prioritised portability. A Rochon prism, 2 linear polarisers, and a charge-coupled device (CCD) camera were part of the system. The sample was positioned into an imaging cell and illuminated by a 532 nm monochromatic laser. Subsequently, the laser beam entered the optical microscope objective, which used oil immersion to increase the magnification factor by a factor of 100. As light passes through a material, its refractive index becomes part of its spectrum. Two light beams with marginally different propagation angles were produced when the light was linearly polarised by the first polariser, and then split by the Rochon prism. Once the beams travel through the second linear polariser, they become parallel in shape. To obtain an optically focused image, the surface of the CCD camera was covered with an interference pattern created using linear polarisers. Using a standard field retrieval method with the given data, a quantitative phase picture is generated, and further investigation might investigate ways to enhance this design for measuring water quality, such as integrating it with flow cytometry tools to automate the analysis to a greater extent. Michelson Coherence Tomography. **Figure 2(d)** shows quantitative phase images of cancer cells acquired

using Michelson interferometry. This allowed for an accurate evaluation of the biophysical cellular characteristics, such as dry mass and radius. The sample solution was automatically processed and data were collected using integrated flow cytometry, which involved moving it through an image flow cell. The photographs were processed according to the steps outlined in **Figure 3(a)**. Pras and Mamane [39] developed a QPIU that uses phase-shifting quantitative phase imaging. An interference pattern can be generated on the lens of a complementary metal-oxide-semiconductor (CMOS) camera using Michelson interferometry. Method 1's portable QPIU also makes use of a 532 nm monochromatic laser for illumination. The imaging flow cell was angled perpendicular to the propagation path of the laser beam to direct its path. Owing to their different refractive indices, the cells in the sample solution underwent phase changes when they crossed the beam as they passed through the flow cell. A nonpolarizing beam-splitter cube was used to separate the beam into 2 identical beams after it was magnified using a microscope objective and tube lens. When these beams recombine after being reflected by the 2 mirrors at slightly different angles, an interference pattern is generated. At a steady rate of 100 fps (i.e. frames per second), the camera recorded the interference pattern as it focused on the surface of the camera.

Using a reconstruction process, such as the Fourier transformation approach, quantitative phase information can be recovered from the pictures. As shown in **Figure 3(b)**, the image processing technique used by Elechi *et al.* [40] was used to train a neural network to recognise cells. Igwaran *et al.* [41] assessed several biophysical cellular properties. Measurements of tumour cell radii in flow were made with precision within 50 - 60 nm, whereas estimates of dry mass were made within 4 - 6 pg. Because of the dependability of the method and the absence of a notable loss of accuracy in flow cytometry experiments, the dry mass and radius measurements of cells in the flow alignment were almost equal to those obtained from static cells. WaterScope Scientists at Hungary's Lake Balaton built WaterScope, a microscope for studying microbes, using the QPI technology. Scientists captured images of microorganisms with sizes varying from 10 to 200 µm using a prototype that did not include a lens. Predictions indicate that this technology will be 50 - 100 times faster than the traditional microscopy for material examination. In contrast to conventional QPI instruments, WaterScope can capture colour pictures of the studied specimens because it uses 3 distinct coloured lasers as the incident light source. Researchers have placed a premium on colour imaging for taxonomic identification. With a success rate of 90 - 95 %, researchers have been able to detect microorganisms in photographs using a basic pattern recognition algorithm. WaterScope demonstrates the potential and importance of Quantitative Phase Imaging (QPI) for water quality; however, it is not currently available for sale. This enabled automated sample counting and identification. Among the submersible microscopes, the 4Deep S7 stands out. One of the first pieces of commercially accessible quantitative phase imaging (QPI) equipment was the S6 submersible microscope, which was mainly intended to measure water quality in its native environment. Underwater holographic microscopes allow researchers to identify bacteria in natural habitats. Environmentally dangerous areas such as the high Arctic Ocean are ideal for development. Instead of polarisers or mirrors, this gadget uses digital inline holographic microscopy to gather images more efficiently. A space between the light source and camera was built into the gadget so that water could be moved by natural convection. The device can capture still images of aquatic subjects as they float past without requiring any movement on their part. This paves the way for a qualitative

assessment of the taxonomy of the sample as well as an approximation of cell counts. Because the amount of water moving through is unknown, this passive monitoring approach restricts the ability to perform a quantitative study by making it impossible to accurately quantify the number of cells.

One major benefit of using QPI to analyse cells, instead of imaging cytometry or bright-field microscopy, is the amount of information it can supply. The endogenous refractive index of each specimen may be associated with its specific biochemical and structural characteristics, as mentioned above. The dry mass and volume of individual cells can be determined by analysing the refractive index data. This is because there is a correlation between cell density and refractive index. These data are essential for training a neural network to identify the cells. The utilisation of a label-free analytical approach is a significant benefit of QPI analysis over imaging and nonimaging cytometry. Cell viability indicators, fluorescent markers, and contrasting agents are examples of labels that can alter the physical characteristics of a sample, thereby compromising its analytical precision. To facilitate a more dispassionate morphological analysis, QPI makes the label discretionary because it provides quantitative phase data based on both wave amplitude and phase. The QPI allows the investigation of specimen sizes ranging from minute to large. Less light is emitted by larger specimens because they absorb or disperse more of the light that comes into them. Against the background of brilliance, this helps them stand out. It is possible that smaller specimens such as single-celled creatures do not absorb or scatter a significant amount of visible light. Consequently, bright-field microscopy may be difficult to perform. These small material samples are called ‘phase objects’ because they induce a phase shift, yet barely alter the amplitude of the light that passes through them. By incorporating the phase interference approach into the QPI, a wider variety of organisms can be observed without chemical tagging, and a more in-depth examination of phase objects can be accomplished. Although the images are generally black and white, QPI provides more information about the target cells than brightfield microscopy. Algorithms for recognition can categorise species based on cellular information such as colour. Access to localised dry mass data by quantitative phase imaging (QPI) is preferable because it provides algorithm information about the inside and outside of cells. For microscopic imaging applications, QPI’s inability of QPIs to achieve high sample throughput rates is a major drawback. The highest analysis rate of quantitative phase imaging (QPI) is 2 times lower than that of non-imaging flow cytometers, which can attain throughputs of approximately 100,000 cells/s. Similar to imaging flow cytometers, QPI imaging rates are constrained by camera technology, which creates an inherent trade-off between the speed and sensitivity. It is common practice to sacrifice subcellular image resolutions to achieve higher sample throughput rates; however, this compromises the picture quality and introduces motion blur. Images often only allow for tens or hundreds of cells owing to the processing limitations imposed by massive amounts of data. It is essential to reduce the resolution of these photographs to make up for it. The development of ultra-high-throughput QPI methods and deblurring algorithms, discussed at length in this article, could potentially overcome this limitation. Creating processes for quality performance improvement.

The research and development of QPI, a technology with several innovative uses, is continuously underway. The goal of the synthetic aperture method is to double the maximal spatial resolution so that more data can be extracted from the same specimen. Combining QPI with fluorescence microscopy may

increase the amount of data collected from the specimens, which in turn improves the accuracy of machine learning and identification. To obtain the current maximum throughput limits of QPI technology, Patriarca *et al.* [42] developed a new method called multiplexed asymmetric-detection time-stretch optical microscopy (multi-ATOM). This method allows for massive categorisation of individual cell samples by merging flow cytometry with integrated quantitative phase imaging (QPI). Its analytical capabilities and sample throughput were 100 times better than those of conventional QPI. Scientists were able to capture images at 10,000 fps while achieving sheath fluid flow rates of up to 2.3 m/s. The phase gradient of light passing through each cell was measured using a knife edge that partially obstructed the beam. A detailed picture was produced because of the ability to quantify intensity variations throughout the cell dimensions. After collecting phase and amplitude contrast data, grayscale brightfield pictures were generated for every cell, which is a novel optical imaging method makes use of. Encoding spatial cell information into the spectrum of a broad-spectrum light pulse is possible with the help of a light beam splitter and dispersive medium, such as a lengthy fibre optic cable. The frame rates achievable with time-stretch photos are several million frames per second, which is far higher than the 100 fps achieved using the conventional techniques. Because this real-time imaging method enables quicker data gathering, it is highly beneficial for thorough imaging of individual cells.

Stochastic integrations

The European Union reported 1,876 instances of listeriosis in 2020 caused by the bacterium *Listeria monocytogenes*, which is common in food. Among zoonotic diseases in the European Union, it has a worst-case fatality rate of 10 %. Global food safety procedures place a premium on preventing listeriosis and its associated severe neurological consequences. People, animals (both domestic and wild) and plants are among the many hosts that *L. monocytogenes* can infect. The most common way these hosts obtain *L. monocytogenes* is by ingesting tainted feed or food. The presence of *L. monocytogenes* in the environment, the fact that hosts excrete the bacteria in their faeces, and the fact that the bacteria can adapt to different environmental pressures all point to the same possible origin for the bacteria in feed and food. The food chain connects the agricultural ecosystem to the people consuming it. Raw materials from animals (milk, etc.) or tainted products (dirt, dung, etc.) are the main entry points for *L. monocytogenes* in food-processing plants. Poor hygiene management raises the possibility that the disease may spread from unintended sources. Some clonal complex (CC) 9 and 121 *Listeria monocytogenes* strains have demonstrated superior survival rates in food production settings. Because *L. monocytogenes* is resistant to quaternary ammonium compounds and other disinfectants, it may persist in some areas of food processing plants for decades. Primary production on the farm, processing, retail and consumer are the 4 stages of food production where *L. monocytogenes* contamination can occur as a result of inadequate hygiene practices. The circumstances necessary for a listeriosis epidemic are satisfied if *L. monocytogenes* can proliferate in consumable feed or food matrices without inactivation by heating. A broad variety of clinical manifestations can be caused by infection in both humans and animals. Some people may not show any symptoms, but can still transmit the virus. However, more serious illnesses may develop, such as blood poisoning, brain inflammation, or miscarriages. Despite extensive research into bacterial virulence factors and host pathomechanisms, the

methods by which some individuals become asymptomatic carriers of *Listeria monocytogenes* are largely unknown. The inIA gene encodes the surface protein internalin A. However, some strains derived from asymptomatic human carriers have shortened versions of this gene. It is possible to house bacteria without experiencing any symptoms if the inlA gene is truncated, as this variation has been linked to a decrease in the disease-causing capacity. It is possible that consumers were exposed to these germs because these shortened variations were more common in food isolates than in clinical isolates. Additional evidence of asymptomatic transmission of viable but non-culturable (VBNC) *L. monocytogenes* has been found. Similar to other bacteria, *L. monocytogenes* has been shown to move from a growing state to a resting, inactive state. However, the presence of *L. monocytogenes* in its VBNC (viable but nonculturable) form poses a problem for diagnostics. This is because most current diagnostics require a culture phase, which means that they cannot identify nongrowing VBNC cells. Several PCR and qPCR techniques, together with DNA intercalating dyes, have been developed recently to differentiate between live and VBNC cells, which is a relief. A major obstacle to guaranteeing food safety is for those who do not show symptoms, but can transmit illnesses. However, recent research has shown that the gut microbiota is an important defence mechanism against food poisoning. The authors stressed that a particular microbiota pattern is linked to the asymptomatic release of *L. monocytogenes*. Researchers have found that this harmful germ is present in faeces, even in healthy people, and that the build-up of bacteria in the gut is a major factor. Strict regulations are in place to ensure that sick animals are not used to produce milk or meat. Sporadic passage of stool by asymptomatic individuals is usually undetectable in animals and humans. Three potential forms of contamination can arise from *L. monocytogenes* in the faeces of farm animals, all of which could affect food safety: (i) the risk of illness in other animals in the barn area is increased because the incidence of *L. monocytogenes* is increased; (ii) there is an increased risk of *L. monocytogenes* infection in feed and crops because of the use of animal waste as fertiliser and the damage that excess water from farms can cause to water systems; and (iii) proper hygiene measures are not taken when milking or killing animals may contaminate raw milk and meat. Finally, due to inadequate hand cleanliness, *L. monocytogenes* may contaminate food or the area where food is processed, even in symptom-free individuals. At the end of this study, the subject matter in the Discussion section provides a better understanding of the risk factors associated with faecal matter excretion in different species and a more up-to-date analysis of the 2012 review of the frequency of asymptomatic carriers in different species. Some domestic animals may have diseases without any symptoms. Ecological studies on Listeria have shown that the bacteria can live in any mammal, including house pets, such as dogs and cats, and that many of these animals carry the germs latently. On rare occasions, *L. monocytogenes* may pass through the faeces of these animals. In addition, tonsil samples collected from domestic animals in good health showed a high frequency of *L. monocytogenes* according to prevalence statistics. However, it is unclear how domestic animals contribute to L. monocytogenes transmission. Currently, there is no evidence of pets transmitting *L. monocytogenes* to humans. Raw meat-based diets (RMBDs) are becoming increasingly popular among pet owners as alternatives to dry or canned pet food for dogs and cats. An investigation by a Dutch group found that RMBDSs could potentially infect companion animals with *Listeria monocytogenes*. People may be in danger of spreading the disease. The researchers considered a total of 35 raw meat-based diets (RMBDs)

made by 8 separate companies. The presence of *L. monocytogenes* in 54 % of the analysed samples was surprising. *L. monocytogenes* can be found in large quantities in pigs. Nger animals are particularly susceptible to the asymptomatic transmission of the virus. Tonsils from fattening pigs contained 22 % more *L. monocytogenes* than those from sows, which only contained 6 % more *L. monocytogenes*. The pig tonsils contained CC6 and other extremely dangerous *Listeria monocytogenes* strains. There is extensive contamination or recontamination of meat from a given source at the slaughterhouse, as evidenced by the presence of closely related strains throughout the production process. The identification of *L. monocytogenes* in otherwise healthy pigs is highly dependent on their living conditions. Pöttker *et al.* [43] found that *L. monocytogenes* was more common in animals raised on organic farms than in those raised on conventional farms. The European Union Regulation on Organic Farming (EU 2018/848) specifies that pigs raised in this manner must have access to straw bedding and free range space. Organic production techniques sometimes include housing pigs in larger groups, which may potentially expose the animals to more *L. monocytogenes* through their surroundings or fellow group members. However, pigs reared in intensive indoor farming experience chronic social, environmental, and metabolic stress, which may increase *the release of L. monocytogenes*. Experiments on healthy dairy cattle have revealed a wide range of faecal sample frequencies, from as little as 1.9 % in individual animals to as high as 46 % in beef herds. Swiss dairy cows tested positive for antibodies against various *L. monocytogenes* pathogenicity proteins, including internalin A and listeriolysin O, accounting for 11 % of the healthy cows. Therefore, dairy cows in Switzerland are likely to come into contact with *L. monocytogenes* regularly. To monitor the prevalence of *Listeria spp.* in dairy farms over 3 consecutive seasons, a large-scale longitudinal study was conducted in Spain. According to previous research, the season and age of cattle affect the prevalence of *L. monocytogenes*. More cases occurred throughout the winter, and second-lactation cows were more prone to excretion of *L. monocytogenes*. External variables, such as feed contamination and time of year, significantly affected the likelihood of *L. monocytogenes* shedding in cow faeces. There appeared to be a strong correlation between shedding and dietary habits. Farms fed animal-polluted feed had a higher prevalence of *L. monocytogenes* in their faeces. Indoor seasons often have a higher prevalence of *Listeria spp.* and *L. monocytogenes* than grazing seasons. This subtype affects *L. monocytogenes* shedding in research farm cows. The frequency of *L. monocytogenes* in cow faeces was also significantly correlated with the presence of contaminated silage. In addition, there is evidence linking stress levels in animals to the shedding of L. monocytogenes. Ruminant animals are at a high risk for listeriosis when fed low-quality silage, which has fermentation problems and pH levels that encourage the development of *Listeria monocytogenes*. Ruminants shed *L. monocytogenes* at different times of the year, which may explain this. Sheep and goats on farms that were fed silage had a 3- to 7- increased chance of isolating *L. monocytogenes* compared to those on farms that were not fed silage. It should be mentioned that the most common clonal group associated with human listeriosis is *L. monocytogenes* clonal complex 1, which is strongly linked to dairy products and animals. Owing to its ability to infect a wide variety of vertebrates, *L. monocytogenes* may have adapted to the specific conditions of the preferential ruminant forestomach by associating with ruminants. Before exposure to acidic conditions in the abomasum, the large rumen enables fast growth of *Listeria monocytogenes* in the pH range of 6.5 to 7.2 and body temperatures of 38 - 40.5 °C. Scientific

investigations corroborate this concept by showing that after *L. monocytogenes* leaves the sheep gastrointestinal tract (GIT), it produces a short infection with no symptoms and is then eliminated in small quantities through faeces. A bacterial reservoir is present in rumen digesta. This study found that *L. monocytogenes* in asymptomatic sheep is caused by both bacterial transit through the body and short-lived development in the rumen. The extent to which this enlargement occurred was influenced by the age of the animal and quantity of *L. monocytogenes* consumed.

Asymptomatic carriers of *L. monocytogenes* are also found in ducks, geese, turkeys and poultry. A recent study revealed that cloacal swabs only tested positive at 1.3 %, whereas carcase rinses tested positive at 11 %. Additionally, there is a plethora of data showing how often contamination occurs in poultry processing plants, which in turn contaminates chicken and meat products. Transportation stress is one of the most important causes of increased shedding and the subsequent contamination of manufacturing lines. In summary, *L. monocytogenes* can be isolated from healthy domestic animals without causing symptoms. Faecal shedding is infrequent, although it can be increased by some husbandry practices, including feeding silage, and by stresses related to housing, group size, and transportation. Studies on animals that shed *L. monocytogenes* do not usually consider their health. This might be due to a lack of assessment of the animals' health, collection of animal droppings in an unsupervised environment, or unfamiliarity or difficulty in identifying listeriosis symptoms in that particular species. Another potential issue with catch-and-release trials is that they may capture sick animals favourably. Therefore, it is difficult to determine whether wild animals are 'asymptomatic'. The assumption that wild animals contribute to the transmission of *L. monocytogenes* in different contexts, regardless of whether they show symptoms or spread excretion, is important for this investigation. Consequently, they need to be considered when determining potential dangers to food safety. **Figure 4(a)** shows that several bird species, such as pheasants, pigeons, gulls, crows, rooks and sparrows, carry *L. monocytogenes*, but show no symptoms. The 996 birds, representing 18 different species, were part of a massive Japanese prevalence study that examined faecal and intestinal samples. Crows were the most frequently affected species, with *Listeria spp.* being detected in 13.4 % of the samples. The research also found that 33 % of urban rooks had *L. monocytogenes* in their stool. Faeces from wild birds in Finland, including pigeons, sparrows, and gulls, contain 36 % *L. monocytogenes* [45-49]. The pulsotypes identified in birds were similar to those in the food web, suggesting that birds may be involved in the transmission of human pathogenic strains of *Listeria monocytogenes*. Many different types of animals, including avian species, carry *L. monocytogenes*. **Figure 4(b)** shows the prevalence of *Listeria spp.*, including *L. monocytogenes*, in several vertebrates, including reptiles, and mammals, including deer, rats and wild boars. *Listeria spp.* were found in 38 (6.1 %) of the 623 wild mammals (8 different species) studied in a Japanese study that examined faecal or intestinal samples. The highest number of *Listeria spp.* isolates (16) was found in monkey samlest (38, 20 % (16 of 80) of monkey samples met this criterion. Deer, moose, otters and raccoons were among the many species whose faeces were among the 268 examined by Canadian researchers. For 29 % of the total, 35 samples (*L. monocytogenes*) yielded positive results. 19 (42.2 %) of the 52 wild boars and 45 red deer caught in Germany and Austria in 2011 and 2012, respectively, tested positive for *L. monocytogenes*. In addition, bacteria were found in 12 boars (25.5 % of the total) and 4 out of 22 combined feed samples (18.2 %). The tonils and ruminal or caecal contents of

many samples were positive for *L. monocytogenes*, even though no faeces were observed. This proves that *L. monocytogenes* may be present in game animals even if it is not detected in their excrement. Rumens of deer and tonsils of boars were the most frequent sites of *L. monocytogenes* infection. Pulsed-field gel electrophoresis revealed a diverse range of strains, with 1/2a and 4b as the most prevalent serotypes. An investigation examining the possibility of free-living carnivores serving as reservoirs for *L. monocytogenes* was conducted in Poland. Five % of the examined animals tested positive for *L. monocytogenes*. This group includes raccoons, beech martens and red foxes. The number and configuration of virulence genes varied across the remaining isolates, with 35 % harbouring all of them.

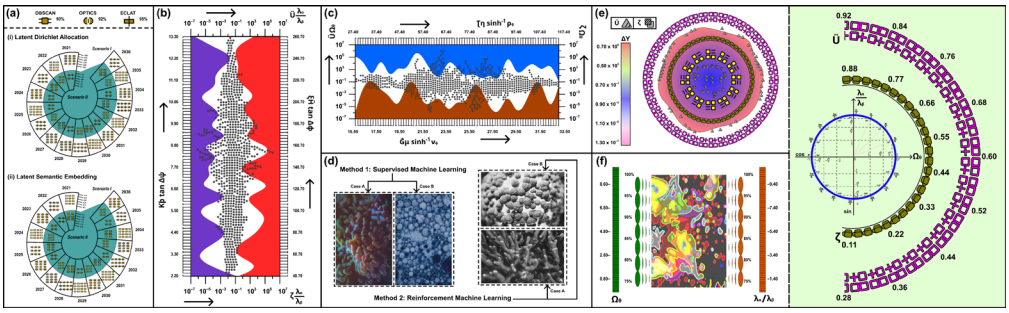

Figure 1 (a) Bayesian mixed models that describe geographical and temporal trends in $\log_{10}$ FIB levels based on the presence of human and ruminant Multilocus Sequence Typing (MST) markers (which are represented as a distribution of pre-existing machine learning algorithms in 2 distinct situations over a period of 15 years, as determined by principal component analysis of the data mentioned in Section 3.3. Non-blind deblurring utilises methods tailored to certain optical setups, whereas blind deblurring uses a Convolutional Neural Network (CNN) or Generative Adversarial Network (GAN) that directly learns the transformation from a blurred picture to a crisp one); (b) Fibonacci sequences depicting highest total cell counts of cyanobacteria species present in both early- and advanced-stage harmful algal blooms (HABs) throughout many freshwater systems in Alberta, Canada, that were determined using FlowCam (8000 Series, Yokogawa Fluid Imaging Technologies) and Amnis ImageStream (MkII, Luminex); (c) Projected cytometry of ion mobility spectrolyser data (of L. platensis PCC7345 pre-cultured in 30 °C in 250 mL shaking flasks with an LED lamp; pre-culture rediluted with fresh modified Zarrouks culture medium twice a week) with significant Marks-Kendall metric, alongside scatter plots of Trellis-density correlations between 3 waterways sampled for microbial source-tracking marker (MST) analysis, representing 5, 10 and 49 % of the Onondaga Lake drainage area, respectively (Note: MSTs collected only from sites where faecal coliforms were consistently high during routine monitoring efforts); (d) Multispectral CDOM sensors can be integrated into existing supervisory control and data acquisition (SCADA) systems to track rapid changes in water quality, represented via clockwise schematics of case studies incorporated for the machine learning analysis, wherein the release of volatile organic compounds (VOCs) is monitored in real time using ion mobility spectrometry (IMS) in combination with a membrane inlet; (e) Polar Akima Interpolation of average $\log_{10}$ concentrations of E. coli, Enterococcus, total coliform and faecal coliform, rectified via Python t-SNE Dimensionality Reduction for which HoloConvNet automatically determines a ‘cell

fingerprint' of important biological traits within images provided to it, allowing direct training of the CNN from raw QPI data; (f) Normalized Laplace transform of PCR discriminating human and ruminant faeces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA (identified via a culture-independent approach based on molecular methods detected *Listeria monocytogenes* in MG-RAST human microbiome datasets).

