Water Research 206 (2021) 117695

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

The value of human data annotation for machine learning based anomaly detection in environmental systems

Stefania Russo [a,b,*], Michael D. Besmer [c], Frank Blumensaat [a,e], Damien Bouffard [a], Andy Disch [a], Frederik Hammes [a], Angelika Hess [a,e], Moritz Lürig [a,h,i], Blake Matthews [a,h], Camille Minaudo [f], Eberhard Morgenroth [a,e], Viet Tran-Khac [g], Kris Villez [a,d]

[a] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[b] ETH Zürich, Ecovision Lab, Photogrammetry and Remote Sensing, Zürich, Switzerland
[c] onCyt Microbiology AG, Zürich, Switzerland
[d] Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[e] ETH Zürich, Institute of Environmental Engineering, 8093 Zürich, Switzerland
[f] École Polytechnique Fédérale de Lausanne, Physics of Aquatic Systems Laboratory, Margaretha Kamprad Chair, Lausanne, Switzerland
[g] INRAE, Université Savoie Mont Blanc, CARRTEL, 74200 Thonon-les-Bains, France
[h] Eawag, Department of Fish Ecology & Evolution, Centre for Ecology Evolution and Biogeochemistry, 79 Seestrasse, 6047, Luzern
[i] Department of Biology, Lund University, 22362 Lund, Sweden

ARTICLE INFO

Keywords:
Machine learning
Anomaly detection
Environmental systems
Labels

ABSTRACT

Anomaly detection is the process of identifying unexpected data samples in datasets. Automated anomaly detection is either performed using supervised machine learning models, which require a labelled dataset for their calibration, or unsupervised models, which do not require labels. While academic research has produced a vast array of tools and machine learning models for automated anomaly detection, the research community focused on environmental systems still lacks a comparative analysis that is simultaneously comprehensive, objective, and systematic. This knowledge gap is addressed for the first time in this study, where 15 different supervised and unsupervised anomaly detection models are evaluated on 5 different environmental datasets from engineered and natural aquatic systems. To this end, anomaly detection performance, labelling efforts, as well as the impact of model and algorithm tuning are taken into account. As a result, our analysis reveals the relative strengths and weaknesses of the different approaches in an objective manner without bias for any particular paradigm in machine learning. Most importantly, our results show that expert-based data annotation is extremely valuable for anomaly detection based on machine learning.

1. Introduction

Over the past decade, there has been a drastic increase in the amount of data collected for environmental monitoring in both engineered and natural systems. In order to manage this vast quantity of data, dedicated tools are required; this includes implementing standard procedures for sensor validation, data importation and storage, as well as data retrieval. Furthermore, it also becomes necessary to ensure high data quality. Anomaly Detection (AD) aims to identify unusual patterns in the data

* Corresponding author at: ETH Zürich, Ecovision Lab, Photogrammetry and Remote Sensing, Zürich, Switzerland.
E-mail address: stefania.russo@geod.baug.ethz.ch (S. Russo).

https://doi.org/10.1016/j.watres.2021.117695
Received 28 April 2021; Received in revised form 7 September 2021; Accepted 20 September 2021
Available online 27 September 2021

that do not conform to the typical behaviour of the system.

Anomalies can also be categorised according to their context: a *point anomaly* is a single measurement that is considered anomalous, irrespective of the time or location of the measurement. A *contextual anomaly* is a single data point that is considered anomalous when taking the time or location of measurement into account. E.g., the data point could be acceptable considering the sensor measurement range but may still be atypical for the considered location or the considered time of day or time in the year. A *collective anomaly* is a collection of measurements, produced by different instruments, at different locations or at different times, which individually appear normal but are found anomalous when analysed jointly. Usually, point anomalies can be detected automatically using some basic quality methods as out-of-range checks (Chandola et al., 2009). Collective and contextual anomalies on the other hand, may need application-specific rules (Ramanathan et al., 2006) or visual inspection of the data by a domain expert. Both can be tedious and time consuming, if not impossible, to implement in many cases. To solve this problem, there has been an increasing interest in automated AD techniques for environmental monitoring, as well as for other domains such as medical imaging, intrusion detection, and so on. These techniques, based on Machine Learning (ML), are more flexible than the classical mechanistic models approaches (Ciavatta et al., 2004) as they can take advantage of data-rich environments. Specifically, ML models support domain experts by automatically detecting unusual patterns or samples thus removing the need for manual inspection.

ML models can be broadly divided in supervised and unsupervised models. Supervised models are calibrated (i.e. trained) using the data with associated labels. Most typically, the labels are the ideal typical/anomalous classification provided by a human expert. After calibration, the models can be used to predict the target label for each new unseen data point. A drawback of supervised methods is that they require large and representative datasets for which both input data and the associated labels are available. In practice, obtaining target labels is difficult and time-consuming because it requires domain expert reviewing the each data point and assigning a binary (normal or anomalous) label to it. For this reason, unsupervised models, which do not require labels, are commonly studied as an alternative to supervised models (Aguado et al., 2007; Aguado and Rosen, 2008; Aguado et al., 2005; Alferes et al., 2013; Baert et al., 2012; Lee et al., 2005; Lee and Vanrolleghem, 2003; Rosén and Lennox, 2001; Villez et al., 2008).

In recent years, academic research has applied a wide variety of tools for anomaly detection using both ML settings. For supervised ML, techniques such as support vector machines (Ni et al., 2011) and artificial neural networks (Hill and Minsker, 2010) for wind speed sensor data streams have been used. For unsupervised ML, self-organizing maps (Postolache et al., 2005), and clustering techniques (Bezdek et al., 2011) have been applied. In (Inoue et al., 2017), unsupervised deep neural networks were evaluated. One-class support vector machines (OCSVM) are used as well and are often referred to as an unsupervised learning method (Amer et al., 2013; Liu et al., 2014). However, OCSVM is calibrated with an anomaly-free dataset for model calibration, thus requiring an expert-based separation between anomalous and normal data records in the data used for calibration. For this reason, we consider the calibration of one-class models a special case of supervised classification (Hendrycks et al., 2019; Khan and Madden, 2014; Sillito and Fisher, 2008).

ML models for AD in environmental monitoring are often presented as a single one-size-fits-all tool (Trilles et al., 2017). Unfortunately, most of these methods have been evaluated only on one dataset or with a limited number of models (Fuente et al., 1996; Lee and Vanrolleghem, 2003; Rosén and Lennox, 2001; Yu, 2012), thus inviting scepticism with respect to the reported benefits. For example, in (Muharemi et al., 2019) the authors test several supervised models on one dataset and, based on performance results, choose one to solve the challenge of AD for water quality. The burden of data labelling for supervised AD is not described. In (Candelieri, 2017) the authors use supervised learning (specifically support vector regression) for water demand forecasting and AD. The approach is validated on one urban water network in the city of Milan. In (Inoue et al., 2017), a deep neural network is proposed for AD in a water treatment system, however the approach was tested only against a one-class support vector machine model and on one dataset. In, (Miau and Hung, 2020), a supervised deep learning approach is proposed to detect abnormal water levels and perform river flooding forecasting. However, the method was only tested on one dataset.

