HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An end-to-end real-time pollutants spilling recognition in wastewater based on the IoT-ready SENSIPLUS platform

Luca Gerevini [a], Gianni Cerro [b], Alessandro Bria [a], Claudio Marrocco [a], Luigi Ferrigno [a], Michele Vitelli [c], Andrea Ria [d], Mario Molinara [a]

[a] Dept. of Electrical and Information Engineering, University of Cassino and Southern Lazio, 03043 Cassino, Italy
[b] Dept. of Medicine and Health Sciences "V. Tiberio", University of Molise, 86100 Campobasso, Italy
[c] Sensichips s.r.l., 04011 Aprilia, Italy
[d] Department of Information Engineering, 56122 Pisa, Italy

ARTICLE INFO

Article history:
Received 3 September 2022
Revised 23 December 2022
Accepted 24 December 2022
Available online xxxx

Keywords:
Machine learning
Smart sensors
Wastewater
Anomaly detection
IoT
Supervised learning

ABSTRACT

The problem of detecting illegal pollutants in wastewater is of fundamental importance for public health and security. The availability of distributed, low–cost and low–power monitoring systems, particularly enforced by IoT communication mechanisms and low-complexity machine learning algorithms, would make it feasible and easy to manage in a widespread manner. Accordingly, an End-to-End IoT-ready node for the sensing, local processing, and transmission of the data collected on the pollutants in the wastewater is presented here. The proposed system, organized in sensing and data processing modules, can recognize and distinguish contaminants from unknown substances typically present in wastewater. This is particularly important in the classification stage since distinguishing between background (not of interest) and foreground (of interest) substances drastically improves the classification performance, especially in terms of false positive rates. The measurement system, i.e., the sensing part, is represented by the so-called Smart Cable Water based on the SENSIPLUS chip, which integrates an array of sensors detecting various water-soluble substances through impedance spectroscopy. The data processing is based on a commercial Micro Control Unit (MCU), including an anomaly detection module, a classification module, and a false positive reduction module, all based on machine learning algorithms that have a computational complexity suitable for low-cost hardware implementation.

An extensive experimental campaign on different contaminants has been carried out to train machine-learning algorithms suitable for low-cost and low-power MCU. The corresponding dataset has been made publicly available for download. The obtained results demonstrate an excellent classification ability, achieving an accuracy of more than 95% on average, and are a reliable "proof of concept" of a pervasive IoT system for distributed monitoring.

1. Introduction

Water, which covers more than 70% of the Earth's surface and is involved in almost all life activities, is a primary factor influencing life on the Earth. Consequently, water quality monitoring is a crucial task, and ways to address it are widely spreading in the scientific literature (Ighalo et al., 2021; Budiarti et al., 2019; Saravanan et al., 2018; Akhter et al., 2022; Ferdinandi et al., 2019). Particularly critical is the issue related to wastewater (Trubetskaya et al., 2021), i.e., water having suffered pollution due to domestic, industrial, or hospital processes. Its monitoring has been a hot topic for two years as the COVID-19 pandemic spread throughout the world (Bogler et al., 2020; Farkas et al., 2020). Capabilities to get detailed and accurate monitoring and detect possible contaminants are related to three distinct components: sensing systems, geographical pervasiveness, and data processing.

As for sensing systems (Tyszczuk-Rotko et al., 2022; Kamaruidzaman and Rahmat, May 2020; Vikesland, 2018; Alam

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

E-mail addresses: luca.gerevini@unicas.it (L. Gerevini), gianni.cerro@unimol.it (G. Cerro), a.bria@unicas.it (A. Bria), c.marrocco@unicas.it (C. Marrocco), ferrigno@unicas.it (L. Ferrigno), michele.vitelli@sensichips.com (M. Vitelli), andrea.ria@ing.unipi.it (A. Ria), m.molinara@unicas.it (M. Molinara)

https://doi.org/10.1016/j.jksuci.2022.12.018
1319-1578/

et al., 2020), different costs and performance levels can be experienced. Issues to be faced are related to sensitivity, selectivity, and miniaturization. Most solutions prefer adopting a sensor array to increase the capability to discern between different substances. To get widespread monitoring and ensure water quality estimation and pollutant detection in a distributed way, the adoption of high-cost systems appears unsuitable.

IoT-ready, low-cost platforms enable the achievement of geographical pervasiveness that could benefit from a high level of energy autonomy, low computational burden, and high data transfer capabilities. Their flexibility allows spreading devices in the area of interest by creating a monitoring network. The IoT capabilities adopted for water monitoring are widespread (Junior et al., 2021; Dupont et al., 2018; Overmars and Venkatraman, 2020).

In terms of data processing, acquired measurements are generally processed to become features to feed Machine Learning (ML)/ Deep Learning (DL) algorithms adopted for classification (Lowe et al., 2022; Koditala and Pandey, 2018; Bansal and Geetha, 2020; Dilmi and Ladjal, 2021; Bria et al., 2021; Bria et al., 2020). Major challenges regard finding pathways to have fast data exchange, classify with acceptable computational complexity in nearly real-time and be able to discriminate among different pollutants that can be found in the flowing wastewater.

The paper's goal is to present an end-to-end system for spilling detection in wastewater that includes a complete chain from sensing to final classification. The proposed system is:

- real-time because it can respond on a single sample basis generating a classification for each set of ten measures: the total time needed for a single acquisition/classification is equal to about 1.6 s;
- low power, low cost, and IoT ready, thanks to the coupling of the SENSIPLUS (Ria et al., 2022; Manfredini et al., 2021) (discussed in the following) with a commercial MCU;
- able to tackle unknown substances, thanks to the anomaly detection module;
- can monitor a single polluting source at a time, considering that spilling of the considered pollutants in wastewater is a rare event.

Based on this concept, the paper is structured as follows. Section 2 contains a complete review of state of the art in pollutant detection in water and wastewater through machine learning. Section 3 highlights the main contributions provided in this paper. Section 4 describes the measurement set-up as well as the data processing stage to get data ready for classification. Section 5 shows the obtained experimental results. A discussion of the obtained results is reported in 6. Conclusions and future directions are finally discussed in Section 7.

2. Related works

In recent years several sensor prototypes for monitoring the composition of wastewater in the context of WasteWater Treatment Plants (WWTP) systems have been proposed in the literature (Ferdinandi et al., 2019; Bourelly et al., 2020; Betta et al., 2019; Molinara et al., 2020; Bria et al., 2020; De Vito et al., 2018; Sewage monitoring system for tracking synthetic drug laboratories, 2022; Hoes et al., 2009; Lim, 2012; Lepot et al., 2017; Ji et al., 2020; Drenoyanis et al., 2019; Pisa et al., 2019; Desmet et al., 2017). The sensors presented are based on different technologies such as electrochemical sensors, optical sensors, based on mass or ion spectrometry, etc. and can be mounted inside wells with the aim of detecting the presence or concentration of certain pollutants. In Ferdinandi et al. (2019), Bourelly et al. (2020), Betta et al. (2019), Molinara et al. (2020), Bria et al. (2020) the application of the SENSIPLUS as air and water monitoring system is presented and its effectiveness preliminary demonstrated. In De Vito et al. (2018) the authors describe a distributed sewage monitoring system based on low-cost technologies. In this case, the authors do not carry out recognition of specific substances but limit themselves to carrying out the detection of generic pollutants. In Sewage monitoring system for tracking synthetic drug laboratories (2022) a drug detection system is described in the sewage system to identify the presence of drug factories. The weak point of this solution lies precisely in the fact that it deals with a very specific problem, not designed for the detection of generic pollutants. In Hoes et al. (2009), a technique to find illicit household sewage connections to storm-water systems in the Netherlands using Distributed Temperature Sensing has been developed. In Lim (2012), a generic system for detecting pollutants in wastewater is presented. The system lacks the ability to discern between different substances and is based on outdated technologies. In Lepot et al. (2017) a system for detecting illicit connections to the sewage system is presented. The solution, based on the use of an infrared camera, is not designed for the detection of specific substances. In Ji et al. (2020), a system for measuring the amount of wastewater based on image analysis is presented. In this case, the distinction between the different substances is completely missing, and in general, the vision systems, although immune to corrosion phenomena and deposits of material on the sensors, are generally characterized by high energy consumption and, therefore, not very suitable for continuous monitoring systems at low power. In Pisa et al. (2019), the authors propose a system specifically designed to detect nitrogen-derived components, specifically ammonium and total nitrogen, without any attention to the pervasiveness and low power/low cost. In Drenoyanis et al. (2019), a standalone, portable radar device allowing non-invasive benchmarking of sewer pumping station pumps is presented. The system is designed to generate timely alarms in the event of anomalies in wastewater flows near WWTP. It does not include any pollutant detection system. In Desmet et al. (2017), a system for detecting explosive precursors is presented, i.e., those substances that terrorists could use to manufacture rudimentary bombs. In this work, sensors functionalized with gold, palladium, and platinum are used, and voltammetry is used to detect substances.

3. Main contribution of this paper

From the review of the scientific literature on the application of machine learning to water analysis, it emerges that the problem of anomaly detection is neglected. Not considering anomalies in real systems means making them unusable in a context other than the laboratory, as the system would not be able to react correctly to substances not taken into consideration during the training phase, potentially generating false positives.

To summarize, the open issues in wastewater analysis are mainly related to the complex and expensive equipment often required, unsuitable for the IoT and pervasive paradigm, and to the lack of an anomaly detection step. This paper proposes a solution to both of these issues.

In terms of IoT readiness, is proposed the adoption of the SENSIPLUS chip, a proprietary device developed by the Italian company Sensichips s.r.l., which has been proven to be effective in reliable measurements for pollutant detection in air and water (Ferdinandi et al., 2019; Bourelly et al., 2020; Betta et al., 2019; Molinara et al., 2020; Bria et al., 2020). The SENSIPLUS chip, together with a commercial Micro Control Unit (MCU), becomes a low-power, low-cost, and IoT-ready miniaturized sensing platform. The MCU is needed to run the C++ API supplied with the SEN-

SIPLUS chip and to equip the system to communicate with external systems (for example, with USB or with MQTT over TCP/IP) and for the inference phase of various machine learning algorithms.

The second issue is tackled with a double-stage classification system: an anomaly detector and a multiclass classifier, starting from the idea that some pollutants are interesting while others are simply interferants and do not need to be classified. The anomaly detection allows stating if the analyzed substance can be one of interest or something else (for simplicity, unknown). Whenever such a module declares that the substance is not an anomaly, the multiclass classifier module is activated, and its computational burden is included in the system load. The combination of both modules permits having a substantial false positive reduction while keeping a very high accuracy value for the substances of interest.

The combination of the developed platform and the new concept of supervised double-stage classification represents the main contribution of this work to the state of the art.