Listeria are commonly associated with seafood and prepared fish, particularly cold-smoked salmon, in humans. The majority of cases are probably caused by *Listeria monocytogenes* strains that persist in food processing facilities after harvest. The prevalence of *Listeria monocytogenes* in farmed fish and their surroundings is influenced by several variables, including water pollution, agricultural runoff and seagull poop. Fish, especially salmon, that have been taken in the wild are another option to consider. A recent Norwegian study found that *L. monocytogenes* was likely introduced and maintained at a salmon-processing firm by recently deceased fish that were contaminated with the bacterium. However, *L. monocytogenes* does not cause as much damage to the fish. The pathogen was successfully eradicated from the stomachs of live salmon within 3 days of infection, without harming the animals. Although *L. monocytogenes* was present in the water of fish farms where rainbow trout were raised, the infection rate was minimal. According to recent research, infected tilapia have a frequency of 5.8 % in the domestic market and of up to 1.2 % during capture. Although fish are not expected to transmit the disease for an extended period, the results indicated that L. monocytogenes can be obtained from contact without showing symptoms. *L. monocytogenes* can infect protozoa, insects, and reptiles. By 2021, the number of reptiles kept as pets in Europe is projected to reach approximately 11 million. The potential danger of infection in reptile owners requires further research. It was recently discovered that Arion vulgaris, often known as Spanish slugs, carries *L. monocytogenes*. While 43 % of the 710 slug samples returned positive results, 16 % had an average of 405 CFU/g of slug tissue. Only 11 % of 62 slug cultures were surface- or mucous-positive. In addition, 22 days after receiving *L. monocytogenes*, the slugs continued to release live germs into their excrement. It has been demonstrated that ants can occasionally harbour *L. monocytogenes* and are capable of carrying harmful diseases from polluted areas to food, indicating that many invertebrates and vertebrates (including birds, mammals and reptiles) may serve as *L. monocytogenes* carriers, allowing the asymptomatic spread of the disease from one ecosystem to another. In humans, asymptomatic carriage of *Listeria monocytogenes* means that the bacteria are present in the body but do not cause any symptoms. The faeces of both wild and domestic animals can carry *L. monocytogenes*. The frequency of excretion in healthy individuals is often less than 5 %, as established by culture (**Figure 4(c)**). However, humans occasionally secrete *L. monocytogenes*. Surprisingly, subsequent research in Austria examined the faeces of 3 persons over a 3-year period, even though the United States and Austria reported low rates of bacterial carriage in healthy individuals. The percentage of samples found to be positive for *Listeria monocytogenes* was 1.2 % (10 out of 868). Serotypes 1/2a and 1/2b were all represented in the positive samples. Additional research found 5 separate cases of bacterial contact, each of which occurred asymptomatically. This amounts to 2 exposures per individual on a yearly basis. There has been an increase in the number of

reported clinical cases among women aged 25 - 44 years and individuals aged 75 years and older, according to a scientific assessment of the contamination of ready-to-eat foods with *L. monocytogenes* and its effects on human health in the European Union. Ingestion of ready-to-eat (RTE) food containing more than 2,000 CFU/g is the suspected culprit in over 90 % of invasive listeriosis cases, as revealed by quantitative modelling. Organisms in the consumer phase of development account for one-third of these cases.

Even in individuals who do not exhibit any visible symptoms of illness, *L. monocytogenes* may still be present because of underlying medical conditions. The prevalence of the bacteria may be higher in those undergoing renal dialysis who also use the antacid ranitidine, which is an H_2 receptor antagonist. However, pregnant women with HIV who were taking antiretrovirals did not have similar levels of *Listeria spp.* or *L. monocytogenes* in their faecal samples. The prevalence of asymptomatic carriers in humans does not change during pregnancy. One study found that faecal carriage occurred in only 2 % of 51 pregnant women (measured between weeks 10 and 16). Compared with the pregnant group, only 3.4 % of the 59 control subjects showed signs of faecal carriage. In addition, during the Los Angeles epidemic, researchers compared the faecal carriage rates of pregnant women with listeriosis to those of non-pregnant controls with comparable characteristics. According to a previous report, both the groups had similar carriage rates. According to a recent study that found an imbalance in the microbiota of breast milk, mothers whose infants suffered from severe acute malnutrition (SAM) may have had higher levels of *Listeria monocytogenes* in their milk. Of the 3,338 human microbiome datasets obtained using MG-RAST with 16S sequencing, 173 (5.2 %) included *L. monocytogenes* using an impartial molecular method-based approach. In addition, out of 900 stool samples (10 %) collected from healthy individuals, PCR analysis showed the presence of *L. monocytogenes* in 90. It should be noted that DNA-based detection methods are unable to distinguish between live and dead organisms when analysing these data. According to the aforementioned study, the prevalence of *L. monocytogenes* is associated with certain gut flora. Unexpectedly, research in mice has suggested that a shift in the typical microbiota due to age might lead to an increase in the prevalence of *L. monocytogenes* in the digestive tract. In addition, those whose jobs put them in close proximity to animals, manure, meat, or who are constantly exposed to germs on the job are more likely to experience subclinical diseases. For instance, a study found that 16 % of hand swabs from agricultural workers and 6 % of hand and garment swabs from slaughterhouse workers had *L. monocytogenes*. This rate was higher than that often observed in faecal samples collected from healthy individuals (**Figures 1(a) - 1(c)**. The exposure of the general human population to *L. monocytogenes* was determined using baseline studies conducted throughout the European Union in 2010 and 2011. A total of 13, 088 food samples were collected to identify *the L. monocytogenes* isolates. The prevalence rates of contamination in meat samples were 2.07 and 0.47 %, respectively, whereas fish samples had a frequency of 10.4 % across the European Union. Mermans *et al.* [50] analysed ready-to-eat foods in Austria. Among the 946 food samples gathered from Vienna's food vendors, 124 (13.1 %) contained *Listeria spp.* and 45 (4.8 %) contained *L. monocytogenes*. Contamination levels were 19.4 % higher in shellfish and fish. The contamination rates in raw beef sausages were 6.3, 5.5 and 4.5 % for soft cheese and cooked meat, respectively. In addition, homes in the same area provided the samples. While 1.7 % of the 640 items tested positive for *Listeria monocytogenes*, 5.6 % tested positive for *Listeria spp.* Surprisingly, mixed faecal samples from the same household revealed the same strains as

those in the aforementioned food items. Faeces can still be released from meals even when the contamination level is low (< 100 CFU/g). The influence of asymptomatic carriers on *the prevalence of L. monocytogenes* in agricultural settings and food-processing facilities is crucial. *Listeria monocytogenes* is excreted in the faeces of asymptomatic agricultural animals. Therefore, the virus is becoming more common in agricultural settings, which may compromise feed and food quality. A part of a comprehensive plan to guarantee food safety is to examine the ecological factors linked to *L. monocytogenes* in farming settings. To reduce the prevalence of pathogens, it is essential to first determine where *L. monocytogenes* is present on farms and then to eliminate it. The updates to this field's research are summarised in **Figures 1(b) - 1(d)**. According to the study findings, asymptomatic faecal shedding is more common in farm animals whose diets contain silages. *L. monocytogenes* was found in silages at a frequency ranging from 2.5 % in clamp silage to 22.2 % in large bales, with a significantly higher prevalence of 44 % in moldy samples. It is believed that *L. monocytogenes* enters silage when soil enters raw grass. An insufficient acidity level caused by insufficient silage fermentation permits the growth of *L. monocytogenes* at potentially harmful levels. The low bacterial load required for initial contamination likely explains why *L. monocytogenes* was rarely found on preprocessed grass and vegetables. Animal feed (in the case of damaged grass) and food (in the case of contaminated crops) may be at risk if bacterium-infected plant material can persist for weeks. A total of 27 % of the water samples and as much as 99 % of the wastewater from stabilisation ponds in northern Canada tested positive for *Listeria monocytogenes*. In addition, *L. monocytogenes* may be able to adapt, survive for a long time, and spread to many different ecosystems because of its capacity to enter VBNC conditions. *L. monocytogenes* is commonly found in feed bunks, water troughs, bedding and silage. Its widespread presence in the soil and subsequent transmission through feed and animals are consistent with this. According to recent research, dairy farms may promote the development of more aggressive clones of *Listeria monocytogenes*. This clone can contaminate food supply. Ruminant farms may encourage the persistence of *L. monocytogenes* in the environment through a chain reaction that begins with contaminated feed, continues with reproduction inside the animal hosts, and ends with environmental contamination through excretion. Despite the rarity of listeriosis symptoms in pigs, pork consumption has long been associated with human illness. Even seemingly healthy pigs can be carriers of diseases due to contamination in processing and slaughterhouses. In in-out or clean-and-wait pig facilities, the risk of *L. monocytogenes* infection is greatly increased if there is less than a one-day gap between the facility being emptied and the growing pigs being introduced to the fattening area. According to the same study, a group of animals that almost finished eating wet food had a higher chance of colonisation with *L. monocytogenes*. The presence of biofilms in pipes and valves, as well as the leftover food layers, are likely to be responsible for this. The presence of *L. monocytogenes* in otherwise healthy pig faeces tends to increase during transportation from farms to processing facilities. However, the slaughter and processing steps seem to be the main culprits of *L. monocytogenes* contamination in food. Bacteria may live quite a few times in these environments.

The persistent nature of infection in agricultural settings, coupled with the fact that animals and people can shed the *L. monocytogenes* virus asymptomatically, poses a threat to the health of both animals and humans. At this point in the food production process, it is crucial to prioritise sanitary milk production,

effective farming techniques and maintaining a high level of feed and animal cleanliness because these are the main entry points for *L. monocytogenes*. The danger to human consumers increases when animals infected with *L. monocytogenes* are raised owing to inadequate sanitation measures on farms. Such practices include not washing hands, cleaning boots, wearing protective clothes and neglecting the silage quality. The colonisation of food processing equipment and facilities by microorganisms includes contamination from raw food sources, inadequate hygiene standards and pests. Commonly found strains of *Listeria monocytogenes* in food-processing settings can cling to surfaces more quickly, increasing their chances of survival and even dominance in such facilities. Sodium chloride, an element often used in food preparation, promotes self-aggregation and enhances the adherence of *L. monocytogenes* to plastic surfaces. Researchers have found that clinical strains are more capable of producing illnesses, whereas persistent strains may be less capable. Some studies have shown that occasional strains of *L. monocytogenes* differ from those that cause chronic infections. However, the origins of these persistent strains remain unclear. Disinfectants may alter the destructive power of the infection. Asymptomatic cows can pass *L. monocytogenes* through their faeces. Contaminating raw milk is 1 way that *L. monocytogenes* can enter dairy-processing facilities. Processed goods are at a risk of contamination when bacteria establish biofilms on surfaces that are resistant to disinfectants. Re-contamination of dairy products can occur even after short-term, high-temperature pasteurisation, which can effectively kill *L. monocytogenes*. This is because the bacterium can survive and multiply under processing conditions that follow pasteurisation. Due to its tendency to attach to surfaces, its special characteristics that allow it to proliferate and survive and other factors, the complete eradication of *L. monocytogenes* is challenging. A particular strain of *Listeria monocytogenes* was found to be prevalent on food contact surfaces, floors and drains in a chicken processing plant in Northern Ireland. This strain is thought to have originated from birds that were brought in and are often found in raw meat processing areas. After a year, the only strain found in cooked chicken products and the processing area for cooked poultry belonged to this genotype; the raw and cooked meat sections shared a third genotype. This shows how processed goods in the same plant may be infected by persistent pathogens. It is crucial to implement control measures starting on farms because there is much evidence that animals and people may host *L. monocytogenes*, which means it can contaminate processed goods. People, animals and the planet should be included as part of this comprehensive one-health strategy. When people in the food industry and others who interact with animals learn about asymptomatic carrying, their cleanliness procedures are influenced. By lowering the bacterial burden on crops and animal feed, proactive farm hygiene practices can reduce critical factors that allow *L. monocytogenes* to infect both animals and humans. Preventing the transmission of *L. monocytogenes* to plants irrigated and treated with sewage treatment systems is particularly important. To stop the cycle of *L. monocytogenes* transmission from asymptomatic ruminant carriers to grass via manure and subsequently to other animals via contaminated feed, it is necessary to provide proper feed hygiene and manage fermentation during silage production. The following are some examples of effective preventative measures: keeping raw and processed foods in separate areas, thinking carefully about whether animals need silage, not watering crops immediately before harvest, keeping all surfaces and equipment that touch food clean, keeping food handlers' personal hygiene strict, and checking the processing area for *L. monocytogenes* on a regular basis.

Experiment frameworks

The *L. platensis* PCC7345 strain was obtained from PCC (Paris, France). Pre-cultures of *L. platensis* were stored at 30 °C in 250 mL tubes. The flasks were lit daily for 18 h using an LED lamp (PGLE 18 RBW; Venso EcoSolutions AB, Sweden). The light intensity in the bioreactor was determined using these settings. The pre-cultures were supplemented with a newly modified Zarrouk culture medium twice a week until the optical density at 750 nm (OD_{750}) reached a range of 0.2 to 0.6. Before dilution in the primary reactor, the optical density of the pre-culture at 750 nm (OD750) was measured and found to have an initial value of 0.11 ± 0.01. Using the prescribed formula, the culture medium was combined with the indicated concentration in L^{-1}. The chemicals used in the experiment were as follows: 12.0 g of sodium bicarbonate (99.5 %, VWR Chemicals, Darmstadt, Germany), 0.5 g of potassium bicarbonate (99 %, Carl Roth, Karlsruhe, Germany), 2.5 g of sodium chloride (99.9 %, VWR Chemicals, Darmstadt, Germany), 1.0 g of potassium bicarbonate (99.5 %, VWR Chemicals, Darmstadt, Germany), 1.0 g of sodium chloride (99 %, Carl Roth, Karlsruhe, Germany), 0.2 g of magnesium sulfate ($7H_2O$, 99 %, Carl Roth, Karlsruhe, Germany) and 0.04 g of calcium chloride (Cl_2). A laboratory vessel composed of Duran GLS80 (DWK Life Sciences GmbH, Wertheim, Germany), with a capacity of 1 L, was used. The culture medium was maintained at 30 °C with agitation at 300 rpm. The culture was fed at a flow rate of 100 mL/min of filtered air using MI (**Figure 1**). Samples of *Chlorella vulgaris* SAG 211-11b were obtained from SammLung von Algenkulturen Göttingen (SAG), Göttingen, Germany. The organisms were cultivated in a modified medium containing Bristol. To prepare 1 L of BMM, a solution was created by combining 10 mL each of mineral salt, nitrate and phosphate solutions with 1 mL of trace element solution in a 1 L volumetric flask. Viscosity increased with the addition of filtered water. Sodium hydroxide or hydrochloric acid was used to adjust the pH to 6.6 to 6.8. After the transfer, a 1 L Schott container was used to sterilise the combination by autoclaving at 121 °C for 20 min. Cultivation occurred at least 3 times between Dresden's Technical University and Leipzig's Helmholtz Center for Environmental Research. Online measurements were conducted during the culture phase, enabling the transfer of molecules from the humid reactor gas stream to the input gas stream of the ion mobility spectrometer. Symmetric channels are present on both sides of the membrane module. Each channel had a length, radius, depth and width of 112 mm, and they were all 1 mm wide and 1 mm deep. The floor space for the exchange was 112.8 mm^2. The gas streams were channelled into a counterflow configuration using a membrane module. The regulation of a gas flow rate of 100 mL/min from the bioreactor was achieved using membrane pumps (3003 series; Gardner Denver Thomas, Inc., Fürstenfeldbruck, Germany). The internal pump of the ion-mobility spectrometer extracted a gas stream at a flow rate of 10 mL/min from the sample. The 127 μm PDMS membrane utilised in this study was provided by J-Flex Rubber Products (Nottinghamshire, UK). A 20 Ω power resistor from the 35-Watt PWR220T-35 Series was employed to raise the temperature of the membrane cell to 70 °C.

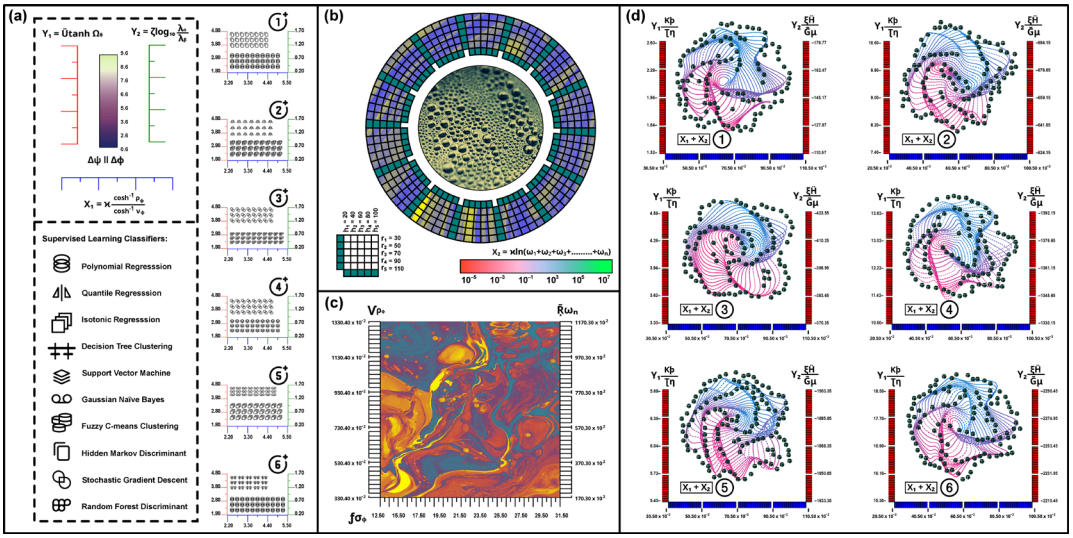

Figure 2 (a) Correlation between human indicators and agricultural land use and rural regions implicating instances of human faecal pollution in the rural hinterlands, represented by a stacked hierarchical band structure of supervised learning classifiers within the Trellis-density correlations, simulated in 6 sets using nominal range sensitivity analysis and clustered by double Y-axes' offsets in standard XY scale; (b) Sankey-visualization of stacked histogram plots that show a normal distribution of *Chlorella vulgaris* SAG 211 - 11b (obtained from SAG, Göttingen, Germany, cultivated in Bristol Modified Medium (BMM); pH adjusted between 6.6 and 6.8 with sodium hydroxide or hydrochloric acid); (c) Effect of test temperature on critical and sprint swimming speed of juvenile *Maccullochella peelii*, acclimated to 14 °C (gradual, intermediate and rapid exposure) or 24 °C (n = 8 per treatment per test temperature) expressed as the Schoeller-Contour profile of correlations between the peak intensity of OD750 datasets (PS > 0.50, PD > 0.75 and ROPE 0.25) pertaining to cell concentration in MI-IMS growth phase, depicted alongside a spectroscopic packing of Bland-Altman density data (in Bayesian regression, PD and ROPE measures are independent above a given threshold, representing a negligible effect in the median's direction); (d) Stochastic cylindrical resonances emphasizing the covariance of *E. coli* levels in Onondaga Creek (New York, United States), increasing by 0.30 $\log_{10}$ CFU/100-mL (Note: When the frequencies of data vary amongst sensors, a decision is taken to either interpolate low-frequency data or discard high-frequency data).

Ion mobility spectrometers (IMSs) categorise the ions produced into different groups according to their transit duration to the detector during the drift phase. For a specific situation, the initial action involved temporarily stimulating the ion gate for 100 μs to generate a cluster of ions. By subjecting the drift tube to an electric field, the ions were driven towards the detector at a uniform and ion-specific speed. The ions in the drift tube decelerated as a result of their encounters with neutral gas. Therefore, ion masses, charges, and collision cross sections were employed for ion separation. The time required for an ion to travel a distance L from the ion gate to the detector is referred to as the drift time and is commonly denoted as t_d.

The electric field E governs the individual motion durations of all ions. Ion Mobility Spectrometry (IMS) systems can be calibrated to operate at a certain temperature and pressure by assigning a distinct ion mobility, K, to each ion. As a result, the reduced ability of ions to move, known as the ion mobility (K_0), can be computed. The IMS system employed in this study included a tritium (3H) ionisation source. The operating parameters used throughout the measurement process are as follows: Ionisation originates from the removal of 1 or more electrons from an atom or molecule, resulting in the formation of ions. This study used tritium, which had an activity of 50 MBq. The shutter was then opened for 100 μs. The length of the drift tube was 56 mm. The drift voltage was 2,470 volts. The discovered ions were positively charged. The drift gas, which consisted of air, flowed at a rate of 340 mL/min. The drift tube was maintained at a constant temperature of 80 °C. The pressure was maintained at ambient temperature between 950 and 1,050 hPa throughout the experimental period. The spectra were captured and stored in Smellmaster 2 at 10-minute intervals. An internally developed program was used to analyse the data. Another technique for standardising the measurement is to compare the drift time with that of a well-defined peak. The reactant ion peak, which signifies the final outcome of ionising the air components and then transferring protons to the target analytes through reactions, is often the most prominent peak in the ion mobility spectrum. Positive-polarity ion mobility spectrometry (IMS) uses an ion called $(H_2O)_nH^+$, which is a protonated water cluster. The presence of ionised air components generates a reactant ion peak, which, when calibrated with the drift time, effectively removes any potential effects caused by the instrument. The standard deviation of the measurement was lower for the relative drift time t_{rel} owing to its independence from instrumental influence. Photometry was employed for offline data analysis, whereby Dr. Lange CADAS 200, a spectrometer manufactured by Dr. Bruno Lange GmbH of Berlin, Germany, was used to perform offline photometric measurements on a regular basis, often 5 or 6 times daily. A small quantity (approximately 4 mL) was extracted from the reactor and further diluted until the optical density (OD_{750}) decreased below 0.2. Subsequently, a quartz cuvette with a length of 1 cm was used to quantify the OD_{750}. The reactor was refilled with new culture medium to maintain a steady culture volume. The algae were agitated before each of the 6 OD_{750} readings. By capturing the entire wavelength range from 400 to 900 nm, it was possible to determine the amount of pigment present. The data were gathered using an IFU GmbH Privates Insti-Content Smellmaster 2 ion-mobility spectrometer. **Figure 1(c)** shows the correlation between algal dry weight (c_x) and OD_{750}.

Tools and techniques

Water treatment plants rely heavily on automation to ensure that their operations are smooth. In both automatic and human systems, data obtained from online sensors forms the basis for a significant share of operational choices. To conduct real-time water quality evaluations, sensors are increasingly utilised in the production of potable water. This helps with process management decisions and allows early contamination detection. Conductivity, pH, temperature, dissolved oxygen, turbidity and flow cytometry are just a few examples of physical, chemical and biological factors that sensors can directly measure. However, proxy measures linked to these parts may be evaluated. Numerous applications exist for sensors that measure water absorption, including those in the drinking water, wastewater, environmental and industrial sectors.

