Water Science & Technology

Water Science & Technology Vol 90 No 11, 3123 doi: 10.2166/wst.2024.387

Domain adaptation through active learning strategies for anomaly classification in wastewater treatment plants

Francesca Bellamoli [a,b,*], Marco Vian[b], Mattia Di Iorio[c] and Farid Melgani[a]

[a] Department of Information Engineering and Computer Science, University of Trento, via Sommarive 9, Trento 38123, Italy
[b] ETC Sustainable Solutions Srl, via dei Palustei 16, Trento 38121, Italy
[c] D-3 Srl, via dei Palustei 16, Trento 38121, Italy
*Corresponding author. E-mail: francesca.bellamoli@unitn.it

FB, 0000-0001-8774-3242

ABSTRACT

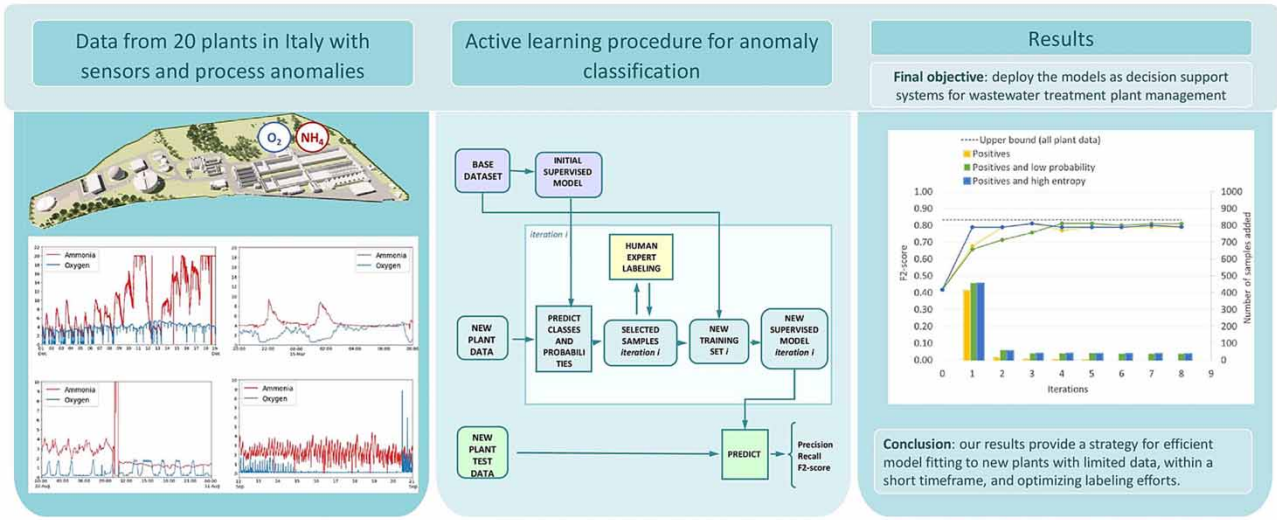
The increasing use of intermittent aeration controllers in wastewater treatment plants (WWTPs) aims to reduce aeration costs via continuous ammonia and oxygen measurements but faces challenges in detecting sensor and process anomalies. Applying machine learning to this unbalanced, multivariate, multiclass classification challenge requires much data, difficult to obtain from a new plant. This study develops a machine learning algorithm to identify anomalies in intermittent aeration WWTPs, adaptable to new plants with limited data. Utilizing active learning, the method iteratively selects samples from the target domain to fine-tune a gradient-boosting model initially trained on data from 17 plants. Three sampling strategies were tested, with low probability and high entropy sampling proving effective in early adaptation, achieving an F2-score close to the optimal with minimal sample use. The objective is to deploy these models as decision support systems for WWTP management, providing a strategy for efficient model adaptation to new plants, and optimizing labeling efforts.

Key words: active learning, domain adaptation, gradient boosting, intermittent aeration, multiclass classification, wastewater treatment plants

HIGHLIGHTS

- Exploring an active learning approach to adapt a pre-trained LGBM model to new WWTPs.
- Comparison of three active learning sampling strategies for anomaly classification.
- Detection of NH_4 sensor drift or offset, O_2 sensor fouling, and process inhibition.
- Training a decision support system for a new wastewater plant with very few data.
- Procedure to guide and optimize the labeling effort of wastewater process experts.

GRAPHICAL ABSTRACT



1. INTRODUCTION

The need for advanced control systems in wastewater treatment plants (WWTPs) has grown due to strict effluent quality regulations and cost reduction goals. Intermittent aeration controllers are effective in reducing aeration and recirculation costs by managing nitrification and denitrification within a single basin using continuous ammonia and oxygen measurements. However, they face challenges in detecting sensor failures and process anomalies.

Data from ammonia, oxygen, and other sensors used by intermittent aeration controllers have been shown to be valuable for training machine learning methods to classify sensor and process anomalies in WWTPs (Bellamoli *et al.* 2023), which can alert operators to take corrective action.

While numerous anomaly detection methods have been proposed (Corominas *et al.* 2018; Newhart *et al.* 2019), most focus on binary classification problems, where the goal is to distinguish normal data from sensor faults or generic anomalies. Few research deals with multiclass scenarios, like differentiating between low and high load or rain events or pinpointing multiple reasons for undesired discharge (Moon *et al.* 2008; Zhao *et al.* 2012; Chow *et al.* 2018). With the view of suggesting to the plant operator what action to take, an algorithm that classifies anomalies by clearly identifying the problem is certainly of great value.

Furthermore, many studies mentioned above use multiple input measurements, including influent flows and concentrations, which can effectively identify inconsistencies when combined with in-tank or outflow measurements (Elsayed *et al.* 2022). However, it is challenging to find such inlet sensors in real plants due to the practical difficulty of making reliable measurements on raw sewage.

Finally, research often lacks validation on real plant data. While some studies validate their anomaly detection systems on a simulation model of a standard WWTP (as in Ghinea *et al.* (2023)) or on pilot-scale plants, applications that apply the model to real plants are scarce; in particular, no applications can be found in which the model is tested on plants not included in the initial training dataset.

Collecting enough data from a WWTP could be a very long process, particularly for an unbalanced classification problem in which the normal class is highly prevalent compared to anomaly classes. For this reason, in our previous work (Bellamoli *et al.* 2023), instead of training a model for each WWTP, we trained a model using data from various plants, paying close attention to data normalization. The results of multiclass classification on data from new plants that were not included in the training dataset, however, were not satisfactory because it can be difficult to distinguish the correct type of anomaly based on a plant's unique characteristics, which can vary greatly from one plant to another. Examples of these characteristics include seasonality and oxygen regulation.

The solution to this issue can be domain adaptation, a branch of machine learning that minimizes the difference between the distributions of one or more labeled source domains in order to learn a model that can be generalized to a target domain

that is closely related (Farahani *et al.* 2020). In the case of WWTP anomaly classification, the different domains are the various plants, which differ in the distribution of features and classes, and the problem takes the form of a multi-source domain adaptation problem with single or multiple target plants. Active learning (AL) is a special case of domain adaptation where the classifier is adapted by repeatedly selecting a small number of carefully selected labeled samples from the target domain (Tuia *et al.* 2016). This approach addresses the challenge associated with the acquisition of labeled data, necessitating domain expertise and substantial time investment for the annotation process.

The prevailing findings in the existing literature and the growing adoption of AL in industry scenarios strongly indicate the effectiveness of AL methodologies (Settles 2012). Tharwat & Schenck (2023) performed a survey of recent studies on AL for classification, pointing out that AL has been widely used in remote sensing image classification (Persello & Bruzzone 2012; Tuia *et al.* 2016; Berger *et al.* 2021), handwritten text recognition and text classification (Romero *et al.* 2018; Schröder & Niekler 2020), and multimedia annotation and retrieval (Wang & Hua 2011).