Although one particular model may outperform another for a specific domain (see e.g. Corominas et al., 2011; Leigh et al., 2019), there is no guarantee that reported performance levels can be extrapolated to other systems or other tasks. This challenge has been discussed in several works (Corominas et al., 2018; Garrido-Baserba et al., 2020; Gibert et al., 2018; 2010). In response to this need, (Corominas et al., 2011) benchmark 5 different univariate techniques for fault detection in wastewater treatment processes, however only in a simulation environment. In this sense, benchmarking refers to the evaluation and comparison of ML methods on different datasets; benchmarking results can be then used as standards for subsequent studies. One benchmarking study at full scale is reported in (Lee et al., 2008), although it is limited to 5 cases of the same unit process. To this day, these two studies remain the most comprehensive benchmarking studies concerning data validation in the urban water cycle.

It is therefore evident that: *i)* current works are almost exclusively limited to results within a single case; *ii)* the usage of supervised, unsupervised, one-class setups is often performed casually without mentioning the burden of data labelling and/or model tuning, and *iii)* a comprehensive, comparative analysis of commonly applied ML models for different environmental applications is missing. This is especially true in environmental datasets which may present a number of important challenges not encountered elsewhere. In fact, environmental datasets are characterised by a high level of heterogeneity in data, since they are produced by a variety of instruments, with noted differences in format, resolution, and quality. Also, environmental systems present high levels of complexity because of interactions between several components (climate, humans, animals etc.) producing effects that are not understood well. The processes are often claimed to be highly nonlinear, and can exhibit stochastic or cyclic behaviours. These appear at different scales, depending on the specific application. As a result, case-specific results are difficult to generalise.

With this paper we compare different supervised, one-class, and unsupervised AD models on several environmental datasets from engineered and natural aquatic systems. With this evaluation, we specifically aim to assess the relative advantages of data annotation by human experts, while also accounting for the strengths and weaknesses of the different model structures, their performance, as well as their sensitivity to manual parameter tuning. In addition, this paper also enables effective benchmarking of data-driven anomaly detection models by publishing both the code and the datasets. This is the first time such a step is done in environmental applications, where, to make real progress, benchmarking of the already existing models is direly needed.

2. Methods

2.1. Machine learning

In this section, we describe the different ML setups and the procedures for calibration and testing applied to all models under study. Finally, we explain the performance metrics used to evaluate these models.

2.1.1. Data preprocessing

Data preprocessing is an important step in ML practice for a number of reasons: (Kotsiantis et al., 2006) first, these operations can improve the chances and the rate of convergence to optimal parameter values; second, they can affect the modelled relationship when using models

based on distance measures (like k-Nearest Neighbours); third, this ensures that the data matches the expected format for a particular model (see e.g., Gurden et al., 2001) fourth, data preprocessing allows to incorporate expert knowledge (through feature engineering) into the ML model. In this work, we center and scale each input variable separately to zero mean and unit variance prior to model calibration. In order to ensure that we evaluate the utility of ML methods, and not the utility of domain expertise, we have also carefully eliminated any knowledge-based feature engineering. The main reason is that such feature engineering could improve performance of the ML models, thus confounding our evaluation of the ML setups. We only applied feature engineering in one case study (onCyt) where each sample was transformed into structured data by means of binning. This step was needed to transform the data into a format suitable for the considered ML models and performed in accordance with the domain expert's instruction. This is described in detail in the Supporting Information.

2.1.2. Model calibration setups for anomaly detection

The main goal of calibration is to identify a model's parameters that enable the prediction of the class, y, for any data point represented as a d-dimensional input vector $\mathbf{x}$ containing the values for d variables measured at the same time. In AD applications, the class y can only

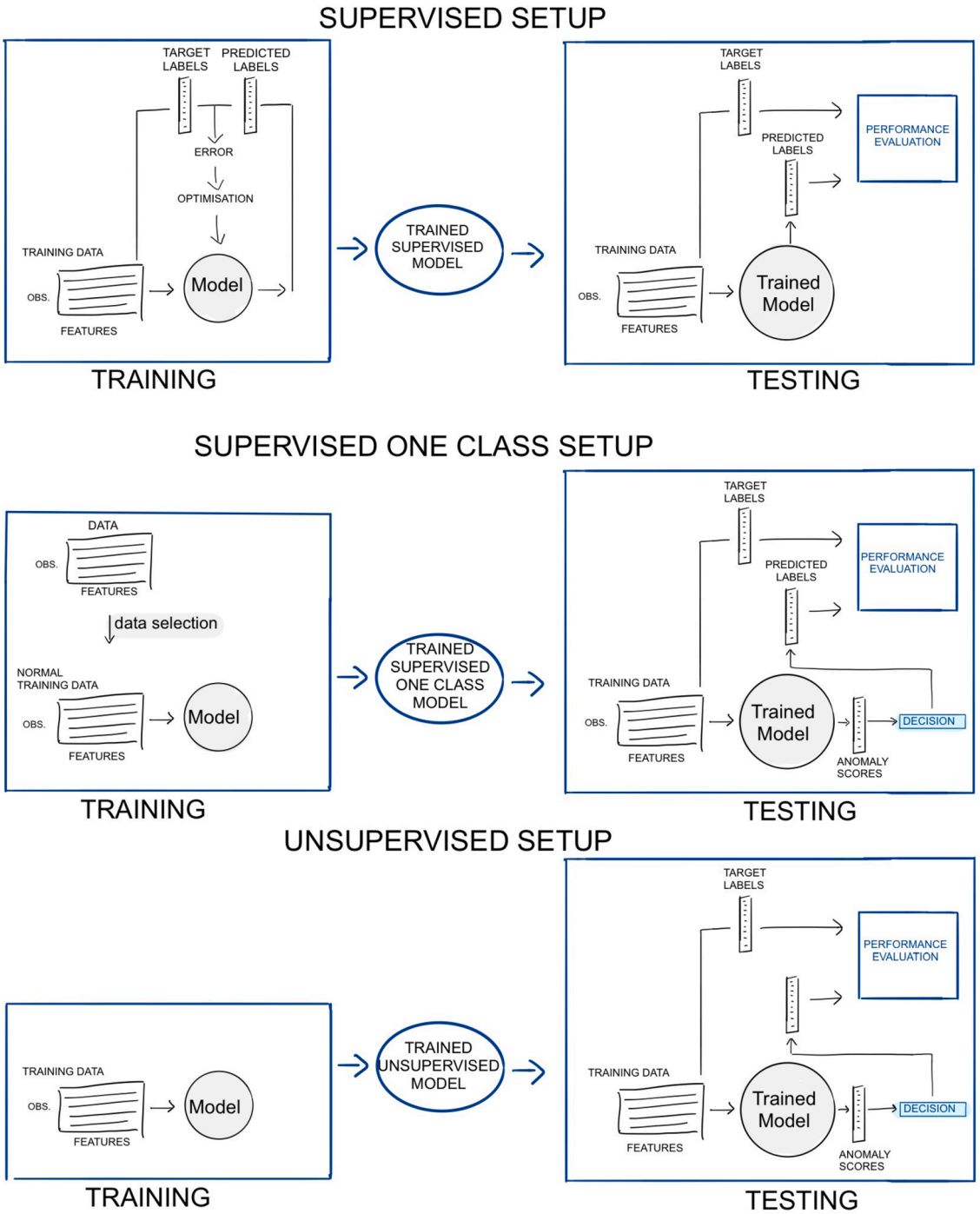

Fig. 1. Different anomaly detection setups depending on the availability of labels in the dataset. The supervised setup uses a labelled dataset for calibration. The one-class supervised setup uses an anomaly-free calibration dataset for calibration while the unsupervised setup does not require any labels. The test dataset, which has to be labelled in all cases, is used to compute performance metrics.

acquire two label values, that is: $y \in \mathscr{Y} = \{0,1\}$. Here, the normal data is generally named as negative class, with label value 0, and the anomalies are named positive class, with label value 1.