These detectors measure the total light intensity loss due to particle scattering or the absorption of dissolved organic molecules over a certain distance in water. One of the main ways in which natural water bodies absorb light is through the chromophoric percentage of dissolved organic matter (CDOM). Strong linear relationships ($r > 0.9$) have been reported between absorption coefficients and dissolved organic carbon (DOC) in various water bodies. However, these relationships do not account for non-absorbing DOM fractions such as labile fractions, which are crucial for biostability and absorbance measurements cannot quantify these fractions. The quality of treated water is negatively affected in several ways when there is a high concentration of natural organic matter (NOM) in the water supply. As the amount and fluctuations of NOM in surface waters in northern and boreal European regions continue to increase, the significance of this issue is growing. Variations in weather, less acid rain, more primary production and standing biomass were all factors contributing to these shifts. The following problems can arise from drinking water treatment processes that fail to adequately remove natural organic matter (NOM): (i) an off-putting flavour and smell; (ii) the potential for bacterial regrowth due to inadequate elimination of viruses, bacteria and parasites; and (iii) an increase in the production of disinfection by-products (DBP) that could cause cancer due to reactions between NOM and disinfectants such as chlorine. The treatment actions were less effective in the presence of NOM. The disinfection of water with chlorine becomes more labourious when NOM is present. Fouling, which occurs when chlorine builds up on the membrane surfaces and/or pores, increases the demand for chlorine. Both reversible and irreversible foulants can cause this type of fouling. Only expensive chemical cleaning treatments such as clean-in-place (CIP) can eradicate them, rendering the physical cleaning and backwashing procedures ineffective. Humic substances (HSs) and biopolymers, which are organic matter components, cause fouling that cannot be reversed. UV absorbance data from online sensors with a wavelength of 254 nm can be used for accurate monitoring of HSs, the main component eliminated during coagulation. This allows the coagulant dose to be adjusted in real time. In addition, the differential UV absorbance at particular wavelengths, such as 272 nm, is strongly correlated with the amount of disinfection byproducts (DBPs) generated during chlorination. Therefore, absorbance-based sensors may be a good way to monitor the DBP levels. S_λ was calculated by applying the sliding window method to the logarithm of the absorbance spectra using linear regression at wavelengths. The slope of the linear regression as a function of the wavelength, also known as the spectral slope curve, is represented by the symbol S_λ. Biogeochemical processes and sources of CDOM can be studied with their assistance. Because the usefulness of different metrics varies among studies, it is important to examine how different metrics behave when exploring data. By seamlessly integrating existing supervisory control and data acquisition (SCADA) systems, high-resolution sensors enable monitoring of rapid fluctuations in water quality. The use of membranes in DWTPs is on the rise, and keeping them in good working order requires more frequent and precise data (in seconds) than older treatment methods, such as coagulation-flocculation. Owing to the massive amounts of data produced, DWTPs are limited to storing datasets that have been compressed or summarised. Physical and chemical characteristics (such as turbidity and DOC) calculated using specialised algorithms are typically used instead of the original data when absorbance-based sensors are considered. However, there is always a chance that crucial information will be misunderstood or rejected when using this approach. The spectrolyser (scan Messtechnik GmbH), ProPS-UV (TriOS) and Viper (MultiOS) are

among the multispectral CDOM sensors that are now available for purchase. At regular intervals, the spectrophotometre examined the attenuated light as a probe for UV-vis spectrophotometry. The measurement considered the absorption and scattering of light in the visible and ultraviolet spectra. Spectral data are often used to forecast dissolved organic carbon (DOC), nutrients or turbidity in published studies employing these devices, instead of analysing spectral chromophoric dissolved organic matter (CDOM) data. This investigation served a triple purpose: Determining the main obstacles that make it difficult to understand and using high-frequency data collected by online absorbance sensors. Develop a set of open-source algorithms for quick processing and visualisation of data from absorbance sensors. The metrics provided by the toolbox should consider specific issues such as redundancy, random error and drift while retaining relevant data. A dataset of sensors was used to identify outliers and provide explanations for variations in plant efficiency and to demonstrate how these approaches are applied in drinking water treatment facilities. An open-source Python library called AbspectroscoPY can conduct specific spectrum analysis and CDOM preprocessing. Its goal is to work with existing commercial and open-source toolboxes that perform tasks such as computing metrics from CDOM absorption spectra and preprocessing and visualising data from non-spectral sensors. Very few input parameters must be entered by the user because the method is mostly automated. Additional instruments, such as turbidity and other sensors, may simply convert their data outputs to a vector format from the matrix form, and the toolbox can then handle this. This adaptability is useful in many fields of environmental management and research, such as watershed studies, wastewater analysis, colour analysis of aqueous solutions and water quality monitoring. With the addition of 13 new features, AbspectroscoPY can now import, prepare, examine and analyse data from absorbance-based sensors. This document serves as a guide for making most of the AbspectroscoPY Toolkit. An example dataset pertaining to potable water is used as a case study. Light attenuation measurements taken for nearly a year (2017 - 2018) at VIVAB's Kvarnagården DWTP in western Sweden, using 3 online spectrophotometres, were included in the drinking water dataset. The entire treatment technique and the locations of the 3 spectrometer units are shown in **Figures 1(d) - 1(f)**. The points shown here matched the locations of the grab samples collected in 2018 (March - December). Raw attenuation values in the UV-vis wavelength range, taken from one of the spectrometer units, are included in the fingerprint file sample shown in **Figures 2(a) - 2(c)**.

The surface water supply for the DWTP comes from Lake Neden, which is a 3 km^2 oligotrophic lake. The lake has a slightly acidic pH of 6.7 and a conductivity of 60 μS cm^{-1}. It was surrounded by a combination of different types of wood. The lake undergoes a turnover process every 5 years. Lake Neden stands out among other lakes in the region due to its transparent water, low levels of total and dissolved organic carbon (TOC and DOC, 3.5 mg L^{-1}), and moderate specific ultraviolet absorbance (SUVA, 3.2 L mg^{-1} m^{-1}). This indicated the presence of both hydrophobic and hydrophilic components with different molecular weights. The water from the lake is mixed with the water from an alkaline groundwater well (GW) that has a pH of 8, a conductivity (σ) of 60 μS cm^{-1}, and a total organic carbon (TOC) concentration of 0.6 mg L^{-1}. This mixing occurs in a conduit that transports water to drinking water treatment plants (DWTP). The ratio between the 2 was 20 % groundwater (GW) and 80 % surface water (SW), with a 5 % margin of error. In other words, the ratio can range from 15 % GW and 85 % SW, to 25 % GW and 75 %

SW. As a result, the DWTP received water with relatively low quantities of dissolved organic carbon (DOC), around 2.9 mg L^{-1}, and a specific ultraviolet absorbance (SUVA) of roughly 3.10 L mg^{-1} m^{-1} (refer to **Figures 2(b) - 2(d)**. The treatment process at the plant consisted of several steps: Rapid sand filtration, a polyethersulfone hollow fibre ultrafiltration membrane process with in-line coagulation using pre-polymerised polyaluminum chloride, pH adjustment by adding $Ca(OH)_2/CO_2$, disinfection with UV irradiation and the addition of NH_2Cl. Calibrating and regularly validating sensor data against captured samples is essential because of the common occurrence of systematic drift that affects sensors. Further details of the treatment process at the Kvarnagården DWTP can be found in other publications. The grab samples used in this study were analysed in 2 different laboratories. The DWTP's private laboratory examined the samples for unfiltered UV absorbance using a Hach DR 5000 instrument, and for temperature and turbidity using a Hach 2100N IS instrument. In contrast, the Swedish University of Agricultural Sciences, SLU, analysed the samples for TOC/DOC, filtered UV absorbance and fluorescence. Prior to analysis, the samples were filtered using pre-combusted glass microfiber filters (GF/F) with a nominal pore size of 0.7 μm. Total organic carbon (TOC) and dissolved organic carbon (DOC) were measured using a TOC-V_{CPH} carbon analyser (Shimadzu). The mean coefficient of variation (CV) for duplicate measurements was 0.7 % for dissolved organic carbon (DOC). The AvaSpec-ULS3648, a spectrophotometer with high resolution, manufactured by Avantes, was used to quantify the UV absorbance at a wavelength of 254 nm. The measurement was performed in a quartz cuvette with a length of 5 cm, and the coefficient of variation (CV) was less than 1 %. The quantity of dissolved organic carbon (DOC) was used to standardise the absorbance at 254 nm (UV_{254}) to calculate the specific ultraviolet absorbance (SUVA) values. A 1 cm quartz cuvette was attached to an ASX-260 auto sampler (CETAC) for fluorescence measurements using an Aqualog spectrofluorometer (Horiba Jobin Yvon). Fluorescent excitation-emission matrices (EEMs) were preprocessed using the methods described by Ravindran *et al.* [51]. A set of samples containing ethylenediaminetetraacetic acid (EDTA) at a concentration of 10 mg L^{-1} for total organic carbon (TOC) and dissolved organic carbon (DOC) analysis, and K-phthalate at a concentration of 10 mg L^{-1} for absorbance analysis were examined for quality assurance purposes. The analysis was conducted using external standards.

The median and interquartile ranges of the water quality values collected from the grab samples in 2018 are shown in **Figures 3(a) - 3(c)**. Surface water (SW) was sampled 11 times, rapid sand filtrate (RSF) 16 times and ultrafiltration (UF) 16 times. It is essential to consider the effect of groundwater dilution when comparing the water quality disparities of surface water (SW) and riverside filtration (RSF). Across the wavelength range utilized to compute the fluorescence indices, the groundwater injection had no discernible impact on the water's fluorescent dissolved organic matter (fDOM) composition. Spectrometer readings were used to adjust the coagulant dosage in real time. Attenuation, colour, and turbidity were measured by inserting Lyser units into the sand filtrate and the permeate. Consequently, the water quality of the permeate is guaranteed to be more consistent and reliable owing to this management method. The intensity decrease data were recorded by the sensors at 2.5 nm intervals throughout a spectrum of light wavelengths from 200 to 750 nm. It is possible that the particles affected the reported absorbance levels, particularly in the surface water where turbidity was the greatest, even if all the sensors remained in their original locations. Intervals

of 2 min were used for SW measurements, whereas RSF and UF used 3-minute intervals. The numbers were adjusted internally to accurately reflect the voyage duration. It was appropriate to change the route length of the sensors at the water source and before ultrafiltration to 35 mm and that at the end of the process to 100 mm. Both DWTP sensors underwent local calibration throughout the sample period. Regular cleaning and maintenance were performed on all the sensors. It is important to discuss the process of using the AbspectroscoPY Toolkit. In the first part of this section, typical difficulties encountered during data analysis are discussed. Subsequently, it delves into particular toolbox features that were developed to address these challenges. The paper concludes by explaining how the case study dataset was analysed using these algorithms. The massive datasets produced by real-time measurements pose significant challenges in terms of preprocessing, visualisation and comprehension. It is a common practice to detect and eliminate or significantly reduce inaccurate data, such as outliers and extreme results, before beginning therapy. Additional problems occur when the time axes do not coincide when data from several sensors are combined. Importing, processing and analysing sensor data are all parts of AbspectroscoPY's toolbox, and the program also displays spectrum parameters to help with interpretation. Regular data downloads from sensors are necessary to prevent data overwriting, because memory is quickly depleted by frequent observations. Data may be stored as text files or, better yet, as csv files from the device.

The files can be imported using the function 'abs_read', which combines a series of successive measurement files into 1 dataset. Data may be obtained from the Spectrolyser by utilising either the Anapro software or a spreadsheet application, such as Microsoft Excel. The datasets used in this study were around 0.4 - 0.5 GB per sensor, which is approximately equivalent to 3×10^5 measurements multiplied by 200 wavelengths. The preprocessing procedures in the toolbox may be utilised to obtain data ready for visualisation. Methods for enhancing data quality and converting data to the appropriate data type for analysis (convert2dtype) are included in the toolbox. Missing data can be handled using nan_check, dropna, and NaN entries. Similarly, 'dup_check' and 'drop_duplicates' can be used to address duplicates. Identifying and removing duplicates and missing data. If the sample frequency was higher than the frequency of major water quality events, removing the missing data would not cause a huge loss of data. Because certain sensors take daylight saving time (DST) changes into consideration, while others do not, dealing with duplicates requires careful interpretation of timestamps. Therefore, it is crucial to thoroughly examine the dates on which copied information is removed based only on timestamps to prevent data loss by accident. The time axis was rotated. It is critical to align time-series data from many sources before comparing their signals. Even if the 2 instruments employ the same type of sensor, it is essential to ensure that their time axes are identical. Continuous timing irregularities may result from DST processing or instrument settings. In this manner, it is advisable to check whether the sensor can automatically adjust the timestamp to reflect the DST. To create a continuous time series, the time shift (tshift_dst) can be used to change the time axis. If there are any other persistent departures from the local time in the dataset, mistakes made when changing the internal clock of the instrument should be fixed (using time delta). Important considerations when comparing sensors include using a time delta to adjust the time axis of each sensor in accordance with the reference sensor, which will bring all clocks into sync. If there is a time lag (time delta) between the 2 sensors when measuring the water flow, it can be fixed by adjusting the timestamps of the

other sensors relative to the time axis of the first sensor. Even if the degree of lag changes with time, the user may achieve time alignment using the toolkit. The ability to align times is essential for determining whether an event in 1 part of the treatment plant is linked to a previous stage. For instance, if the attenuation data recorded by the SW sensor causes a change in the coagulant dosage, then the attenuation data observed by the RSF sensor can shift accordingly.

Figures 4(a) - 4(c) illustrate how the difference in time between the internal clock (t_{scan}) of the 3 spectrometer units and the local time (t_{CEST}) can be adjusted during periods of Daylight-Saving Time (DST) and Standard Time (ST). The constant time difference between the internal clock and local time during the DST and ST periods indicates that the 2 DWTP sensors separately compensated for the DST, in contrast to the SW sensor. In the first phase, the time axis of the DWTP sensors moved ahead by 1 h when the DST ended. In addition, none of these 3 sensors displayed any signs of daylight-saving time-related fluctuations in the local time. The time axes of these sensors were adjusted in the same manner as in step 2. Time-shifted data were used to measure the time lag between the 3 sensors. In the third phase, the time axis of the SW sensor is 1 h. Furthermore, the user's knowledge of the water parcel's journey time between the SW and DWTP locations was used to push the time axis of the SW sensor for 11 h. In accordance with Step 4, this modification was made, and one must decide whether to ignore certain high-frequency data or interpolate low-frequency data when the data frequencies differ among the sensors. Sweden's Kvarnagården DWTP tracks absorbance every 3 min and measures membrane permeability (TMP) every 5 s. Given the measurement frequency in relation to the time scale of significant changes in the observed data, the decision to interpolate or delete the data was made. In most cases, spectral data can be safely disregarded if the measurement frequency is significantly higher than the data change rate. In such cases, interpolation would have been the superior choice. Applying interpolation to the data with constant or predictable changes yielded the most precise results under all scenarios. This holds true when the data exhibits a predictable cyclical pattern that can be precisely anticipated throughout the interpolation process. It is possible for the signal output to change gradually over time, even after extensive sensor calibration, which might affect the interpretation of information. Therefore, after calibration, it is essential to check the devices for unusual behaviours, particularly if the user finds a pattern. In this investigation, absorbance spectrophotometres were used to internally calibrate the signal using a dual beam. Although this calibration technique significantly reduces the electrical drift, it does not fix the optical drift that can be caused by factors such as dirty windows or damaged equipment. Baseline correction may be used to resolve this situation. When working with clean or aligned datasets, the AbspectroscoPY toolkit provides several options for fixing attenuation data. Unless the optical path length (abs_pathcor) is used automatically, as it is with spectro, data may need to be normalised before analysis. Light up. The median was determined by averaging the absorbance values within a certain range of wavelengths, such as 700 - 735.5 nm. The absorbance measurements were adjusted for instrumental baseline fluctuations by subtracting the median value. The median and noise levels, measured in 3 standard deviations, were made easy to observe in the toolbox. The absorption of chlorophyll and CDOM was negligible at wavelengths greater than 700 nm, and the signals were mostly caused by turbidity and random electronic noise. The turbidity remained after calculating the mean value across a range of wavelengths, which removed random interference. Plotting the attenuation

spectra for several samples and evaluating their divergence from 0 might help determine the appropriate wavelength range for the baseline, taking into consideration any temporal variability in the data. This approach allows one to regulate the changes to the baseline. has the option of applying the technique to the entire dataset or selecting portions thereof. In addition, this function can be used to multiply, add or remove a specified value from an entire dataset or its subset. This feature is crucial when performing crucial calibrations or adjusting for the presence of interfering anions and cations, such as iron and nitrate. For the DWTP example, it is essential to compare the apparent absorbance of the sensor with the apparent absorbance (unfiltered) measured using a desktop spectrophotometre, which will help to establish whether there are any consistent biases in the results.

The scatterplot in **Figure 1(b)** illustrates the UV_{255} data obtained by spectrophotometry. It compares the unfiltered UV_{254} data from the captured samples on the x-axis with the SW spectrometer data on the y-axis. The SW spectrometer data represent the wavelength closest to UV_{254}, with a resolution of 2.5 nm. There seems to be a slight bias in the sensor data, considering the instrumental error in the laboratory analyses; after the reliability of the data has been assessed, visualisation can be performed. **Figure 1(c)** shows a graphical representation of the 3 spectrometers' preprocessed time-series data for UV absorbance values at 255 nm; the measurement units are shown in **Figure 1(d)**. Using the SW time series, 5 distinct periods were identified due to the dimictic nature of Lake Neden. Two phases (P2 and P5) with large temporal fluctuations during fall circulation, a rather constant phase (P1) around the end of summer stagnation, and 2 periods (P3 and P4) with varying absorbance trends were observed. P3 occurs after winter stagnation and circulation end, whereas P4 occurs when spring and circulation begin. **Figure 1(e)** contrasts with **Figure 1(f)** in that it shows 3 occurrences related to changes in lake levels and coagulant dosage adjustments in the DWTP. Two years in a row, the lake circulated in the fall, which caused Events 1 and 3. The second event was challenging for the DWTP because of the spring lake circulation, which caused the membrane permeability to drop for a long time and ultimately required the cleaning-in-place (CIP) of the UF membrane. Data smoothing and noise reduction were performed easily using Python's built-in functions, such as rolling and lowness. Thus, the rolling function can be used to apply median filtering. A simple and long-lasting method of data smoothing, median filtering, performs well, even in the presence of rare outliers. The user selects the window size for the median filter based on the data frequency and the filtering objective. To choose the best size for the smoothing window while using median filtering, it is essential to visually inspect the data. While narrow spikes are desirable, data becomes noisier when the window size is reduced. However, increasing the size of the window would make trends, not oscillations, more prominent by reducing the prominence of cyclical peaks. Therefore, it is appropriate to undersmooth rather than oversmooth to keep important details intact, and things that require further investigation could be the source of outliers in sensor datasets. It is best to keep or disregard outliers, such as sudden changes in coagulant dosage or recognised artefacts, such as maintenance operations on sensors and membranes. Additional methods for handling outliers are discussed in section 3.3. **Figure 2(b)** shows the results of applying the smoothing function to the data shown in **Figure 2(c)** for P1. The 60-minute timeframe was selected because it adequately captured the trends and fluctuations. Variations in flow rate caused by demand variations are likely responsible for the daily cycles shown in the raw RSF data, which exhibit a

twin-peak pattern. Owing to the backwashing operations, which happen every 1 h or so, the UF data display a periodic pattern. A rolling median filter applied over a 60-minute period mitigates the narrow spikes in the UF signal. These characteristics are preserved in the filtered signal when a 15-minute frame is used.

Sampling procedures

The dataset may be examined with the help of the toolbox's capabilities to determine which data points deviate significantly from the average and eliminate them. User-defined events can be utilised to classify data outliers, and the outlier_id_drop function can automatically remove outliers that belong to specified event categories. It is essential to eliminate times when performance variations might be caused by external factors, such as power interruptions or unplanned maintenance, while benchmarking membranes. To identify artefacts and anomalies in data, it is important to meticulously record all WTP operations in a logbook. This includes sensor maintenance and plant activities. The outlier_id_drop function was used for the SF and UF absorbance data presented in **Figure 2(d)**, and the results are shown in **Figure 3(a)**. When the RSF and UF systems did not receive feed water, it was shown in the plot as a 'no feed event', and no relevant data were provided. Importantly, RSF system data were unreachable prior to June 2018. **Figure 3(b)** also shows 'Al dose events' that occur when the coagulant dose is adjusted with their respective symbols. To better visualise an event, symbols provide an approximate idea of chronological placement. The user must provide the beginning and ending dates, event type and label reference in a CSV file so that known events may be assigned labels. An outlier ID drop iqr is a function that can be used to identify and perhaps remove outliers and unexplained occurrences. The user must choose the time intervals (such as P1-P5 in **Figure 3(c)**) before outliers are found. Therefore, it is vital to use the interquartile range (IQR) thresholding method to identify outliers. The IQR was used one-and-a-half times. The slope ratio data were subjected to the IQR technique because of the heightened susceptibility of the slopes to outlier effects. The data used to calculate the slope ratio in this case were obtained from SW absorbance measurements that were preprocessed according to the procedures outlined in Section 3.2, with the exclusion of median smoothing and baseline correction. The data were examined using a pre-processed dataset [52-58]. Period P1's slope ratio is statistically different from that of periods P3 and P4, according to the data. The kernel density estimate (KDE) can be used to determine the underlying probability density function of the dataset (KDE). Similar to a histogram, it can be constructed in a variety of ways using multiple kernel types, making it more versatile. Each data point was assumed to have a Gaussian distribution using the Python kde-plot function. The effect of varying the observation wavelength on the density distribution and absorbance values (the height of the curve at each point) is shown in **Figure 4(b)**. Compared to SW, whose KDE plots displayed less variation in absorbance values, the RSF and UF data had more pronounced peaks. There were 3 distinct areas in the distribution of wavelengths below 327.5 nm. This is an inevitable consequence of coagulant injections performed automatically at the DWTP to meet specific ultrafiltration (UF) permeability goals. Significant differences in water quality were observed when the 3 different permeability objectives were implemented in the DWTP. Tools for analysing and investigating spectral changes were provided by AbspectroscoPY after the data were processed and

displayed. The aim was to find shared patterns and individual differences in spectral characteristics due to changes in the organic material composition.

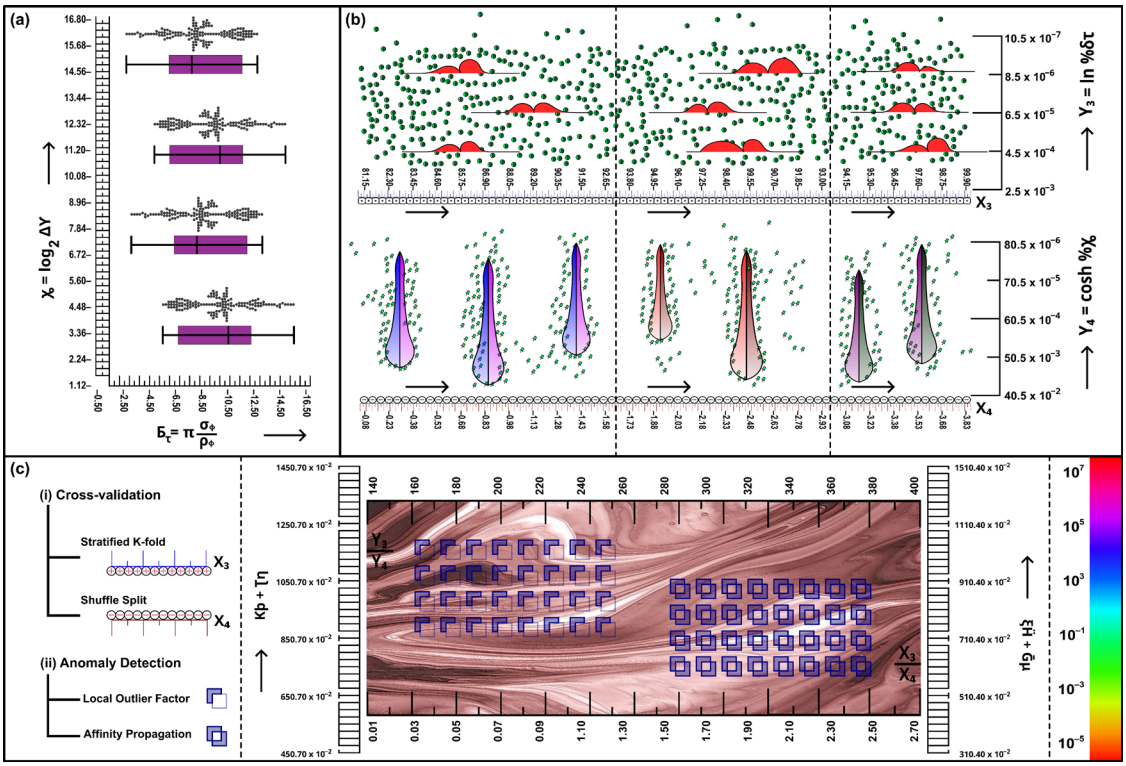

Figure 3 (a) Notched box-chart (with outliers) of Excitation-mission matrix spectroscopy (EEM) combined with parallel factor analysis (EEM-PARAFAC) used to analyse the simulation results of a254 values in CKDu and non-CKDu groundwater in Sri Lanka, revealing significantly lower BIX values of CKDu groundwater than non-CKDu groundwater (Note: Fluorescence index of dissolved organic matter (DOM) in all the samples ranged from 1.26 - 1.88); (b) Verification of experimental results (with 4 fluorescent components extracted from wastewater (sample obtained from Kvarnagården DWTP, Varberg, Sweden) and passed half-split validation, whereby: C1 is terrestrial humic-like components, C2 is microbial humic-like components with low molecular weight and C3 is autochthonous protein-like components) illustrated as raincloud and grouped-violin plots; (c) Posthoc verification of the data plots pertaining to FDOM was predominated by C1 and C2 in CKDu groundwater, wherein the C1 was significantly higher in CKDu groundwater than in non-CKDu groundwater (the 2-component principal component analysis show that the positive direction of PC1 is dominated by humic-like substances, and the negative direction is dominated by protein-like components related to microbial activities), illustrated as cross-validation and anomaly detection (Note that, recognizing applicability of HIX, DOC and C4 % is not as good as C1 %, because their DPN or DPC curves fluctuate much with changes in thresholds).