In the context of anomaly detection, Pelleg & Moore (2004) used an AL model to find anomalies of special interest within the context of noisy data. More recently, Pimentel *et al.* (2020) tested an AL method that can be built upon existing deep learning models to separate outliers from normal data, and Pan *et al.* (2020) applied AL to anomaly detection of UAV sensor time series, which have few labeled data and many unlabeled data.

AL has been applied to environmental monitoring anomaly detection by Russo *et al.* (2020). They highlight the unique challenges of environmental data, including seasonal variations at various scales, non-stationarity, non-linearity, and the presence of diverse sensor faults. These distinctive characteristics pose greater challenges for applying anomaly detection in environmental contexts, as anomalies can be both infrequent and highly varied. These considerations also apply to wastewater data, for which no work has been published so far to study the performance of AL.

The aim of this work is to construct and train a machine learning algorithm capable of identifying and classifying the main anomalies in a WWTP operating with intermittent aeration, using only ammonia and oxygen data in the biological tank, and limiting the labeling cost. Since the model must be able to be trained on only a few data points available for the plant, it has to be adaptable for a new plant, starting from an initial model trained on a Base Dataset of multiple WWTPs.

Various AL strategies will be tested for applying a base model to three new real WWTPs, assessing the performance improvement gained by adding data to the model and the cost savings resulting from the reduction in the amount of data that needs to be labeled. A comparison will be conducted between the performance of the initial model, those achieved through AL, and those obtained with other approaches, such as using only plant data or labeling all available data.

2. METHODS

2.1. Problem and data description

This section gives a brief overview of the problem and of the data used. For more details, please refer to our previous work, described by Bellamoli *et al.* (2023).

The dataset underlying this study is derived from 51 trains of 20 WWTPs in Northern Italy. The data consist of minute-by-minute measurements of oxygen and ammonia levels in the biological compartment, which operates as an activated sludge system with intermittent aeration. In addition to the measurements, the aeration state is recorded every minute, indicating whether the system is in an aerated phase, where the air is given to nitrify, or a non-aerated phase, where the air is not given to denitrify. The study was conducted in collaboration with ETC Sustainable Solutions, a company that offers process control systems and support services to WWTP operators and conducts monitoring activities for over one hundred plants in Italy.

The choice to use measurements of ammonia and oxygen stems from the fact that these instruments are commonly found in all intermittent aeration systems based on ammonia. Other measurements, such as treated flow rate and suspended solids in the tank, are not available in all plants. Compressors' power measurements, on the other hand, can be readily obtained from the machines themselves. Still, in the present study, it was decided to exclude them, as features based on power do not enhance the model's applicability to new plants, given the significant variations in aeration systems from one plant to another (some have fixed-power systems, others have systems with inverters, others automated valves) and the difficulty in finding some parameter to be used for the feature normalization (normalizing with respect to the nominal power of the blowers, for example, is not advisable because there are plants that always operate at the nominal power, while others never reach it due to over-dimensioning of the blowers). This decision is supported by the analysis of feature importance presented in

the Supplementary Materials of Bellamoli *et al.* (2023), which underscores that power-based features rank among the least influential ones.

Single measurements of ammonia and oxygen, usually collected every minute, do not represent the state of the process and are not suitable for identifying anomalies. For this, the data are aggregated in an interval equal to the length of the intermittent aeration cycle, which can last from approximately one to 8 h. This allows the calculation of typical features of an intermittent aeration cycle, such as the anoxic fraction or the increasing and decreasing slope of ammonia.

The model presented in this paper is trained to classify each aeration cycle into one of the following classes:

- No anomaly
- Low ammonia probe drift
- Oxygen probe fouling
- Offset of the ammonia probe
- Inhibition of the biological process
- The high drift of the ammonia probe.

We remark that it is important to separate inhibition and high drift because the actions to be implemented as a consequence of one or the other are very different: both of them lead to extended aeration phases and high values of both ammonia and oxygen, but when the process is inhibited, the correct action to take is to aerate more, and conversely, when there is an ammonia probe drift, the correct action is to ignore the ammonia measurement and use a controller based only on oxygen. Due to the operators' greater experience in distinguishing between the two events and performing calibration tests on the ammonia probe when identifying a possible inhibition or drift, it was possible to label them separately, significantly increasing the system's usefulness.

In this work, the identification of high and low load events is not included in the model output. They are considered non-anomalies and are subsequently discriminated against by post-processing. This allows the machine learning algorithm to train specifically on the classes of anomalies that really matter.

In intermittent aeration processes, the trends in oxygen and ammonia are related to each other both by the biological processes taking place in the tank and by the action of the controller itself. The controller may be impacted by inaccurate ammonia or oxygen probe measurements, which could alter the tank's actual ammonia and oxygen concentrations. This can make anomaly identification extremely challenging, even for highly skilled technicians. Different plants have varying capabilities for regulating oxygen levels. Some have a fixed oxygen setpoint, while others can adjust dynamically. Some plants can regulate the power of the blowers, while others operate at a fixed-power level. Additionally, some plants use mixers, while others use pulsed air to mix the sludge during non-aerated phases (Bellamoli *et al.* 2023).

From these two points, it is easy to understand the challenge of adapting a trained model to a new plant with different characteristics, even though it still falls under the category of intermittent aeration activated sludge systems. Additionally, the difficulty in collecting data for the new plant is clear: the time required to obtain examples for each event class can be extensive, and the data labeling process can be labor-intensive and time-consuming. This highlights the need to develop AL algorithms that allow a model to be adapted to a new plant without having to label a large amount of data.

For the dataset of our interest, data from different plants were collected and labeled for varying periods from 2016 to 2023. For some plants, up to 2 years of data were collected; for others, only 6 months. For better model training, only the first 24 h of each anomaly were selected; cycles that started more than 24 h after the beginning of the anomaly were discarded. In this way, the model can focus on the beginning of the event and train itself to identify anomalies in a time that is useful for the plant operator (24–48 h). Besides, after 24 h, there is a risk of data pollution as the plant operator may have implemented compensatory measures that affect the data trend.

For the purpose of this work, it was decided to select three plants (those for which more data and more anomalies were available) for the testing of AL techniques. All other plants were grouped into a 'Base Dataset', which was used for training a starting model to be specialized later for each of the three test plants.

In Figure 1, we show for each plant the available number of cycles (samples), divided by event class. The plants of Calderara di Reno, Canegrate, and Castiglione Torinese were selected due to the presence of a higher number of cycles with anomalies. Canegrate was chosen instead of Pero to represent more types of anomalies. The characteristics of the three chosen plants are as follows:

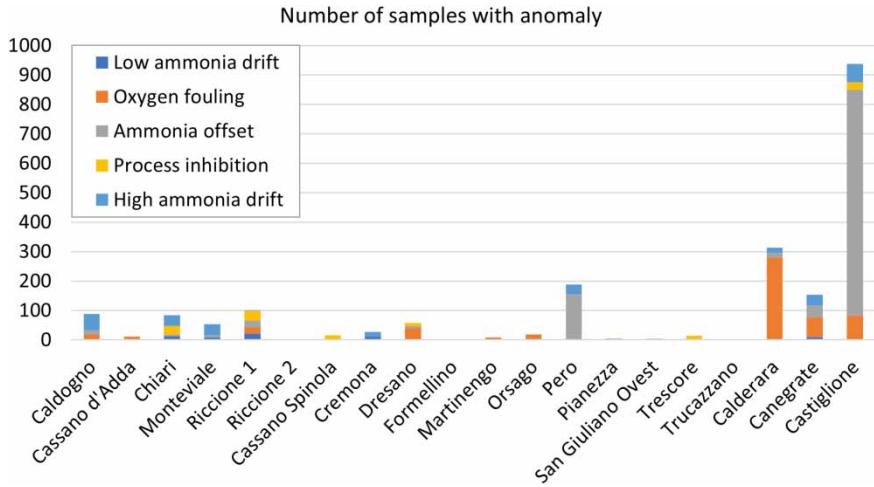


Figure 1 | Number of anomaly cycles per plant divided by type of event. The number of non-anomaly cycles is not visualized here.