For our purposes, we group the selected models into three ML setups based on the data and target labels y_i that are available for calibration. These setups are illustrated in Fig. 1 and described next.

1. **Supervised learning (SUP)** is the setup where each calibration data point is paired with a target label y_i. It is also known as binary classification. During calibration of a supervised AD model, the model ϕ is calibrated with the calibration dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ to predict the label. These predictions are compared with the target labels and the parameters of the model are adjusted to minimise discrepancies. During testing, the calibrated model ϕ predicts a probability value between 0 to 1 for each new unseen test data point in the test dataset $\mathscr{T}$, indicating how likely the data belongs to the anomalous class. A threshold, α, is used to convert the probability to a binary prediction $\hat{y}$ (normal/anomaly). If the predictive probability is higher than this threshold, the test data point is considered anomalous, otherwise normal.
2. **One-Class supervised learning (OCSUP)** is a special case of SUP where the calibration dataset only consists of anomaly-free data (Chalapathy et al., 2018; Ruff et al., 2018). While labels are not strictly required as an input during calibration, this setup assumes that all data points in the calibration dataset belong to the negative class. Once the model is identified, it is used to evaluate how much a new data point, $\mathbf{x}^*$, deviates from the calibration data. The measure of deviation is known as the anomaly score. Similarly to the SUP setup, a threshold is also set to convert the anomaly score to a binary prediction. Setting this threshold is typically more subjective as the anomaly score cannot be interpreted as a probability.
3. **Unsupervised learning (UNSUP) setup** is a setup where the AD model scores the data solely based on the patterns in the calibration dataset $\mathscr{D}$ and uses the complete calibration dataset without any target label (Hastie et al., 2009; Leung and Leckie, 2005). This eliminates a tedious examination of the data set available for calibration. Similarly to OCSUP, the model produces an anomaly score for every new data point and a threshold is set to convert this to a binary prediction (normal/anomaly). The UNSUP is expected to work best if the calibration data are contaminated with a negligible or non-existent amount of anomalies.

As noted in the introduction, the distinction between UNSUP and OCSUP learning is rarely provided in the AD literature (Amer et al., 2013; Sabokrou et al., 2018). Quite often, this implies that the task of removing anomalies from the calibration data has been executed perfectly. With the above definitions, we stress that the use of a supervised model (being either binary or one-class) requires the input of a domain expert.

2.1.3. Anomaly detection models

In this work, we benchmark a wide range of ML model structures. The main purpose of this is to ensure that the quantified utility of expert-provided labels does not hinge on the choice of a particular model structure. The following criteria were used for selection of these paradigms:

- The selected model structures should be fairly popular in the ML community.
- The selected model structures should be available in on-the-shelf softwares (e.g. scikit-learn, Matlab ML Toolbox, Spark MLlib, Weka) and the selected implementation should be compatible with the Apache License 2.0.
- The selected model structures should be available for each of the above learning setups. As an example, we include a random forest as a binary classifier and an isolation forest for the corresponding one-class models.

Five model structures were found that satisfy these requirements. They are based on five distinct paradigms, which are *(a)* mixture-of-Gaussians, *(b)* nearest neighbours, *(c)* ensembles of decision trees, *(d)* models based on the kernel trick and support vectors, and *(e)* feed-forward artificial neural networks. These model structures are listed in Table 1. Each model type can be configured with different hyper-parameters. Hyperparameters are those parameters that determine the detailed structure and flexibility of the models. These do not change during model calibration and are typically set based on prior knowledge, experience, and/or exploratory analysis of the calibration data. The chosen models and their hyperparameters are discussed in detail in the Supporting Information.

2.2. Performance evaluation

Evaluating the performance of AD models in real-world practice is not as straightforward as in typical text-book examples. The main challenge is that we are dealing with an unbalanced dataset that is, the classes are not represented equally. Indeed, the fraction of anomalies in the studied datasets is fairly low. As a result, detection accuracy metrics (e.g. ratio of correctly identified data samples vs. total number of data samples) should not be used (Géron, 2019). The selected performance metrics are described next.

ROC curve It is important to note that AD models, as defined above, deliver a class membership or an anomaly score, which needs to be compared to a threshold value to determine the predicted class. This means that the resulting model performance critically depends on the choice of threshold. In order to evaluate all models fairly, irrespective of the chosen threshold, we evaluate the model performance for every candidate threshold, rather than picking one. This results in multiple values for the true positives rate (TPR) and the false positives rate (FPR), which change monotonically with the threshold. These values are computed as: $TPR = TP/(TP+FN)$ and $FPR = FP/(FP+TN)$. Where true positives (TP) is the number of correctly identified anomalies ($y = 1, \hat{y} = 1$), true negatives (TN) is the number of correctly classified normal data ($y = 0, \hat{y} = 0$), false negatives (FN) and false positives (RP) are respectively the number of incorrectly classified normal data ($y = 0, \hat{y} = 1$) and anomalies ($y = 1, \hat{y} = 0$).

A Receiver Operating Characteristic (ROC) curve is then computed for each model by plotting the TPR against the FPR (Maxion and Roberts, 2004). In AD applications, we would like our models to learn a complete separation of the two underlying distributions (normal and anomalous). A ROC curve corresponding to better separation is located closer to the upper left-hand corner in the ROC space and thus enables visualisation of the model performance. One advantage of ROC curves is that they enable visual and quick assessment of the performance of the classifier over its entire operating range (Fawcett, 2006).

We remind the reader that a single model corresponds to a unique combination of calibration setup, model structure, and model

Table 1
Anomaly detection models used for benchamarking, arranged according to their learning setup.

Supervised	One-class supervised	Unsupervised
Linear, Quadratic Discriminant Analysis and Naive Bayes (DANB)	Gaussian Mixture Model (GMM)	Gaussian Mixture Model(GMM)
k-NearestNeighbours (kNN)	Local Outlier Factor (LOF)	Local Outlier Factor (LOF)
Random Forest (RF)	Isolation Forest (IF)	Isolation Forest (IF)
Support Vector Machine (SVM)	OCSVM	Unsupervised OCSVM
Artificial Neural Network (ANN)	Autoencoder(AE)	Autoencoder(AE)

hyperparameters. The list of evaluated hyperparameters can be found in the Supporting Information.

Pareto ROC curve In addition to the ROC for single models (i.e., unique set of hyperparameter values), we also compute a Pareto ROC by selecting those points for which the FPR and TPR values are the highest, and therefore optimal among all models sharing the same ML setup. This new ROC is the upper envelope of the individual ROC and thus presents the best results one can obtain through systematic hyper-parameter tuning. This leads to 15 Pareto ROCs (5 case studies, 3 setups) and gives a general insight into the utility of the three calibration setups but ignores differences in sensitivity to the selected model structure and the tuning of the hyper-parameters.

Area under the ROC curve Additionally, the Area Under the Curve (AUC) is chosen to summarize the model performance into a single measure (Fawcett, 2006), which ranges from 0 to 1. Excellent models, with a good measure of separability, have an AUC near 1. For our purposes, the AUC is computed first for each model. Then, the mean and standard deviation for these AUCs is reported for each of the 15 model structures. By reporting the averaged AUC as well as the standard deviation for each model we can assess both average performance and sensitivity of the model structures to hyperparameter tuning. Although the AUC is a helpful indicator of model performance, it is still an aggregate measure, which excludes details shown in the ROC. For this reason, our main conclusions will be based on the ROCs and Pareto ROCs described above.