4. Methodology

4.1. The detection chain

Fig. 1 shows the overall detection system that is based on the Smart Cable Water (SCW) visible in Fig. 2, a proprietary IoT-ready smart sensor system by Sensichips s.r.l, composed of InterDigitated Electrodes (IDEs) and based on SENSIPLUS (Ria et al., 2022).

The latter is a tiny analytical sensing platform of 1.5 mW power absorption, with communication capabilities like SPI, I2C, and SENSIBUS (a proprietary 1-wire communication protocol). The SENSIPLUS needs an MCU to run its C++ API, which includes the engine for inference of machine learning algorithms. The ESP32/ESP8266 with USB and WiFi communication capability has been selected as an MCU. ESP32/ESP8266 can also guarantee data transfer through (for example) MQTT on TCP/IP. In this configuration, the MCU can act as a simple bridge to transmit the data collected through the sensors to the cloud (via MQTT, for example) and as a device that performs local processing for detecting substances of interest through the execution of suitable machine learning algorithms. During the operational time, SCW can be flooded in the water, and communication and control signals are conveyed through a suitable cable. SENSIPLUS is a micro-chip capable of interrogating on-chip and off-chip sensors with its versatile and accurate Electrical Impedance Spectrometer (EIS) in the frequency range comprised between 3.1 mHz and 1.2 MHz. With the SENSIPLUS, it is possible to perform measurements working with multi-

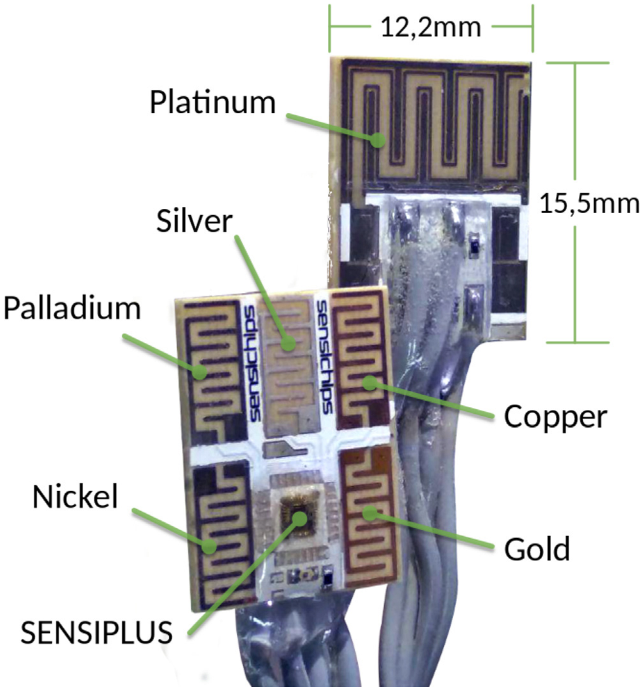


Fig. 2. Smart Cable Water (SCW) with InterDigitated Electrodes functionalized by coating them with six different metals.

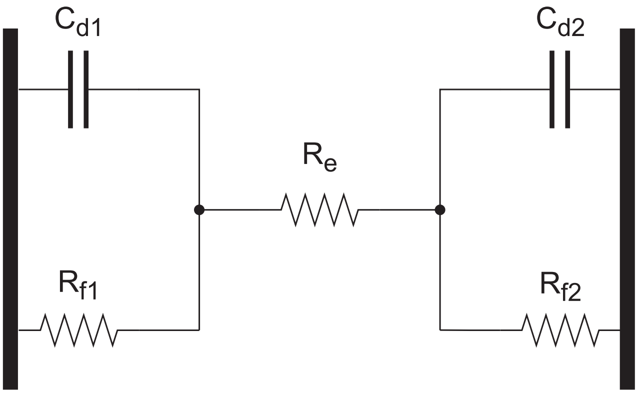


Fig. 3. Randles equivalent circuit.

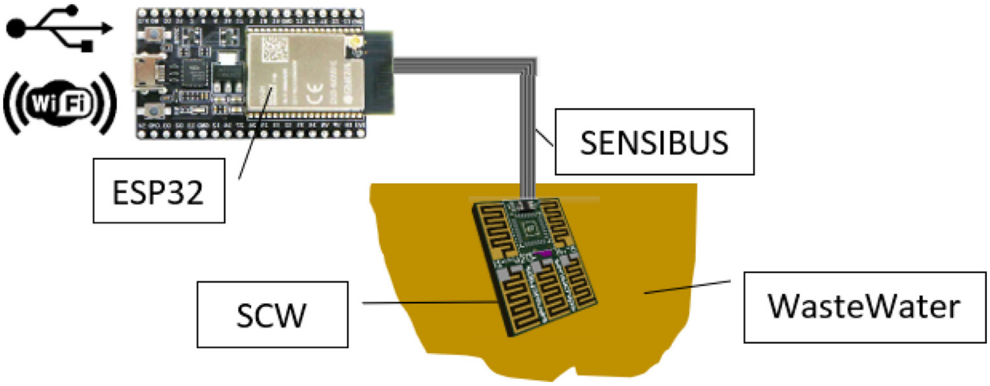


Fig. 1. The overall detection system deployed.

ple sensors; in particular, the SCW system has 6 IDEs. The physical principle adopted to detect and recognize a given set of substances is related to the RedOx dynamics of catalytic noble metals. Such a phenomenon can be observed as an electrical behavior. Fig. 3 shows the modeled electrical equivalent circuit of two electrodes flooded in a water solution, known in the literature as the Randles circuit (Alavi et al., 2017).

As can be seen from the electrical circuit, each electrode is modeled through a double-layer capacitance C_d and a faradic resistance R_f, which takes into account the interface between the water solution (called bulk) and the electrode itself. The model values depend on the electrode composition, geometry, bulk composition, etc. The parameter R_e is the equivalent resistance of the bulk, so it mainly depends on the bulk composition and the electrode area.

To maximize sensitivity to the substances of interest and the RedOx dynamics, the 6 IDEs of the SCW have been functionalized by coating them with six different metals: (M1) Gold, (M2) Copper, (M3) Silver, (M4) Nickel, (M5) Palladium and (M6) Platinum. (M1) to (M5) IDEs are 3 mm by 7 mm each, while (M6) is 12 mm by 8 mm (see Fig. 2).

4.2. Dataset acquisition for training

The proposed system is intended to detect and recognize substances spilled in wastewater. Consequently, the best solution to build a good dataset for the training phase would be to acquire all the measurements directly in a controlled drain of a sewage network. However, this is not a viable solution mainly for two reasons:

- Measurements point of view: all measurements should be taken from the same and reliable conditions; however, due to the instability typical of the sewage background environmental composition, it is impossible to reach an acceptable level of reliability conditions.
- Heath point of view: due to the presence of viruses, bacteria, and other dangers, operating directly in the sewage network would represent biological hazards.

To solve the listed problems, we create Synthetic WasteWater (SWW) to simulate the sewage composition and a measurement setup as described in Fig. 4 to create a suitable dataset. The adopted recipe for the SWW is inspired by a simplified version of the one created in Nopens et al. (2001). Moreover, to better reproduce a real wastewater scenario, the pH of every batch of the SWW has been corrected according to Janna (2016), where measurements on the real wastewater are reported. For the more detailed chemical composition of the SWW refer to Table 1.

Fourteen substances have been spilled in the SWW background: (1) Acetic Acid; (2) Acetone; (3) Ethanol; (4) Ammonia; (5) Formic Acid; (6) Phosphoric Acid; (7) Sulphuric Acid; (8) Hydrogen Peroxide; (9) Synthetic Waste Water; (10) Sodium Hypochlorite; (11) Sodium Chloride; (12) Dish Wash Detergent; (13) Wash Machine Detergent; (14) Nelsen.

The listed substances can be split into two groups: substances 1–9 (group 1) and 10–14 (group 2). Group 1 includes only substances that our system should be capable of recognizing, while group 2 includes only the outlier samples that our system should be able to reject.

The measurement procedure for each substance for dataset creation is composed of two phases:

- *Warm-Up phase:* to let all sensors stabilize, 600 samples at 0.5 Hz rate (total warm-up time: 900 s) in pure SWW are acquired.

Table 1
Synthetic waste water chemical composition.

Compounds	Concentration [mg/l]
Fertilizer	91.74
Ammonium Chloride	12.75
Sodium Acetate Trihydrate	131.64
Magnesium Hydrogen Phosphate Trihydrate	29.02
Monopotassium Phosphate	23.4
Iron (II) Sulfate Heptahydrate	5.80
Starch	122.00
Milk Powder	116.19
Yeast	52.24
Soy Oil	29.02

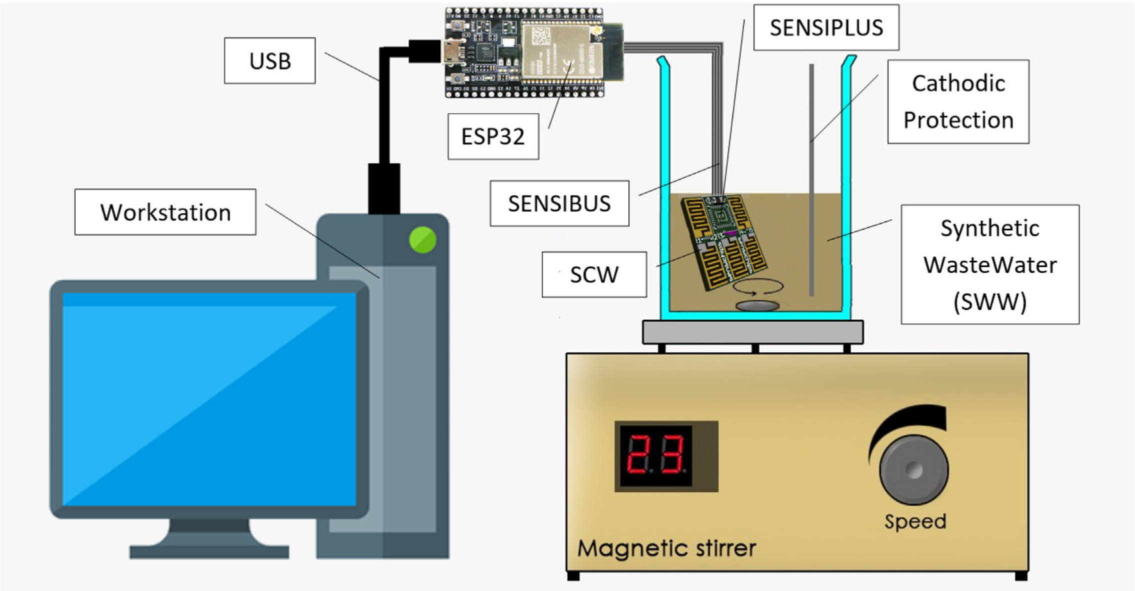


Fig. 4. Measurement Set-Up for dataset acquisition.