- *Castiglione Torinese WWTP*: the biggest WWTP in Italy, it treats 3,800,000 population equivalent (PE) and is divided into four sub-plants and made up of 24 trains in parallel. In this work, we have data of only 18 trains. Since the influent is unique and the structure of the trains (volume, shape, and equipment) is similar, we can consider data from all 18 trains as originating from a unique plant.
- *Canegrate WWTP*: 139,000 population equivalent, divided into three biological trains.
- *Calderara di Reno WWTP*: 24,000 population equivalent, only one treatment train.

All these plants have a biological section based on an intermittent aeration process, which uses one ammonia and one oxygen measurement for each train. The trains operate in parallel and are independent of an aeration point of view but can be considered part of a single plant (domain) from the model training point of view.

Figure 2 shows the location of the chosen plants.

2.2. Features

Starting from ammonia and oxygen measurements, some features are calculated for each intermittent aeration cycle and are specifically designed for this type of process, based on the theoretical knowledge of the process and of the intermittent aeration controller and on the operators' experience in identifying anomalies. We implemented specific features to enhance the identification of a particular anomaly (e.g., ammonia skewness for the offset anomaly). Subsequently, the importance of each added feature in the model and the resultant improvement in model performance were thoroughly evaluated. Features that were unimportant or that did not bring improvement were discarded.

Developed features consider the length of the aerated and non-aerated phases compared to the maximum times set for that plant, and new features have been introduced to better distinguish the typical ammonia trend during the aerated phase in offset events.

The final list of features is outlined as follows:

- Average ammonia in the cycle, normalized (Ammonia_norm)
- Average oxygen during the aerated phase, normalized (OxSetPoint)
- Anoxic fraction of the cycle (fxt)
- Average of ammonia in the 24 h before the beginning of the cycle, normalized (Ammonia_r)
- Average of oxygen in the 24 h before the beginning of the cycle, normalized (Ox_r)
- Average ammonia slope during the non-aerated phase, which is indicative of the plant load. In fact, during the non-aerated phase, the rate at which ammonia rises is only determined by the incoming load (Ammonia_Up)
- Average ammonia slope during the aerated phase; during this phase, the slope is a composition of the inlet load and the bacteria ammonia uptake rate (Ammonia_Down)

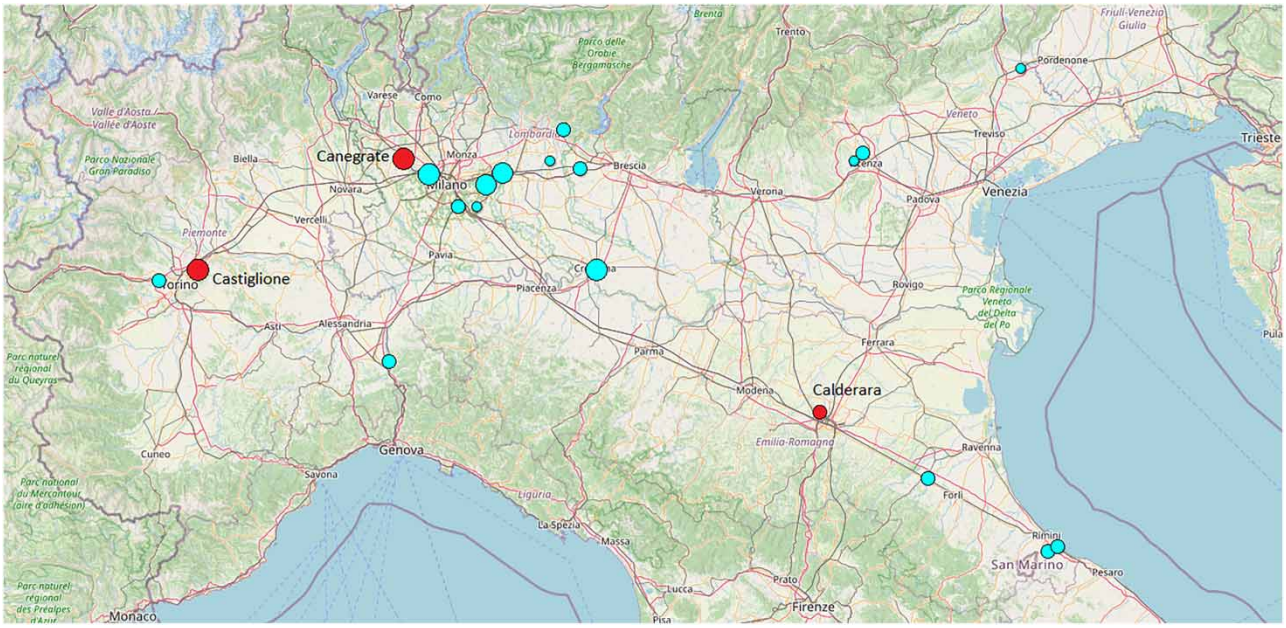


Figure 2 | Location of plants in Northern Italy. Big circles represent plants over 100,000 PE, medium circles represent plants between 10,000 and 100,000 PE, and small circles represent plants under 10,000 PE. Red circles are the plants chosen for the AL test.

- The slope of the ammonia in the last 15 min of the cycle. This feature has been introduced to help in the detection of offset events, which show, in the last part of the cycle, an ammonia slope near zero and a minimum ammonia greater than zero (Ammonia_Down_end)
- Maximum and minimum value of ammonia in the cycle (normalized and not normalized)
- Skewness of ammonia during the aerated phase. In normal cycles, the distribution of ammonia is symmetrical, as ammonia decreases at an almost constant rate. In cycles with ammonia offset, the distribution is highly unbalanced toward lower values because ammonia decreases rapidly at the beginning of the cycle and then remains stable near the offset value
- Length of the aeration phase with respect to the minimum and maximum duration of an aeration phase set in the controller (AerPhase)
- Length of the mixing phase with respect to the minimum and maximum duration of a mixing phase set in the controller (MixPhase)
- Difference between the maximum and minimum value of oxygen in the cycle, useful to recognize the oxygen sensor fouling (OxMaxMin)
- Difference between the minimum ammonia in the last 5 h and the high ammonia threshold (Ammonia_ThrHigh)
- Difference between the maximum ammonia in the last 14 h and the low ammonia threshold (Ammonia_ThrLow)
- Average of ammonia during an aerated phase when between the two ammonia thresholds set in the controller (normalized); this feature helps to identify an offset of the ammonia probe when this average is lower than in normal cycles (Ammonia_Thr)
- Day of the week
- Hour of the day: as this is a cyclical feature, it is converted into two features with a sine-cosine encoder
- Season.

As indicated above, some features are normalized so that their range of variation is comparable between the different plants. In Bellamoli *et al.* (2023), normalization is carried out based on historical measurement data for each plant. However, this approach is not easily applicable in practice, as it would require collecting a large amount of data from periods without anomalies each time the system is installed on a new plant. Instead, the algorithm proposed in this work bases the normalization on the parameters set in the intermittent aeration controller, collected continuously during its operation. These

parameters are: maximum and minimum ammonia threshold, oxygen setpoint, maximum and minimum length of aeration phase, and maximum and minimum length of mixing phase. This approach allows the model to better generalize across different plants by emphasizing deviations from the plant's desired behavior rather than absolute differences.