2.3. Case studies

The datasets used for this study stem from a set of five engineered and natural systems that are currently operational at Eawag, Dübendorf (Switzerland). These infrastructures exhibit several common challenges such as mechanical sensor faults, the presence of biological processes with their nonlinearity and seasonality and the presence of natural stochastic events. On the other hand, they also present a diverse set of measured variables, time resolutions, and spatio-temporal correlations. To the best of our knowledge, this is the first time that such a diverse set of domains, covering a wide variety of data types present in the water sector (seasonal/non-seasonal, urban/environmental, water quantity/quality, etc.) are used for benchmarking of ML models within one study. A summary about the datasets characteristics is given in Table 2. All datasets have been made publicly available at https://doi.org/10.25678/0002WJ. The code developed for labelling the data has also been made available at https://doi.org/10.25678/0002PC. The datasets are described next with particular focus on their type of anomalies and variables used for model calibration. Specific details regarding data collection and labelling for each case can be found in the Supporting Information. Note that a sixth domain (LéXPLORe) was also studied within this study. However, the provided labels do not reflect an unbiased opinion of the domain experts since they were provided by an automated tool. Consequently, we have ignored this dataset for benchmarking and discuss the results in the Supporting Information.

Eawag ponds This is a multivariate time series data of high spatial and temporal resolution that was collected as part of a long term ecological experiment described in (Lürig et al., 2020; Narwani et al., 2019). The experiment used pond ecosystems (hereafter ponds). 16 of such ponds were set up at Eawag Dübendorf (Switzerland) with macrophytes (Myriophyllum spicatum) and dreissena mussels (Dreissena polymorpha) were added progressively to the ponds together with inorganic nutrients. Each of these ponds was equipped with sensors for eight variables: conductivity, chlorophyll and phycocyanin fluorescence, dissolved organic matter fluorescence, dissolved oxygen (saturation and concentration), pH, and temperature. Measurements of these water parameters were recorded simultaneously in each multi-sensor, with a fixed time interval of 15 min. As for the variables used for model training, our domain expert has supplied us with the anomaly labels for specific conductivity, and we have trained our models based on this sole information. The domain expert manually labelled anomalies for specific conductivity presenting both spiky and long term anomalous events. This resulted in a labelled data set containing $N = 22464$ data records with $d = 8$ variables (the measured water quality variables). The data set covers a period of 234 consecutive days and includes 2% anomalies.

Table 2
Statistics from the five datasets used to benchmark the different machine learning models.

Domain	Observations	Variables	Anomalies	Anomalies (%)
Ponds	22,464	8	2117	2.34
onCyt	1148	81	117	10.2
UWO S1	14,545	3	2551	16.2
UWO S2	20,979	3	4793	27.0
WaterHub	436,320	4	58,249	13.3

UWO S1 mainly shows (simpler), point and contextual anomalies, while UWO S2 presents (more complex), collective type of anomalies.

Online flow cytometry This dataset is a time series of high temporal resolution and was collected as part of multiple applied research projects on microbial monitoring (Besmer et al., 2016; 2014). The measurements were carried out on water samples directly extracted from water streams by an automated flow cytometry system. Flow cytometric analysis of water samples was typically carried out every 15 min after sample preparation including tagging of bacteria with fluorescent dyes. Based on experience, 7 categories of potential anomalies were defined and individual flow cytometric measurements were labelled manually by the leading expert user. The domain expert reported that most of the time he was confident about the labelled anomaly type. After labelling, all anomalous labels were grouped as one single anomaly to enable anomaly detection with binary classification, resulting in a total of 10.2% anomalies.

Urban water observatory (UWO) This case study presents data from a long term in-sewer process monitoring initiative in the municipality of Fehraltorf, near Zurich in Switerland (www.eawag.ch/uwo). The data used in this study are in-sewer flow rates which had been recorded during a period of 1.5 years with a temporal resolution of 5 min. Two datasets covering the periods 26 March - 25 April 2017 and 1 September - 12 November 2017 were selected and manually labelled by two domain experts. This resulted in two labelled data sets (namely UWO S1 and UWO S2). It is important to note that, differently to UWO S1, the UWO S2 data set was filtered to eliminate simple, point anomalies before labelling to shift the focus to collective anomalies. Therefore, UWO S1 mainly shows simple, contextual anomalies, while UWO S2 presents complex, collective type of anomalies.

Water Hub In the Water Hub, researchers from Eawag investigate sustainable, decentralized, and source separated waste water treatment. Greywater collected from wash basins and showers within the building is treated with a two-step process: a membrane bioreactor (MBR) is followed by a biological activated carbon (BAC) filter. We focused on the data measured with four pressure sensors placed in the MBR and in the BAC (respectively two in each process). These sensors are installed to monitor the water levels to control the system and to keep track of the hydraulic performance of the system. Data from all four sensors was analysed to detect process anomalies, which were: *i)* foam in the MBR, *ii)* clogging of the membrane, *iii)* blocked level switch in the MBR, *iv)* pressure loss drop in the BAC, and *v)* lack of data. The domain expert manually labelled process anomalies for the MBR and the BAC separately resulting in a labelled data set containing $N = 436320$ data records, of which 13.3% are anomalies, with $d = 4$ variables (the pressure variables). The data set covers a period of 10 months.

3. Results

In this section we show the benchmarking results obtained for each model and domain. First, we focus on the Pareto ROCs which give a

broad overview of the results of each ML setup. This is followed by a detailed discussion of the AUC results for every model in every domain. We conclude the section with the results related to model selection. Note that all test were run without the use of a GPU.

3.1. Pareto ROC curves

The Pareto ROCs in Fig. 2 are obtained by selecting the Pareto-optimal combinations of the FPR and TPR values obtained for all models for a given calibration setup, i.e. across *(a)* all model structures, *(b)* all applied hyper-parameter settings for each model structure, and *(c)* each threshold for a given model. In most cases, the SUP models perform better than the OCSUP and UNSUP models. In one case (onCyt), the OCSUP models show the same performance as the SUP models. OCSUP models show only slight improvements over the UNSUP models in all cases, except onCyt and UWO S2.

It can also be seen that, although the UWO S1 and S2 domains belong to the same domain of application (i.e. the measured variables are the same), the presence of different types of anomalies, which was discussed with our domain experts, makes a substantial difference in the performance of the AD methods.

3.2. AUC and sensitivity to hyperparameters

To evaluate whether it is important to select a particular set of hyperparameter values, we have summarised each ROC by means of the average AUC and the standard deviation for each model structure and for each domain under study. These are visualised in Fig. 3. In general, SUP models outperform the others in all domains, which corroborates our findings on the importance of having labels to achieve the best possible performance in AD applications. The same results are quantitatively reported in the form of summary statistic in the Supporting Information. Note that some results related to the SVM implementation are not present due to the high computation time of the models (calibration time $>$ 10h). This is due to the large size of the kernel matrix. Where feasible, only the linear (lin.) kernel functions for SVM was implemented. Inspecting Fig. 3 makes clear that ML models may suffer from a high degree of intra-domain variability, in addition to inter-domain variability. Indeed, the UWO domain cases (S1 and S2) are very similar in the nature of the data but produce dissimilar model performance levels. We speculate that this is due to different types of anomalies in the two datasets: after discussing with our domain expert, we concluded that UWO S1 contains more simpler point and contextual anomalies, while UWO S2 contains a high number of collective anomalies. Importantly, this means that extrapolating the performance of ML models from one domain to the next is likely to remain difficult, even if such domains are considered similar.