- *Measurement phase:* after the first 600 samples, the substance of interest is spilled in the SWW, and, to record the entire sensor's evolution after the injection, another 1000 samples at the same sample rate are acquired (total measurement phase time: 2000 s).

The obtained dataset has been made publicly available here (Public link for downloading the acquired dataset, 2022).

One of the main problems related to Machine Learning is related to feature identification, i.e. the choice of informative properties derived from sensors, able to maximize the classification accuracy.

In our case, according to the electrical equivalent circuit described in the previous section, we choose to record the following features:

- Resistance measured at 78 kHz frequency, for the Gold and Platinum IDEs.
- Resistance and Capacitance measured at 200 Hz frequency, as concerns Gold, Platinum, Silver, and Nickel.

obtaining a feature vector of size ten (6 resistance and 4 capacitance). The Palladium and Copper IDEs have not been used in this experimental campaign.

The cited features have been chosen because of the different behavior of the equivalent circuit at low and high frequencies. In particular, both C_d exhibit a high impedance at the low frequencies and can be represented as an open circuit (see Fig. 5a). So the measurements depend either on the faradic or bulk resistance (R_e). On the other hand, at the high frequencies, the two C_d present a low impedance and can be seen as a short circuit (see Fig. 5b): the measurements mainly depend on the bulk resistance.

4.3. Dataset structure and usage

For each substance, ten acquisitions of 1600 samples obtained through the measurement procedure as mentioned above have been collected, obtaining 16000 samples overall.

For evaluation purposes, the k-Fold Cross-Validation procedure has been adopted. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. Its application generally results in a less biased or less optimistic estimate of the model efficiency than other methods, such as a simple train/test split. Usually, the first step in k-fold Cross-Validation is the random shuffle of the collected data. In our case, taking into account that measures belonging to the same experiment are strongly correlated, we preferred to assume as a unit for k-fold an entire acquisition (1600 samples) of all the substances.

In order to find the best anomaly detection and multiclass classifiers model to use for the entire system, the entire Data set has been organized in ten Fold (*Fold 0*, *Fold 1*, ..., *Fold 9*). Each *Fold* contains nine additional Split (*Split 0*, *Split 1*, ..., *Split 9*) and one Test. The given *Split* are organized like the following:

- Training data: used to train both anomaly detection and multiclass classifier model.
- Test data: used to find the best model's hyperparameters for anomaly detection and multiclass classifier.

For the final evaluation concern, it is composed of whose samples are not contained either in the Training data nor in the Test data of all the *Splits* related to the given *Fold*.

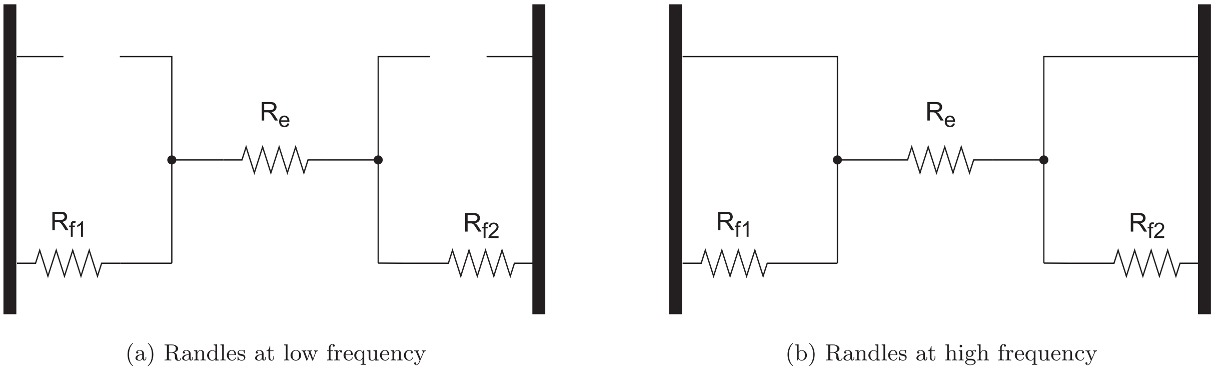
In order to keep things clear we used a fixed nomenclature: the number inside the given *Fold*'s name, indicates the experiment (data acquisition) used to perform the final evaluation, while the number inside the given *Split* indicates the experiment used for the related Test data. For the Training data concern, it is composed of all the experiments except the one used for the related Test Set and the one used for the final evaluation that, as said before, contains data that is unseen from both the Training and Test data of the related *Fold*. For example, the Fold0 contains the *Split* from 1 to 9, excluding the *Split 0* since the experiment 0 of all the substances is used to build the related Test set.

The Test data of the *Split 1* is made of experiment 1 of all substances while the Training data is made of all the remaining experiments (excluding experiment 1 used for the Test and experiment 0 used for the final evaluation). In this way, the Test data of the *Split 2* is made by experiment 2, while the related Training data will exclude experiments 2 and 0, and so on. The final evaluation of *Fold 0* is made by experiment 0 of all substances. See Fig. 6 for a graphical representation of the Data Set splitting. It is worth specifying that in Fig. 6 Exp 0, Exp 1, ..., Exp 9 means respectively acquisition 0, 1, ..., 9 of all substances.

Finally, it is worth specifying that regards the multiclass classifier and anomaly detection model the training, test, and final evaluation set ratio during the learning phase was respectively: $80\%, 10\%, 10\%$. Furthermore, in order to properly validate and test the learned anomaly detection models, the validation and test sets have been polluted with outliers points taken from the substances belonging to Group2.

4.4. Classification

The classification system is organized in two phases: (i) Data Preprocessing; (ii) Classification. As can be seen in Fig. 7, the Data Preprocessing phase (i) normalize the raw data coming from sensors and discriminates through a Finite State Machine (FSM, see



(a) Randles at low frequency

(b) Randles at high frequency

Fig. 5. Randles at different frequencies.

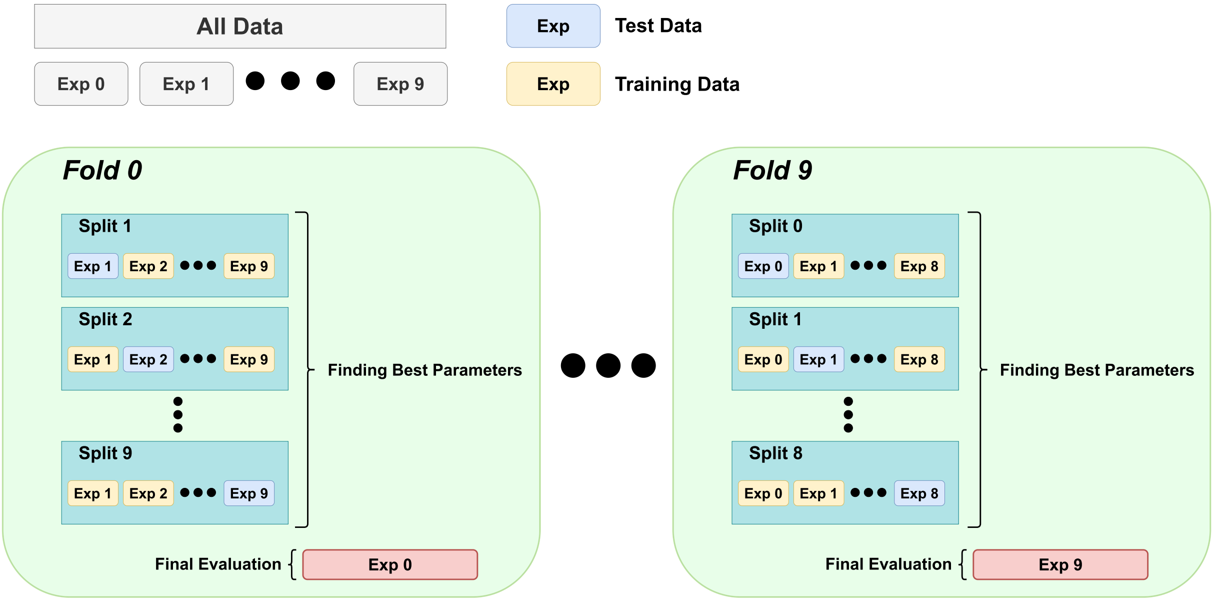


Fig. 6. Data Set Structure. Exp 0, Exp 1, . . ., Exp 9 means respectively acquisition 0, 1, . . ., 9 of all substances.

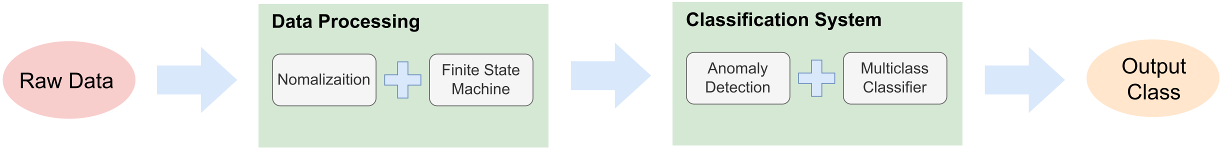
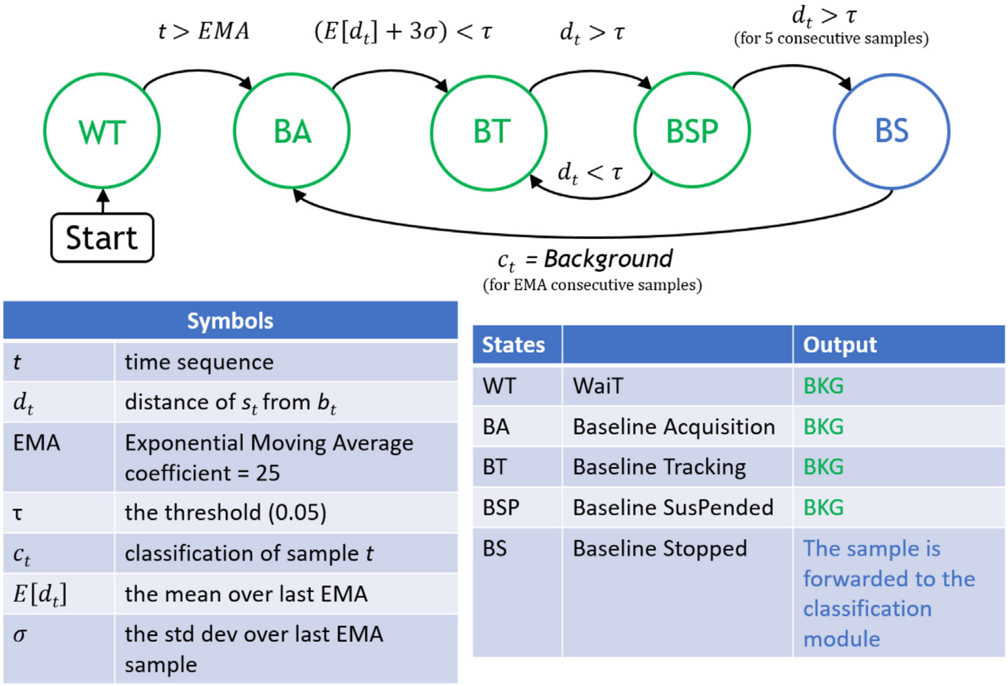


Fig. 7. An overall view of the system.