In Figure S5 of the Supplementary Materials, we show the feature importance calculated using SHapley Additive exPlanations values (Lundberg & Lee 2017), a model explanation method that can produce feature attributions for each individual instance and thus provide a more detailed and accurate understanding of model behavior.

The order of importance of features by anomaly confirms what is expected from process biology and knowledge of plant behavior, demonstrating the model's goodness and robustness. For instance, we observe that Ox_maxmin is crucial in identifying oxygen sensor fouling events (class 4) and that Ammonia_skew is essential in detecting ammonia offset events (class 2).

2.3. Machine learning method: gradient boosting

Since the problem is highly unbalanced (anomaly cases are much fewer than non-anomaly cases), the distribution of most features is skewed, and the use of ensembles based on decision trees is particularly suitable because each member of the decision tree can represent a part of the data distribution. In contrast, models with a generalizability objective would tend to annul these sparsely occurring values as outliers (Ching *et al.* 2021).

In accordance also to the results obtained by Bellamoli *et al.* (2023), we decided then to base our algorithm on a gradient boosting method (see Table S8 of Supplementary Materials for a resume of results obtained with different methods). These methods have been recently applied in a similar research field, water quality detection, by Xin & Mou (2022), with good performances. They have also been applied successfully in other environmental fields, such as hydrological modeling, groundwater management, scouring and sediment transport estimation, reservoir modeling, flow predictions, and hydroclimatic forecasting (Niazkar *et al.* 2024).

In the present work, we use the LightGBM method (Ke *et al.* 2017), which combines advantages like sparse optimization, parallel training, multiple loss functions, regularization, and bagging with a very fast training speed. We implemented the models in Python 3.9 using the following libraries: Scikit-Learn (Pedregosa *et al.* 2011), LightGBM Python API (Ke *et al.* 2017). For hyperparameters tuning, we used the Optuna library (Akiba *et al.* 2019), which automatically finds the optimal hyperparameter values by trial and error with a Bayesian optimization algorithm.

2.4. Evaluation protocol

In order to evaluate the model performances and compare the results of the pretrained model and of AL, we use the evaluation protocol developed by Bellamoli *et al.* (2023). A final test on the new plants is included, along with the selection of hyperparameters using a Bayesian optimization cycle and cross-validation. A cross-validation on the Base Dataset is also used to evaluate some performance metrics. Additionally, it establishes a performance metric that is in line with the algorithm's scope of application by combining cycles into events and evaluating whether the entire event is detected and whether it is detected within a timeframe that is helpful to the WWTP operator.

2.4.1. Performance metrics

It is advised to consider precision and recall instead of the accuracy metric when dealing with unbalanced classification problems. From a binary perspective (anomalous events versus normal data), the recall metric shows the likelihood of predicting true positives (TP) over the total events that occurred (true positives + false negatives (TP + FN)), while the precision metric represents the rate of positively predicted classes (true positives + false positives (TP + FP)) that are actually positive. The F_β-score (Van Rijsbergen 1975) is widely used to gauge model performance for imbalanced classification problems since it shows the balance between recall and precision.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

Other possible metrics include the Precision Recall Area Under the Curve (PR-AUC), representing the area under the precision-recall curve (focusing exclusively on TP and false negatives), and the Matthews correlation coefficient (MCC), which measures the quality of binary predictions by considering all values of the confusion matrix. In unbalanced datasets, the PR-AUC gives a clearer picture of how well the model can identify positive cases, whereas the Receiver Operating Characteristic

Area Under the Curve (ROC–AUC) can be misleading as it considers true negatives, which are often abundant in unbalanced data.

For the same reason, in unbalanced problems, the F2-score is preferable to the F1-score (the harmonic mean of precision and recall), as it emphasizes recall, which measures the model's ability to capture the minority class. The F2-score allows for a certain number of false positives to be tolerated in order to ensure the correct identification of as many TP as possible. While PR-AUC and MCC are useful for evaluating performance in imbalanced contexts, the F2-score specifically focuses on cases where recall is of major importance, making it a preferable choice when the priority is minimizing false negatives.

In the case under examination, the anomaly classification method must recover the majority of anomalies (recall of at least 75%) while producing an acceptable number of false positives. Given the data imbalance, where the number of anomalous events is much lower than in normal periods, a precision of 50% was considered adequate. This means that twice as many alarms are generated, which the operator will need to review compared to actual anomalies.

In light of all these considerations, the evaluation metric identified is the F2-score, which places a greater weight on recall than precision.

In multiclass classification problems, a way to calculate the model F2-score representing all classes is the macro-F2 value, calculated by averaging the F2-score of N categories. The same averaging method can be adopted for precision and recall.

2.4.2. Cross-validation

The LightGBM model has been cross-validated on the whole dataset, excluding the three plants chosen for AL, and the model's hyperparameters have been optimized to obtain the best F2-score.

In our analysis, we utilized a stratified k-fold cross-validation method with non-overlapping groups. This was done to ensure that cycles from the same event were not used in both the training and testing phases, while also maintaining the distribution of classes. We used the Base Dataset to tune the hyperparameters, dividing them into four folds, as done by Bellamoli *et al.* (2023). For each of the k-folds, a model is trained using $k - 1$ of the folds and validated on the remaining part of the data, calculating a performance metric. Afterward, the cross-validation reports the performance measure as the mean of the values that were calculated during the loop.

2.4.3. Confusion matrix based on number of events

The method's capacity to recognize events and the algorithm's value to plant operators would not be indicated by a performance evaluation based on the number of cycles found.

In Bellamoli *et al.* (2023), a methodology that considers an event detected only if it is detected in its first 48 h has been established in order to calculate a confusion matrix (matrix collecting TP, TN, FP, and FN) on events rather than on individual cycles. Once calculated, the confusion matrix, macro-averages of recall, precision, and the F2-score can be calculated. Recall and precision provide information about how many events the algorithm would miss and how many times it would raise a false alarm.

2.4.4. Final training and testing

Finally, after setting the hyperparameters through cross-validation, the model is trained using the entire Base Dataset and tested on the new plants. This test can evaluate the model's ability to adapt to data from different plants than those on which the training was carried out. Performance is calculated using the procedure described above.

In order to compare the results with those obtained through the application of AL techniques, we decided to divide the datasets of the three test plants into two parts: one part is set aside to be used as training through AL, and the other part is used for testing, both for the model trained on the Base Dataset and for each model obtained after every step of AL. In Table 5, the division between the train and test set for each plant is shown. We aimed to maintain a balanced number of events in both sets while dividing the dataset on a temporal basis, with the first period designated for training and the second for testing.

2.5. Active learning

AL is an incremental learning method that does not use all the new available samples to train the model but selects a part of them. The fundamental assumption of AL is that predicted probabilities of the classes contain information about the usefulness of a label for model training before that label becomes available. AL techniques aim to intelligently select the most

informative or uncertain training examples to enhance model performance. These techniques are valuable when collecting and manually labeling large amounts of training data, which can be costly or time-consuming.