3.3. Model related ROC curves

Finally, in Fig. 4 we show the Pareto ROC curves in a unique combination of domain and model structure. These plots visually highlight the difference in performance between each model for the different domains and confirm the above results, that is, SUP models perform

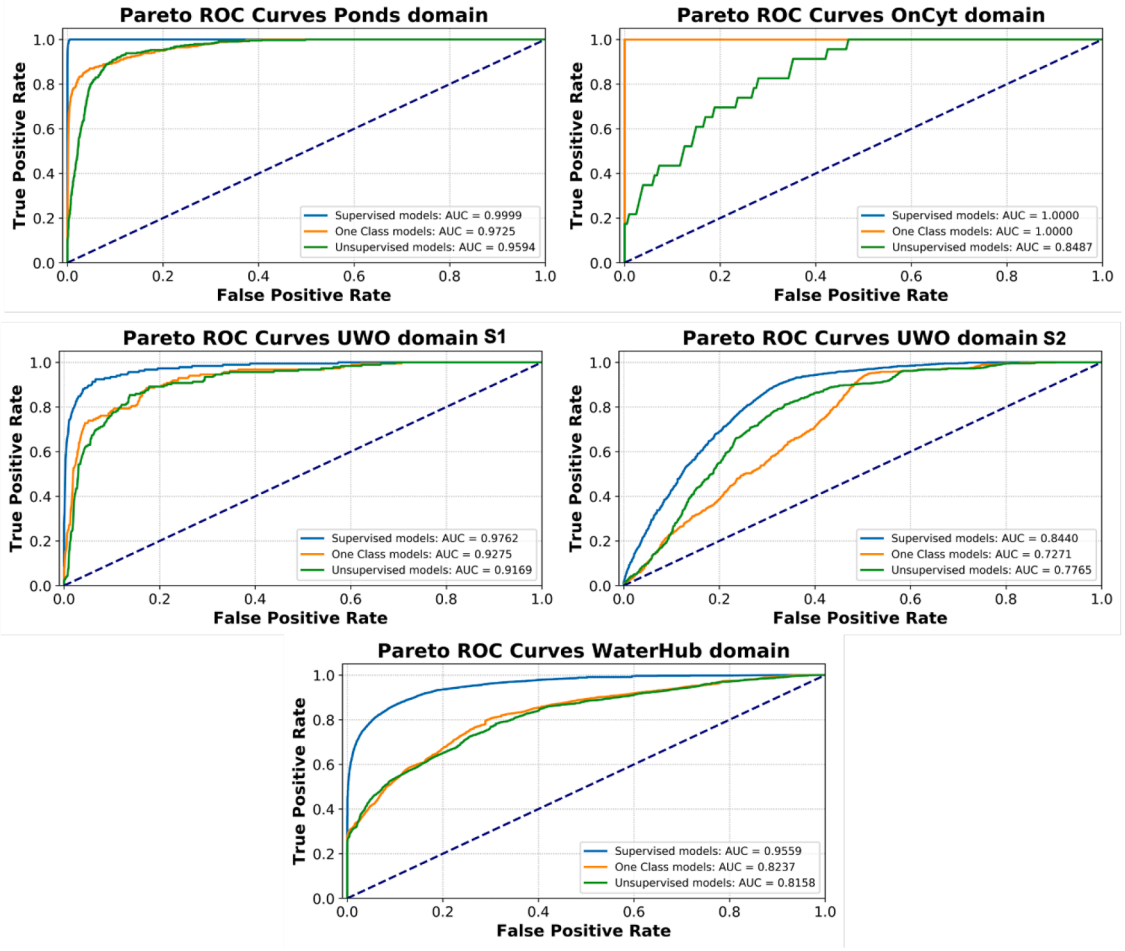

Fig. 2. Pareto ROCs for each domain. Each Pareto ROC is computed showing the different ratios of True Positives Rate (TPR) and False Positives Rate (FPR). This is obtained by selecting the Pareto optimal models among the models with the same calibration setup. For each combination of calibration setup and domain, the area under the curve (AUC) is computed. A dotted diagonal line is added to reflect the performance of a model that is no better than chance level.

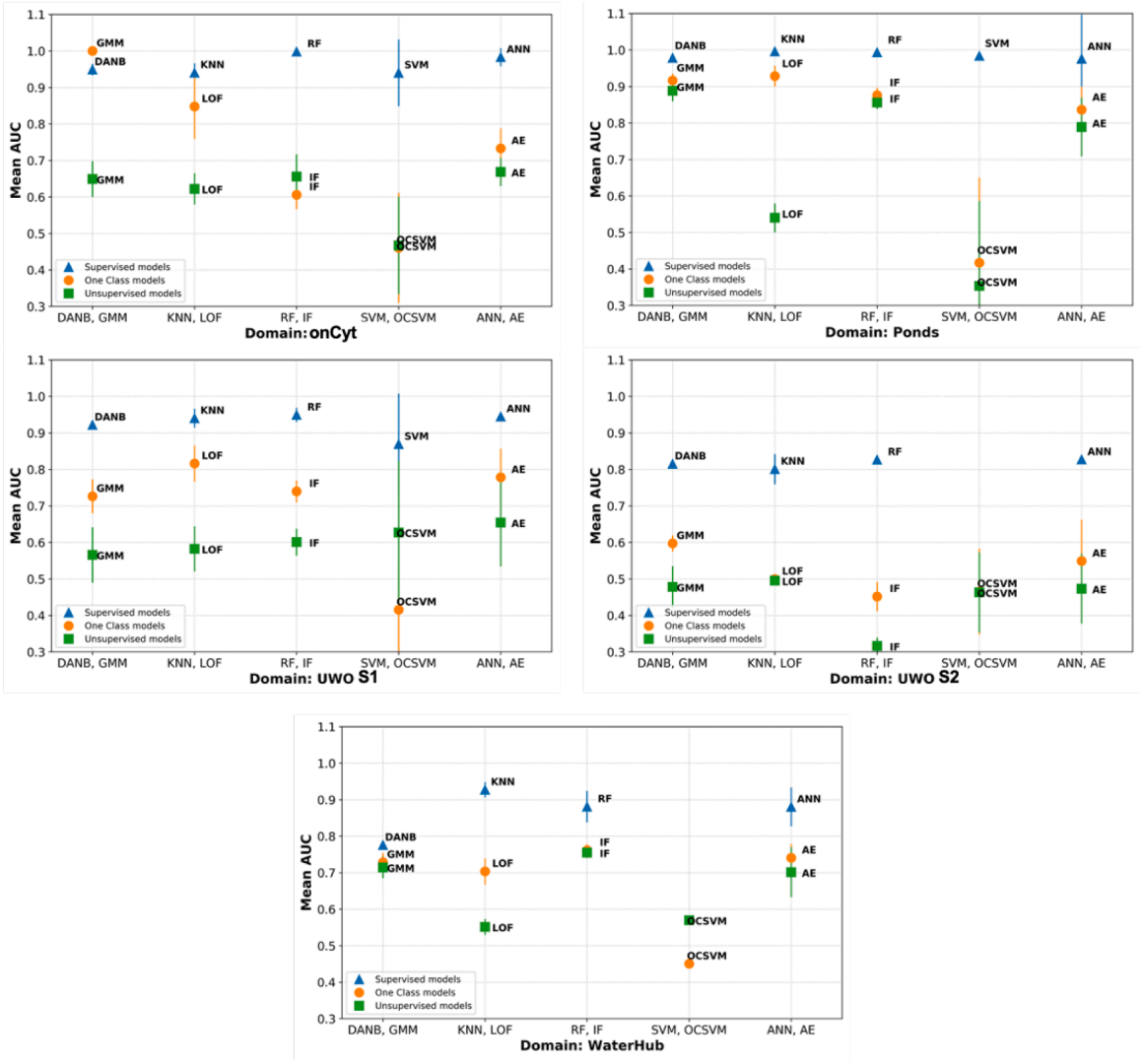

Fig. 3. AUC-ROC mean and standard deviation for each domain under study. Each plot shows the three Machine Learning setups, with the models arranged in a way that the supervised, one-class and unsupervised analogs are aligned vertically.