$t > EMA$
$(E[d_t] + 3\sigma) < \tau$
$d_t > \tau$
$d_t > \tau$
(for 5 consecutive samples)
WT
BA
BT
BSP
BS
$d_t < \tau$
Start
c_t = Background
(for EMA consecutive samples)

Symbols	
t	time sequence
d_t	distance of s_t from b_t
EMA	Exponential Moving Average coefficient = 25
τ	the threshold (0.05)
c_t	classification of sample t
$E[d_t]$	the mean over last EMA
σ	the std dev over last EMA sample

States		Output
WT	WaiT	BKG
BA	Baseline Acquisition	BKG
BT	Baseline Tracking	BKG
BSP	Baseline SusPended	BKG
BS	Baseline Stopped	The sample is forwarded to the classification module

Fig. 8. Finite state machine.

Fig. 8) if there should be submitted to the Classification phase (ii) or not.

4.4.1. Data preprocessing

The Data Preprocessing phase is realized in two steps:

- normalization of the raw data coming from sensors, through the creation of a robust baseline signal;
- takes a decision if the normalized sample should be forwarded to the anomaly detector or classified directly.

The baseline signal $\boldsymbol{b}_t$ is generated by the union of the FSM with the application of an Exponential Moving Average (EMA) according to the following equation:

$$\boldsymbol{b}_t = \begin{cases} \boldsymbol{s}_t & t = 0 \\ \boldsymbol{b}_{t-1} & t > 0, S \in \{BS, BSP\} \\ \alpha \boldsymbol{s}_t + (1-\alpha) \cdot \boldsymbol{s}_{t-1}, & t > 0, S \in \{WT, BA, BT\} \end{cases} \quad (1)$$

where s_t are the sensors' raw data, Wait (WT), Baseline Acquisition (BA), Baseline Tracking (BT), Baseline Suspended (BSP), and Baseline Stopped (BS) are the states of the FSM.

The FSM aims to build a robust baseline capable of coping with: the variability between sensors/chips, the sensor drift, the environmental noise, interferences, etc. The first two states (WT and BA) guarantee that the baseline is not affected by noise and/or interferences. Once the FSM reaches the BT state, the system tries to detect the injection of a substance, revealed by a peak in the raw data with respect to the baseline generated with EWA (see below).

The BSP state is an intermediate state between BT and BS that try to filter out signal spike by waiting if a cluster of samples confirms the presence of a spilled substance. Once the system is confident there is the spilling of a substance (after 5 acquisitions), the FSM moves to the BS states. Consequently, the normalized samples are passed as input to the classification phase algorithm.

Regarding the EMA, the α parameter is the reciprocal of EMA$_c$ (coefficient empirically set to 25). The normalization value is given by the following formula:

$$\boldsymbol{f}_t = \boldsymbol{s}_t / \boldsymbol{b}_t \quad (2)$$

where the $\boldsymbol{f}_t$ is the normalized feature vector, while $\boldsymbol{s}_t$ is the raw sensor data and the $\boldsymbol{b}_t$ is the baseline signal computed as described by the Eq. 1.

Fig. 8 shows the entire FSM system. In particular, t is the current time sample, while τ is a threshold that, in our case, has been empirically set equal to 0.05. Regards the $\boldsymbol{d}_t$ parameters, it represents the Euclidean distance between the normalized features vector f_t and the unit vector $\boldsymbol{u}$ (a vector of ones) in a 10-dimensional space that is the size of the vector $\boldsymbol{s}_t$ (see Eq. 3).

Starting from the euclidean distance $\boldsymbol{d}_t$ evaluated between $\boldsymbol{s}_t$ and $\boldsymbol{b}_t$ in the feature space, there is a peak that reveals an injection when $\boldsymbol{d}_t$ is greater than a threshold τ (empirically established to 0.05).

Looking at the Eq. 2 it is clear that the vector $\boldsymbol{f}_t$, when $\boldsymbol{b}_t$ is equal to $\boldsymbol{s}_t$ is equal to the unit vector. For this reason, the Euclidean distance has been computed with respect to the unit vector and so when $\boldsymbol{d}_t$ is equal to zero means that the baseline signal $\boldsymbol{b}_t$ is perfectly tracking the sensors signals s_t.

$$\boldsymbol{d}_t = \|\boldsymbol{f}_t - \boldsymbol{u}\| \quad (3)$$

As seen in the Fig. 8, the current state of the FSM can change according to a given rule. In particular, the FSM starts with the WT state. In this state the classification system will simply compute, and store into a vector, the first EMA_c distance computed over the $\boldsymbol{s}_t$ measured samples, as reported in the Eq. 2. Once the distance vector has been filled, the FSM can pass in the BA state. Here the system will keep updating the distance vector and, once the variability of the vector (computed as the mean plus three times the standard deviation) is below a given threshold, the system can move to the next state. At this point, the system will check if a substance has been spilled in the water, and this is done by checking when the current distance is major of a given threshold. Once the FSM moves to the BSP state, in order to not confuse the spill of a substance with a measurement spike or simple noise, the system will check that the current distance remains above the threshold for five consecutive samples (BSP), otherwise, the system comes back to the BT state. Finally once the FSM is in BS state the current normalized samples are given in input to the detection system. In this state, if the sample classification is equal to the background substance, the FSM will return to the BA state.

The entire system depicted so far is shown in Fig. 9.

Where S indicates the state of the FSM, C_t is the classification of the sample at time t and BKG is the background substance.

4.4.2. Detection phase

In a real scenario, there are plenty of substances that flow in the sewerage network, therefore it is crucial to be able to distinguish between the substances of interest and the other ones.

In this sense, the main goal of this phase is to determine if the given flowing substance is one of the substances of interest, in order to be then able to predict its name correctly.

The detection phase is basically divided into two main parts:

- Anomaly Detection
- Multiclass Classification

Anomaly Detection

Regards the anomaly detection algorithms, we can mainly distinguish them in two approaches:

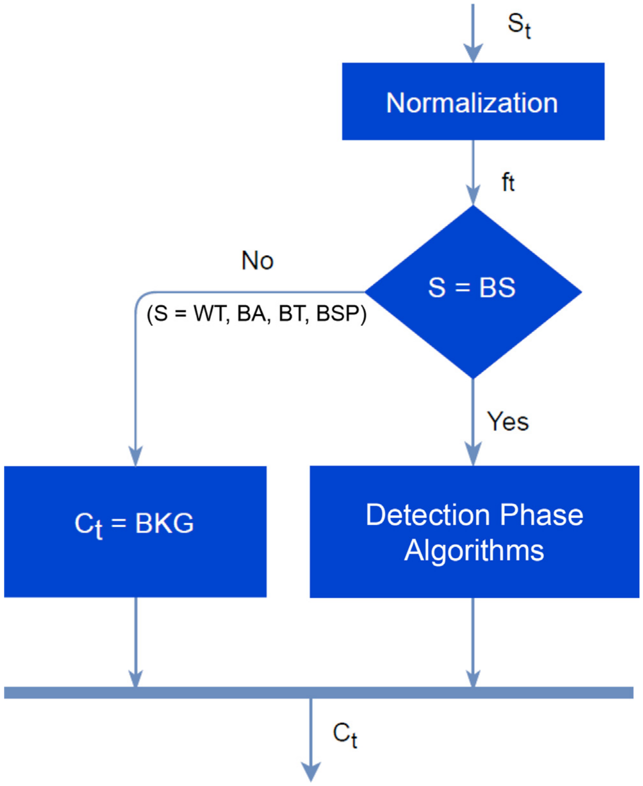


Fig. 9. Finite state machine flow chart.

- Outlier Detection
- Novelty Detection

In the outlier detection algorithms, the training data contains outliers samples. In this case, the estimators try to fit the regions where the training data is the most concentrated, ignoring the deviant observations. The training data is not polluted with outliers samples in the novelty detection algorithms. In this context, we want to determine whether a new observation is an outlier. In this sense, an outlier is also called a novelty.

Our case is better represented by the novelty detection approaches according to our data set and the application field. This is because, in our application field, we want to discard all those substances that are usually present in the sewage system, and we want to recognize only the substances of interest that represent a minimum part of the substances that can be found in wastewater. To have as complete a point of view as possible, we trained and tested anomaly detection models built with either novelty or outlier approaches:

- Novelty Detection: One-class SVM, Local Outlier Factor, and KNN
- Outlier Detection: Elliptic Envelope and Isolation Forest

All the algorithms have been taken from the sci-kit learn library (Pedregosa et al., 2011) except for the KNN, which has been taken from the Python Outlier Detection (PyOD) library (Zhao et al., 2019).

As described in Section 4.3, we have divided the entire data set into ten cross-validations folders, each of which contains additional nine folders containing training and a validation set. For the outlier detection data set concern, it is the same depicted in Section 4.3 with the addition of some outlier samples in the training set (about 10%).

Multiclass Classification

Starting with the results obtained in previous work, we have trained and optimized the accuracy of a KNN on the described data set. It is noted that, unlike anomaly detection, the training and validation sets are formed with only the samples of the substances of interest.

In both cases, anomaly detection and multiclass classification, the grid search approaches have been chosen to optimize the models' accuracy. All the models parameters are detailed in the Table 2 and Table 3.

Table 2
Anomaly detection models parameters.

Classifier	Parameters	
KNN	contamination	[0.01, 0.05, 0.1, ..., 0.5]
	N neighbors	[10, 100, 200, ..., 500]
SVM	ν	[0.01, 0.05, 0.15, ..., 1.0]
	Kernel	Radial basis function
	γ	[auto, scale, 0.01, 0.05, 0.15, ..., 1.0]
Local Outlier Factor	contamination	[0.01, 0.05, 0.1, ..., 0.5]
	N neighbors	[10, 100, 200, ..., 500]
Elliptic Envelope	contamination	[0.01, 0.05, 0.1, ..., 0.5]
Isolation Forest	contamination	[auto, 0.01, 0.05, 0.1, ..., 0.5]
	N estimators	[50, 100, 150, ..., 500]

Table 3
Multiclass classification model parameters.