The AbspectroscoPY toolbox comprises methods for calculating standard metrics from the absorbance spectra of CDOM, including S, S_R and S_λ, as well as the ratios between the absorbance values

at specific wavelengths. These functions are comparable to those offered by the 'cdom' module in the R computing environment. The fitting of the absorbance spectra derived from the spectro is illustrated by the exponential curves in **Figure 4(c)**. A singular function (abs_fit_exponential) was used to characterise the exponential decay of the data at a specific date. The wavelength range employed in the analysis affects this function, which is derived from Eq. (1). The reference wavelength used for the analysis was 350 nm. The time series of the slope ratio for the SW (abs_slope_ratio) is plotted in **Figure 1(a)**. The decrease in solar reflectance (SR) compared to P1 during P2 and P3 implies that surface water (SW) was primarily composed of terrestrial coloured dissolved organic matter (CDOM) with a higher molecular weight (MW). Throughout the periods P2, P3, P5 and the initial phase of P4, the S_R and time series of the absorbance ratio A_{250}/A_{365} in **Figure 2(a)** demonstrated comparable patterns. Nevertheless, S_R experienced only a slight increase during period P1, whereas it began to steadily increase during period P4 from April 2018 until it reached its zenith in August 2018. The ratio A_{250}/A_{365} demonstrated a substantially greater growth rate than the remainder during both P1 and P4. In addition, incremental increases were observed, particularly during P4. In accordance with prior research, this investigation implemented a sliding window technique with a width of 21 nm, which was implemented across wavelengths from 220 to 697.5 nm with a resolution of 1 nm. The spectrometer yields absorbance data with a resolution of 2.5 nm. The data were resampled at 1 nm intervals using cubic spline interpolation to enhance accuracy. Data (abs_spectral_curve) were filtered using a correlation coefficient threshold of $R^2 = 0.98$. For clarity, it is important to present the absolute values of the slopes, instead of the initial negative slopes. This trait is uncertain because absorbance slopes often have negative values. The spectral slope curve data rapidly reflected changes in the shape and amplitude of the absorbance curves, as the absorbance at high wavelengths remained constant throughout the experiment. Identifying the most variable wavelength areas is more effective when using spectral slope curves instead of evaluating the absolute changes in absorbance. For events 1 and 3, the researchers aimed to compare the normal profile with the circulation of an autumn lake. **Figure 3(a)** shows that the spectral slope changed most noticeably at 290.5 nm. Two lake circulation events were used to assess changes in the spectral slope at a particular wavelength. The research was run again at time intervals where no events occurred during the year to establish a standard profile. Over the course of 5 weeks in 2017, Incident 1 decreased the slope at 290.5 nm by 7.4 %. When compared, Event 3 in 2018 accelerated the slope increase by 8.3 %, approximately 2.5 weeks. A 3-week period without occurrence resulted in an average slope change of less than 1 %. Not only does the spectral slope vary from 270 to 350 nm during circulation episodes, indicating amplitude and shape changes in absorbance, but the overall alterations in the profiles with wavelength also remain constant over time. The lack of variation in the profile shape is probably caused by Lake Neden's long-term existence in the same spot. The spectral slope increased by 1.3 % from the end of March to mid-June 2018 (Event 2). By adjusting the spectral slope, researchers can determine when to collect grab samples for more precise studies designed to address certain issues. The local spectral slope data can be built into a time series using the AbspectroscoPY Python toolkit. To do this, one must be proficient in solving Eq. (2) to determine the negative spectral slope at a certain wavelength, say 290.5 nm. The cdom package, which is based on R, uses statistical methods and these characteristics enhance these approaches. For each wavelength, the program calculated the percentage change relative to the average

spectral slope on the reference day. The % variations in the spectral slopes during the 2018 lake circulation event, namely in the SW, RSF and UF sectors, are shown in **Figure 4(a)**. Except for a plateau in the UF data between November 12th and 17th, 2018, the SW, RSF and UF profiles were essentially identical in the present dataset. Spectral slope variations were observed at various angles and wavelengths. A time series of the spectral slope at a wavelength of 254.5 nm is shown in **Figure 1(b).** There was significantly less evidence of 290.5 nm oscillations compared to the figure. This demonstrates that different wavelengths may have different effects on the elimination of the organic molecules. Both 272.5 and 290.5 nm showed very similar changes in the spectral slope over time. Analysing time-series data can be helpful for monitoring DBPs and functioning as an early detection strategy, given the known association between the 272 nm wavelength and DBPs.

Engineering specifications and calculations

Samples obtained from above the bioreactor throughout the development of *L. platensis* were analysed using a series of ion mobility spectra, as shown in **Figure 1(c)**. The relative drift times of 1 are characterised by the reactant ion peak, whereas lower relative drift times show a smaller peak, according to Eq. (2). Based on these results, it seems reasonable to purge the ionisation zone of the ion mobility spectrometer with filtered air and to sample the headspace above the algae. **Figure 1(d)** shows 7 distinct peaks that were discovered by comparing the growth medium baseline measurements with measurements made throughout the culture. The detected peaks were more than 3 times the noise in the background signal, which was the limit of detection (LOD). In addition, the strength of these peaks changed when plants were grown. These peaks could be easily identified using the Python method scipy.signal.find_peaks, as their amplitude exceeded the limit of detection (LOD). The central peaks (c, d and e) tended to strengthen as the developmental phase progressed, whereas peak g exhibited a rapid increase when the cell concentration in the culture reached its maximum. Ion mobility spectra showed that the concentration of VOCs changed with growth. Ion mobility spectra exhibited unique patterns at each culture stage. Ion mobility spectra of the culture medium were measured before the start of the experiment. If the culture media became contaminated between trials owing to inadequate reactor cleaning, this method would have prevented the ion mobility spectra from showing a signal that was distinct from the typical blank spectrum. A comparison of the ion mobility spectra from *L. platensis* cultivation at the Technical University of Dresden and the Helmholtz Centre for Environmental Research in Leipzig confirmed the stability of the MI-IMS system. In both the experiments, the bioreactor, sample location, MI and ion mobility spectrometer were identical. By meticulously collecting the airspace above the 2 Smellmaster 2 ion mobility spectrometers, the methodology could reproduce the exact placement of peaks a-g. The cultures were identical to those used for *L. platensis*. **Figure 1(e)** in Supporting Information shows these results. Similarly, the pattern of peaks changed during the growth phase, with peaks c, d and e increasing, and peak g increased dramatically towards the conclusion of the cultivation period. During an 8-day cultivation period, parallels may be observed in the fluctuations of the peak intensity. This demonstrates that IMS can continuously record changes in the volatile organic compound (VOC) profile of a microalgal culture, regardless of the location, user or equipment.

The significance of peak d in the growth-related peak pattern and the efficient functioning of the MI-IMS-based system at low cell densities are highlighted by this connection. There are other types of online sensors, including those that detect dispersed light, but this one sticks out. Despite their usefulness in measuring culture development, these sensors sometimes require large cell densities before a direct correlation can be demonstrated. The backscattered signal is typically interrupted by the absorption and cell clumping. A parametric t-test was used by a student to determine whether the intensities of peaks a-g varied significantly during the development stages. To accomplish this, it is crucial to examine the experimental data from 4 separate cultivation methods. During the culture phase, the intensities of peaks a-g were compared with their intensities at the beginning of the experiment. Variations in peak intensities were statistically significant, as shown in **Table S2(a)**. To determine how well IMS tracked the levels of VOCs generated by microalgae, a simple experiment was conducted. Collecting samples from the air space above the 2 separate cultures was essential because of the identical metabolic activities of the microalgae. Spirulina (*L. platensis*) and chlorella (*C. vulgaris*) have quite different ion mobility spectra, as seen in **Figure 1(f)**. Within the relative drift time range of 1.2 to 1.6, the ion mobility spectra of Spirulina displayed distinct peaks labelled a - g. The 2 ion mobility spectra of Chlorella, on the other hand, had prominent peaks prior to 1.2. This proves that the ion mobility spectrometer can distinguish between the profiles of 2 distinct microalgae's volatile organic compounds. The purpose of this study was to determine how to get the most out of IMS by using it as a process management tool for online monitoring. Nevertheless, the findings demonstrate that IMS is capable of detecting contaminants, such as grazers, in an agricultural context, and cultivation of diverse microalgae results in radically diverse spectra. (2) The relative strengths of the ion mobility spectra peaks changed considerably as algae were cultivated. (3) The ion mobility spectra and intensity patterns were constant and repeatable throughout the culture phase when the same species of microalgae were cultivated. These results have several real-world consequences when using IMS with bioprocessing data, such as (1) when the specific peaks are strongest, the growth rate is at its maximum; (2) the highest intensities are directly proportional to the cell concentration, especially at low concentrations; (3) it is often indicative of a day/night cycle when all development-related peaks decrease, without new peaks appearing; and (4) when the ion mobility spectrum shows the emergence of new peaks, it signals a shift in biological processes such as cell lysis or nutritional restriction.

While changes in cell concentration can be successfully associated with certain peaks, prior studies have discovered a substantial correlation between peaks c and e. In addition, there were more than thousand data points in the ion mobility spectrum, covering the relative drift durations of 0.7 - 3.0. A peak is an accumulation of data points that, in a perfect world, would fit into a normal distribution on the graph. Consequently, 2 distinct approaches to data processing are available, both of which seek to optimise the utilisation of data points for analysis while simultaneously minimising the impact of signal noise. Multiple Gaussian peaks with constant relative drift durations were used to represent data. However, one may find associated peaks and produce principal components (PC) using principal component analysis (PCA). Every data point following the reactant ion peak is included in the dataset. The spectra were not based on the maximum intensities of the 7 peaks, but rather on the intensity collected at over thousand relative drift time points. The data matrix was compared with a model consisting of peaks with a defined centre point and

width in the Gaussian peak model. The intensity measurements are significantly less affected by noise, and the time series now has 7 additional trustworthy data points for each spectrum. In this case, principal component analysis (PCA) is used to decrease the dimensionality of the dataset from almost a thousand times the number of spectra in the time series. Four PC Vectors or 7 Gaussian model peaks can be connected to a process management proposal using reference data. The sklearn package in Python was used to develop the custom code. Therefore, this study used a conventional scaler to preprocess the data before performing a 4-principal component analysis (PCA). The spectra from the 6 different cultivations were analysed using principal component analysis. Of the 6 cultures, 2 were subjected to low-level light on the initial day of development before being switched to high-level illumination (60 → 200 % intensity compared to cultures 1 - 4), whereas the other 4 cultures were subjected to the same conditions. The initial growth phase persisted for several days in all 6 cultures, followed by a stable phase, and finally, a decline in cell viability. The substantial shift in the cell concentration and subsequent VOC generation observed across several IMS spectra led to the initial assumption that the first PC signalled the development phase. A new peak, g, appeared during the second PC cycle, which was thought to indicate the time of stasis or cell death. With an increase in cell concentration, PC 1 scores go up from −30 to 25. There is presently very little variation in the PC 2 scores. From −10 to 30, the PC 2 scores increased as the cell concentration decreased. In addition, as can be seen in **Figure 2(b)**, the activation and deactivation of the light source caused the PC 1 scores to fluctuate. **Figure 2(c)** shows that the data points for low cell concentrations (coloured green) were densely packed, even though the change in PC1 for Culture 1 over time differed from that of the other cultures. A small change in the PC1 concentration over time can be tolerated if it is assumed that VOC generation is proportional to the cell concentration and growth stage. **Figure 2(d)** illustrates that the variation in PC 1 scores was explained by the impact of peaks c, d and e. A high PC 1 score is a consequence of the strong peak intensities seen in peaks c, d and e. According to previous studies, these peaks are related to development, and are less intense under dark conditions. The negative correlation between PC 2 and growth suggests that peaks c - e are likewise adversely related to PC 2. In addition, peak g had a positive correlation with PC 2 despite its earlier association with cell death, which can lay the groundwork for controlling processes. The major component transformation acquired from the training dataset can be used to describe freshly collected spectra. Automatic initiation of control events or proposals can be achieved using a process-control system. (1) An increasing cell concentration and growing culture are shown by static PC 2 and growing PC 1, respectively. Consequently, the system did not require any changes. When PC 1 was at its maximum and PC 2 stays the same, the cell concentration in the culture was at its highest. Either harvest the culture or provide extra culture medium if this occurs. (3) The cell concentration drops as PC 2 increases. Harvesting, adding more growth media, and checking for contamination are the 3 choices presented by the process-control system. (4) Verifying the system is necessary if PC 2 returns a strongly negative result because this implies a system condition that is not obvious. Light is crucial for the photosynthetic growth of cultured algae. The development of L. platensis was enhanced in all the cultures investigated because it utilises carbonate ions in the culture medium as a carbon source to produce sugars through photosynthesis. Previous evidence has shown that in the absence of light, both PC 1 and the highest levels of intensity in the IMS spectra decrease. However, it is not always the case that more light causes

plants to grow faster. *Arthrospira sp.* photosystems reach saturation at species-specific light intensity, meaning that the growth rate cannot be increased further. In addition, CPC levels decrease, and less of this vital antioxidant is synthesised when exposed to strong light. The light level was carefully regulated during testing in a controlled environment. In contrast, sunlight is essential for large-scale microalgal cultivation. This may cause either too much or too little light to reach the reactor depending on its design. To enhance the growth conditions without incurring the extra expense of artificial lighting, a mechanism was devised to shade the reactor during periods of intense light, as low light retards development. This shading method requires process control, which may be automated in a bioreactor with predetermined dimensions by monitoring the light intensity and concentration of cells in the solution. Living things rely on these characteristics for defence, which is why they are significant. Consequently, microalgae observe varying intensities of light based on cell concentration, even when the illumination levels remain constant.

Another option for process management is to use fluctuations in the composition of volatile organic compounds (VOC) caused by excessive light. Consequently, it is necessary to measure the effects of low- and high-light cultivation on the IMS spectra and PC scores. Light intensification accelerated the development of Cultures 5 and 6. Consequently, the c - e peaks stand out much more than those in the IMS spectra shown in **Figure 3(b)**. Furthermore, a distinct pattern of peaks was observed under strong illumination when the peak denoted as f was identified. Because PC 3 reacts rapidly to ambient light levels, its fluctuating scores may be used to track peak light periods. **Figure 3(c)** demonstrates the innovative idea of process control, that is, rapid sensitivity to high light intensity. When PC 3 is below −5, a shadow culture was used. The degradation of CPC into a nitrogenous substrate may explain peak f and the marked alterations in metabolic activity in the presence of strong light. The extent to which the addition of a nitrogen source to the growth medium inhibits peak formation and permits substantial development under intense light without CPC degradation requires further research. Machine Intelligence (MI) in conjunction with Intelligent Monitoring Systems (IMS) has great potential as a tool for managing and monitoring microalgal development in a consistent and accurate manner. Ion mobility spectrometry (IMS) can automatically sample volatile organic compounds (VOCs) and gather data every 10 min when connected in-line. Operator interactions were not required in this process. Regular monitoring allows for efficient process control. Automatic, economical, and consistently accurate VOC spectra were obtained using MI-IMS. These spectra can be utilised to evaluate the state of the crop and provide recommendations for managing the process. Principal Component Analysis (PCA) was used to examine a training dataset consisting of 6 cultures. Using the first 3 PCs, one may determine if the culture is expanding, if the maximum cell concentration has been reached, if the cell concentration is decreasing, or if the amount of light is too high. By utilising a single-sensor system to provide several process-control signals, the MI-IMS deviates from the analytical approaches outlined in **Figure 4(b)**. Utilising MI-IMS process control in commercial production can enhance the yield, facilitate contaminant identification and modify the growth rate.

The rapidity, low cost and high sensitivity of ultraviolet/visible spectroscopy make it an ideal tool for monitoring NOM in water-treatment processes. By enabling constant monitoring, sensors can enhance this method to detect sudden changes in the water quality. However, time-axis correction, filtering and outlier identification are only a few of the tedious preparatory procedures that require enormous datasets. Utilising

spectrum measurements to guide and support interpretation is of utmost importance, and Python's AbspectroscoPY package fixes fundamental issues with processing sensor data, such as outlier identification, baseline correction, systematic temporal fluctuations and duplication management. Spectral slope curves, absorbance ratios, exponential fits and slope ratios were obtained from the data analysis. In addition, it includes tools for making time-series visualisations of data. The parts of the complete process include the following: to gain a complete view of the time periods with significant changes in CDOM sources and molecular characteristics, plot the absorbance ratio, find the wavelength ranges where the absorbance changes significantly over time and determine the derivative of absorbance with respect to wavelength, also known as the spectral slope. Periods of interest to the user, such as when lake circulation or membrane permeability drops, or periods based on specified criteria (such as option a) can be utilised to concentrate the research, and the time series of the percentage changes in the spectral slope for the specific wavelength ranges listed in (b) can be used to examine fluctuations in the absorbance curve slope over time. By finding correlations with important occurrences, the time series can be utilised as a system for early warning. The AbspectroscoPY toolkit integrates these methods with a powerful open-source Python framework, making them applicable to numerous data sources in various industries such as the food industry and water or sewage treatment. Data acquired from optical sensors at Kvarnagården WTP, which draws water from Lake Neden, was used to demonstrate the toolbox's capabilities. There were 5 distinct periods in the attenuation dataset that were associated with the lake's natural processes, such as seasonal circulation, and were measured during a 15-month period. These occurrences, together with changes in the water treatment plant such as a decrease in membrane permeability or variations in the quantity of coagulant used, can be detected using the spectral metrics included in the toolkit. Even when water quality remains constant, irrespective of the type of situation that calls for continuous turbidity measurements, the open-source design of the toolbox makes it straightforward to add additional features such as particle compensation methods. Surface water absorbance measurements were made more precisely by using turbidity adjustments. Methods for removing the spectra of substances that interfere with dissolved organic matter (DOM) by sharing the same wavelength range. Advanced anomaly detection equipment. To aid in decision-making, water producers may utilise algorithms such as the absorbance slope index (ASI).

Results and discussion

Key observations

The data presented herein were collected in 3 distinct stages. From June 11th to 8th October 2014, 189 samples were collected during the first sampling period. At this point, weather, both present and past, had no bearing on the sampling process. However, in phases 2 and 3, only dry weather was considered. This means that, in the 48 h leading up to sampling, the rain gauge had to register less than 0.20 cm. Three streams were the primary focus of the sampling efforts: Ley Creek (N = 291 samples, S = 58 locations), Onondaga Creek (N = 1,700 samples, S = 109 sites) and Harbor Brook (N = 612 samples, S = 38 sites). Sanders Creek, a branch of Ley Creek, is also tested. More specifically, Cold Brook, Hopper Brook, Kimber Brook and West Branch, all of which are tributaries of the Onondaga Creek, were investigated. In parentheses, can see the total number of samples taken from each tributary; the letter 'N' stands for the total

number of samples, whereas the letter 'S' refers to the total number of species. Starting in rural, largely agricultural, and forested regions and ending in the urban core of Syracuse, NY, near the outflows of Onondaga Lake, all the streams investigated exhibited a steady change in characteristics. The surface areas of Harbor Brook, Ley Creek, and Onondaga Creek are 35, 76 and 285 km^2, respectively. The entire drainage area of Onondaga Lake is composed of 3 regions that account for 5, 10 and 49 % of the total area. To ensure that samplers could easily reach all areas of the country, and that urban and rural areas were fairly represented, sampling sites were strategically positioned along each canal. Cold Brook (8 samples), Harbor Brook (10 samples), Hopper Brook (10 samples) and Onondaga Creek (3 samples) were the sites of 31 samples collected for microbial source-tracking marker (MST) analysis between August 2015 and August 2017. Human and nonhuman faecal sources, such as agricultural inputs, marshes, and ponds with large waterfowl populations, were considered when selecting the MST data collection sites. Regular monitoring also showed consistently high levels of faecal coliforms, and earlier source-tracking efforts failed to pinpoint the sources of faecal contamination; therefore, these locations were selected. If the faecal coliform count in a sample obtained at the same time was equal to or greater than 200 cfu/100 mL, only samples from these sites were analysed for the presence of MST. To analyse the influence of land use close to each sample site, the percentage of land within 122, 366 and 1,098 m, which was pasture-hay, cropland, forest-wetland or developed (> 20 % impervious) cover, respectively, was calculated.

To summarise, data on land use were extracted from the National Land Cover database for the selected year and the distance buffer. Buffer distances were calculated according to the buffers specified in the Leafy Green Marketing Agreement. Depending on the proximity of the surrounding land uses, which might compromise food safety, buffers were set up. The required buffer distance between farmland and concentrated animal feeding operations (CAFOs) housing thousands of animals is approximately 1,200 feet (366 m). To collect land use data, this buffer size was chosen, along with 2 others that were 1/3 smaller (122 m) and 1/3 larger (1,098 m). Locations were classified as urban if they were within Syracuse's geographical boundaries (**Figure 1(d)**) to compare the water quality in rural and urban regions. Despite being in areas with mostly developed land cover, 4 of the 5 sites north of the city limits were in areas where there was significant construction. However, the complex ecology of Syracuse is oversimplified when the area is grouped into rural and urban areas. Consequently, further studies were conducted to determine the extent to which the surrounding regions were covered by developed properties. Every other month, from May to November, samples were collected at each location. Except for 2010 and 2011, samples were collected every year from 2008 to 2017. To prevent the sampling of the same water region, each watercourse was sampled consecutively, moving downstream before moving upstream. All microbiological, physiochemical and nutrient/sediment samples were collected at the same time, but in different volumes, from different bottles. a sterile 150 mL plastic container was submerged along the middle of the river to perform the microbiological test. Sodium thiosulfate was used to pre-preserve 125 mL of each water sample in a 150 mL coliform vial that had been sterile. Chloride analysis was conducted by transferring 10 mL of each water sample to a 25 mL glass vial. The direct grab approach was used to gather an extra 1-L grab sample during sample collection for MST analysis. Within 8 h of collection, a 125 mL sample was tested to assess the levels of total coliforms, faecal coliforms, *E. coli*, and *Enterococcus*. Following the procedures

laid down by the New York State Department of Health, the Onondaga County Health Department, an accredited laboratory, determined total coliform and FIB levels. Shortly after samples were collected, they were sent to the Wadsworth Laboratory in Albany, NY, which is a division of the New York State Department of Health, to be tested for MST detection. After Bacteroidales members were preferentially selected by anaerobic enrichment, avian, canid, human and ruminant faecal indicators were detected using individual PCR screening. Meteorological conditions and physicochemical markers of water quality were recorded during each sampling session. Data on water temperature, pH, dissolved oxygen, specific conductance and turbidity were collected *in situ* using a YSI 650 MDS portable equipment (YSI Inc., Yellow Springs, Ohio, USA). The device was equipped with a multiparameter water-quality probe (Model 6600 or 6820-V2).