Various types of AL methods exist, depending on the circumstances under which data are generated and how a domain expert is expected to provide labels. In this work, we use pool-based AL methods, which query from a complete set of unlabeled data samples already available before querying begins (Lewis & Gale 1994). AL methods may also differ in how the utility of a still-invisible label is estimated (Settles 2012): the most used techniques are uncertainty sampling (select the samples that the current model finds most uncertain), query by committee (select instances on which an ensemble of models disagrees the most), expected model change (choose instances expected to cause the most considerable change in the model's predictions), density-based sampling (select instances from regions of the feature space with low data density), cluster-based strategies, Bayesian methods. In this work, we use uncertainty sampling, a selection strategy where the AL model selects the samples that the current model finds most uncertain. There are several variants of uncertainty sampling, including capturing examples with high entropy (the uncertainty of the model's predictions between the classes), capturing examples with low confidence (lowest predicted class probability), and margin maximization (chose samples where the difference between the top two predicted class probabilities is small).

In this study, we decided to explore three labeling strategies:

- *AL with positive sampling*: since we are dealing with an anomaly detection problem and the anomalies are much fewer in number than the non-anomaly class, it is considered useful to let an operator check all the anomalies identified by the model; moreover, in the case of an online implementation of the system, the operators will necessarily have to check each anomaly reported by the implemented model in order to apply any necessary countermeasures. This is why we decided to always include the positives identified by the model among the samples to be added to the training set.
- *AL with low probability sampling*: in addition to the samples of the previous point, the η% of samples that have the lowest predicted class probability are selected; this strategy is also called the least confident approach: the active learner queries the point with a low posterior probability (Tharwat & Schenck 2023).
- *AL with high entropy sampling*: in addition to the samples that are predicted as anomalies, the η% of samples with the highest entropy of the predicted classes probabilities are selected.

In all three strategies, all samples for which the model predicts an anomaly are used. This choice is because the method is designed for the real use of the model on a plant, where the operator must necessarily check every anomaly alarm and record whether it is confirmed or not. This choice also makes it possible both to add as many anomaly samples as possible to the training set (which is especially important in the case of rarer anomalies) and to correct the model in the case of a high number of false positives.

The last two strategies are absolutely similar from an operational point of view. From the point of view of sample selection, the probability-based strategy only considers information about the most likely class and neglects the information about the rest of the distribution, while the entropy method takes all classes into account (Tharwat & Schenck 2023); however, the probability-based strategy also selects cases in which the model is uncertain over a restricted subset of classes (often two classes), in contrast to the entropy-based strategy, which selects samples with a higher entropy (indecision among a large number of classes).

The AL procedure can be summarized as follows: the initial set of data samples (Base Dataset) is used to train the initial supervised model. This model is applied to the new pool of samples of the chosen plant, and the classification results are used to select a small number of samples from this pool (with one of the above-mentioned strategies). The domain expert is asked to provide a label for the selected samples. The model is then retrained on the original dataset with the new labeled samples and applied to the new pool of unlabeled samples. The process is repeated, and the performance is evaluated on the test set at each iteration. The process stops when the model reaches satisfactory performance or no further performance improvement is observed. Another stopping criterion could be a cost criterion: the procedure is repeated until the labeling budget is reached (Settles 2012). In our case, the maximum cost was identified as 4 days of work by an expert operator for large plants (at least 200,000 PE, at least 8 trains) and 1 day of work for small plants. The whole AL procedure is outlined in Figure 3.

During real-time operation, the system will send an alarm to the plant operators whenever an anomaly is identified. Then, at defined time intervals (e.g., 3 months), the system will choose the η% of samples with the highest entropy (or lowest probability) and send the operator a request to label them; it will then retrain the model, make a new prediction, and select new samples for labeling.

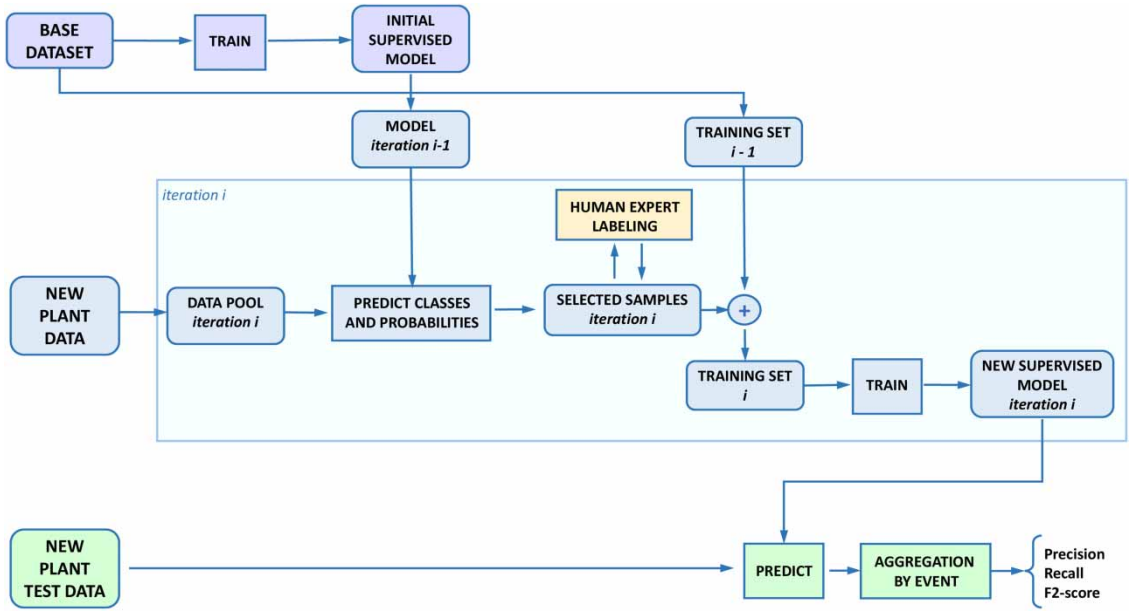


Figure 3 | Scheme of the AL procedure.

3. RESULTS AND DISCUSSION

3.1. Cross-validation and hyperparameters tuning on Base Dataset

The LightGBM model described in Section 2.3 was tested on Base Dataset by cross-validation, evaluating the events with the procedure described by Bellamoli *et al.* (2023), and optimizing the hyperparameters to obtain the best F2-score. Subsequently, the resulting model, retrained on all Base Dataset samples, was tested on the test parts of the single Calderara, Canegrate, and Castiglione datasets.

The hyperparameter tuning was performed with the Optuna library (Akiba *et al.* 2019), using a Bayesian optimization algorithm. The final hyperparameters are the following (the names are those used in the Scikit-Learn library): estimators: 308; max depth: 9; num leaves: 29; min data in leaf: 72; feature fraction: 0.6; learning rate: 0.13; lambda 12: 41; min gain to split: 0.0025.

The results of cross-validation are shown in Table 1. Macro-recall results 0.74, macro-precision 0.64, and macro-F1-score 0.71. In addition to these metrics, recall, precision, and F2-score values were also calculated in 'binary' mode, thus summing up all TP found for the classes (regardless of misclassification between the different anomaly classes), all false negatives (first column of Table 1), and all false positives (first row of Table 1). This results in a recall of 0.77, a precision of 0.56, and an F2-

Table 1 | Confusion matrix for model cross-validated on Base Dataset

		Predicted								
		No anomaly	Low ammonia drift	Oxygen fouling	Ammonia offset	Process inhibition	High ammonia drift	Recall	Precision	F2-score
Actual	No anomaly	419	1	8	20	10	7	0.9	0.96	0.91
	Low ammonia drift	3	3	0	0	0	0	0.5	0.75	0.54
	Oxygen fouling	2	0	8	0	0	0	0.8	0.5	0.71
	Ammonia offset	7	0	0	19	0	0	0.73	0.49	0.66
	Process inhibition	3	0	0	0	12	1	0.75	0.5	0.68
	High ammonia drift	3	0	0	0	2	14	0.74	0.64	0.71

score of 0.71. The inhibition and high drift classes, which were not differentiated in Bellamoli *et al.* (2023), are only confused by the algorithm in three cases.