better than the OCSUP and UNSUP models. Moreover, the OCSUP and UNSUP models produce more variable performance levels, as already shown previously. Consequently, this means that choosing the best model structures and hyperparameters is important for the OCSUP and UNSUP setups. In contrast, the variability of model performance among the SUP models is far lower, suggesting that model structure selection and hyperparameter tuning is far less important in this case.

4. Discussion

We discuss below the main findings of our study. We include our practical recommendations for the implementation of AD strategies and in particular data labelling, and conclude with an outlook at future research.

Machine learning setups and the value of labels for anomaly detection

Our study offers a clear indication that the SUP learning setup is preferred over the OCSUP and UNSUP setups. Clearly, this means the collection of trusted data labels are critical to maximise the performance of data-driven AD. OCSUP or UNSUP models, while requiring less effort in the labelling step, need a higher effort in hyperparameter tuning to obtain a performance that is as close as possible to the SUP model performance levels.

Moreover, the dominance of the SUP setup over the OCSUP setup suggests that having access to labelled anomalous data are especially valuable. This corroborates the major conclusion of Russo et al. (2020).

Model type selection for anomaly detection

A key notion in model assessment is model selection, that is, given two or more ML models, choosing the best one for future deployment. Model selection can be performed based on a specific set of performance metrics dictated by the problem domain, e.g. which FPR is tolerated or which TRP is needed. We discuss below the specific performance for the tested models to provide help for model selection.

In Fig. 2 it is shown that the studied models performed poorly when the dataset contained complex anomalies, as in the case of UWO S2. On the contrary, it appears that models with a simpler structure (specifically DANB) performed well on the Ponds, onCyt and UWO S1 datasets. These results may be explained by the nature of the anomalous data points. Visual inspection of the data, although subjective, suggested that the anomalies are very distinct from normal data points in these domains and easy to spot, even for non-experts. ML models falling in the DANB category are generative models that make assumptions on the distribution of the data. In cases where the data is well distributed, such models could reach high performance possibly because their generative assumptions prevent overfitting (Ng and Jordan, 2002). However, that is not always true in environmental applications, where normal data might not be generated from the same distribution as it presents baseline

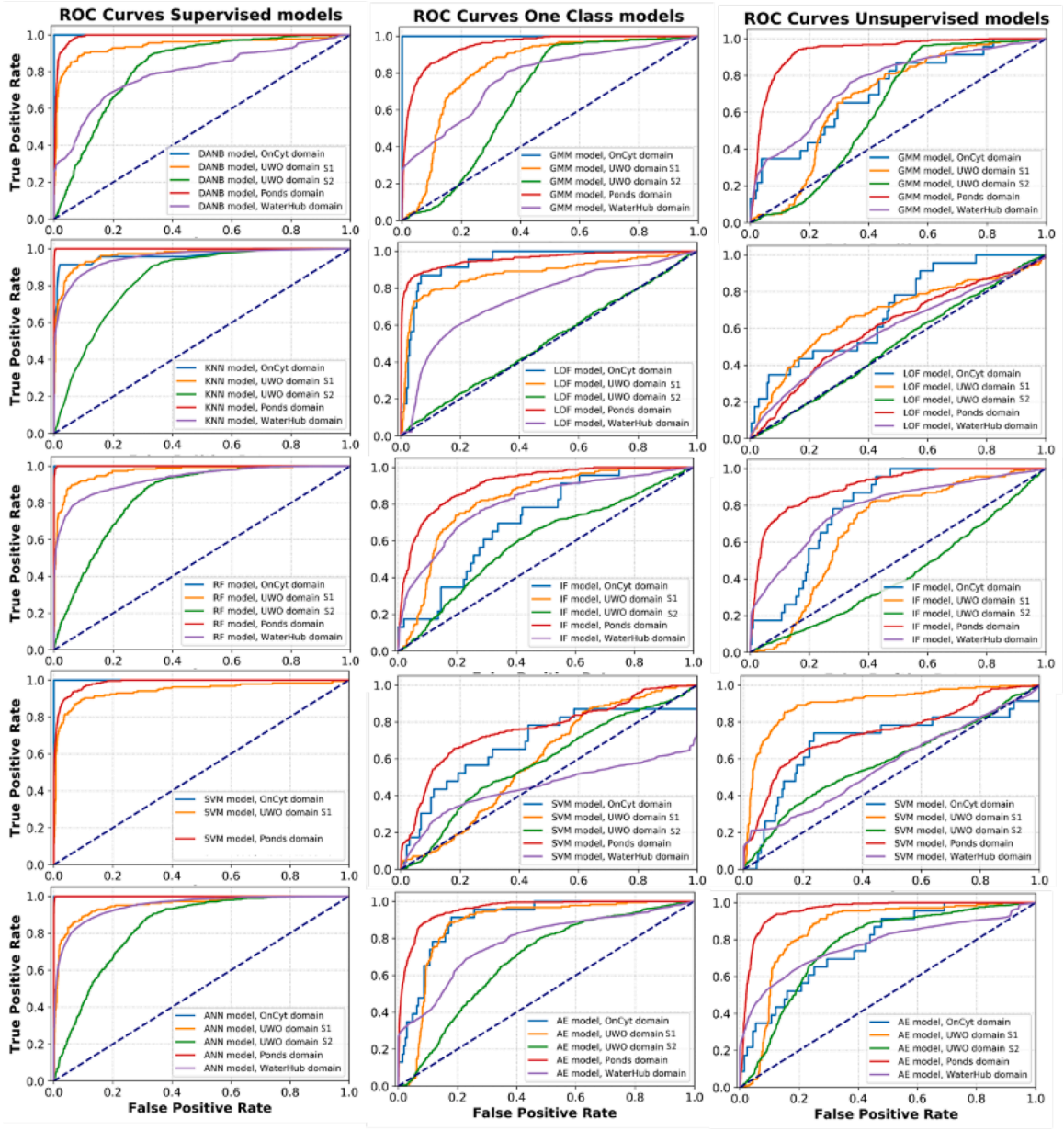

Fig. 4. Pareto ROC curves for each model for all datasets. Left: supervised setup. Centre: one-class setup. Right: unsupervised setup. A dotted diagonal line is added in each plot to reflect the performance of a model that is no better than chance level.

changes (due to different seasons), and anomalies may be generated by different events which do not show the same pattern. In fact, as a general detection performance result, more flexible models like ANN and AE performed well across the different datasets, with AEs being sensitive to noisy data. That is, in the UNSUP setup, where the calibration data contains anomalies, they lead to lower performance compared to their OCSUP setup implementation. This suggests that the generative version of such models (that is, denoising or variational AE (An and Cho, 2015)) could be valuable in such cases. These flexible models also have a higher sensitivity to hyperparameters, which implies that hyperparameter tuning - and so an increased effort in the design of the model architecture - is necessary for achieving such performance.