Classifier	Parameters	
KNN	algorithm	ball tree
	N neighbors	[10, 100, 150, ..., 500]
	weights	[uniform, distance]

Finally, the entire system, composed of the data processing and the detection systems, is shown in Fig. 10.

Eventually, to find out the best anomaly and multiclass classifier model, the cross-validation technique over the ten sub-set of the data set (see Section 4.3 for more details) has been used. Once the best model of each classifier has been found, the entire system has been tested over the test data.

It is important to point out that the proposed detection system does not relay over any pattern/trajectory recognition, or time series, or, in other words, it is time-independent. This feature allows us to build an IoT-ready system capable of detecting and recognizing, the given spilled substance based only on the current samples, as shown in Fig. 10. In this sense, we can refer to our system as an IoT-ready platform for real-time pollutant spilling detection.

Algorithm 1. Training procedure

```
input: A dataset $\mathscr{F}$ representing a single Fold, list of classifiers
   to train, hyperparameters
output: Best Classifier
begin
  $\mathscr{F}_n = normalizeDataSet(\mathscr{F})$;
  for $clf$ in $classifiers$ do
    $X_{train}, Y_{train} = loadTrainigData(\mathscr{F}_n)$;
    $X_{validation}, Y_{validation} = loadValidationData(\mathscr{F}_n)$;
    for $param$ in $hyperparameters$ do
      $clf.set\_params(param)$;
      $clf.fit(X_{train})$;
      $Y_{pred} = clf.predict(X_{validation})$;
      $Accuracy.append([clf, evaluate(Y_{pred}, Y_{validation})])$;
  $clf_{best} = getBestClf(Accuracy)$;
  return $clf_{best}$;
end
```

Algorithm 2. Test procedure

```
input: A TestSet $\mathscr{T}$, best anomaly model ($anly$), best
   multiclass model ($clf$), doAnomaly
output: [Accuracy, CM]
begin
  groundtruth $= getGroundtruth(\mathscr{T}_n)$
  for $sample$ in $\mathscr{T}_n$ do
    if $doAnomaly$ then
      $outClass.append(onlineClassidication(sample)$;
    else
      $sample_n = normalize(sample)$;
      $state = getFsmState(sample_n)$;
      if $state \neq BS$ then
        $outClass.append(\mathrm{BKG})$;
      else
        $outClass.append(clf.predict(sample_n))$;
  $ConfusionMatrix = evaluate(outClass, \mathrm{groundtruth})$;
  return $[Accuracy, CM]$;
end
```

Algorithm 3. Online Classification procedure

```
input: Sample $\mathscr{S}$, anomaly detection classifier (anly),
   multiclass classifier (clf)
```

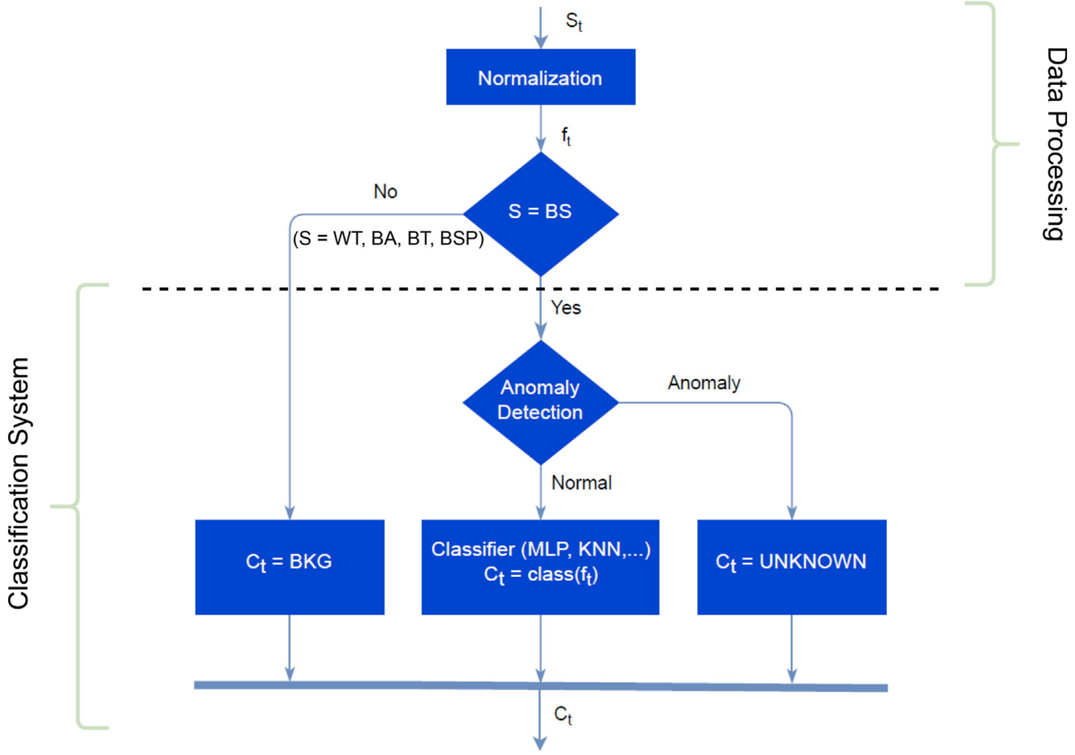


Fig. 10. Entire system flow chart.

```
output: predicted class (outClass)
begin
  $\mathscr{S}_n = normalize(\mathscr{S})$;
  $state = getFsmState(\mathscr{S}_n)$;
  if $state \neq BS$ then
    $outClass = \mathrm{BKG}$;
  else
    if $anly.predict(\mathscr{S}_n) = \mathrm{inlier}$ then
      $outClass = clf.predict(\mathscr{S}_n)$;
    else
      $outClass = \mathrm{UNKNOWN}$;
  return $outClass$;
end
```

The Algorithm 2 and 1 shows the pseudo code regarding the Training and Test procedure. As described in the previous sections, the training procedure is the same for both anomaly detection and multiclass classifiers. For the test procedure concern, instead, it has been built to be able to test either the entire system (anomaly detection and multiclass classifier) rather than only the multiclass classifier one. For that reason, the Test procedure takes as input an extra parameter "*doAnomaly*" that serves to decide if the test must be performed over only the multiclass model (case *doAnomaly* = FALSE) or on both anomaly detection and multiclass models (case *doAnomaly* = TRUE). In the latter case, the online classification procedure (Algorithm 3) is called. It is worth specifying that the Algorithm 3 represents the procedure implemented on the end-to-end system to perform online tests of the entire system. The time complexity of the overall chain is, in the worst case, the sum of the time complexity of a kNN with an SVM at inference time that is compatible with the computational capability of the selected MCU (Ray et al., 2021). The time complexity of the kNN algorithm is $O(n * d)$, where n is the number of samples in the training set, and d is the total number of features. The time complexity of our kernel SVM is linear with the number of support vectors n_s and the number of features d and can be represented as $O(n_s * d)$. The elapsed time for both steps is around 500*ms* on the ESP8266 allowing a complete evaluation between two acquisitions.

5. Experimental results

For each case (anomaly and multiclass classifier) the best model to verify the entire system over the test set has been selected. In the following subsections, the obtained results are reported.

5.1. Anomaly detection results

The best results have been obtained in the Fold0 case. In the case of Fold0, all the experiments between 1 to 9 have been used as training and validation sets following the cross-validation technique, while experiments 0 of all substances have been used as the test set.

The best results are reported in Table 4, while the average plus the standard deviation (STD), obtained over all the splits of the Fold0 data set, are reported in Table 5. It is worth noting that the obtained results show almost the same performance across the used algorithms. Thus, it is impossible to easily declare a winner. For that reason, since the application field of the proposed system best fits the novelty detection approaches, to test the entire system has been used the One-class SVM classifier.

Furthermore, to statistically validate the obtained results, we performed the Wilcoxon rank-sum test ($\alpha = 0.05$). Indeed, Table 5 also shows the p–value of the Wilcoxon test. From the table, it is possible to see that the performance differences between the three algorithms that best perform (One-Class SVM, Elliptic Envelope and Isolation Forest) are not statically significant (p-value > 0.05). Regarding the Local Outlier Factor and the KNN algorithm, it is possible to notice that the p-value is < 0.05, highlighting a statistical difference between the obtained results. Finally, it is worth

Table 4
Best results anomaly detection.

Approach	Algorithm	Accuracy	F1 Score	MCC	Parameters		Split
Novelty	One-Class SVM	0.9546	0.8868	0.8675	ν	0.01	6
					Kernel	rbf	
					γ	0.45	
	KNN	0.5982	0.0938	−0.2485	contamination	0.45	1
					N neighbors	10	
	Local Outlier Factor	0.8624	0.7639	0.7123	contamination	0.01	7
					N neighbors	400	
Outlier	Elliptic Envelope	0.9545	0.8864	0.8671	contamination	0.05	6
	Isolation Forest	0.9547	0.8872	0.8679	contamination	0.1	4
					N estimators	350	

Table 5
CrossValidation0 results.

Algorithm	Accuracy	F1Score	MCC	*p*-value
One-Class SVM	0.9358 ± 0.0171	0.8474 ± 0.0352	0.8115 ± 0.0497	–
KNN	0.5651 ± 0.0226	0.0902 ± 0.0018	−0.2762 ± 0.0188	5.6e − 6
Local Outlier Factor	0.8201 ± 0.0346	0.7137 ± 0.0382	0.6519 ± 0.0463	5.6e − 6
Elliptic Envelope	0.9325 ± 0.0157	0.8448 ± 0.0311	0.8069 ± 0.0445	0.4860
Isolation Forest	0.9463 ± 0.0130	0.8689 ± 0.0277	0.8423 ± 0.0391	0.7317

noticing that the Wilcoxon test has been performed by evaluating all chosen figures of merit (Accuracy, F1Score and MCC).

Regards the reported figure of merit they have been computed by the following formulas:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{F1Score} = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad \text{where} \quad \begin{cases} \text{precision} = \frac{TP}{TP+FP} \\ \text{recall} = \frac{TP}{TP+FN} \end{cases} \tag{5}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

where the True Positive (TP) are all the outlier samples classified as an outlier, True Negative (TN) are all the inlier samples classified as inlier, False Positive (FP) are all the inlier samples classified as an outlier, and False Negative (FN) are all the outlier samples classified as an inlier.