These results are good and comparable to those obtained in Bellamoli *et al.* (2023), considering that here we are focusing on the most difficult event classes, that are the probe fault classes and the inhibition class, excluding high and low load classes, and that the Base Dataset contains more plants than the Dataset number 1 used in our previous paper. Our goal here is to have a pre-trained and not overfitted model to use as a base model on which to construct a plant-specialized model through AL.

Finally, we report in Table 2 the results obtained for the three test plants with the model trained on the whole Base Dataset.

As can be seen in Tables 2, S1–S3, the performance is rather poor, especially for Calderara, where 10 oxygen fouling events are not identified, and for Canegrate where, conversely, eight false positives of oxygen fouling are reported and two low drift events of the ammonia probe are not detected. In the case of Canegrate, the oxygen measurement is often very low, not due to probe fouling but due to aeration system issues. On the contrary, in Calderara, oxygen levels are typically high, so an operator would notice a dirty probe when he observes values slightly lower than the norm but still high compared to the Base Dataset average. Furthermore, there were only six events of low ammonia probe drift in the Base Dataset, which suggests that the algorithm trained on this dataset might need more data to recognize this anomaly effectively. Last, the Castiglione Torinese plant exhibits numerous events of ammonia probe offset. The higher frequency of these events compared to the other plants is attributed to the unique operation mode of this plant, which operates with very low ammonia levels and prolonged aerated phases, allowing even small offsets to be identified.

The idea behind the specialization of the model with new plant data arises precisely from this basis, i.e., the differences in behavior between one plant and another due to the plant's intrinsic characteristics. However, given the onerousness of the labeling procedure, adding months of labeled data from a new plant would not be sustainable. In this work, we therefore test techniques for selecting the most meaningful data to be labeled and added to the dataset to train a model for the new plant.

3.2. Results of AL

In this paragraph, we summarize the results obtained from the application of the three AL strategies described in Section 2.5 to the plants of Calderara, Canegrate, and Castiglione. We performed eight AL iterations, which correspond to 4 days of labeling work for the Castiglione plant and one day for the other two plants. However, it can be observed that, after only 3–4 iterations, no further performance improvement is observed.

In Figure 4, we report the improvement on the F2-score obtained along the AL iterations for the three methods tested for the Calderara WWTP. We also report the quantity of labeled samples (aeration cycles) added in each iteration. The F2-score increases from 0.31 to 0.75 with the 'low probability' method and to 0.78 with the 'high entropy' method. The strategy of labeling only the positives detected by the algorithm is not effective in this case because the number of anomaly events detected by the initial algorithm is very low, so little information is added to the model.

In Table 3, we report the new confusion matrix obtained after all the iterations with the 'higher entropy' method. If we compare it with that in Table 2, we can see that the recall of oxygen fouling events has definitely improved, with only a slight worsening in precision. An example of a fouling event that was not detected by the initial algorithm and is detected by the final Calderara model is shown in Figure S1.

In Figure 5 and Table 3, we show the results for Canegrate WWTP. We can see that the F2-score increases from 0.42 to 0.79 with the 'high entropy' and 'only positives' method and to 0.81 with the 'low probability' method. Here, the initial model detects many false positives, so the difference between the three methods is not so high. From Table 3, we can see a reduction of false positives of oxygen sensor fouling (an example is shown in Figure S2 of Supplementary Materials) and the detection of two low ammonia drift events that were not detected by the initial model (Figure S3).

Table 2 | Performances for model trained on Base Dataset and tested on Calderara, Canegrate, and Castiglione

	Calderara	Canegrate	Castiglione
Recall	0.27	0.50	0.56
Precision	0.80	0.25	0.59
F2-score	0.31	0.42	0.57

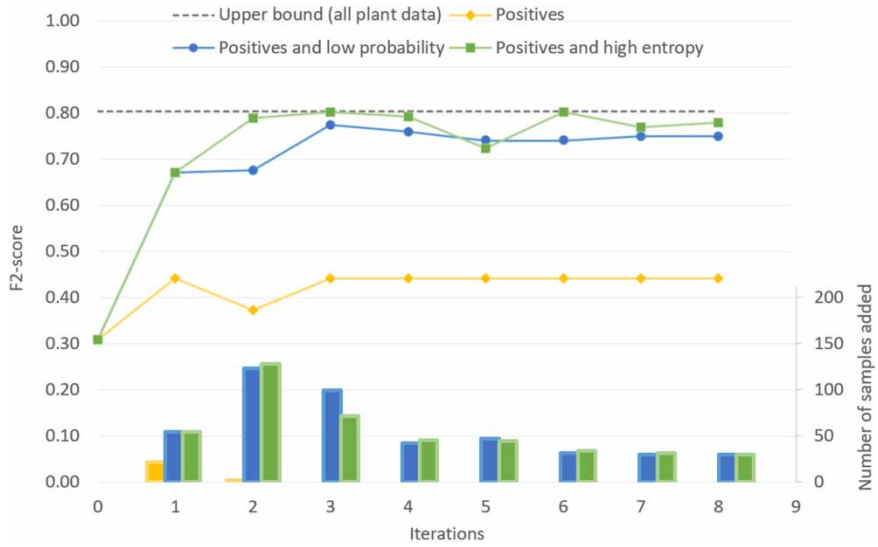


Figure 4 | Results of AL on Calderara WWTP. Lines and dots represent the F2-score along iterations, and bars represent the samples added at each iteration.

Table 3 | Performances for models tested on Calderara, Canegrate, and Castiglione before and after AL with the high entropy method

	Calderara		Canegrate		Castiglione	
	Before AL	**After AL**	**Before AL**	**After AL**	**Before AL**	**After AL**
Recall	0.27	0.80	0.50	1.00	0.56	0.77
Precision	0.80	0.71	0.25	0.43	0.59	0.51
F2-score	0.31	0.78	0.42	0.79	0.57	0.70
F1-score	0.40	0.75	0.33	0.60	0.57	0.61
PR-AUC	0.38	0.61	0.24	0.53	0.51	0.59
MCC	0.39	0.67	0.33	0.64	0.53	0.58

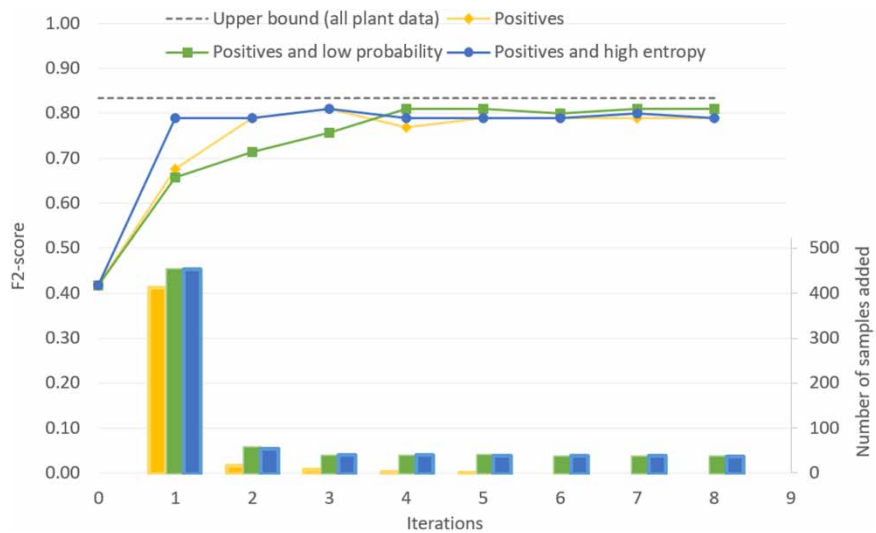


Figure 5 | Results of AL on Canegrate WWTP. Lines and dots represent the F2-score along iterations, and bars represent the samples added at each iteration.