Finally and more importantly, the results of this study provide insights on the utility of expert-based data labelling. Given the above results, we recommend to perform data labelling and SUP modelling if the nature of anomalies in the dataset is complex, that is if the anomalies are contextual, not well defined and varying over time. If however, the application domain presents more simple, collective anomalies, OCSUP and UNSUP modelling could be performed choosing flexible models. We envisage that more complex models are needed for the OCSUP and UNSUP setups. These can be: an ensemble of models (Baruque and Corchado, 2011); models with Bayesian priors (Baimuratov et al., 2019) reflecting generally applicable domain knowledge; or models that use a background or garbage class, e.g. an UNSUP two-class setup where only class is learned and the other one is fixed. For more details on this and related concepts, please see (Dhamija et al., 2018).

Limitations of the study and future research

In this work, 5 selected data sets have been used to benchmark several ML setups and models. It is important to note that it is rather difficult, if not impossible, to evaluate whether these could be completely representative of typical data in water research. We do however believe that such a combination of full-scale natural and engineered environmental monitoring systems, and careful labelling procedure is unprecedented. Additionally, we have assumed that the domain expert provides labels without error or uncertainty. This is unlikely and should be accounted for in future work. Several methods to

account for imperfect labels during calibration exist (Donmez and Carbonell, 2008; Du and Ling, 2010; Liu et al., 2013), but require much higher computational requirements. One can also employ a higher number of domain experts for the same dataset to quantify uncertainty in the provided labels. This increases the labelling effort significantly. Active learning (Russo et al., 2020) and semi-supervised learning (Zhu and Goldberg, 2009) could however be combined with labelling by multiple experts to minimize this effort. Note that a sixth case was included within this study at first. Unfortunately, we found that the provided labels could not be considered an unbiased opinion of the domain experts. This is explained particularly by the fact that this dataset is so complex that anomalies are not easily spotted even by application experts (Russo et al., 2019). When so, one may chose to incorporate domain knowledge during data preprocessing (Zheng and Casari, 2018) or model construction, as was applied in this case. While this prevents an objective study of the merits of expert-based labelling, this can lead to useful hybrid models for AD. The embedded expertise can be based on temporal correlation/redundancy, hardware redundancy, physical-chemical relationships, spatial redundancy, etc. OCSUP or UNSUP models that incorporate temporal dynamics explicitly are also a specific way to include the existence of temporal correlation into the model. This would include classical time series analysis models such as ARIMA (Pena et al., 2013), or recurrent neural networks or long short-term memory networks, which exist in the family of the deep learning models (Malhotra et al., 2015). The utility of domain expertise in AD, either through knowledge-based feature engineering or hybrid modelling, could not be quantified yet. This should be considered for evaluation both with or without target labels and can be benchmarked with the provided data sets.

5. Conclusions

A comprehensive evaluation of 15 distinct combinations of model structure and calibration setup was executed with datasets produced by 5 full-scale environmental monitoring systems. The primary interest was to evaluate the utility of expert-based data labelling for anomaly detection purposes. This is the first time this has been studied at this scale for aquatic applications. Our most important conclusions are:

- Supervised models are better than models based on unsupervised learning. This is true for all of the 5 domains. This means that access to labelled data is critical for effective use of machine learning for anomaly detection in environmental systems data. Additionally, we have shown that if the anomalies are not complex, then one-class models which require less time for labelling, could be used.
- The comparison of one-class and unsupervised models underlines the value of expert labelling as these models only differ in the provided data for calibration (and not in the objective function). Across all studied domains, anomaly detection performance with one-class models is better than with unsupervised models. This is true for all models, with exception of the one-class support vector machine models, which however perform poorly always.
- The quantified anomaly detection performance of any particular machine learning algorithm depends strongly on the particular domain. Indeed, our results show that evaluated classification accuracy metrics, like the area-under-the-curve (AUC), cannot be extrapolated from one domain to the next. This is especially true for the one-class and unsupervised models, where anomaly detection performances were shown to vary between 0.39 and 0.93. Even in the SUP setting, the reported AUC varies between 0.77, which is unlikely to be acceptable for online deployment, and 0.99, which we expect to exceed minimum requirements. Unfortunately yet true, this means that broad statements on the expected utility of machine learning algorithms should not be trusted.
- Our results do not indicate a strong preference for any particular model structure. The best available supervised models were: artificial neural networks (UWO S2), k-nearest neighbours (Ponds, WaterHub), and random forest (onCyt, UWO S1). Among the one-class models, one-class support vector machines perform particularly poorly in all domains. It is therefore reasonable to exclude it as a suitable candidate model in future work. In absence of data labels (unsupervised models), the models with best AUC include auto-encoders (onCyt, UWO S1), Gaussian mixture models (Ponds), isolation forest (WaterHub), and local outlier factor (UWO S2). This shows again that a general preference for a particular modelling paradigm cannot be based on anomaly detection performance alone.

Author contribution

A.H. and E.M. provided data for the WaterHub domain. A.H. conducted the labelling and curation of the data and contributed to the interpretation of the results. B.M. and M.L. provided data for the Eawag ponds domain. M.L. conducted the labelling and curation of the data and contributed to the interpretation of the results. F.H. and M.B. provided data for the flow-cytometry domain. M.B. conducted the labelling and curation of the data and contributed to the interpretation of the results. D.B., C.M. and V.T.-K. provided data for the Léxplore domain. C.M. and V. T.-K conducted the labelling and curation of the data and contributed to the interpretation of the results. F.B. and A.D provided data for the UWO domain and helped in the design of the labelling tool. Both authors conducted the labelling and curation of the data and contributed to the interpretation of the results.

S.R. and K.V. conceived the research concept, methodology and software experiments, and carried out writing of the manuscript. All authors provided critical feedback.

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.

Acknowledgements

The authors would like to thank Ccile Bettex, Juan Pablo Carbajal, Anita Narwani and Piet Spaak for their contributions to the work presented in this paper and Ben LaRiviere for his useful feedback. The study has been made possible by the Eawag Discretionary Funds (grant number: 5221.00492.012.02, project: DF2018/ADASen).

References

Aguado, D., Ferrer, J., Seco, A., 2007. Multivariate SPC of a sequencing batch reactor for wastewater treatment. Chemom. Intell. Lab. Syst. 85, 82–93.
Aguado, D., Rosen, C., 2008. Multivariate statistical monitoring of continuous wastewater treatment plants. Eng. Appl. Artif. Intell. 21 (7), 1080–1091.
Aguado, D., Zarzo, M., Ferrer, J., Seco, A., 2005. A multivariate methodology for detecting operational shifts: application to a sequencing batch reactor. IWA Conference on Nutrient Removal in Wastewater Treatment Plants and Recycle Streams (BNR2005), Krakow, Poland, pp. 755–764.September 19–21, 2005
Alferes, J., Tik, S., Copp, J., Vanrolleghem, P.A., 2013. Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. Water Sci. Technol. 68 (5), 1022–1030.
Amer, M., Goldstein, M., Abdennadher, S., 2013. Enhancing one-class support vector machines for unsupervised anomaly detection. Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, pp. 8–15.
An, J., Cho, S., 2015. Variational Autoencoder based anomaly detection using reconstruction probability. Spec. Lect. IE 2 (1), 1–18.
Baert, A., Villez, K., Steppe, K., 2012. Functional unfold principal component analysis for automatic plant-based stress detection in grapevine. Funct. Plant Biol. 39 (6), 519–530.
Baimuratov, I., Shichkina, Y., Stankova, E., Zhukova, N., Than, N., 2019. A Bayesian information criterion for unsupervised learning based on an objective prior. International Conference on Computational Science and Its Applications. Springer, pp. 707–716.
Baruque, B., Corchado, E., 2011. Fusion Methods for Unsupervised Learning Ensembles, vol. 322. Springer.