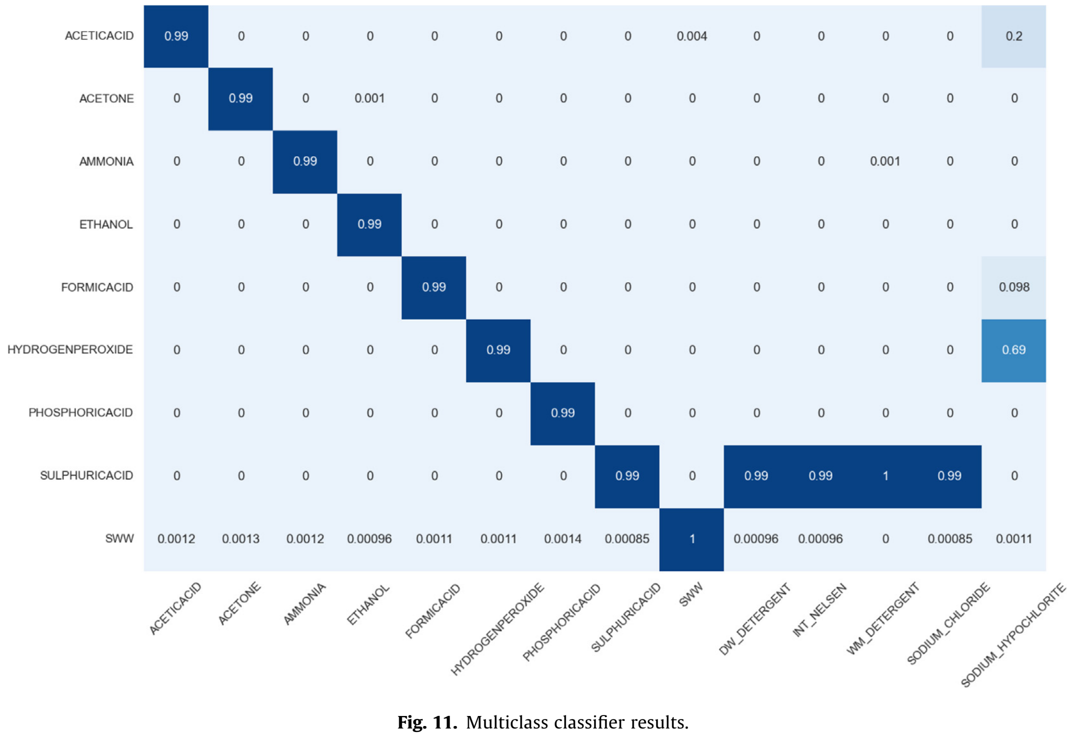


Fig. 11. Multiclass classifier results.

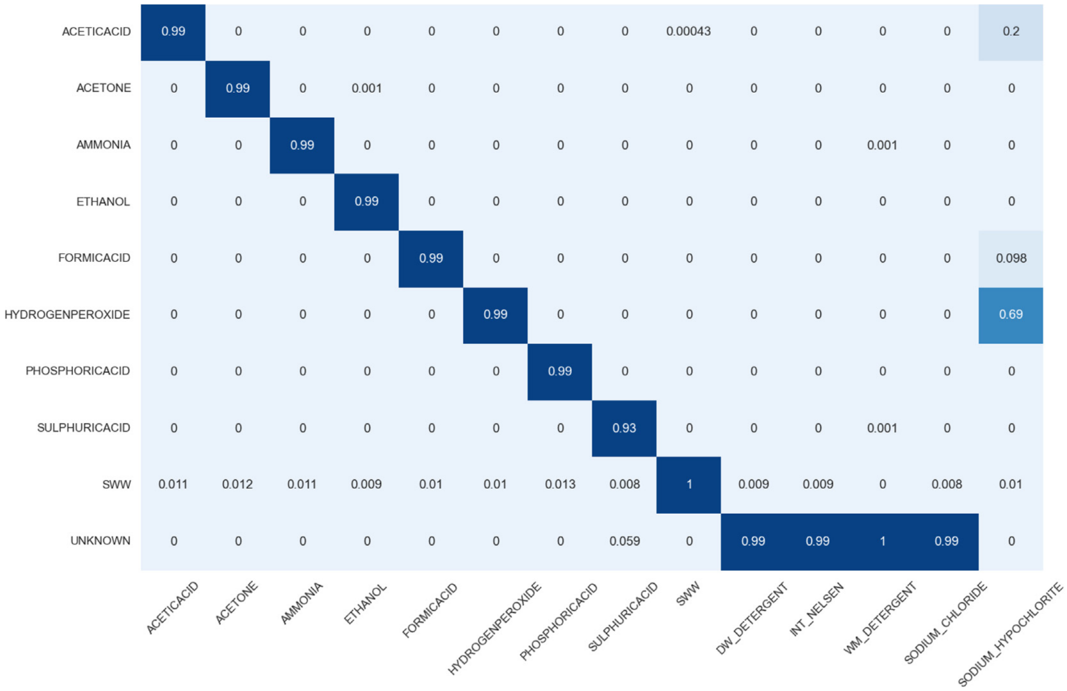


Fig. 12. Entire system results.

5.2. Multiclass classifier results

The best result obtained in the Fold0 is obtained by the KNN algorithm, using a number of neighbors (N) equal to 10 and adopting uniform weights. The obtained accuracy is equal to 99.37%. In terms of mean accuracy and standard deviation over the nine folds contained in Fold0, the obtained result is A%= (91.0 ± 5.7)%.

5.3. Entire system results

Two main tests have been done to highlight the benefits obtained from the usage of anomaly detection followed by the multiclass classifier for the entire system concerns.

Once with only the multiclass classifier, and once with anomaly detection plus multiclass classifier. The obtained results are shown in Figs. 11 and 12. As can be seen from the two confusion matrices, the outlier substances used are:

- Dish Wash Detergent (DW_DETERGEMT)
- Nelsen (INT_NELSEN)
- Washing Machine Detergent (WM_DETERGENT)
- Sodium Chloride (SODIUM_CHLORIDE)
- Sodium Hypochlorite (SODIUM_HYPOCHLORITE)

In the case of the Multiclass classifier only, the outlier substances get erroneously confused with one of the known substances generating many false positive alarms. To solve this problem, as described in the previous sections, before the multiclass classifier has been added an anomaly detection system capable of working as a false positive reduction filter.

As reported in Fig. 12, with the addition of the anomaly detection system, most outlier samples get correctly labeled as “UNKNOWN”. More precisely, 79.4% of the outlier samples has been correctly labeled as “UNKNOWN”, while the remaining 20.6%, which represents all the sodium hypochlorite samples, gets mostly confused with the hydrogen peroxide (according to what is shown in Fig. 11).

Finally, as already depicted in Section 5.1, since the obtained results (see Table 4 and 5) show almost the same performance across the used algorithms, and given the application field of the proposed system, in the reported result the One-class SVM classifier has been used.

6. Field tests

In order to obtain field tests, two preliminary experiments were conducted on real scenarios: one at the Acqualatina treatment plant in Borgopiave (Latina, Italy) (see Fig. 13), and the second on a series of wells located in Via Castelbottaccio (Rome, Italy) in collaboration with ACEA S.p.A. (Azienda comunale energia e ambiente, 2022) (see Fig. 14).

A flotation system has been designed to allow the immersion of the SCW in water at the proper depth, as shown in Fig. 15.

To be able to install the sensing system completely inside the manhole, a prototype of the measurement system has been developed, as shown in Fig. 16. As can be seen, the measurement system is composed of an IP56 waterproof certificate box, a Raspberry Pi4, a GSM hat for Raspberry Pi based on SIM7600E-H with two external antennas, a 20000mAh power bank, an ESP8266 board connected to the SCW via a 10 m SENSIBUS cable. This configuration can ensure a continuous measurement and transmission of about 1 week. The prototype presented here was designed for a field test without paying attention to energy consumption. A solution based exclusively on MCU (without Raspberry) and on the LoRaWAN (Long Range Wide Area Network) standard could work continuously for several months of continuous monitoring with periodic transmissions, exploiting an adequate local memory.

In this context, many substances have been tested: Phosphoric Acid, Sodium Hypochlorite, Acetic Acid, Formic Acid, Ammonia, and Hydrogen Peroxide. Unlike laboratory tests, in the real context, they encountered many problems, such as the accumulation of scale or air bubbles near the sensors of the SCW. After some preliminary tests, part of the problems encountered was solved, and

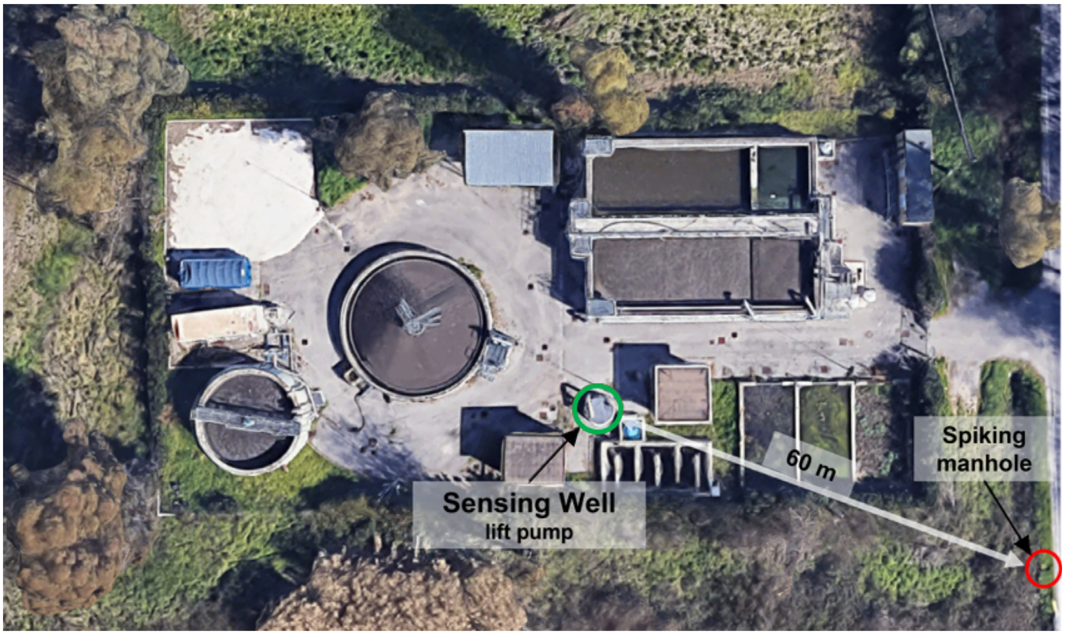


Fig. 13. Borgopiave. Green circle represents the sensing manhole, while red circles represent the spiking manhole positioned at 60 m from the sensing manhole.