A Hach^{TM} portable colorimeter was used to conduct on-site analysis of the chloride concentration. A 7.6 L sample was taken from the same spot as that used to assess faecal indicator bacteria (FIB) to estimate sediment and nutrient levels. For the analysis of nutrients, a 1 L portion of the 7.6 L sample was separated, and a 0.5 L portion was divided for sediments. Note that not all samples had all parameters assessed because of factors such as personnel availability, schedule conflicts, equipment problems and sample collection feasibility. Air temperature, rainfall, relative humidity and wind speed were some of the meteorological variables supplied by the Network for Environmental and Meteorological Applications (NEWA) for every occurrence in the sample. The weather station that was geographically nearest to each sample site was used to collect data. The R program (R Foundation for Statistical Computing, Vienna, Austria) was used for all the analyses. The version used was Version 3.6.2. A logarithm base 10 transformation was employed in all analyses. Some analytes were not included in the 2,816 collected samples. Consequently, 5-point summaries and missing data percentages were constructed for each variable. Bayesian mixed models were used to evaluate the presence of human and ruminant MST markers as well as the regional and temporal patterns of $\log_{10}$ FIB levels. Environmental factors linked to $\log_{10}$ FIB levels and probability of MST detection were determined. To evaluate the $\log_{10}$ values of *Escherichia coli*, *Enterococcus*, faecal coliform and total coliform, Bayesian mixed models were constructed. Alternatively, Bayesian linear mixed models have been developed to evaluate the likelihood of detecting ruminants or human MST. Using fixed factors such elevation, latitude, longitude, rurality (i.e. whether the sample was gathered within or outside of Syracuse, NY city boundaries) and canal, models were created to analyse spatial and temporal trends. The models also consider months and years. The monthly variable was handled as a random effect in models with geographical fixed effects, whereas the waterway variable was treated as a random component in models with temporal fixed effects. The associations between faecal markers and environmental variables, including the random effects of month and canal, were explored using Bayesian mixed models. When simulating land use, climate and water quality, the models only considered a single fixed influence. Agrarian, pasture-hay, forest-wetland and developed cover fractions were used to assess land use within specified distances from the sampling site. Wind speed 0 - 1 day before, rainfall 0 - 1, 1 - 2 and 2 - 3 days before, and air temperature at the time and 0 - 3 days before were the weather parameters. Log_{10} chloride, conductivity, TDS, nitrate, TOC, TPO, salinity, TSP, turbidity, pH and total organic carbon were the water quality parameters. Because not all analytes were detected in every sample, models were only employed

when there were a minimum of 20 paired observations for the microbiological target and associated environmental components. The faecal coliform models were relevant to all the parameters explored in this investigation because of the high number of samples tested (N = 2,658). Total phosphorus, total organic carbon, total dissolved solids and chloride were all potential explanatory factors which the Enterococcus models failed to account for, and uninformative priors were used to fit the models using the brms package. A thinning setting of 10 µm was applied to all 3 chains. Some models exhibited problems with convergence, even though most had a fixed number of iterations per chain (5,000 with a burn-in of 2,500). Following the methodologies of relevant research attempts [59-64], the burn-in and iterations per chain were increased to fix models that had trouble achieving convergence. The effect estimates were obtained using the bayestestR program, which also provided the median, mean, maximum a posteriori (MAP) and 89 % integrity interval. Confidence intervals, as used in frequentism and the Bayesian credibility theory, are not interchangeable. The current information suggests that the most likely range of effect estimates is the 89 % credibility interval. The data suggest that the effect estimate is likely to fall somewhere between x and y, with an 89 % probability.

Alternatively, Bayesian models offer a more flexible perspective by providing a probabilistic view of parameters and associated uncertainties. The omission of significance from the Bayesian models distinguishes them from the frequentist regression models. Compared to frequentist methods, Bayesian investigations are better able to withstand small sample sizes, do not depend on the researcher's chosen *p*-value threshold, and are less likely to make Type I errors. Determining whether the parameter falls beyond a range that is regarded to have minimal influence is preferable to depending only on a *p*-value below a defined threshold to indicate a link. The ROPE % quantifies the impact. The range of practically no influence and the 89 % believability interval were used to determine the proportion of overlap. The FIB levels or the probability of MST detection become more specific as ROPE % approaches zero. More specifically, to comprehend the ROPE, the following criteria can be selected: if the value is greater than 99 %, then the impact is negligible; if it is greater than 97.5 %, then the effect is probably insignificant; if it is between 2.5 and 97.5 %, then the effect is uncertain; if it is less than 2.5 %, then the effect is non-negligible; if it is less than 1 %, then the effect is considerable. One way to evaluate connections in Bayesian regression is to examine the Probability of Direction (PD). Regardless of whether the effect is small or large, PD can determine whether it is good or negative. Numerous frequentist p-values strongly correlated with the Pearson correlation coefficient (r). With PD values close to 1.0, it can be said, with more certainty, that a component has a positive or negative influence. This indicates trust in the link's intended path. The 2-sided frequentist *p*-values for PD values of 0.95, 0.975, 0.995 and 0.9995 were 0.10, 0.05, 0.01 and 0.001, respectively. Because ROPE and PD are not complementary metrics, a component may rank highly in PD, but poorly in ROPE. This result strongly indicated an insignificant impact. This inverse connection indicates that it is a significant issue. In summary, the probability that the described component has a negative or positive influence is PD, and its importance may be rated as significant or non-significant by coupling PD with ROPE. In addition, statistical significance (PS) measures were computed to determine the likelihood that the impact of the parameter exceeded a specific threshold, indicating a small impact on the median. To determine whether the effect is significant in only 1 direction, this test uses unidirectional

equivalence. For the values to be considered practically significant, they must be greater than 0.5. However, in this case, a more cautious threshold of 0.75 was employed out of an abundance of caution. Reporting a significant amount of data is necessary when PS is greater than 0.50, PD is greater than 0.75, and ROPE is less than 0.25 after looking at the suggested criteria and reasons for PD, PS and ROPE.

Statistical inference

Data on the physicochemical and microbiological water quality of the 2,816 samples are shown in **Table S2(b)**. A total of 2,658 samples were tested for faecal coliforms, and 281, 288 and 96 samples were tested for *Enterococcus*, total coliforms and *E. coli*, respectively. As shown in **Table S3**, the average $\log_{10}$ concentration (CFU/100-mL) of *E. coli*, *Enterococcus*, total coliform and faecal coliform levels were 2.2 (SD = 0.9; Range = −0.3, 4.2), 2.0 (SD = 0.9; Range = −0.3, 5.0), 3.0 (SD = 1.0; Range = −0.3, 4.2) and 2.5 (SD = 0.8; Range = −0.4, 7.2), respectively. To illustrate this point, total suspended solids had a mean of 7.0 mg/L and a standard deviation of 110.5. The measured values ranged from 1.0 to 2,077.0. The fact that MST markers were not associated with FIB suggests that there are several potential contaminants and that the levels of *Enterococcus* and total coliforms correlate favourably with faecal coliforms. There was a correlation between *E. coli* and other faecal indicator bacteria (FIB), land use and geographic variables. **Table S4** shows that, whereas rural areas were previously linked positively to faecal coliforms and *Enterococcus* bacteria, they were shown to be negatively correlated with *E. coli* bacteria. Of the 31 samples examined for host-specific microbial source tracking markers (MST), 22 (71 %) tested positive for human markers and 11 (35 %) tested positive for ruminant markers (MST). The detection of ruminant markers was significantly related to that of human markers (PD = 0.91; PS = 0.80). However, the precise strength of this connection remains debatable (ROPE = 0.16). In addition, the levels of additional microbial targets examined in this study (PS < 0.50, PD < 0.95 and ROPE > 0.025) were not linked to any marker. This suggests that ruminant and human faecal contamination may have the same causes, or that faecal indicator bacteria (FIB) originate from different places. In addition, this study did not consider additional contamination sources specific to certain hosts. Previous evidence suggests that similar to E. coli, FIBs may survive and even flourish under conditions that are not typically conducive to their growth, including algal mats, water and soil. Consequently, there may not be a correlation between FIB levels and host-specific markers, as FIB levels may not accurately represent recent faecal contamination. It should be noted that a previous study conducted on Onondaga Creek (NY, USA) contradicts the results presented here. Prior research has found a strong negative connection between human and ruminant markers, and a strong positive association between *Enterococcus* levels and human marker detection. The current study incorporates other waterways and a longer monitoring period (2008 - 2017), which may explain why the results differ from previous work that used Onondaga Creek as a sample site in 2015. There is consensus that water quality varies not only between locations and times, but also across different analytical scales. This could also mean that the processes that cause faecal contamination vary among the watersheds studied, which would make it harder to find contamination patterns specific to each watershed (which was the main aim of the prior study on Onondaga Creek, NY, US).

Nevertheless, the study did not examine particular patterns of connectedness related to the watershed because of the limited sample size (only 3 from Onondaga Creek out of 31). Researchers studying the Onondaga Lake and similar watersheds should consider this component in the future. The relationship between FIB levels and nearby land use was affected by the buffer size used to establish the land-use features. In Syracuse, New York, FIB (faecal indicator bacteria) distribution was not uniform. Although *E. coli* levels were lower outside the city, faecal coliforms and *Enterococcus* levels were usually higher. A disparity was identified when comparing the samples from urban and rural areas, as shown in **Figure 2(a)**. No correlation was found between rural regions and prevalence of ruminant or human microbial source tracking (MST) markers. With an 89 % confidence interval (CI) of −0.61, 0.00; PF of 0.94; and ROPE of 0.10, the average *E. coli* level in the urban samples was 0.35 $\log_{10}$CFU/100 mL lower. The levels of *Enterococcus* and faecal coliforms were higher in these samples than in those obtained from rural regions (CI = 0.13, 0.52; PD = 0.99; ROPE < 0.01) and other places. It is expected that faecal coliforms and *Enterococcus* will correlate favourably with latitude and negatively with elevation along the rural-urban gradient, which is characterised by an increase in latitude and a decrease in height. This is due to the fact that compared to the surrounding rural regions, Syracuse, NY is situated lower and farther north. **Table S2(a)** shows the inverse trends for the *E. coli*, human and ruminant MST markers. These tendencies are in line with the observed land use correlations. As shown in **Table S2(b)**, pasture cover was the only land use type that was positively correlated with *E. coli* levels. The positive correlation between pasture cover and the likelihood of detecting bovine and human MST markers is consistent with this result. **Table S3** shows that the developed cover was favourably connected with *Enterococcus* and faecal coliform levels, whereas agricultural land use and forest-wetland cover were adversely correlated. Regardless of the buffer size, a surprising correlation between *E. coli* levels and pasture cover persisted. **Table S4** and **Figure 1(b)** show that connections between *Enterococcus* and faecal coliform levels, ruminant detection, and pasture cover were only observed when buffer sizes of 1,098 and 122 m were utilised. When looking at the correlation between total coliform levels and factors, including cropland presence, human marker detection, ruminant marker detection, degree of developed regions and forest-wetland cover, it appeared that the size of the buffer zone had an effect. Consequently, links were found between particular land uses and each faecal signal. However, it appears that the level of analysis determines the strength of the correlations. It is not unexpected that the size of the research region affected the model results of a prior investigation of nutrient poisoning in water bodies in Maryland. Therefore, this finding was logical. A related study investigated the correlation between riparian land use and *Escherichia coli* levels in Arkansas, USA. According to the research, the change-point - the exact spot within the land use parameter when *E. coli* levels changed dramatically - was affected by the area and size of the riparian buffer. Because the buffers used in this study were selected to resemble the recommended distance between agricultural water sources and large livestock operations (> 1,000 head), as stated in the Leafy Green Marketing Agreement (exactly 366 m), this finding has significant consequences for guaranteeing the safety of the produce. It stresses that in coming up with methods or suggestions to identify potential risks of faecal contamination, one must consider the geographical scale (such as utilising buffers of 122, 366 or 1,098 m). Consistent with previous research in the northeastern US and elsewhere, this study found robust relationships between adjacent land use and

faecal pollution. For every percent increase in pasture cover within a 1,098 m radius of the sample location, there was a 0.30 $\log_{10}$ CFU/100-mL rise in *E. coli* levels (89 % CI = 0.17, 0.39; PD = 1.00; ROPE < 0.01). The results of a previous study conducted in Arkansas, USA, were corroborated by this association. As pasture cover increased, faecal coliforms and *Enterococcus* levels decreased; however, as the cover increased, the opposite was true (**Table S1**, in the Supporting Information document enclosed with the article). It can be inferred that the origins of *Enterococcus* and faecal coliform contamination are distinct from those of *E. coli* or MST indicators based on evidence of land-use correlations. Whereas the latter denotes rural areas, the former denotes urban centres. The presence of human markers in agricultural areas and rural regions suggests the presence of sources of human faecal contamination in rural hinterlands, although these markers are more commonly found in metropolitan areas and developed properties. For instance, although most of the southern rural sections of Syracuse use septic systems, other parts of the city have access to water and sewer infrastructure. Therefore, faecal pollution in rural communities may be caused, in part, by malfunctioning sepsis systems. This analysis did not investigate the age, location, or density of rural faecal pollution sources, such as septic systems, because of data restrictions. However, prior studies in Queensland, Australia, Georgia, USA, Michigan and NY, USA, lent credence to this idea. There is a pressing need for further study of the causes and effects of rural faecal pollution in the Onondaga Creek Watershed. Among the rivers examined, Onondaga Creek consistently showed greater levels of all indicators, except for the correlations between land use and other faecal markers, which varied. The average *E. coli* level in Kimber Brook was over 2 $\log_{10}$ units lower than that in Onondaga Creek, with a 95 % confidence range ranging from −3.29 to −0.31. The odds of the difference being less than 0.01 (ROPE) were very low, whereas the odds of the difference being practically different (PD) were 0.97. According to earlier research, the majority of the debris and germs that end up in Onondaga Lake come from the Onondaga Creek. The reliability of faecal indicators, such as host-specific markers, allows us to infer that regardless of land use at each sample location, the faecal contamination patterns in the rivers studied here varied. This deterioration was more pronounced in Onondaga Creek than in others. This finding may not be surprising, but it is essential to direct efforts to reduce the amount of sewage in the surface waterways that drain into Onondaga Lake. The rivers considered in this research were all empty into Onondaga Lake; therefore, fixing the problem should ideally focus on Onondaga Creek instead of Ley Creek, Harbor Brook, or Hopper Brook.

FIB contamination is linked to nutrient and sediment pollution. This suggests that the pollutants may have come from the same places or may have been transported to the river in the same way, and that there was a positive correlation between nutritional (nitrate and phosphate), chemical (salinity and conductivity), or sediment (turbidity and total suspended solid) pollution for all faecal indicator bacteria (FIB) that were examined. The concentration of *E. coli*, *Enterococcus*, and faecal coliforms respectively increased by 0.16 (CI = 0.06, 0.26; PD = 0.99; ROPE = 0.11), 0.68 (CI = 0.08, 1.24; PD = 0.96; ROPE = 0.01) and 0.33 (CI = 0.26, 0.41; PD =1.00; ROPE < 0.01) $\log_{10}$ CFU/100-mL for every $\log_{10}$ increase in nitrate levels. Table S5 (in the Supporting Information document) shows consistent trends in conductivity, salinity, total organic carbon, total phosphorous, total dissolved solids, total suspended solids and turbidity. The levels of *Enterococcus* and faecal coliform bacteria were enhanced by approximately 1 logarithmic unit for every

logarithmic unit increase in the total suspended solid levels (89 % CI = −0.48, 2.39; PD = 0.89; ROPE = 0.03) and total organic carbon levels (89 % CI = 0.71, 1.48; PD = 1.00; ROPE < 0.01), respectively. This work adds credence to earlier research by showing a good correlation between physicochemical and microbiological water quality parameters. The results showed that nutrient levels explained 48 % of the variation in *E. coli* levels in the 64 rivers across the US state of Michigan. In addition, research in a watershed in Ontario, Canada, linked high levels of total phosphorus and conductivity to the presence of faecal parasites, most notably Giardia. The physicochemical and microbiological properties of Hawaiian streams have been studied by Vawda *et al.* [65]. They found that many physicochemical and microbiological factors are positively correlated. They found a connection between enterococci, such as E. coli and total phosphorus, and turbidity. Several studies conducted in various locations and with different types of water (e.g. ponds, canals and reservoirs) have shown that turbidity levels are positively correlated with microbiological water quality. These studies have also been conducted in Ecuador and in the southeastern and northeastern United States of America. Various study strategies, including microbiological targeting and sampling methodologies, have also been utilised in these studies. Several studies have suggested adding turbidity to the list of indicators used to identify streams that may be contaminated with FIBs because of the long-standing correlation between these 2 variables and microbiological water quality. Similar processes may account for microbial, sediment and nutrient contamination, as indicated by the correlation between microbial water quality (e.g., faecal indicator bacteria levels), nutrient levels (e.g., nitrate) and sediment levels (e.g., turbidity) in this and previous research. Prior research on microbial, nutritional and sediment pollutants as well as their origins, destinations, and migrations lends credence to this finding. The scientific community generally agrees that rainfall plays an important role in the transport of sediments, nutrients and microorganisms from their original locations to surface rivers. Wind speed, which is used to evaluate storm occurrence and rainfall prior to sampling events, is associated with high levels of faecal indicator bacteria (FIB), according to previous research. Even though the second and third phases were only sampled during dry weather, the study emphasised that the studied rivers were consistently contaminated with germs. Particles frequently harbour bacterial clinging. Nearly half of the enterococci identified in the study were found to be bound to particles of dirt or manure, according to previous research. Sedimentary materials on streambeds are known to harbour faecal indicator bacteria (FIBs) in a water channel. Disruption of these sediments increased the FIB content in the water. The lack of correlation between the FIB types may be attributable, in part, to the fact that FIBs' persistence in stream sediments varies. These findings suggest that implementing measures to control faecal pollution can reduce sediment and nutrient contamination. Land managers in the Syracuse (NY, US) area are working hard to restore the water quality of Onondaga Lake; thus, this is essential for them. In addition, other locations with similarly degraded waterways could benefit from these results.

Onondaga Creek, Ley Creek and Harbor Brook are waterways in the Syracuse, NY area. This study aimed to identify the probable causes of faecal contamination in both the urban and rural areas of these creeks. In addition, this study aimed to examine the relationships between environmental factors, including weather and nutrition levels and microbiological indicators of faecal contamination. Finally, this study aimed to provide recommendations for maximising restoration efforts by prioritising mitigating initiatives.

Finally, there was no correlation between FIB levels and host-specific MST markers in the present study. Faecal coliform and *Enterococcus* levels, as well as geographical variables such as rurality, were positively correlated with land use. However, they were inversely related to *E. coli* levels and vice versa. These results indicate that there are several potential sources of faecal contamination in the streams being tested. In contrast, this study showed that FIB levels were strongly correlated with nutrient and sediment levels in the analysed rivers. Therefore, it is more likely that these rivers were contaminated from the same locations or were contaminated using the same procedures. This finding has great practical significance in light of the current focus on enhancing the water quality flowing into Onondaga Lake. This implies that nutrient and sediment contamination might decrease as a result of attempts to reduce faecal pollution. Regarding the latter point, this study showed that FIB levels differed geographically, with the highest amounts found in Onondaga Creek. In addition, there was a considerable correlation between these levels and the nearby land use. These results highlight the importance of considering land use in the area and the heterogeneity between streams when planning strategies to decrease the faecal risk in water used for agriculture and leisure. When planning methods for managing watersheds, this is of utmost importance. Based on this analysis, it appears that Onondaga Creek is the best choice for mitigation activities compared with Ley Creek, Harbor Brook, and Hopper Brook. Focusing on the most severely affected stream that flows into Onondaga Lake, rather than all rivers, allows this strategy to maximise both financial resources and water quality.

Empirical validation

Chronic kidney disease (CKD) is an important global public health problem that affects approximately 15 % of the world's population. The root causes of chronic kidney disease include hypertension, diabetes and chronic nephritis. Chronic kidney disease (CKDu), the cause of which is unknown, has been reported in several parts of the world, including Central America, India and Sri Lanka. However, this is unrelated to the aforementioned factors. With a frequency of 15 - 23 % in North Central Province, Chronic Kidney Disease of Unknown Cause (CKDu) threatens the health of more than 400,000 individuals in Sri Lanka. Living in dry rural areas, most people with CKDu have a very low standard of living. According to recent studies, chronic kidney disease (CKDu), which has an unclear cause, is associated with poor groundwater quality. Additionally, the infrastructure for supplying filtered water is inadequate. One study stated that the availability of high-quality drinking water from underground sources is a determinant of CKDu development in regions where it is widespread. Additionally, the groundwater sources associated with CKDu were not evenly distributed. Residents of underdeveloped nations such as Sri Lanka, where chronic kidney disease of unknown aetiology (CKDu) is common, are thus very concerned about finding clean groundwater supplies and dissolved organic carbon (DOC) levels in groundwater have been associated with renal disorders. According to research conducted in Balkan kidney disease-endemic areas, a buildup of humic chemicals can induce the death of human kidney cells. However, there is a dearth of research on how CKDu-affected and unaffected water sources differ in terms of the primary components of dissolved organic matter (DOM). Dissolved organic matter (DOM) in groundwater from CKDu-affected areas was found to be less bioavailable, more resistant to degradation, and higher in

aromatic compound concentration, as compared to non-CKDu areas, according to the available research. However, the exact role of the sensitivity of DOM components in identifying water sources linked to CKDu remains unknown, and the benefits of excitation-emission matrix spectroscopy (EEM) include easy pre-processing, higher sensitivity and rapid measurement. Their use in describing FDOM or fluorescent dissolved organic matter has been extensive. Pairing EEMPARAFAC with parallel factor analysis makes it a powerful tool for understanding the changes in dissolved organic matter (DOM), determining its fluorescent components and sources, and clarifying its optical properties. It may also be used to examine how DOM evolves over time as a result of both environmental and human influence. Twenty variables were analysed for correlation using the EEM-PARAFAC method, and studies have demonstrated that DOM and general water quality may be affected by the concentration of inorganic chemicals in groundwater. Chronic Kidney Disease with an Uncertain Cause (CKDu) has been linked in many studies.

The fluoride (F^-) concentrations in groundwater samples collected from regions in Sri Lanka where CKDu is prevalent frequently surpass the threshold established by the World Health Organization (0.6 mg L^{-1}). Groundwater in regions affected by CKDu also exhibited greater hardness and a higher Ca^{2+}/Na^+ ratio than those in non-CKDu areas. CKDu has also been linked to elevated levels of nephrotoxic heavy metals (Cd, Pb and Si) in groundwater. A recent study found that organic compounds can pose a risk to human kidneys by creating complexes with Ca^{2+} and SO_4^{2-}, potentially causing injury. Focusing on samples linked to chronic kidney disease of unknown aetiology (CKDu) and unrelated samples, this study examined the optical properties, content and origin of fluorescent dissolved organic matter (FDOM) in groundwater samples from Sri Lanka. The EEM-PARAFAC method was used for this purpose. The study also aimed to determine whether it would be possible to use inorganic-sensitive compounds with specific fluorescent components to determine the origin of CKDu-related water. This study also aimed to develop a screening and early warning system for CKDu-related groundwater sources by including crucial fluorescent components. Contributing to the prevention and surveillance of CKDu, this work aims to provide a reliable approach for safeguarding drinking water quality, with the highest point on the black rectangle representing the most valuable thing and the lowest point at the bottom. The 75th percentile is at the top of the black rectangle and the 25th percentile is at the bottom. The average value of the central white square can be observed. The likelihood of the density is shown in the density diagram outside the box. The 2 asterisks indicate significant variations in spectral indices between the 2 groups, with *p*-values less than 0.01 and less than 0.05, respectively.