In Figure 6 and Table 3, we also show the results for Castiglione WWTP: the F2-score increases from 0.57 to about 0.70 with all three methods. We can see an improvement in the recall of offset events (with an acceptable reduction in precision) and a general improvement in distinguishing an anomaly from one another (ammonia offset, high ammonia drift, and inhibition). An example of an offset event that was not detected by the initial algorithm and is detected by the final Calderara model is shown in Figure S4 of Supplementary Materials.

In Table 3, we also report the values of metrics such as the F1-score, the PR-AUC, and the MCC.

3.2.1. Comparison with other re-training strategies

Finally, we present in Table 4 the comparison between F2-scores obtained with (1) a pre-trained model (on Base Dataset); (2) an AL model; (3) a model trained only on the plant data; and (4) a model trained on Base Dataset plus all the plant's data. The results show that with AL we can obtain performances very similar to those achievable by adding all the plant's data to the training set, but with considerable time and resource savings in labeling. It can also be seen that using only single plant data does not yield such good results; moreover, not all anomalies may be present in the training set (see Table 5), thus preventing the identification of missing anomalies in the future.

As a final point, we would like to remark that an alternative strategy, not tested in this work, could be to continuously calculate the probability in real-time and send the operator an alarm and a request for labeling whenever the probability is lower than a fixed threshold. This method would allow the operator to label the samples the same day with a fresher memory of what happened in the plant. However, it is not possible to have control over the number of samples that will be requested to be labeled, risking overloading the operator. Furthermore, also in this case, a period of time must be defined for re-training

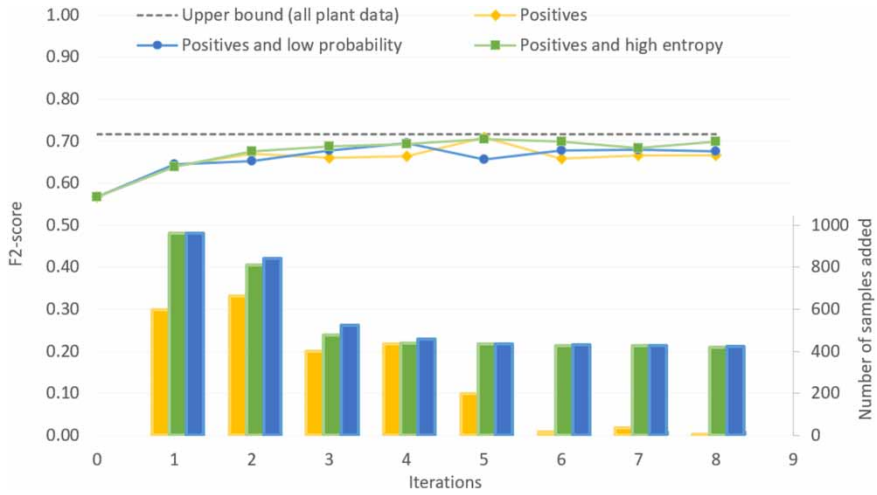


Figure 6 | Results of AL on Castiglione Torinese WWTP. Lines and dots represent the F2-score along iterations, and bars represent the samples added at each iteration.

Table 4 | F2-scores obtained for plants under investigation using only Base Dataset as training, or only plant data, or all data, or a selection of data based on AL (with the high entropy method)

	Calderara		Canegrate		Castiglione	
Training dataset	**F2-score**	**Training samples**	**F2-score**	**Training samples**	**F2-score**	**Training samples**
Base Dataset	0.31	45,134	0.42	45,134	0.57	45,134
Only plant data	0.51	3,262	0.74	4,731	0.67	46,331
Base Dataset + all plant's data	0.80	48,396	0.83	49,865	0.72	91,465
Base Dataset + plant's cycles selected with AL	0.78	45,569	0.81	45,893	0.70	49,654
Percentage of plant's samples added with AL	13%		16%		9%	

Table 5 | Number of events for each class in training and test sets

	Calderara		Canegrate		Castiglione	
	Train (11 months)	Test (11 months)	Train (6 months)	Test (8 months)	Train (8 months)	Test (6 months)
Low ammonia drift	0	0	1	2	0	0
Oxygen fouling	11	12	3	1	5	2
Ammonia offset	0	3	5	2	51	38
Process inhibition	0	0	0	0	3	5
High ammonia drift	2	0	1	1	3	3

the model; re-training after each new labeling would be burdensome in terms of computational resources. For these reasons, we decided not to include this strategy in the test.

3.2.2. Discussion on rare events

Dealing with imbalanced classes is a significant challenge. In the case of AL, waiting for rare events to occur in order to improve prediction ability is not always feasible. Therefore, we need initial models that already ensure satisfactory handling of those classes; in our case, we start with a model with an F2-score for the 'low ammonia drift' class of 0.54, as shown in Table 1.

In fact, in the case of our gradient-boosting model (using ensembles based on decision trees is particularly suitable for highly unbalanced problems) and the dataset in question, it was possible to obtain an acceptable baseline. In addition, we remark that we used balanced weighting between the classes in the LightGBM model.

In cases where this could not be achieved, the literature refers to various techniques, besides the class weighting, to overcome the problem, such as:

- Modification of the loss function to favor rarer classes (using a weighted cross entropy loss function)
- Data augmentation techniques, e.g., SMOTE, ADASYN, Imbalance Generative Adversarial Network, Gaussian mixture models (Babu & Narasimha Rao 2023). Additionally, Hancock & Khoshgoftaar (2021) proposed a fraud detection method that uses CatBoost and LightGBM to encode categorical data.

3.2.3. Limitations of AL approach

The AL approach is highly adaptable, allowing the model to easily adjust to different plant setups, including changes in the aeration system. Such modifications do not present a challenge, as the model focuses on data patterns rather than specific aeration types, and oxygen trends do not vary significantly with different aeration systems. Moreover, our experience shows that the type of sensors used does not significantly affect model performance. The use of different ammonia sensors (analyzers instead of ion-selective probes) can change measurement frequency, but this does not impact the model if at least one reading is taken every 10 min, as overall features of the entire aeration cycle are calculated. Fewer offset events occur with ammonia analyzers, but this does not limit the model's ability to detect anomalies.

Conversely, domain adaptation might fail when there are fundamental differences between plants in terms of instruments used by the process controller and of the operational logic of the controller itself. Basically, the domain adaptation fails when the changed conditions make the initial model not applicable, e.g., when there is a change of features because of changed measurements. For example, the model is not applicable in plants with continuous aeration or in plants where the intermittent aeration is based only on oxygen or on other measurements like oxidation-reduction potential (ORP) or NO_3. Still, with such a radical change in conditions, the anomalies to be detected could also change completely. The study of these other types of processes is certainly interesting and could be the subject of future work.

4. CONCLUSIONS

We tested three AL techniques for applying a base, pre-trained, model to three new real WWTPs. The model is based on the LGBM gradient-boosting method and has been initially trained on a dataset of over 45,000 samples collected from 17 plants. Then, this model has been adapted using data of 6 11 months from the three plants of interest. We tested AL with positive

sampling, AL with low probability sampling, and AL with high entropy sampling, always keeping a part of the plant's dataset as the test set for the evaluation of the results.