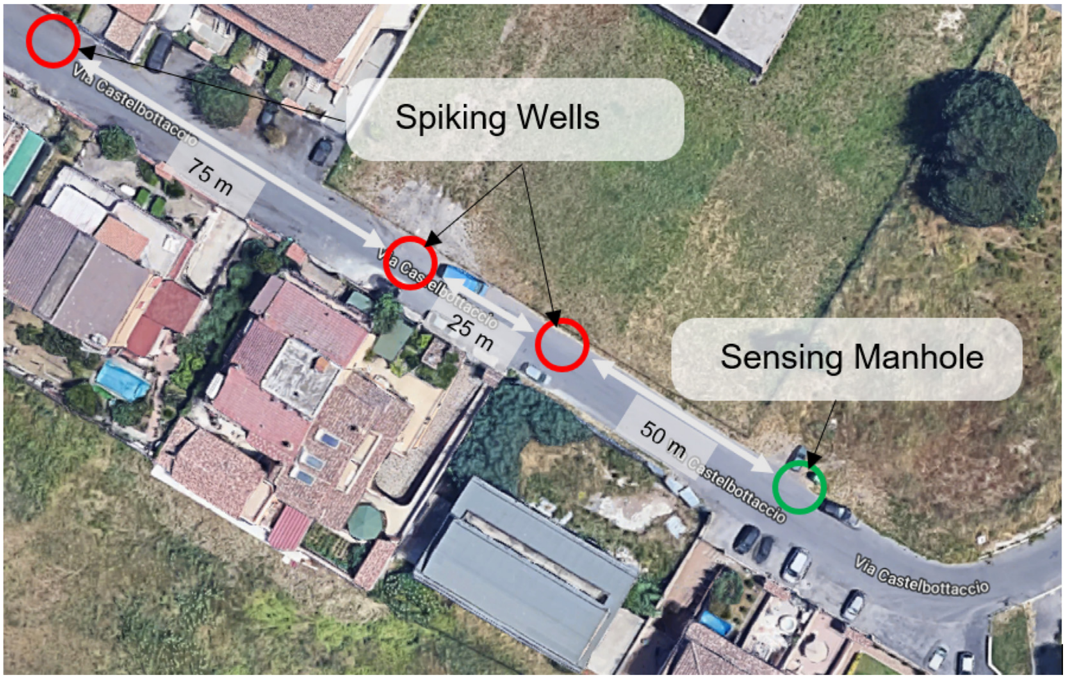


Fig. 14. Castelbottaccio. The green circle represents the sensing manhole, while red circles represent the spiking manholes respectively positioned at 50 m, 75 m, and 150 m from the sensing manhole.

although other problems were still present, it was still possible to reach an accuracy of more than 80%.

7. Discussion

From the Experimental Results section, given an outlier sample as input to a multiclass classifier, the output will fall under one of the known classes. This behavior leads to generating a number of false alarms equal to the number of outlier samples, making the system useless in a real scenario application. The results shown in Fig. 11, clarify the drawbacks of using only the multiclass classifier system to recognize a given substance. In this case, indeed 100% of the outlier samples represented by the dish wash detergent, Nelsen, washing machine detergent, sodium chloride, and sodium hypochlorite has been mainly confused with the sulphuric acid and the hydrogen peroxide generating a great number of false alarms. For sure, a multiclass classifier, used alone, cannot reject any outlier sample. For that reason, an anomaly detection module as a false alarm filter has been introduced to solve this kind of behavior.

Table 4 and 5 show the results obtained by the anomaly detection system. As can be seen, the performance of the One-Class SVM, Elliptic Envelope and Isolation Forest are pretty similar, this means

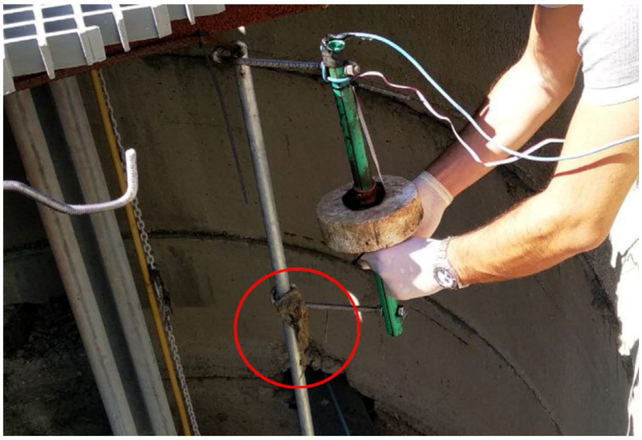


Fig. 15. The flotation system for the immersion of the SCW. In red are highlighted some garbages.

that the three algorithms can be equally used. Moreover, to statistically validate the obtained results, the Wilcoxon rank-sum test ($\alpha = 0.05$) has been performed, and the results, shown in Table 5, didn't show any significant differences. Finally, the chosen figures of merit show that the anomaly detection algorithms are able to distinguish between the outlier samples and the normal ones correctly.

At this point, putting the anomaly detection and multiclass classifier systems together has been possible to reach the results shown in Fig. 12. The results show that the anomaly detection system has rejected most outliers samples (around the 80%) by labeling it as "UNKNOWN". As can be seen, all the sodium hypochlorite samples have been mainly confused with hydrogen peroxide. Even though this is a behavior that worsens the system's performance, it can still be considered an acceptable behavior. Indeed, even if the two substances are chemically different (i.e., sodium hypochlorite is a polar compound while hydrogen peroxide is nonpolar), they have a similar oxidation potential: $1.6V$ for the sodium hypochlorite and $1.75V$ for the hydrogen peroxide (Vanýsek, 2010). Moreover, among all the substances of interest, hydrogen peroxide is the only compound that can be considered a strong oxidant (in a range that goes from $+3V$ for the oxidants to $-3V$ for the reducers). This similarity is particularly evident with the measurements at $78kHz$. For all those reasons, the confusion between sodium hypochlorite and hydrogen peroxide can be considered acceptable.

For the substance of interest concern, instead, Fig. 12 shows as the multiclass classifier results (see Fig. 11) are substantially maintained. Indeed, the overall accuracy reached by the entire system over the substances of interest is around 98.44% with a 0.93% of accuracy loss with respect to the multiclass classifier system alone (99.37%).

With the consideration made, we can say that the improvement made by putting an anomaly detection system before the multiclass classifier one has been proven. The entire system has been able to reject around the 80% of outliers samples and correctly recognize around 98.44% of inlier samples. Finally, given a real scenario application, it is clear that the presence of an anomaly detection module is of vital importance for the utility of the system itself.

In conclusion, after all the discussed analysis, we would like to report what we believe represents the major system's weaknesses. In particular, from the classification system point of view, as already discussed, one problem is related to all those substances which share some chemical properties, as stated before. Another weakness is related to sensor poisoning. More in detail, from some laboratory tests, it has been noted that SCW's sensors are particularly sensitive to acids substances, capable of poisoning SCW's sensors, inhibiting their ability to distinguish the different substances for a while. For sure, the poisoning problem is a complex one that can be caused by countless other substances, but, to the best of our knowledge, acids are the most difficult to be recovered. Eventually, another problem regarding the classification system is related to the computational complexity that tends to increase linearly with the number of learned substances, so as consequence, a suitable MCU has to be selected.

As the tests on the field concern, another system weakness related to the solid waste emerged, which can get stuck over the SCW's sensors altering all the measurements made, leading, in the worst cases, to a degradation of the overall system's performance.

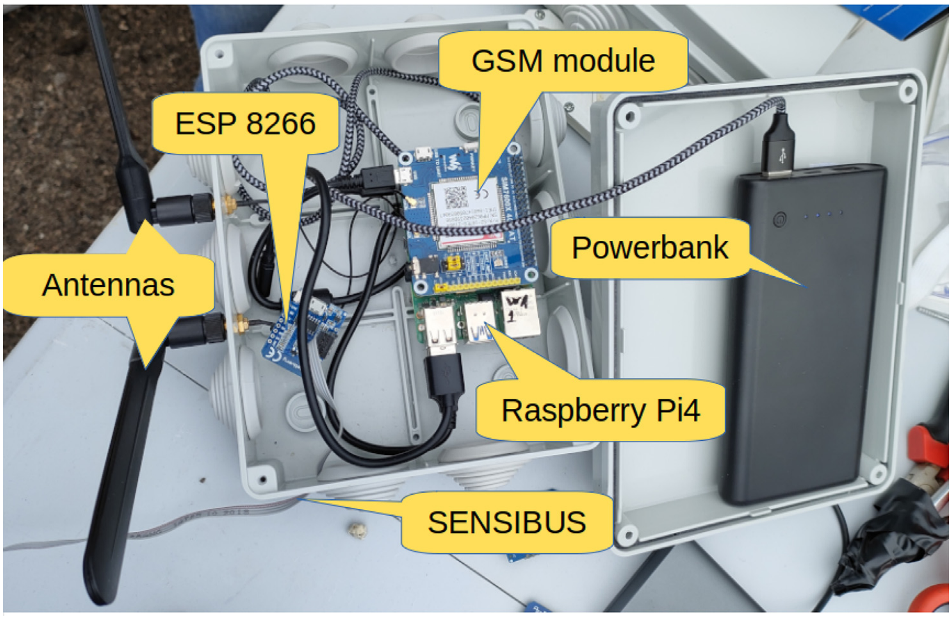


Fig. 16. Measurement system prototype.

Although the proposed algorithm could be applied to an array of sensors for real-time spilling recognition, to the best of our knowledge, there are no similar publicly available data sets and neither papers that follow a similar approach. This makes it impossible to provide other sets of experimental results other than those obtained from our data. At the same time, in the recent scientific literature, many papers face the problem of wastewater analysis. Still, nothing of these use an array of generic IDE in real-time recognition of different substances.

We also discarded the possibility of considering the signals during the injection as time series because their pattern over time would be highly dependent on the rate of injection of the substance (quantity per second), rather than its nature. We preferred an approach where the orthogonality of the adopted sensor, read "instantaneously", gives us the right information, not considering what happens in the time dimension.

As a final remark, it is known that Deep Learning (LeCun et al., 2015) Algorithms (DLA) are well-established in many fields. However, we don't have considered such an approach in this paper because the feature space that we obtained from sensors is very small. DLAs are well-established for time series, images, Natural Language Processing, etc., but their effectiveness is questionable when the feature space is too small. This assumption is also demonstrated in a previous paper, where Convolutional Neural Network (CNN) and Long-Short Term Memory models (LSTM) have been considered (Molinara et al., 2020), and where CNN and LSTM have been outperformed by traditional machine learning techniques like Multi-Layer Perceptron or kNN.