This study investigated the correlation between chronic kidney disease (CKD) of uncertain aetiology and the presence of dissolved organic carbon (DOC) and extractable elemental markers (EEMs) in groundwater. **Figure 2(d)** demonstrates that the groundwater with CKDu had lower levels of dissolved organic carbon (DOC) at 3.25 ± 0.73 mg C L^{-1} compared to surface water at 4.65 ± 0.89 mg C L^{-1}. However, the groundwater without CKDu had even lower levels at 2.81 ± 0.76 mg L^{-1}. The differences were statistically significant ($p < 0.05$). Although this study did not observe elevated levels of dissolved organic carbon (DOC), Panda *et al.* [66] found that areas with a high incidence of chronic kidney disease of unknown aetiology (CKDu) had significantly higher concentrations of DOC (6.4 mg C L^{-1}) than areas without CKDu (3.7 mg C L^{-1}). The disparity in dissolved organic carbon (DOC) levels between

groundwater samples that exhibited chronic kidney disease of unknown aetiology (CKDu) and those that did not might be ascribed to 2 possible factors. Initially, the concentration of dissolved organic carbon (DOC) in the CKDu groundwater likely increased because of an influx of surface water recharge. **Table S2(a)** (in the Supporting Information document) provides further data indicating that microbial activity is more pronounced in groundwater that is unaffected by CKDu. This results in the decomposition of organic compounds into inorganic chemicals, which is discussed in more detail later in this paper. The CKDu groundwater had a higher a_{254} value of 5.07 m^{-1} than non-CKDu groundwater, which had an a_{254} value of 4.38 m^{-1}. CKDu groundwater appears to have a greater abundance of unsaturated dissolved organic matter (DOM). The concentration of unsaturated dissolved organic matter (DOM) with an absorbance at 254 nm (a_{254}) is typically less than 1 m^{-1} in natural groundwater. Based on **Table S2(b)**, the investigation revealed that the a_{254} value in all groundwater samples was as low as 1.15 m^{-1}. This indicates that both CKDu and non-CKDu groundwater may have been contaminated to some extent. The a_{254} values of groundwater (median: 10.48 m^{-1}) were lower than those of surface water, perhaps because of human intervention. a_{254} was shown to have a positive correlation with DOC concentrations, indicating that human activities may be responsible for the rise in DOC levels. **Figure 2(b)** shows that the properties of dissolved organic matter (DOM) in surface water and groundwater affected by chronic kidney disease of unknown aetiology (CKDu) were distinct from those in groundwater not affected by CKDu. In the non-CKDu groundwater EEM study, the intensity of the T-peak signal was significantly higher, but that of the A-peak signal was considerably lower. This demonstrates the presence of a greater number of bioactive tryptophan-like chemicals in non-CKDu groundwater. When Chronic Kidney Disease of Unknown Etiology (CKDu) impacted both surface water and groundwater, the DOM EEMs (Excitation-Emission Matrices) exhibited a flat region between peaks M and C in the peak A region. This illustrates the substantial impact of organic substances with humic characteristics on the surface and groundwater affected by the CKDu. Furthermore, the resemblance between the DOM of CKDu groundwater and surface water in terms of EEMs patterns implies a correlation, indicating that the groundwater was replenished by surface water. **Table S3** (in the Supporting Information document) demonstrates that there was no significant difference in the level of DOM aromaticity ($SUVA_{254}$) between groundwater treated with CKDu and groundwater not treated with CKDu (mean: 0.78 ± 0.37; $p > 0.05$). Nevertheless, the a_{254} values in many groundwater samples from CKDu were elevated because of the presence of aromatic dissolved organic matter (DOM). The majority of unsaturated dissolved organic matter (DOM) consists of aromatic compounds, as indicated by the significant positive relationship between a_{254} and $SUVA_{254}$. Based on these data, it can be concluded that CKDu groundwater contains a higher concentration of high-molecular-weight organic compounds. This is shown by its lower spectral slope ($S_{275\text{-}295}$) with a mean value of 20.5 ± 6.7, in contrast to the unaffected groundwater which has a mean value of 23.7 ± 9.4. The heightened microbial activity seen in the non-CKDu groundwater, as shown by the significantly higher biological index (BIX) values, might be attributed to this phenomenon.

Complex organic molecules, such as lignin, can be broken down into simpler organic components, such as aromatic compounds, by microbes and the dissolved organic matter (DOM) fluorescence indices (FI) varied from 1.26 to 1.88 across all samples, suggesting a combination of microbial and terrestrial sources with a small amount of autochthonous material. The results showed that dissolved organic matter

(DOM) had a higher Fluorescence Index (FI) in non-CKDu groundwater (mean: 1.64 ± 0.12) than in CKDu groundwater (mean: 1.52 ± 0.17), suggesting that a larger amount of organic matter originated from outside sources and a smaller amount from inside the area. The terrestrial impact of groundwater movement is often greater than that of stagnant groundwater. Therefore, groundwater that is affected by CKDu and has a reduced FI can potentially experience improved flow. Groundwater and surface water that did not contain CKDu did not differ significantly in the FI. In the box plot, the maximum and minimum values were represented by the highest and lowest points, respectively. The distribution of the data can be observed at the 75th percentile at the top and the 25th percentile at the bottom of the box. The median value is represented by the horizontal line in the middle of the box, whereas the average value is represented by the square in the centre. Two asterisks denote significant differences between 2 groups when $p < 0.05$, and 1 asterisk denotes significant differences between 2 groups when $p < 0.01$. Average readings for groundwater (mean: 0.81 ± 0.09) and surface water (0.79 ± 0.08) were significantly lower than those for non-CKDu groundwater (mean: 0.86 ± 0.08) ($p < 0.05$). This agrees with the conclusions of the FI that dissolved organic matter (DOM) in groundwater, which does not include CKDu, is more biologically active. Overall, the positive correlation between FI and BIX is consistent with the findings of Laad and Ghule [67]; Karila *et al.* [68]. BIX was the method of choice when comparing the biological activity of DOM in groundwater and surface water. Groundwater that included CKDu had a dissolved organic matter (DOM) humification index (HIX) of 4.36 ± 2.16, which was substantially greater than that of groundwater that did not contain CKDu (mean: 3.15 ± 1.49; $p < 0.01$). Evidently, DOM was humified to a higher degree in the groundwater samples from CKDu. Possible causes for this phenomenon include increased microbial activity in groundwater that does not have chronic kidney disease of unknown aetiology (CKDu) or the intrusion of surface water into groundwater that does have CKDu. Reducing the H/C ratio caused the humification level to increase in direct proportion. Most organic materials with low hydrogen/carbon ratios are unsaturated or contain hydrophobic aromatic compounds. Consequently, hydrophobic aromatic or unsaturated chemicals, known for their durability and persistence, are more likely to be present in Dissolved Organic Matter (DOM) in the CKDu groundwater. Groundwater samples taken from areas with chronic kidney disease of unknown cause (CKDu) showed elevated concentrations of fulvic acid and other large hydrophobic organic compounds, according to research by Díez-Quijada *et al.* [69]; Chen *et al.* [70], and the results of the correlation analysis between FI and HIX were inconclusive. Hence, **Figure 3(c)** shows that HIX source identification was different from FI and BIX source identification. There may be a connection between complicated hydrogeological conditions and human influences and the decoupling of the Fracture Index (FI) and the Hydrogeological Index (HIX). One example is the potential for endogenous organic material to be released from sediments and rocks as a consequence of water-rock interactions. It is important to remember that spectral indices are computed at certain wavelengths that might be affected by the presence of many interacting fluorophores; however, they can provide a rapid evaluation of DOM composition. Complex organic molecules can originate from many different locations, and determining their composition could be a ‘distortion’. Therefore, the fluorescence results should be interpreted with caution and PARAFAC should be used for further research. Contaminated groundwater from CKDu can damage FDOM components. Four luminous components were identified and validated using half-splits, and the

results are shown in **Figure 4(c)**. C1 represents components that resemble terrestrial humic substances, which are created via biogeochemical breakdown of organic materials found on Earth. The shapes of these parts are similar to those of the Coble A and C summits. Composed of low-molecular-weight components resembling humic substances produced by microbes, C2 is similar to the Coble M peak. These components can be used to make up the bulk of the visible dissolved organic matter (DOM) in wastewater. The parts with high molecular weights that resemble humic substances are referred to as C3. These elements are derived from naturally occurring organic substances. The letter C4 stands for components that resemble proteins and are produced by microbes that occur naturally in the environment.C1 (68.7 % frequency) and C2 (62.9 % frequency) were the most common FDOM components in CKDu and non-CKDu groundwater, respectively. This highlights the importance of humic-like chemicals in the FDOM. The proportion of C1 was 35.1 % in the CKDu groundwater and 28.0 % in the non-CKDu groundwater, with a *p*-value of less than 0.01. The proportions of C2 and C3 % in the groundwater samples taken from both CKDu and non-CKDu locations were the same ($p > 0.05$). C2 is probably a byproduct of organic matter created by microbes in the same location or humic compounds that are easily separated by minerals given the comparable geological and irrigation conditions in the study region. In contrast, the introduction of agricultural products is likely to be associated with C3. In addition, C4 % was much higher in DOM from non-CKDu groundwater (23.8 vs. 17.3 %, $p < 0.01$) than in DOM from CKDu groundwater. Organisms can more readily absorb DOM from the non-CKDu groundwater. As protons are produced during the microbial degradation of organic materials, the lower pH values observed in non-CKDu groundwater lend credence to this hypothesis. The C1 % concentration in the surface water was noticeably higher than that in the non-CKDu groundwater ($p < 0.01$), but was the same as that in the CKDu groundwater ($p > 0.05$). This suggests that the increase in C1 % in the CKDu groundwater may have been caused by the inflow of surface water. C1 % loadings were positively correlated with PC1, explaining 43.9 % of the variance, according to the 2-component principal component analysis. **Figure 3(d)** shows that C4 % loadings were negative for PC1. In the PC1 positive group, there was an abundance of humic-like chemicals, whereas in the PC1 negative group there was an abundance of protein-like components associated with microbial activity. Notably, PC1 was positively correlated with over 50 % of the CKDu-containing groundwater samples and only slightly more than 22 % of the CKDu-free groundwater samples. Based on these results, the C1 and C4 % values are good indicators of whether the groundwater is CKDu. However, different screening findings were observed, because C4 is vulnerable to disturbances caused by microbes in aquifers.

Therefore, CKDu groundwater may be detected by utilising C1, a molecule that exhibits both environmental stability and persistence. The presence of positive loadings on PC1 for HIX, a_{254} and $SUVA_{254}$ indicates a correlation with humic-like chemicals (C1). PC2 showed a favourable trend with a percentage of 22.2 %, whereas both C2 and C3 % showed positive trends. Approximately 53 % of groundwater samples showed a declining trend in PC2 for CKDu, whereas the ratio was 28 % for non-CKDu. The use of C2 and C3 % as indicators to identify CKDu groundwater may be misleading because both CKDu and non-CKDu groundwater exhibit similar concentrations of these chemicals. This study investigated the relationship between Dissolved Organic Matter (DOM) and water chemistry in relation to chronic kidney disease of unknown aetiology (CKDu). Fluoride ions (F^-) can harm the human kidneys.

Chronic Kidney Disease of unknown origin may be associated with elevated fluoride (F^-) levels in water sources. Analysis of the groundwater samples revealed a significant positive correlation ($r = 0.62$, $p < 0.05$) between the C1 % and F^- values. This indicates that C1 is a highly responsive biomarker for the detection of probable groundwater sources that contribute to CKDu. The correlation between Ca^{2+} and the C1 % measure was significantly positive, with a correlation coefficient (r) of 0.60 and a *p*-value less than 0.05. The carboxyl groups present on C1 molecules can bind to Ca^{2+} ions, resulting in the formation of a complex that can potentially harm human kidneys. This complex plays a significant role in the development of Chronic Kidney Disease with an Uncertain Cause (CKDu). This supports the finding that places affected by CKDu had a greater abundance of Ca^{2+} in their groundwater than those that were not affected. The study discovered a significant positive association ($r = 0.60$, 0.61, respectively, $p < 0.05$) between the levels of Silicon (Si) content and hardness, along with an increase in the percentage of carbon (C1 %). The consumption of groundwater with elevated levels of hardness and silicon is believed to be linked to CKDu because of its detrimental effects on human embryonic kidney cells. Ultimately, the inorganic chemical constituents frequently associated with CKDu showed a substantial correlation with the C1 %. Based on the available information, C1 % may be a viable method for identifying groundwater sources associated with CKDu. **Figure 2(d)** shows that the elevated C1 % level in the CKDu groundwater, as opposed to that in the non-CKDu groundwater, can be attributed to the intrusion of surface water. As shown in **Table S3**, the concentrations of the indicated inorganic chemical components were much lower in surface water than in groundwater. The process of recharging groundwater from surface water is slow because of the predominant reliance on weathered fissure water as the primary source of groundwater in the research region. The progressive buildup of inorganic compounds in groundwater is a result of complex interactions between water and rocks, including ion exchange and leaching, which naturally occur throughout this gradual process. The recharging process is expected to result in a greater release of F^- ions from minerals through ion exchange, owing to the elevated pH levels in the surface water. This is because the F^- and OH^- ions are chemically unreactive. The groundwater affected by CKDu had elevated levels of F^- and a lower pH than the surface water.

Sensitivity analysis

A comparison of groundwater affected by chronic kidney disease of undetermined origin (CKDu) to that unaffected by the condition revealed significant differences in dissolved organic matter (DOM). The popular CKDu recognising threshold assessment (CRTA) method can be used to detect and swiftly identify water sources associated with CKDu based on these alterations, especially in terms of C1 % and HIX. Investigating the relationship between the assumed DOM indicator threshold and the detection probability of CKDu groundwater (DPC) and non-CKDu groundwater (DPN) in 54 groundwater samples allowed us to determine the appropriate recognition threshold (RT) for various DOM indicators and evaluate the efficacy of the CRTA method. The results showed that DPC and DPN had nearly perfect S-shaped curves and were very sensitive to changes in HIX, C1 %, DOC and C4 %. Considering C1 % as an example, when the anticipated C1 % threshold was higher than the junction value (28.8 %), the DPC continued to grow and was consistently larger than the DPN. In contrast, the DPN increased with decreasing C1 % and

remained higher than the DPC when the predicted C1 % threshold was below the intersection value (28.8 %). With C1 % greater than RT, DPC attains a detection probability of 70.1 % at the intersection if the predicted junction threshold of C1 % is utilised as RT. This indicates that CKDu was associated with at least 70.1 % of the identified water sources. The identified water source will be at least 70.1 % disconnected from CKDu if the assumed threshold of C1 % is smaller than RT because the DPN will be greater than 70.1 %. The excellent identification capacity of C1 % RT is demonstrated by the minimum 70.1 % likelihood of distinguishing between water sources connected with CKDu and those that are not. Additionally, the association between the assumed C1 % threshold and probability of detection was identical to that between the assumed HIX, DOC and C4 % thresholds and probability of detection. C1 % was more effective than HIX, DOC and C4 % for recognition but only because the likelihood of detection at the junction was much lower. In reaction to changes in the threshold, the DPN or DPC curves of the other DOM markers exhibited significant oscillations. In addition, the correlation between the anticipated PC1 score threshold and the likelihood of detection suggests that PC1 score recognition was 63.8 %. An empirical Boltzmann equation was employed to compare the supposed C1 %, C4 %, HIX and PC1 score thresholds with detection probability. **Figure 4(a)** shows that the expected C1 % criterion yielded the highest R^2 value. The accuracy of the fitting findings was validated by the verification results, which demonstrated close proximity between the anticipated and observed values (**Figure 5**). Furthermore, regardless of whether the C1 % or PC1 score was used as an indicator of recognition, the paired t-test indicated no statistically significant difference ($p > 0.05$) between the projected and observed values. C1 % had the best accuracy because of its minimal root mean square error (RMSE). DPN is indicated by the blue lines and DPC by the red lines. The likelihood of solid line identification was computed using mathematical sequence as the cutoff. The solid lines' fitting results are reflected by the dashed lines, and the 95 % confidence intervals for the fitting curves are shown by the pink bands. This study's data was in agreement with the observed C1 % outcome. When the CRTA approach was used to differentiate between the water sources associated with CKDu and those unrelated to CKDu, the optimum and most efficient DOM index was C1 %. The CRTA approach achieved an RT of 28.8 % for C1 % by using curve-fitting methods. This study analysed the confusion matrix of 54 groundwater samples using receiver operating characteristic (ROC) and precision-recall (P-R) curves to learn more about the advantages and predictability of the ideal threshold of the CRTA method. If the AUC was greater than 0.7, the curve model was deemed predictable in the ROC and P-R analyses. A more predictable model often exhibits higher AUC values. C1 % had the highest area under the receiver operating characteristic (ROC) curve (0.779), outperforming all other DOM indices. C1 % also had a higher area under the curve (AUC) for precision-recall (P-R) than the other DOM indices, reaching 0.816. The results indicate that C1 % is the best recognition indicator, which aligns with the findings of CRTA. The optimal value for C1 % on the ROC curve was 36.3 % based on Yonden's index, which is strongly related to the performance of the screening models. Similarly, the optimal value for C1 % in the P-R curve, which represents the ability to produce accurate predictions while considering recall and accuracy, was determined using the maximum F1-score. Using C1 % as the recognition indicator, 21 groundwater samples were checked to assess the predictability of the P-R curve, ROC curve and CRTA. Predictability using the CRTA method for CKDu groundwater was 75.0 %, which was more than the 50.0 % predicted by the ROC curve

but in line with the P-R curve. However, compared to the ROC curve's 88.9 % forecast and the P-R curve's 55.6 %, the CRTA approach's 66.7 % prediction for non-CKDu groundwater was lower. The geometric mean of predictability for both CKDu and non-CKDu groundwater samples was used for comparison to provide a comprehensive assessment of predictability. Predictability was best attained by the CRTA method (70.7 %), followed by the ROC curve (66.7 %) and P-R curve (64.5 %). This demonstrates that while screening groundwater for CKDu and non-CKDu, the CRTA method achieved better C1 % RT values than the ROC and P-R curves. Despite a receiver operating characteristic (ROC) curve area under the curve (AUC) for PC1 scores of less than 0.7, the curve (AUC) of the precision-recall (P-R) curve for PC1 scores was 0.700, suggesting that PC1 scores may also be utilised as trustworthy indications. A comparison between the ideal RT of the PC1 score obtained using the P-R curve and the RT of the PC1 score created using the CRTA approach showed that both methods had the same level of real predictability for non-CKDu groundwater. Nevertheless, the actual predictability was worsened using the P-R curve method. Therefore, the CRTA method seems to offer greater benefits than the P-R and ROC curves for distinguishing CKDu groundwater from non-CKDu groundwater.

This study focused on groundwater dissolved organic matter (DOM) features associated with chronic kidney disease (CKDu). DOM optical indices were used to examine the groundwater properties. In addition, an acceptable indication level was determined using probability-based guidelines to distinguish between CKDu-related groundwater and CKDu-related groundwater. Locating clean groundwater supplies in low-income neighbourhoods will be easier using this tool. Some researchers have suggested using reverse osmosis and nanofiltration to clean polluted water supplies and reduce the likelihood of CKDu. Be that as it may, these technologies are not yet extensively deployed in the real world and necessitate favorable economic conditions. To filter CKDu groundwater, it is practical and cost-effective to use C1 %, which CRTA has already accepted. To determine the appropriate fitting curve functions, it is necessary to compare the measured value with the reference value (RT) for each indication. The DPC or DPN can be calculated using the selected function. Screening water sources for CKDu using the CRTA method, which detects C1 % FDOM, increases the safety of drinking water and decreases the incidence of CKDu in regions where this disease is common.

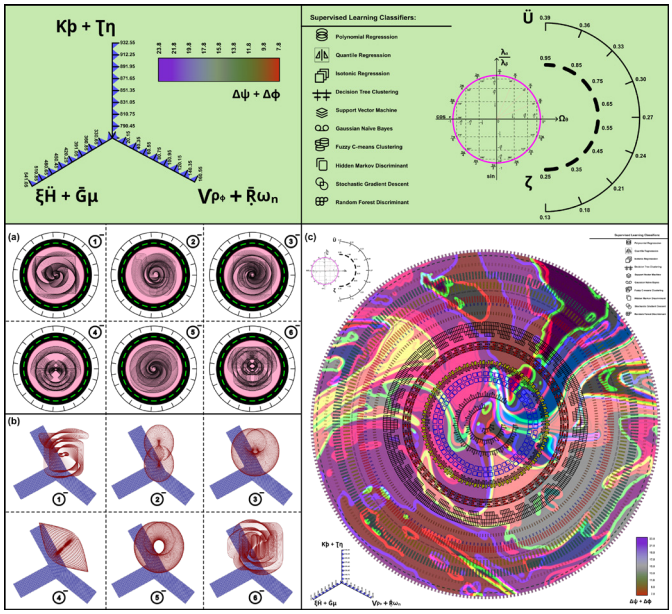

Figure 4 (a) Angular response-surface plots of ion mobility spectra (maximum peak intensities of 7 peaks changes in the extinction measured at 750 nm over an *L. platensis* cultivation) and preprocessed UV absorbance at 255 nm (absorbance per meter) nexus estimated via the 6-set hierarchical bands of machine-learning classifiers are used to illustrate infections with *L. monocytogenes*, including asymptomatic carriers, septicemia, encephalitis and abortions. (b) Stochastic spectral slope curve illustrating the lyser absorbance data in surface water as a function of wavelength (calculated using the abs_spectral_curve function in the AbspectroscoPY toolkit, based on lyser absorbance data at 290.5 nm) with plots for surface water (SW), rapid sand filtrate (RSF), and ultrafilter permeate (UF) across the rolling median smoothing function over a 60-minute window size applied to UV absorbance at 255 nm (absorbance per meter); (c) Dendrogram of stacked hierarchical bands of machine-learning classifiers illustrating the correlation between the water quality of respondents and the objectively assessed water quality (corrected for salinity; accounting for the emotional attachment, gender, and education of respondents) in the spread of *L. monocytogenes* between habitats, with a particular emphasis on food safety.