Our results demonstrate that both low probability sampling and high entropy sampling methods perform well across all three plants, achieving an F2-score very close to the upper bound (score obtainable using all the data) as early as the second iteration, which means using on average only the 6% of available samples. This results in considerable time savings for sample labeling: two labeling iterations correspond to less than 2 days of work for the Castiglione plant (the most articulated, consisting of 18 trains) and about half a day for the other two plants.

This indicates that it is feasible to adapt a model to a new plant by collecting data for only about 6 months and then applying AL with a labeling effort of a few hours or days, depending on the plant size.

Our study is positioned in a landscape of literature in which we find no application of AL in the wastewater field and little regarding machine learning for anomaly classification in WWTPs. By drastically reducing the labeling burden, our study presents a useful approach to overcome one of the main barriers that have hindered the widespread adoption of machine learning in this field.

The ultimate goal of the developed models is to be used by plant operators as decision support systems. Through this work, we have given the basis for a procedure to train a decision support system for a new plant with a very scarce amount of data, defining a simple strategy to guide and optimize the labeling conducted by experienced operators.

In this work, we focused on AL methods, but future work should explore other domain adaptation techniques, which generalize a pre-trained model to a closely related domain by minimizing the difference between domain distributions (e.g., domain invariant feature extraction or adversarial-based adaptation, Farahani *et al.* (2020)). Those methods should still allow model adaptation with limited labeling (semi-supervised or unsupervised domain adaptation). Another interesting future research could be the use of stream-based AL instead of pool-based AL, with the perspective of continuous learning during the use of a decision support system applied to a WWTP.

AUTHOR CONTRIBUTIONS

F. B. conceptualized and investigated the whole process, rendered support in data curation, validated the work, prepared the software, wrote the original draft, wrote the review and edited the article. M. V. supervised the study, wrote the review and edited the article. M. D. I. rendered support in data curation and feature engineering. F. M. supervised the study, developed the methodology, wrote the review and edited the article.

ETHICS STATEMENT

Ethics: Human participants. No human participants. Ethics: Animal testing. No animal testing.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors have a conflict to declare.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. (2019) 'Optuna: A next-generation hyperparameter optimization framework', *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '19*. New York, NY, USA: Association for Computing Machinery, pp. 2623–2631.

Babu, K. & Narasimha Rao, Y. (2023) A study on imbalanced data classification for various applications, *Revue D Intelligence Artificielle*, **37**, 517–524. 10.18280/ria.370229.

Bellamoli, F., Di Iorio, M., Vian, M. & Melgani, F. (2023) Machine learning methods for anomaly classification in wastewater treatment plants, *Journal of Environmental Management*, **344**, 118594.

Berger, K., Rivera Caicedo, J. P., Martino, L., Wocher, M., Hank, T. & Verrelst, J. (2021) A survey of active learning for quantifying vegetation traits from terrestrial earth observation data, *Remote Sensing*, **13** (2), 287.

Ching, P. M., So, R. H. & Morck, T. (2021) Advances in soft sensors for wastewater treatment plants: A systematic review, *Journal of Water Process Engineering*, **44**, 102367.

Chow, C., Liu, J., Li, J., Swain, N., Reid, K. & Saint, C. (2018) Development of smart data analytics tools to support wastewater treatment plant operation, *Chemometrics and Intelligent Laboratory Systems*, **177**, 140–150.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U. & Poch, M. (2018) Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques, *Environmental Modelling and Software*, **106**, 89–103.

Elsayed, A., Siam, A. & El-Dakhakhni, W. (2022) Machine learning classification algorithms for inadequate wastewater treatment risk mitigation, *Process Safety and Environmental Protection*, **159**, 1224–1235.

Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H. R. (2020) A brief review of domain adaptation, Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020. *arXiv*, 877–894.

Ghinea, L., Miron, M. & Barbu, M. (2023) Semi-supervised anomaly detection of dissolved oxygen sensor in wastewater treatment plants, *Sensors*, **23**, 8022.

Hancock, J. T. & Khoshgoftaar, T. M. (2021) Gradient boosted decision tree algorithms for medicare fraud detection, *SN Computer Science*, **2** (4), 1–12.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process Syst.*, **30**, 3149–3157.

Lundberg, S. & Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. 10.48550/arXiv.1705.07874.

Lewis, D. D., Gale, W. A., (1994) A sequential algorithm for training text classifiers. In: Croft, B. W. & van Rijsbergen, C. J. (eds.) *SIGIR '94*, London, UK: Springer London, pp. 3–12.

Moon, T., Kim, Y., Kim, J., Cha, J., Kim, D.-H. & Kim, C.-G. (2008) Identification of process operating state with operational map in municipal wastewater treatment plant, *Journal of Environmental Management*, **90**, 772–778.

Newhart, K., Holloway, R., Hering, A. & Cath, T. (2019) Data-driven performance analyses of wastewater treatment plants: A review, *Water Research*, **157**, 498–513.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P. & Righetti, M. (2024) Applications of xgboost in water resources engineering: A systematic literature review (Dec 2018–May 2023), *Environmental Modelling and Software*, **174**, 105971.

Pan, D., Nie, L., Kang, W. & Song, Z. (2020) 'Uav anomaly detection using active learning and improved S3VM model', *2020 International Conference on Sensing, Measurement Data Analytics in the era of Artificial Intelligence (ICSMD)*, pp. 253–258.

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, *Journal of Machine Learning Research*, **12**, 2825–2830.

Pelleg, D., Moore, A., (2004). Active learning for anomaly and rare-category detection. In: Saul, L., Weiss, Y. & Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 17. Vancouver, BC, Canada: MIT Press.

Persello, C. & Bruzzone, L. (2012) Active learning for domain adaptation in the supervised classification of remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, **50**, 4468–4483.

Pimentel, T., Monteiro, M., Veloso, A. & Ziviani, N. (2020) 'Deep active learning for anomaly detection', *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

Romero, V., Sánchez, J. A. & Toselli, A. H. (2018) 'Active learning in handwritten text recognition using the derivational entropy', *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 291–296.

Russo, S., Lürig, M., Hao, W., Matthews, B. & Villez, K. (2020) Active learning for anomaly detection in environmental data, *Environmental Modelling and Software*, **134**, 104869.

Schröder, C. & Niekler, A. (2020) A survey of active learning for text classification using deep neural networks. arXiv:2008.07267. doi: 10.48550/arXiv.2008.07267.

Settles, B. (2012) *Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning*. San Rafael, CA, USA: Morgan Claypool Publishers.

Tharwat, A. & Schenck, W. (2023) A survey on active learning: State-of-the-art, practical challenges and research directions, *Mathematics*, **11** (4), 820.

Tuia, D., Persello, C. & Bruzzone, L. (2016) Domain adaptation for the classification of remote sensing data: An overview of recent advances, *IEEE Geoscience and Remote Sensing Magazine*, **4**, 41–57.

Van Rijsbergen, C. (1975) *Information Retrieval*. London, UK: Butterworths.

Wang, M. & Hua, X.-S. (2011) Active learning in multimedia annotation and retrieval: A survey, *ACM Trans. Intell. Syst. Technol.*, **2**, 1–21.

Xin, L. & Mou, T. (2022) Research on the application of multimodal-based machine learning algorithms to water quality classification, *Wireless Communications and Mobile Computing*, **2022**, 1–13.

Zhao, L.-j., Chai, T.-y., Diao, X.-k., Yuan, D.-c., (2012) Multi-class classification with one-against-one using probabilistic extreme learning machine. In: Wang, J., Yen, G. G. & Polycarpou, M. M. (eds.) *Advances in Neural Networks – ISNN 2012*, Berlin, Heidelberg, Germany: Springer, pp. 10–19.

First received 8 August 2024; accepted in revised form 12 November 2024. Available online 27 November 2024