8. Conclusions and future directions

In conclusion, the proposed work has been meant to develop a stable and robust detection system capable of working in an aggressive environment such as the one represented by the sewage network. The complex environment implies that many different substances can be present, even those whose danger level is not significant and, therefore, not to be detected by the proposed system. Nevertheless, adopting a classical supervised ML approach, whatever substance would be recognized as one of those belonging to the training set. The important novelty carried out and proven effective in this work is implementing a two-stage scheme to strongly reduce false alarms and keep the classification accuracy very high. To do that, a Finite State Machine was intended to filter, process, and normalize the measured sensors data, and then a detection system was built. The detection system is divided into two main parts, one represented by the One-class SVM classifier, an anomaly detection algorithm with the purpose of rejecting all the samples belonging to the unknown substances, and one represented by the KNN multiclass classifier to recognize the given substance belonging to those of interest.

From the obtained results, shown in Section 5, it can be seen that the developed system work as supposed, drastically reducing the false positive errors given by the outlier samples and keeping accuracy on the substances of interest higher than 0.93 in all considered cases.

As concerns future developments, the authors would like to continue testing in real scenarios to validate in a quantitative way the promising results obtained in the laboratory activity and enforce the generalization property of the proposed system. Furthermore, the sodium hypochlorite confusion shown in Fig. 12, as discussed in Section 7, suggests to us that substances with some common chemical properties could be confused by the anomaly detection system. A possible way to reduce this phenomenon as much as possible would be to investigate an optimum set of orthogonal features that can exploit the chemical differences to maximize the overall system performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement SYSTEM No. 787128. The authors are solely responsible for it and that it does not represent the opinion of the Community and that the Community is not responsible for any use that might be made of the information contained therein. This work was also supported by MIUR (Minister for Education, University and Research, Law 232/216, Department of Excellence).

References

Akhter, F., Siddiquei, H.R., Alahi, M.E.E., Jayasundera, K.P., Mukhopadhyay, S.C., 2022. An iot-enabled portable water quality monitoring system with mwcnt/pdms multifunctional sensor for agricultural applications. IEEE Internet Things J. 9 (16), 14307–14316.

Alam, A.U., Clyne, D., Jin, H., Hu, N.-X., Deen, M.J., 2020. Fully integrated, simple, and low-cost electrochemical sensor array for in situ water quality monitoring. ACS Sensors 5 (2), 412–422.

Alavi, S.M.M., Mahdi, A., Payne, S.J., Howey, D.A., 2017. Identifiability of generalized randles circuit models. IEEE Trans. Control Syst. Technol. 25 (6), 2112–2120.

"Azienda comunale energia e ambiente." https://www.acea.it/, 2022. [Online: accessed 11-November-2022].

Bansal, S., Geetha, G., 2020. A machine learning approach towards automatic water quality monitoring. J. Water Chem. Technol. 42 (5), 321–328.

Betta, G., Cerro, G., Ferdinandi, M., Ferrigno, L., Molinara, M., 2019. Contaminants detection and classification through a customized iot-based platform: A case study. IEEE Instrument. Measur. Mag. 22 (6), 35–44.

Bogler, A., Packman, A., Furman, A., Gross, A., Kushmaro, A., Ronen, A., Dagot, C., Hill, C., Vaizel-Ohayon, D., Morgenroth, E., et al., 2020. Rethinking wastewater risks and monitoring in light of the covid-19 pandemic. Nat. Sustainab. 3 (12), 981–990.

Bourelly, C., Bria, A., Ferrigno, L., Gerevini, L., Marrocco, C., Molinara, M., Cerro, G., Cicalini, M., Ria, A., 2020. A preliminary solution for anomaly detection in water quality monitoring. In: 2020 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 410–415.

Bria, A., Cerro, G., Ferdinandi, M., Marrocco, C., Molinara, M., 2020. An iot-ready solution for automated recognition of water contaminants. Pattern Recogn. Lett. 135, 188–195.

Bria, A., Ferrigno, L., Gerevini, L., Marrocco, C., Molinara, M., Bruschi, P., Cicalini, M., Manfredini, G., Ria, A., Cerro, G., Simmarano, R., Teolis, G., Vitelli, M., 2021. A false positive reduction system for continuous water quality monitoring. In 2021 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 311–316.

Budiarti, R.P.N., Tjahjono, A., Hariadi, M., Purnomo, M.H., 2019. Development of iot for automated water quality monitoring system. In: 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp. 211–216.

Desmet, C., Degiuli, A., Ferrari, C., Romolo, F.S., Blum, L., Marquette, C., 2017. Electrochemical sensor for explosives precursors' detection in water. Challenges 8 (1), pp.

De Vito, S., Fattoruso, G., Esposito, E., Salvato, M., Agresta, A., Panico, M., Leopardi, A., Formisano, F., Buonanno, A., Delli Veneri, P., Di Francia, G., 2018. A distributed sensor network for waste water management plant protection. In: Andò, B., Baldini, F., Di Natale, C., Marrazza, G., Siciliano, P. (Eds.), Sensors, Springer International Publishing, Cham, pp. 303–314.

Dilmi, S., Ladjal, M., 2021. A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. Chemomet. Intell. Lab. Syst. 214, 104329.

Drenoyanis, A., Raad, R., Wady, I., Krogh, C., 2019. Implementation of an iot based radar sensor network for wastewater management. Sensors 19 (2), pp.

Dupont, C., Cousin, P., Dupont, S., 2018. Iot for aquaculture 4.0 smart and easy-to-deploy real-time water monitoring with iot. In: 2018 Global Internet of Things Summit (GIoTS). IEEE, pp. 1–5.

Farkas, K., Hillary, L.S., Malham, S.K., McDonald, J.E., Jones, D.L., 2020. Wastewater and public health: the potential of wastewater surveillance for monitoring covid-19. Current Opin. Environ. Sci. Health 17, 14–20. Environmental Health: COVID-19.

Ferdinandi, M., Molinara, M., Cerro, G., Ferrigno, L., Marroco, C., Bria, A., Di Meo, P., Bourelly, C., Simmarano, R., 2019. A novel smart system for contaminants detection and recognition in water. In: 2019 IEEE international conference on smart computing (SMARTCOMP). IEEE, pp. 186–191.

Hoes, O., Schilperoort, R., Luxemburg, W., Clemens, F., van de Giesen, N., 2009. Locating illicit connections in storm water sewers using fiber-optic distributed temperature sensing. Water Res. 43 (20), 5187–5197.

Ighalo, J.O., Adeniyi, A.G., Marques, G., 2021. Internet of things for water quality monitoring and assessment: a comprehensive review. Artificial intelligence for sustainable development: theory, practice and future applications, pp. 245–259.

Janna, H., 2016. Characterisation of raw sewage and performance evaluation of al-diwaniyah sewage treatment work, Iraq. World J. Eng. Technol. 4 (2), 296–304.

Ji, H.W., Yoo, S.S., Lee, B.-J., Koo, D.D., Kang, J.-H., 2020. Measurement of wastewater discharge in sewer pipes using image analysis. Water 12 (6), pp.

Junior, A.C.D.S., Munoz, R., Quezada, M.D.L.A., Neto, A.V.L., Hassan, M.M., Albuquerque, V.H.C.D., 2021. Internet of water things: A remote raw water monitoring and control system. IEEE Access 9, 35790–35800.

Kamaruidzaman, N.S., Rahmat, S.N., May 2020. Water monitoring system embedded with internet of things (IoT) device: A review. IOP Conf. Series: Earth Environ. Sci. 498, 012068. May.

Koditala, N.K., Pandey, P.S., 2018. Water quality monitoring system using iot and machine learning. In: 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE). IEEE, pp. 1–5.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–44.

Lepot, M., Makris, K.F., Clemens, F.H., 2017. Detection and quantification of lateral, illicit connections and infiltration in sewers with infra-red camera: Conclusions after a wide experimental plan. Water Res. 122, 678–691.

Lim, J, 2012. Mobile sensor network to monitor wastewater collection pipelines. https://escholarship.org/uc/item/0d9813bn, [Online: accessed 11-November-2022].

Lowe, M., Qin, R., Mao, X., 2022. A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. Water 14 (9), 1384.

Manfredini, G., Ria, A., Bruschi, P., Gerevini, L., Vitelli, M., Molinara, M., Piotto, M., 2021. An asic-based miniaturized system for online multi-measurand monitoring of lithium-ion batteries. Batteries 7 (3), pp.

Molinara, M., Ferdinandi, M., Cerro, G., Ferrigno, L., Massera, E., 2020. An end to end indoor air monitoring system based on machine learning and sensiplus platform. IEEE Access 8, 72204–72215.

Nopens, I., Capalozza, C., Vanrolleghem, P.A., 2001. Stability analysis of a synthetic municipal wastewater. Department of Applied Mathematics Biometrics and Process Control, University of Gent, Belgium.

Overmars, A., Venkatraman, S., 2020. Towards a secure and scalable iot infrastructure: A pilot deployment for a smart water monitoring system. Technologies 8 (4), 50.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Machine Learn. Res. 12, 2825–2830.

Pisa, I., Santín, I., Vicario, J.L., Morell, A., Vilanova, R., 2019. Ann-based soft sensor to predict effluent violations in wastewater treatment plants. Sensors 19 (6), pp.

”Public link for downloading the acquired dataset.” https://aida.unicas.it/data/JKSU_2022.zip, 2022. [Online: accessed 11-November-2022].

Ray, S., 2021. An analysis of computational complexity and accuracy of two supervised machine learning algorithms—k-nearest neighbor and support vector machine. In: Sharma, A., Chakrabarti, A., Balas, V.E., Martinovic, J. (Eds.), Data Management, Analytics and Innovation (N. Springer Singapore, Singapore, pp. 335–347.

Ria, A., Cicalini, M., Manfredini, G., Catania, A., Piotto, M., Bruschi, P., 2022. The sensiplus: A single-chip fully programmable sensor interface. In: International Conference on Applications in Electronics Pervading Industry, Environment and Society. Springer, pp. 256–261.

Saravanan, K., Anusuya, E., Kumar, R., Son, L.H., 2018. Real-time water quality monitoring using internet of things in scada. Environ. Monit. Assess. 190 (9), 1–16.

”Sewage monitoring system for tracking synthetic drug laboratories.” http://micromole.eu/, 2022. [Online: accessed 11-November-2022].

Trubetskaya, A., Horan, W., Conheady, P., Stockil, K., Merritt, S., Moore, S., 2021. A methodology for assessing and monitoring risk in the industrial wastewater sector. Water Resourc. Ind. 25, 100146.

Tyszczuk-Rotko, K., Kozak, J., Czech, B., 2022. Screen-printed voltammetric sensors – tools for environmental water monitoring of painkillers. Sensors 22 (7), pp.

Vanýsek, P., 2010. Electrochemical series.

Vikesland, P.J., 2018. Nanosensors for water quality monitoring. Nat. Nanotechnol. 13 (8), 651–660.

Zhao, Y., Nasrullah, Z., Li, Z., 2019. Pyod: A python toolbox for scalable outlier detection. J. Machine Learn. Res. 20 (96), 1–7.