The C1 % retention time (RT) is sensitive to several environmental factors, including hydrology and geology. To evaluate and expand the application of CRTA, further research on dissolved organic matter in groundwater is required in other CKDu-affected locations. Notably, the CRTA method can be used for most FDOM indications, with the exception of the most sensitive ones. Previous investigations have demonstrated substantial differences in the amounts of inorganic components, such as Ca^{2+}, F^{-}, hardness and Si, between groundwater associated with CKDu and groundwater not associated with CKDu. Therefore, the CRTA approach is advantageous for detecting inorganic-sensitive indicators. The investigation determined that CRTA RT, when compared to DPC and DPN, had an 82.4 % chance of detecting Ca^{2+} and 64.2 % likelihood of detecting Si. Regrettably, their real-time analysis exhibited a poor probability of identifying hardness and fluoride ions. **Figures 1(c) - 1(e)** demonstrates a link between DPC

and the hypothesised F-undulated threshold, as well as increased F^- concentrations in the non-CKDu groundwater samples. Therefore, when comparing the DPC curve with the claimed F-threshold using CRTA, the DPN curve with respect to the believed F-threshold proved to be a more reliable approach for identifying water sources. In addition, the combination of FDOM indications and inorganic indicators enhances the accuracy of the recognition findings, enabling the identification of safe water sources using CRTA. Additional groundwater samples should be collected from a broader range of sites that accurately represent the various geological and hydrological conditions for further investigation. The yearly precipitation of Sri Lanka determines whether the country is humid, arid, or semiarid. The study area was situated in a semiarid zone, namely the CKDu-endemic Girandurukotte, Dehiattakandiya and Sewanagala areas (non-CKDu areas). Extremely low water tables and high rates of evaporation characterise the flat terrain of this region. Every year, the geosphere receives approximately 1,000 mm of rain, and the temperature remains between 29 and 33 °C. The southern monsoon (SW) occurs between May and October and the northern monsoon (NE) occurs between November and February. The westernmost part of Sri Lanka, adjacent to the Mahaweli River, was the study area. Precambrian granite and gneiss comprised the majority of the rocks found in this study. Minerals that contain fluorine, such as mica, hornblende and apatite, are abundant in these rocks. The fractures and joints in these rocks act as reservoirs for groundwater even though they are not very porous. Locals rely mostly on groundwater for drinking water. Drilling wells into unconsolidated river sediments or shallow weathered bedrock aquifers is the most common method for of accessing the groundwater. A few houses draw water from deep aquifers in bedrock using tube wells. Control over surface hydrological networks is exerted by man-made reservoir cascade systems, which are mostly utilized for irrigation. Our group obtained groundwater samples from wells used by typical CKDu patients using data supplied by the neighborhood hospital. Additionally, the methodology obtained a separate groundwater sample from wells used by households that did not have CKDu patients; this sample was referred to as non-CKDu groundwater. A total of 83 water samples were collected: 32 samples of non-CKDu groundwater, 8 samples of surface water from tanks and 43 samples of CKDu groundwater. A 0.7 µm quartz filter was used to filter the samples, and then, using high-quality pure hydrochloric acid, they were acidified until their pH was less than 2. To determine DOC, the materials were initially passed through a 0.45 µm membrane and then acidified with pure, high-quality phosphoric acid until their pH was less than 2. The fluorescence properties of dissolved organic matter (DOM) were evaluated using a fluorescence spectrometer (Fluomax-4, HORIBA JboinYvon, Japan). The emission wavelengths (Em) were configured to range from 300 to 550 nm at intervals of 2 nm. The excitation wavelength (Ex) was set between 250 and 400 nm at intervals of 4 nm. The slit width was set to 3 nm. The scanning signal had an integration period of 0.1 s. The illumination source was provided by a xenon bulb with a power output of 150 W. A spectrophotometre (UV1900, Shimadzu, Japan) was used to measure the UV-visible absorbance of the samples in the wavelength range of 200 - 600 nm. Dissolved organic carbon (DOC) was quantified using a TOC analyser (Aurora 1030w, OI, USA) with a detection limit of 0.01 mg L^{-1} and an analytical accuracy of ± 2.0 %. Analyses of anions and cations, aspect ratios and PARAFACFT were performed to calculate additional parameters such as the humification index (HIX), biological index (BIX), fluorescence index (FI), concentration of dissolved organic matter with an unsaturated structure (a_{254}), aromaticity index

($SUVA_{254}$) and spectral slope ($S_{275 - 295}$). The inner filter effect and background fluorescence signal of the ultrapure water were adjusted to remove Rayleigh and Raman scattering. To prepare for the PARAFAC application, the MATLAB graphical user interface's efc toolbox, FDOM correct toolbox and N-way toolbox were used. The model successfully passed both split-half and core consistency tests. A similarity score of over 0.95 was employed to match the PARAFAC components with those from previous investigations. To determine the relative abundance of PARAFAC components (C1, C2, C3 and C4 %), the ratio of the highest peak intensity (Fmax) of each component to the total Fmax of all the components was calculated. Examination of the threshold for detecting Chronic Kidney Disease of unknown aetiology (CKDu) is known as the CKDu detecting threshold evaluation (CRTA).

The assessment included inorganic compounds such as Ca^{2+} and F^{-}, as well as data from DOM (such as C4 %, C1 % and HIX) as indicators. During the evaluation, it was necessary to establish and compute the detection probabilities for CKDu (DPC) and non-CKDu (DPN) groundwater samples. The minimum recorded value was frequently used as the initial threshold for each indicator. Individuals were classified as high if their values exceeded this threshold, and low if their values fell below it. **Figure 1(f)** shows the presence of dissolved organic carbon (DPC) in the high value group. The following indicators contained DPC: C1 %, C2 %, C3 %, HIX, $SUVA_{254}$, a_{254}, DOC, PC1 score and inorganic markers. Eq. (2) was used to calculate the dissolved particulate nitrogen (DPN) for the low-value group. The data in **Figure 2(c)** were used to calculate the DPC in the low-value group and the DPN in the high-value group, together with the supplementary indicators C4 %, $S_{275 - 295}$, BIX and FI. Volume of groundwater samples for Crank-Dependent Chronic Kidney Disease (CKDu) within the specified group. IBM SPSS Statistics-22 was utilized to conduct several statistical analyses, including paired t-tests, 1-way analyses of variance (ANOVA), independent-sample t-tests and Spearman correlation. At a significance level of $p < 0.05$, t-tests and ANOVA were used to assess the disparities and variations among the groups. The analysis also involved Principal component analysis (PCA) considered the following variables: dissolved organic carbon (DOC), fluorescence index (FI), biological index (BIX), humification index (HIX), absorbance at 254 nm (a_{254}), specific ultraviolet absorbance at 254 nm ($SUVA_{254}$), spectral slope between 275 and 295 nm ($S_{275 - 295}$) and the relative abundance of PARAFAC components (C1, C2, C3 and C4 %). Origin 2021 uses the least-squares approach to fit the data. The corresponding DPC and DPN values were obtained by gradually increasing the presumed threshold of each indicator, usually in 100 steps according to a predetermined mathematical sequence, from the lowest to the highest. **Figures 1(d) - 1(f)** show the following plot of the indicator's assumed threshold versus DPC or DPN. The ability of the indicator to identify the CKDu groundwater was used to determine the detection probability of the curve junction. RT, or the recognition threshold, was calculated by taking the value at which the 2 curves met and assuming it to be the threshold. Recognition is easier at room temperature (RT) because of the increased likelihood of detection. The empirical Boltzmann equation was used to determine the DPN and DPC from 54 groundwater samples, accounting for 72 % of the total. To validate the fitting curves, the remaining samples (n = 21) were used in conjunction with the paired t-test and root mean square error (RMSE). In a paired t-test, a *p*-value larger than 0.05 indicates that the expected and observed outcomes are quite close. If the RMSE is small, the expected and measured results are more closely aligned.

Optimisation prospects and challenges

An alternative to conventional Convolutional Neural Networks (CNNs) for image identification is a novel Capsule Neural Network (CapsNet). CapsNet was implemented in 2017. Instead of scalar numbers, CapsNet uses vectors as the input and output, which is different from that of a CNN. Thus, CapsNet can learn image attributes, even when exposed to different viewing circumstances or distortions in the original image. The CapsNet architecture comprises individual capsules that contain clusters of neurones. In a capsule, each neuron's output is associated with a distinct surface quality of an object [71-78]. By decomposing an item into its constituent parts and reassembling them, the network strengthens CapsNet's ability to withstand environmental changes, distortions and situations when some parts of the object are missing or hidden [79-86]. The distortion detection capabilities of CapsNet, which do not require further training, may potentially expedite the evaluation of the water treatment performance [87]. The reason behind because it allows for the identification and comparison of fragmented or non-viable cells with viable cell populations [88]. Customised training utilising pictures of damaged and broken cells is necessary to train a convolutional neural network (CNN) for this task [89-94]. Instead of connecting individual neurones, the capsules were linked to achieve this [95]. For instance, although 2 capsules can form a single connection, 9 neurones can only form a connection with each other through as many as 92 connections [96]. Because training CapsNets online is not feasible with traditional CNNs owing to their high computing resource requirements, this area of study focuses on finding a solution [97]. By delving into AI picture recognition and machine learning, identifying the most important criteria to consider when selecting the best microscopic imaging method is of the utmost importance [98-103]. Despite their ease of use and speed, image-capture neural networks require large and diverse training sets to achieve a satisfactory class classification. Each cyanobacterial species may end up in a plethora of taxonomies owing to the aforementioned polymorphism and the examination of cell fragments during viability testing. Consequently, it is necessary to gather many photos of every class for each conceivable configuration. Therefore, the microscopy method must provide a rapid sample throughput. Information extracted from photographs. When more data are easily accessible, more accurate 'cell fingerprint' data are available, which increases the capacity of the neural network to correctly distinguish between classes. Increasing the spatial resolution, employing 3-dimensional imaging, or utilising an imaging technique that gathers additional biophysical cell properties can improve this information. When targeting a class with few physically differentiating features, such as bacteria, access to additional biophysical cell information is crucial. After considering the many microscopic imaging methods described in the literature in relation to the capabilities of currently available commercial equipment, quantitative phase imaging was determined to be the most feasible choice. Quick sample processing using QPI technology and flow cytometry was possible using the aforementioned settings. The improved object detection accuracy is a direct outcome of the increased data availability compared with pictures obtained using brightfield microscopy [104]. This section elaborates on the methods employed in this study to improve the design of the water quality monitoring system. Considering the training needs, advantages and disadvantages of artificial intelligence picture recognition and evaluation of various microscopic imaging technologies, quantitative phase imaging has become an obvious frontrunner. This section takes a close look at 3 different quantitative phase

imaging (QPI) methods that have been employed effectively in recent cell imaging literature, and compares and contrasts their benefits and drawbacks. After a thorough evaluation of the methods, Method 2 (QPI with Michelson interferometry) was selected as the most feasible choice. As shown in **Figures 2(b) - 2(d)**, a preliminary plan for a water quality monitoring station can be developed. The first uses a portable instrument that can quantify and evaluate phase imaging. Method 1, a small device detailed in the 'Portable QPIU' section, is being tested by a trained convolutional neural network to see if it may aid in cell identification. The output of a standard bright-field microscope can be linked to this device. As mentioned earlier, the trained CNN 'HoloConvNet' was able to attain identification accuracies of approximately 95 %, which proves that this technique is viable.

Method 1's main selling point is that it is a proof-of-concept that has the potential to be both easy and portable. Optical calibration considers only a handful of factors and the picture-gathering method is straightforward. This device is designed to be readily portable between sites because of its easy-to-follow instructions and the fact that it does not require disassembly between data collection instances. As mentioned earlier, a trained Convolutional Neural Network (CNN) was able to successfully identify bacteria using the data generated by Method 1. Bacteria grown using this method were identified with 96.3 % accuracy after adjusting the algorithm parameters. Without modifying the algorithm, an 85 % accuracy rate was achieved by collecting photographs using precise QPIU and CNN to correctly identify *L. monocytogenes*. The main drawback of Method 1 is the time and effort required to manually evaluate and prepare samples. As mentioned previously, materials must be carefully placed between 2 microscope slides in an imaging chamber. Manual focusing and immersion in oil are necessary to obtain high spatial resolution using a microscope. Method 2, an automated operation, can take 100 photos per second of cells in flow; in comparison, the manual sample preparation and analysis technique will severely slow down the sample processing rate. Labour expenses and training requirements are higher for manual sample preparation because they require technical expertise from the microscope operator. Using an imaging flow cell instead of an image chamber makes this method applicable to flow cytometry. Changing the experimental setup and buying new equipment were necessary for this integration. Although Method 1 offers a handheld device for capturing QPI pictures, it is not portable because the highest magnification is only achievable in a controlled laboratory setting using a brightfield microscope. This device is therefore easily transportable between labs but has very limited practical use in real-world settings, yielding high-resolution pictures that are on par with those of Method 2. Nevertheless, it is necessary to use an oil immersion technique to attain 100× objective magnification. This makes the imaging process more tedious because oil must be manually applied to each fresh slide holding a sample before imaging can begin. Biophysical cellular characteristics, such as radius and dry mass, can be investigated in detail using phase-shifting quantitative phase imaging with the help of a quantitative phase imaging unit (QPIU) that employs Michelson interferometry, particularly Method 2, as outlined in the 'Michelson interferometry' section. By mechanising the sample preparation and collection processes, integrated flow cytometry increased the rate of data collection. An important benefit of Method 2 is the high spatial resolution of the images obtained using this technique, which enables the acquisition of accurate data on biophysical cellular features. This level of accuracy does not require oil immersion, which eliminates the preparatory step and speeds up the image-processing

process compared to Method 1. The effectiveness of this method was demonstrated by capturing precise images of tumour cells with a diameter of 7 - 9 μm, which is similar to that of smaller species of cyanobacteria. Because the system is modular, pieces such as the microscope objective may be swapped out with those with a higher magnification if the cyanobacteria species being studied are too small to be captured well by this approach. Combining Method 2 with flow cytometry equipment allows for automated sample analysis, which is another key advantage. Data collection is accelerated compared to manual analysis methods by this integration, which greatly enhances the speed of sample processing. The most economical design was Method 2 according to the cost analysis of the 3 techniques presented in this section. The price is approximately the same as that of Method 1, but much cheaper than that of Method 3. Method 2 suffers from the drawbacks of interferometry design, despite its ability to produce a very accurate interference pattern. Interferometers must be fine-tuned to produce interference patterns with resolutions of tens of nanometres. There are more potential sites of failure in the design owing to the large number of components, which makes it more difficult to discover and fix prospective problems. Of the 3 methods considered, Method 2 is the least portable, suggesting that the design is complex. It is suggested that the interferometer should be dismantled and reassembled before being transported by an automobile. The components of the design are delicate and require special handling to maintain their precision and calibration in the event of an accident. Having somewhat constant conditions, such as those provided by weatherproof sheds, is crucial, although a laboratory is not strictly necessary. The third method, created by 4Deep, is the first example of commercially accessible QPI technology. Regarding the on-site analysis, the portability and small size of the 4Deep S7 submersible microscope. Even in the frigid Arctic, it has been used to effectively capture photographs of small aquatic creatures. Method 3, a small device with more flexibility than Methods 1 - 2, is mostly used for on-site diagnostics. Lake conditions are unlikely to present any major operating challenges for this technology given that it was mainly developed for use in marine settings with extreme temperature and pressure changes. The portable nature of the gadget makes it ideal for use in various environments. The steps in the image processing approach shown in **Figure 3(a)** can be eliminated using technique 3, which has intrinsic quantitative phase reconstruction capabilities. Provided by the manufacturer, pre-calibrated and optimised optics streamline the equipment setup procedure, and the design of the device as a passive probe is an inherent drawback of Method 3, as it makes the quantitative analysis more complicated. Although there is a device with the same design, there is currently no commercially available option that can accurately count cells using a small-scale flow cytometer. The high price of Method 3 is a major turnoff for consumers compared with Methods 1 - 2. The improved picture information it provides is difficult to justify when compared to standard fluorescence monitoring, especially considering that an optimal cyanobacteria monitoring system would require the use of multiple devices placed at various points within a body of water [77,93]. One drawback of Method 3 is that it produces lower-quality pictures in terms of spatial resolution than Methods 1 and 2. The use of Method 3 to evaluate water quality and small cyanobacteria species is restricted because it cannot photograph specimens smaller than 20 μm. The 3 QPI methods discussed in this section are compared in **Figure 3(b)**, according to their performance indicators. Based on the criteria mentioned earlier, Method 2 appeared to be the most promising opportunity for process improvement. The lack of mobility in this technology is probably not a

major drawback if the gadget is placed in a specific monitoring area. Keeping this in mind, this study overcomes some of the drawbacks of this approach and discusses ways to fix them. Method 2 outperformed, or was on par with, the methods listed in **Table S4**, except for its limited mobility.

A possible implementation of Method 2 (QPIU with Michelson interferometry) at a monitoring station for continuous observations is presented in **Table S2(a)**. The QPIU is housed in a tiny, weatherproof, shed-like building located either near a body of water that is under surveillance or inside a water treatment facility. The processing centre is linked to the processing buoys with tubing, which are positioned strategically around the body of the water in certain places. The sample solutions were transported using a small pump attached to the buoy. To maintain accurate cell counts and taxonomic identification, it is necessary to flush or replace the tube regularly to avoid the accumulation of cyanobacteria or algal biomass. Submerging tubing in rivers with high motor activity may make it less likely that it will be damaged or interrupted. OC-300's capability to process 12 simultaneous streams from 12 different sample-collecting buoys was re-studied in Section 3.5. A tube that went to a certain depth in water was attached to the sample buoys. A filter mesh was placed at the terminus of the tube to admit cyanobacteria while excluding unwanted materials, such as plant debris and tiny animals. Clogs in this mesh may be prevented with regular maintenance, and manual repositioning of the sample buoys is necessary during the initial installations if monitoring at many sites is necessary. A GPS-connected buoy that runs on renewable energy and has a small electric motor may be a part of future innovation. This makes it possible to move the buoy remotely to different locations of interest. The second method involves processing water samples by photographing any organisms or debris that passes through the imaging flow cell at a rate of 100 images/s. OC-300 included Lugol's solution during the sample preparation phase to enhance the performance of flow cytometry. Thus, the cell adhesion of the sheath fluid is significantly reduced, and quantitative phase images can be obtained using a Fourier reconstruction approach. To further enhance the overall performance of the network, the reconstructed QPI pictures were then fed into a succession of post-processing neural networks, as explained in the previous 2 sections of this article. The deblurring method, image segmentation strategy, and unfocused classifier are all components of the network. The processing speed of a neural network optimised for cyanobacteria detection in QPI pictures was approximately 1 cell/ms. Using the accurate sample volume from the OC-300, it provides a quantitative evaluation of the taxonomic content of the sample and calculates the cell density per millilitre. A neural network can be programmed to measure the extent to which the treatment process kills cells in treatment plants and determines cell viability. The uncertainty interval for the number of cells and sample composition were then calculated using the number of out-of-focus photos and the number of unidentifiable objects in the sample. Cells that cannot be identified are recognised and tagged by the network, so that an operator can check and identify them. Thus, additional training can be fed into the neural network. The object detection capabilities of the MATLAB-built neural network are initially low. It is advisable to use a dedicated computer for post-processing and recognition activities to allow the continuous processing of samples from many collecting buoys. After data compression, it can be used to build more AI training sets, evaluate water treatment efficacy, and provide a hand to a risk assessment group for tailoring assessments of water quality hazards according to cyanobacteria taxonomy and cell counts.

Conclusions

Based on a systematic exploration of convolutional neural networks (CNN), this study evaluates the efficacy of small imaging sensors in monitoring the real-time presence of pathogenic cyanotoxins and other hazardous aqueous contaminants in peri-urban ecosystems. The methodology first examines 3 quantitative phase imaging (QPI) techniques: A portable QPI unit coupled with a CNN, QPI with Michelson interferometry, and a commercially available submersible microscope. The portable QPI unit (Method 1) is found to achieve a high accuracy rate of 96.3 % for bacterial identification when paired with a trained CNN. However, manual sample preparation and dependence on bright-field microscopy restricts its portability and throughput. Method 2, which uses Michelson interferometry, provides high spatial resolution and allows for the study of biophysical cellular characteristics. Its integration with flow cytometry allows for automated sample analysis, which accelerates data collection. Despite its cost-effectiveness and accuracy, Method 2 suffers from complexity of its interferometer design and limited portability. Method 3, a commercially available submersible microscope, offers the highest portability and flexibility, making it suitable for on-site diagnostics in various environmental conditions. Considering the strengths and limitations of each method, Method 2 (QPI with Michelson interferometry) can be selected as the most feasible option for developing a water quality monitoring station. As such, by incorporating Method 2 in a machine learning based CNN framework, this study subsequently examines the link between the occurrence of hazardous algal blooms (HABs) and faecal indicator bacteria (FIB) in waterways and aquifers of certain semi-arid regions in Sri Lanka, Sweden and New York (United States). Utilising the widely used AbspectroscoPY technique to process the spectrophotometric data of the obtained samples, the formulation reveals strong positive correlations between FIB coliforms and nutrient loads, particularly between nitrate and phosphate loads. Furthermore, a significant association between the incidence of chronic kidney disease of unknown cause (CKDu) among the residents of the studied regions confirms this hypothesis, thereby reinforcing the reliability of the computational methodology. These findings emphasise the importance of considering the geographical and land use characteristics of urban areas in strategies aimed at reducing water-borne health risks. Consequently, this article highlights the potential of QPI techniques for enhancing cell imaging applications and provides valuable insights for researchers and practitioners in the field.

References

[1] A Hojjati-Najafabadi, M Mansoorianfar, T Liang, K Shahin and H Karimi-Maleh. A review on magnetic sensors for monitoring of hazardous pollutants in water resources. *Sci. Total Environ.* 2022; **824**, 153844.

[2] C Pettinari, R Pettinari, CD Nicola, A Tombesi, S Scuri and F Marchetti. Antimicrobial MOFs. *Coord. Chem. Rev.* 2021; **446**, 214121.

[3] R Huang, C Ma, J Ma, X Huangfu and Q He. Machine learning in natural and engineered water systems. *Water Res.* 2021; **205**, 117666.

[4] N Tripathi and MK Goshisht. Recent advances and mechanistic insights into antibacterial activity, antibiofilm activity, and cytotoxicity of silver nanoparticles. *ACS Appl. Biol. Mater.* 2022; **5**, 1391-463.

[5] B Mubeen, AN Ansar, R Rasool, I Ullah, SS Imam, S Alshehri, MM Ghoneim, SI Alzarea, MS Nadeem and I Kazmi. Nanotechnology as a novel approach in combating microbes providing an alternative to antibiotics. *Antibiotics* 2021; **10**, 1473.

[6] C Zhang, L Huang, H Pu and DW Sun. Magnetic surface-enhanced Raman scattering (MagSERS) biosensors for microbial food safety: Fundamentals and applications. *Trends Food Sci. Tech.* 2021; **113**, 366-81.

[7] KS Lee, Z Landry, FC Pereira, M Wagner, D Berry, WE Huang, GT Taylor, J Kneipp, J Popp, M Zhang, JX Cheng and R Stocker. Raman microspectroscopy for microbiology. *Nat. Rev. Methods Primers* 2021; **1**, 80.

[8] C Ongpipattanakul, EK Desormeaux, A DiCaprio, WA van der Donk, DA Mitchell and SK Nair. Mechanism of action of ribosomally synthesized and post-translationally modified peptides. *Chem. Rev.* 2022; **122**, 14722-814.

[9] Y Su, JT Yrastorza, M Matis, J Cusick, S Zhao, G Wang and J Xie. Biofilms: Formation, research models, potential targets, and methods for prevention and treatment. *Adv. Sci.* 2022; **9**, 2203291.

[10] E Molina-Grima, F García-Camacho, FG Acién-Fernández, A Sánchez-Mirón, M Plouviez, C Shene and Y Chisti. Pathogens and predators impacting commercial production of microalgae and cyanobacteria. *Biotechnol. Adv.* 2021; **55**, 107884.

[11] M Meena, A Zehra, P Swapnil, Harish, A Marwal, G Yadav and P Sonigra. Endophytic nanotechnology: An approach to study scope and potential applications. *Front. Chem.* 2021; **9**, 613343.

[12] P Ma, C Li, MM Rahaman, Y Yao, J Zhang, S Zou, X Zhao and M Grzegorzek. A state-of-the-art survey of object detection techniques in microorganism image analysis: From classical methods to deep learning approaches. *Artif. Intell. Rev*. 2023; **56**, 1627-98.

[13] M Awashra and P Młynarz. The toxicity of nanoparticles and their interaction with cells: An *in vitro* metabolomic perspective. *Nanoscale Adv.* 2023; **5**, 2674-723.

[14] SR McCuskey, J Chatsirisupachai, E Zeglio, O Parlak, P Panoy, A Herland, GC Bazan and TQ Nguyen. Current progress of interfacing organic semiconducting materials with bacteria. *Chem. Rev.* 2021; **122**, 4791-825.

[15] KH Cho, J Wolny, JA Kase, T Unno and Y Pachepsky. Interactions of E. coli with algae and aquatic vegetation in natural waters. *Water Res.* 2021; **209**, 117952.

[16] X Wei, Z Huang, L Jiang, Y Li, X Zhang, Y Leng and C Jiang. Charting the landscape of the environmental exposome. *iMeta* 2022; **1**, e50.

[17] Y Wan, C Zong, X Li, A Wang, Y Li, T Yang, Q Bao, M Dubow, M Yang, LA Rodrigo and C Mao. New insights for biosensing: lessons from microbial defense systems. *Chem. Rev.* 2022; **122**, 8126-80.

[18] Y Zheng, L Bonfili, T Wei and AM Eleuteri. Understanding the gut-brain axis and its therapeutic implications for neurodegenerative disorders. *Nutrients* 2023; **15**, 4631.

[19] S Chatterjee and M More. Cyanobacterial harmful algal bloom toxin microcystin and increased vibrio occurrence as climate-change-induced biological co-stressors: Exposure and disease outcomes via their interaction with gut-liver-brain axis. *Toxins* 2023; **15**, 289.

[20] QY Zhang, F Ke, L Gui and Z Zhao. Recent insights into aquatic viruses: Emerging and reemerging pathogens, molecular features, biological effects, and novel investigative approaches. *Water Biol. Secur.* 2022; **1**, 100062.

[21] SJ Lim, M Son, SJ Ki, SI Suh and J Chung. Opportunities and challenges of machine learning in bioprocesses: Categorization from different perspectives and future direction. *Bioresour. Tech.* 2022; **370**, 128518.

[22] P Chakraborty and KK Krishnani. Emerging bioanalytical sensors for rapid and close-to-real-time detection of priority abiotic and biotic stressors in aquaculture and culture-based fisheries. *Sci. Total Environ.* 2022; **838**, 156128.

[23] K Ayhan, S Coşansu, E Orhan-Yanıkan and G Gülseren. Advance methods for the qualitative and quantitative determination of microorganisms. *Microchem. J.* 2021; **166**, 106188.

[24] CC Lim, J Yoon, K Reynolds, LB Gerald, AP Ault, S Heo and ML Bell. Harmful algal bloom aerosols and human health. *eBioMedicine* 2023; **93**, 104604.

[25] F Saleem, JL Jiang, R Atrache, A Paschos, TA Edge and HE Schellhorn. Cyanobacterial algal bloom monitoring: Molecular methods and technologies for freshwater ecosystems. *Microorganisms* 2023; **11**, 851.

[26] NJ Rowan. Current decontamination challenges and potentially complementary solutions to safeguard the vulnerable seafood industry from recalcitrant human norovirus in live shellfish: Quo Vadis? *Sci. Total Environ.* 2023; **874**, 162380.

[27] LS Rösner, F Walter, C Ude, GT John and S Beutel. Sensors and techniques for on-line determination of cell viability in bioprocess monitoring. *Bioengineering* 2022; **9**, 762.

[28] S Kumar, T Arif, AS Alotaibi, MB Malik and J Manhas. Advances towards automatic detection and classification of parasites microscopic images using deep convolutional neural network: Methods, models and research directions. *Arch. Comput. Methods Eng.* 2023, **30**, 2013-39.

[29] BE Igere, AI Okoh and UU Nwodo. Non-serogroup O1/O139 agglutinable *Vibrio cholerae*: A phylogenetically and genealogically neglected yet emerging potential pathogen of clinical relevance. *Arch. Microbiol.* 2022; **204**, 323.

[30] GE Quintanilla-Villanueva, J Maldonado, D Luna-Moreno, JM Rodríguez-Delgado, JF Villarreal-Chiu and MM Rodríguez-Delgado. Progress in plasmonic sensors as monitoring tools for aquaculture quality control. *Biosensors* 2023; **13**, 90.

[31] D Çimen, N Bereli and A Denizli. Advanced plasmonic nanosensors for monitoring of environmental pollutants. *Curr. Anal. Chem.* 2023; **19**, 2-17.

[32] A Tauseef, F Hisam, T Hussain, A Caruso, K Hussain, A Châtel and B Chénais. Nanomicrobiology: Emerging trends in microbial synthesis of nanomaterials and their applications. *J. Cluster Sci.* 2022; **34**, 639-64.

[33] SA Okaiyeto, PP Sutar, C Chen, JB Ni, J Wang, AS Mujumdar, JS Zhang, MQ Xu, XM Fang, C Zhang and HW Xiao. Antibiotic resistant bacteria in food systems: Current status, resistance mechanisms, and mitigation strategies. *Agr. Commun.* 2024; **2**, 100027.

[34] S Kumari, R Taliyan and SK Dubey. Comprehensive review on potential signaling pathways involving the transfer of α-synuclein from the gut to the brain that leads to Parkinson's disease. *ACS Chem. Neurosci.* 2023; **14**, 590-602.

[35] EW Morgan, GH Perdew and AD Patterson. Multi-omics strategies for investigating the microbiome in toxicology research. *Toxicol. Sci.* 2022; **187**, 189-213.

[36] Y Hou, R Chen, Z Wang, R Lu, Y Wang, S Ren, S Li, Y Wang, T Han and S Yang. Bio-barcode assay: A useful technology for ultrasensitive and logic-controlled specific detection in food safety: A review. *Anal. Chim. Acta* 2023; **1267**, 341351.

[37] S Bej, S Swain, AK Bishoyi, CP Mandhata, CR Sahoo and RN Padhy. Wastewater-associated infections: A public health concern. *Water Air Soil Pollut.* 2023; **234**, 444.

[38] KB Shivaram, P Bhatt, B Applegate and H Simsek. Bacteriophage-based biocontrol technology to enhance the efficiency of wastewater treatment and reduce targeted bacterial biofilms. *Sci. Total Environ.* 2022; **862**, 160723.

[39] A Pras and H Mamane. Nowcasting of fecal coliform presence using an artificial neural network. *Environ. Pollut.* 2023; **326**, 121484.

[40] JOG Elechi, R Sirianni, FL Conforti, E Cione and M Pellegrino. Food system transformation and gut microbiota transition: Evidence on advancing obesity, cardiovascular diseases, and cancers - a narrative review. *Foods* 2023; **12**, 2286.

[41] A Igwaran, AJ Kayode, KM Moloantoa, ZP Khetsha and JO Unuofin. Cyanobacteria harmful algae blooms: Causes, impacts, and risk management. *Water Air Soil Pollut*. 2024; **235**, 71.

[42] M Patriarca, N Barlow, A Cross, S Hill, A Robson and J Tyson. Atomic spectrometry update: Review of advances in the analysis of clinical and biological materials, foods and beverages. *J. Anal. At. Spectrom.* 2024; **39**, 624-98.

[43] M Pöttker, K Kiehl, T Jarmer and D Trautz. Convolutional neural network maps plant communities in semi-natural grasslands using multispectral unmanned aerial vehicle imagery. *Remote Sens.* 2023; **15**, 1945.

[44] V Ahuja, A Singh, D Paul, D Dasgupta, P Urajová, S Ghosh, R Singh, G Sahoo, D Ewe and K Saurav. Recent advances in the detection of food toxins using mass spectrometry. *Chem. Res. Toxicol.* 2023; **36**, 1834-63.

[45] R Trivedi, TK Upadhyay, F Khan, P Pandey, RS Kaushal, M Sonkar, D Kumar, M Saeed, MU Khandaker and TB Emran. Innovative strategies to manage polluted aquatic ecosystem and agri-food waste for circular economy. *Environ. Nanotechnol. Monit. Manage.* 2024; **21**, 100928.

[46] D Sharma, G Teli, K Gupta, G Bansal, GD Gupta and PA Chawla. Nano-biosensors from agriculture to nextgen diagnostic tools. *Curr. Nanomater.* 2022; **7**, 110-38.

[47] H Raki, Y Aalaila, A Taktour and DH Peluffo-Ordóñez. Combining AI tools with non-destructive technologies for crop-based food safety: A comprehensive review. *Foods* 2023; **13**, 11.

[48] B İnce, İ Uludağ, B Demirbakan, C Özyurt, B Özcan and MK Sezgintürk. Lateral flow assays for food analyses: Food contaminants, allergens, toxins, and beyond. *TrAC Trends Anal. Chem.* 2023; **169**, 117418.

[49] H Ruan, Y Huang, B Yue, Y Zhang, J Lv, K Miao, D Zhang, J Luo and M Yang. Insights into the intestinal toxicity of foodborne mycotoxins through gut microbiota: A comprehensive review. *Compr. Rev. Food Sci. Food Saf.* 2023; **22**, 4758-85.

[50] F Mermans, V Mattelin, R van den Eeckhoudt, C García-Timermans, J van Landuyt, Y Guo, I Taurino, F Tavernier, M Kraft and H Khan. Opportunities in optical and electrical single-cell technologies to study microbial ecosystems. *Front. Microbiol.* 2023; **14**, 1233705.

[51] DR Ravindran, S Kannan and M Marudhamuthu. Fabrication and characterisation of human gut microbiome derived exopolysaccharide mediated silver nanoparticles - An *in-vitro* and *in-vivo* approach of Bio-Pm-AgNPs targeting *Vibrio cholerae*. *Int. J. Biol. Macromol.* 2024; **256**, 128406.

[52] SR Law, F Mathes, AM Paten, PA Alexandre, R Regmi, C Reid, A Safarchi, S Shaktivesh, Y Wang and A Wilson. Life at the borderlands: Microbiomes of interfaces critical to One Health. *FEMS Microbiol. Rev.* 2024; **48**, fuae008.

[53] G Zammit, MG Zammit and KG Buttigieg. Emerging technologies for the discovery of novel diversity in cyanobacteria and algae and the elucidation of their valuable metabolites. *Diversity* 2023; **15**, 1142.

[54] G Sharma and P Chadha. Evaluation of haematological, genotoxic, cytotoxic and ATR-FTIR alterations in blood cells of fish *Channa punctatus* after acute exposure of aniline. *Sci. Rep.* 2023; **13**, 20757.

[55] A Rebelo, B Duarte, AR Freitas, L Peixe, P Antunes and C Novais. Exploring peracetic acid and acidic ph tolerance of antibiotic-resistant non-typhoidal salmonella and enterococcus faecium from diverse epidemiological and genetic backgrounds. *Microorganisms* 2023; **11**, 2330.

[56] JL Hernández-Martínez, JA Perera-Burgos, G Acosta-González, J Alvarado-Flores, Y Li and RM Leal-Bautista. Assessment of physicochemical parameters by remote sensing of Bacalar Lagoon, Yucatán Peninsula, Mexico. *Water* 2023; **16**, 159.

[57] GF Neuhaus, AT Aron, EW Isemonger, D Petras, SC Waterworth, LS Madonsela, EC Gentry, X Siwe Noundou, JCJ Kalinski and A Polyzois. Environmental metabolomics characterization of modern stromatolites and annotation of ibhayipeptolides. *Plos One* 2024; **19**, e0303273.

[58] M Manzano, A Cossettini, D Pinamonti, M Maifreni and M Marino. Development of an aptasensor for rapid and specific detection of *Escherichia coli* in water and milk samples. *Eurobiotech J.* 2023; **7**, 52-3.

[59] M Maifreni, M Marino, A Cossettini, D Pinamonti, A Lena and M Manzano. Quartz Crystal Microbalance (QCM): A biosensor to study the microbial biofilm formation in real-time. *Eurobiotech J.* 2023; **7**, 58-9.

[60] ML Liu, XM Liang, MY Jin, HW Huang, L Luo, H Wang, X Shen and ZL Xu. Food-borne biotoxin neutralization *in vivo* by nanobodies: Current status and prospects. *J. Agr. Food Chem.* 2024; **72**, 10753-71.

[61] KC Fickas, RE O'Shea, N Pahlevan, B Smith, SL Bartlett and JL Wolny. Leveraging multimission satellite data for spatiotemporally coherent cyanoHAB monitoring. *Front. Remote Sens.* 2023; **4**, 1157609.

[62] OJ Bassey, JR Gumbo, M Mujuru, A Adeyemi and F Dondofema. Comparative analysis of cyanotoxins in fishponds in Nigeria and South Africa. *Microbiol. Res.* 2024; **15**, 447-56.

[63] A Raina, S Kaul and MK Dhar. Sniffing out adulteration in saffron: Detection methods and health risks. *Food Contr.* 2024; **155**, 110042.

[64] PR Hill, A Kumar, M Temimi and DR Bull. HABNet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*. 2020; **13**, 3229-39.

[65] MI Vawda, R Lottering, O Mutanga, K Peerbhay and M Sibanda. Comparing the utility of Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) on Sentinel-2 MSI to estimate dry season aboveground grass biomass. *Sustainability* 2024; **16**, 1051.

[66] D Panda, BP Dash, S Manickam and G Boczkaj. Recent advancements in LC-MS based analysis of biotoxins: Present and future challenges. *Mass Spectrom. Rev.* 2022; **41**, 766-803.

[67] M Laad and B Ghule. Removal of toxic contaminants from drinking water using biosensors: A systematic review. *Groundwater Sustain. Dev.* 2023; **20**, 100888.

[68] K Karila, RA Oliveira, J Ek, J Kaivosoja, N Koivumäki, P Korhonen, O Niemeläinen, L Nyholm, R Näsi, I Pölönen and E Honkavaara. Estimating grass sward quality and quantity parameters using drone remote sensing with deep neural networks. *Remote Sens.* 2022; **14**, 2692.

[69] L Díez-Quijada, AI Prieto, R Guzmán-Guillén, A Jos and AM Cameán. Occurrence and toxicity of microcystin congeners other than MC-LR and MC-RR: A review. *Food Chem. Toxicol.* 2019; **125**, 106-32.

[70] L Chen, JP Giesy, O Adamovsky, Z Svirčev, J Meriluoto, GA Codd, B Mijovic, T Shi, X Tuo, SC Li, BZ Pan, J Chen and P Xie. Challenges of using blooms of *Microcystis* spp. in animal feeds: A comprehensive review of nutritional, toxicological and microbial health evaluation. *Sci. Total Environ.* 2021; **764**, 142319.

[71] MF Abdallah, JM Recote, CV Camp, WHRV Hassel, L Pedroni, L Dellafiora, J Masquelier and A Rajkovic. Potential (co-)contamination of dairy milk with AFM1 and MC-LR and their synergistic interaction in inducing mitochondrial dysfunction in HepG2 cells. *Food Chem. Toxicol.* 2024; **192**, 114907.

[72] M Yin, R Ma, H Luo, J Li, Q Zhao and M Zhang. Non-contact sensing technology enables precision livestock farming in smart farms. *Comput. Electron. Agr.* 2023; **212**, 108171.

[73] S Yasmin and M Billah. Digital image processing applications in agriculture with a machine learning approach. *Agr. Sci. Tech.* 2023; **15**, 12-22.

[74] M Torky and AE Hassanein. Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges. *Comput. Electron. Agr.* 2020; **178**, 105476.

[75] C Stumpe, J Leukel and T Zimpel. Prediction of pasture yield using machine learning-based optical sensing: A systematic review. *Precis. Agr.* 2023, **25**; 430-59.

[76] A Siddique, K Cook, Y Holt, SS Panda, AK Mahapatra, ER Morgan, JA van Wyk and TH Terrill. From plants to pixels: The role of artificial intelligence in identifying sericea lespedeza in field-based studies. *Agronomy* 2024; **14**, 992.

[77] T Shitanaka, H Fujioka, M Khan, M Kaur, ZY Du and SK Khanal. Recent advances in microalgal production, harvesting, prediction, optimization, and control strategies. *Bioresour. Tech.* 2023; **391**, 129924.

[78] L Shangru, Z Chengrui, W Ruixue, S Jiamei, X Hangshu, Z Yonggen and S Yukun. Establishment of a feed intake prediction model based on eating time, ruminating time and dietary composition. *Comput. Electron. Agr.* 2022; **202**, 107296.

[79] N Sánchez, J Plaza, M Criado, R Pérez-Sánchez, MÁ Gómez-Sánchez, MR Morales-Corts and C Palacios. The second derivative of the NDVI time series as an estimator of fresh biomass: A case study of eight forage associations monitored via UAS. *Drones* 2023; **7**, 347.

[80] S Ruan, R Di, Y Zhang, T Yan, H Cang, F Tan, M Zhang, N Wu, L Guo, P Gao and W Xu. Metric-based meta-learning combined with hyperspectral imaging for rapid detection of adulteration in domain-shifted camel milk powder. *LWT* 2024; **206**, 116537.

[81] R Roy, S Marakkar, MP Vayalil, A Shahanaz, AP Anil, S Kunnathpeedikayil, I Rawal, K Shetty, Z Shameer, S Sathees, AP Prasannakumar, OK Mathew, L Subramanian, K Shameer and KK Yadav. Drug-food interactions in the era of molecular big data, machine intelligence, and personalized health. *Recent Adv. Food Nutr. Agr.* 2022; **13**, 27-50.

[82] J Rooney, E Rivera-de-Torre, R Li, K Mclean, DRG Price, AJ Nisbet, AH Laustsen, TP Jenkins, A Hofmann, S Bakshi, A Zarkan and C Cantacessi. Structural and functional analyses of nematode-derived antimicrobial peptides support the occurrence of direct mechanisms of worm-microbiota interactions. *Comput. Struct. Biotechnol. J.* 2024; **23**, 1522-33.

[83] SSQ Rodrigues, LG Dias and A Teixeira. Emerging methods for the evaluation of sensory quality of food: Technology at service. *Curr. Food Sci. Tech. Rep.* 2024; **2**, 77-90.

[84] X Ren, H Tian, K Zhao, D Li, Z Xiao, Y Yu and F Liu. Research on pH value detection method during maize silage secondary fermentation based on computer vision. *Agriculture* 2022; **12**, 1623.

[85] N Pillai, M Ramkumar and B Nanduri. Artificial intelligence models for zoonotic pathogens: A survey. *Microorganisms* 2022; **10**, 1911.

[86] P Mohseni and A Ghorbani. Exploring the synergy of artificial intelligence in microbiology: Advancements, challenges, and future prospects. *Comput. Struct. Biotechnol. Rep.* 2024; **1**, 100005.

[87] M Meenu, C Kurade, BC Neelapu, S Kalra, HS Ramaswamy and Y Yu. A concise review on food quality assessment using digital image processing. *Trends Food Sci. Tech.* 2021; **118**, 106-24.

[88] LJ Marcos-Zambrano, K Karaduzovic-Hadziabdic, TL Turukalo, P Przymus, V Trajkovik, O Aasmets, M Berland, A Gruca, J Hasic, K Hron, T Klammsteiner, M Kolev, L Lahti, MB Lopes, V

Moreno, I Naskinova, E Org, I Paciência, G Papoutsoglou, …, J Truu. Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Front. Microbiol.* 2021; **12**, 634511.

[89] C Maraveas, D Konar, DK Michopoulos, KG Arvanitis and KP Peppas. Harnessing quantum computing for smart agriculture: Empowering sustainable crop management and yield optimization. *Comput. Electron. Agr.* 2024; **218**, 108680.

[90] Y Li, R Zheng, Y Wu, K Chu, Q Xu, M Sun and ZJ Smith. A low-cost, automated parasite diagnostic system via a portable, robotic microscope and deep learning. *J. Biophotonics* 2019; **12**, e201800410.

[91] RS Kent, EM Briggs, BL Colon, C Alvarez, SS Pereira and MD Niz. Paving the way: Contributions of big data to apicomplexan and Kinetoplastid research. *Front. Cell. Infect. Microbiol.* 2022; **12**, 900878.

[92] G Kapalaga, FN Kivunike, S Kerfua, D Jjingo, S Biryomumaisho, J Rutaisire, P Ssajjakambwe, S Mugerwa and Y Kiwala. A unified foot and mouth disease dataset for Uganda: Evaluating machine learning predictive performance degradation under varying distributions. *Front. Artif. Intell.* 2024; **7**, 1446368.

[93] C Ippoliti, L Bonicelli, MD Ascentis, S Tora, AD Lorenzo, SG D'Alessio, A Porrello, A Bonanni, D Cioci, M Goffredo, S Calderara and A Conte. Spotting *Culex pipiens* from satellite: Modeling habitat suitability in central Italy using Sentinel-2 and deep learning techniques. *Front. Vet. Sci.* 2024; **11**, 1383320.

[94] A Herlin, E Brunberg, J Hultgren, N Högberg, A Rydberg and A Skarin. Animal welfare implications of digital tools for monitoring and management of cattle and sheep on pasture. *Animals* 2021; **11**, 829.

[95] D Handa and JM Peschel. A review of monitoring techniques for livestock respiration and sounds. *Front. Anim. Sci.* 2022; **3**, 904834.

[96] S Ghaffarian, MVD Voort, J Valente, B Tekinerdogan and YD Mey. Machine learning-based farm risk management: A systematic mapping review. *Comput. Electron. Agr.* 2022; **192**, 106631.

[97] E Elbasi, N Mostafa, Z AlArnaout, AI Zreikat, E Cina, G Varghese, A Shdefat, AE Topcu, W Abdelbaki, S Mathew and C Zaki. Artificial intelligence technology in the agricultural sector: A systematic literature review. *IEEE Access* 2022; **11**, 171-202.

[98] W Ekloh, A Asafu-Adjaye, CNL Tawiah-Mensah, SM Ayivi-Tosuh, NKA Quartey, AF Aiduenu, BK Gayi, JAM Koudonu, LA Basing, JAA Yamoah, AK Dofuor and JHN Osei. A comprehensive exploration of schistosomiasis: Global impact, molecular characterization, drug discovery, artificial intelligence and future prospects. *Heliyon* 2024; **10**, e33070.

[99] K Džermeikaitė, D Bačėninaitė and R Antanaitis. Innovations in cattle farming: Application of innovative technologies and sensors in the diagnosis of diseases. *Animals* 2023; **13**, 780.

[100] FRD Santos, CMM Filho, RFLD Cerqueira, RY Yada, PED Augusto, BRD Castro Leite and AAL Tribst. Strategies to extend the shelf life of sheep and goat cheese whey under refrigeration: Nisin, bioprotective culture, and acidification. *Food Biosci.* 2024; **57**, 103495.

[101] CMD Korne, LV Lieshout, FWBV Leeuwen and M Roestenberg. Imaging as a (pre)clinical tool in parasitology. *Trends Parasitol.* 2023; **39**, 212-26.

[102] E Clear, RA Grant, M Carroll and CA Brassey. A review and case study of 3D imaging modalities for female amniote reproductive anatomy. *Integr. Comp. Biol.* 2022; **62**, 542-58.

[103] T Chen, H Zheng, J Chen, Z Zhang and X Huang. Novel intelligent grazing strategy based on remote sensing, herd perception and UAVs monitoring. *Comput. Electron. Agr.* 2024; **219**, 108807.

[104] JO Chelotti, LS Martinez-Rau, M Ferrero, LD Vignolo, JR Galli, AM Planisich, HL Rufiner and LL Giovanini. Livestock feeding behaviour: A review on automated systems for ruminant monitoring. *Biosyst. Eng.* 2024; **246**, 150-77.

[105] S Koley. Augmenting efficacy of global climate model forecasts: Machine learning appraisal of remote sensing data. *Int. J. Eng. Trends Tech.* 2024; **72**, 442-502.

[106] S Koley. Role of fluid dynamics in infectious disease transmission: Insights from COVID-19 and other pathogens. *Trends Sci.* 2024; **21**, 8287.

[107] S Koley. Electrochemistry of Phase-Change Materials in Thermal Energy Storage Systems: A Critical Review of Green Transitions in Built Environments. *Trends Sci*. 2024; **21**, 8538.

Appendix A

List of notations defined in the paper (as per computational models [105-107])

ϕ	Laplace limit
ψ	Time series of observation data
ϕ	Time series of model data
ρ_ϕ	Eigen function for Fourier transform of convolutional datasets
ν_ϕ	Eigen function for Fourier transform of Bayesian datasets
Ƙ	Lagrange polynomial multiplier
Þ	Flajolet-Odlyzko constant
Ḧ	Vandermonde polynomial multiplier
ξ	Khinchin-Lévy constant
Ƭ	Dirichlet integral of cyanobacterial listeriosis
$\bar{G}$	Dirichlet integral of asymptomatic infection
η	Oscillatory integral operator for monocytogenes' prevalence
μ	Oscillatory integral operator for phycocyanin prevalence
Ƴ	Mean squared deviation
Ü	Dirichlet integral of physiological interference
ζ	Riemann zeta function representing Dirichlet series
θ	Lemniscate constant representing the integral covariance of stochastic data
α	Peak frequency density of observation data
β	Peak frequency density of model data
Ω_θ	Hyper-harmonic median of Dirichlet series
λ_α	Peak wavelength density of observation data
λ_β	Peak wavelength density of model data
Ƙ	Kullback-Leibler divergent coefficient
ω_n	Sylvester sequence of Eigen solutions
Ⱶ	Taylor series expansion derivative
$\bar{Ṛ}$	Maclaurin series expansion derivative
f	Bessel corrected variance
σ_ϕ	Fourier integral of feedback components
χ	Bernoulli continuity coefficient for Gregory's series
$Б_\tau$	Recursive Bayesian Estimation of relative conjugate error