Besmer, M.D., Epting, J., Page, R.M., Sigrist, J.A., Huggenberger, P., Hammes, F., 2016. Online flow cytometry reveals microbial dynamics influenced by concurrent natural and operational events in groundwater used for drinking water treatment. Sci. Rep. 6, 38462.
Besmer, M.D., Weissbrodt, D.G., Kratochvil, B.E., Sigrist, J.A., Weyland, M.S., Hammes, F., 2014. The feasibility of automated online flow cytometry for in-situ monitoring of microbial dynamics in aquatic ecosystems. Front. Microbiol. 5, 265.
Bezdek, J.C., Rajasegarar, S., Moshtaghi, M., Leckie, C., Palaniswami, M., Havens, T.C., 2011. Anomaly detection in environmental monitoring networks [application notes]. IEEE Comput. Intell. Mag. 6 (2), 52–58.
Candelieri, A., 2017. Clustering and support vector regression for water demand forecasting and anomaly detection. Water 9 (3), 224.
Chalapathy, R., Menon, A. K., Chawla, S., 2018. Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360.
Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surv. (CSUR) 41 (3), 1–58.
Ciavatta, S., Pastres, R., Lin, Z., Beck, M., Badetti, C., Ferrari, G., 2004. Fault detection in a real-time monitoring network for water quality in the lagoon of venice (Italy). Water Sci. Technol. 50 (11), 51–58.
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. Environ. Model. Softw. 106, 89–103.
Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C., Vanrolleghem, P.A., 2011. Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnol. Bioeng. 108 (2), 333–344.
Dhamija, A.R., Günther, M., Boult, T., 2018. Reducing network agnostophobia. Advances in Neural Information Processing Systems, pp. 9157–9168.
Donmez, P., Carbonell, J.G., 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. Proceedings of the 17th ACM conference on Information and Knowledge Management, pp. 619–628.
Du, J., Ling, C.X., 2010. Active learning with human-like noisy oracle. International Conference on Data Mining. IEEE, pp. 797–802.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.
Fuente, M., Vega, P., Zarrop, M., Poch, M., 1996. Fault detection in a real wastewater plant using parameter-estimation techniques. Control Eng. Pract. 4 (8), 1089–1098.
Garrido-Baserba, M., Corominas, L., Cortés, U., Rosso, D., Poch, M., 2020. The fourth-revolution in the water sector encounters the digital revolution. Environ. Sci. Technol. 54 (8), 4698–4705.
Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
Gibert, K., Horsburgh, J.S., Athanasiadis, I.N., Holmes, G., 2018. Environmental data science. Environ. Model. Softw. 106, 4–12.
Gibert, K., Sànchez-Marrè, M., Codina, V., 2010. Choosing the right data mining technique: classification of methods and intelligent recommendation. Int. Cong. Environ. Model. Softw.
Gurden, S.P., Westerhuis, J.A., Bro, R., Smilde, A.K., 2001. A comparison of multiway regression and scaling methods. Chemom. Intell. Lab. Syst. 59 (1–2), 121–136.
Hastie, T., Tibshirani, R., Friedman, J., 2009. Unsupervised learning. The Elements of Statistical Learning. Springer, pp. 485–585.
Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty. Advances in Neural Information Processing Systems, pp. 15663–15674.
Hill, D.J., Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model. Softw. 25 (9), 1014–1022.
Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C.M., Sun, J., 2017. Anomaly detection for a water treatment system using unsupervised machine learning. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, pp. 1058–1065.
Khan, S.S., Madden, M.G., 2014. One-class classification: taxonomy of study and review of techniques. Knowl. Eng. Rev. 29 (3), 345–374.
Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2006. Data preprocessing for supervised leaning. Int. J. Comput. Sci. 1 (2), 111–117.
Lee, D.S., Park, J.M., Vanrolleghem, P.A., 2005. Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. J. Biotechnol. 116, 195–210.
Lee, D.S., Vanrolleghem, P.A., 2003. Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. Biotechnol. Bioeng. 82 (4), 489–497.
Lee, M.W., Hong, S.H., Choi, H., Kim, J.-H., Lee, D.S., Park, J.M., 2008. Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors. Process Biochem. 43 (10), 1107–1113.
Leigh, C., Alsibai, O., Hyndman, R.J., Kandanaarachchi, S., King, O.C., McGree, J.M., Neelamraju, C., Strauss, J., Talagala, P.D., Turner, R.D., et al., 2019. A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Sci. Total Environ. 664, 885–898.
Leung, K., Leckie, C., 2005. Unsupervised anomaly detection in network intrusion detection using clusters. Proceedings of the Twenty-eighth Australasian Conference on Computer Science, vol. 38, pp. 333–342.
Liu, B., Xiao, Y., Philip, S.Y., Hao, Z., Cao, L., 2013. An efficient approach for outlier detection with imperfect data labels. IEEE Trans. Knowl. Data Eng. 26 (7), 1602–1616.
Liu, W., Hua, G., Smith, J.R., 2014. Unsupervised one-class learning for automatic outlier removal. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3826–3833.
Lürig, M., Narwani, A., Penson, H., Wehrli, B., Spaak, P., Matthews, B., 2020. Non-additive effects of species interactions on aquatic ecosystems responses to nutrient perturbation. Ecology.(in press)
Malhotra, P., Vig, L., Shroff, G., Agarwal, P., 2015. Long short term memory networks for anomaly detection in time series. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), vol. 89. Presses universitaires de Louvain, pp. 89–94.
Maxion, R.A., Roberts, R.R., 2004. Proper use of ROC Curves in Intrusion/Anomaly Detection. University of Newcastle upon Tyne, Computing Science Tyne, UK.
Miau, S., Hung, W.-H., 2020. River flooding forecasting and anomaly detection based on deep learning. IEEE Access 8, 198384–198402.
Muharemi, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. J. Inf. Telecommun. 3 (3), 294–307.
Narwani, A., Reyes, M., Pereira, A.L., Penson, H., Dennis, S.R., Derrer, S., Spaak, P., Matthews, B., 2019. Interactive effects of foundation species on ecosystem functioning and stability in response to disturbance. Proc. R. Soc. B 286 (1913), 20191857.
Ng, A.Y., Jordan, M.I., 2002. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Advances in Neural Information Processing Systems, pp. 841–848.
Ni, J., Zhang, C., Ren, L., Yang, S.X., 2011. Abrupt event monitoring for water environment system based on KPCA and SVM. IEEE Trans. Instrum. Meas. 61 (4), 980–989.
Pena, E.H., de Assis, M.V., Proença, M.L., 2013. Anomaly detection using forecasting methods ARIMA and HWDS. 2013 32nd International Conference of the Chilean Computer Science Society (SCCC). IEEE, pp. 63–66.
Postolache, O.A., Girao, P.S., Pereira, J.M.D., Ramos, H.M.G., 2005. Self-organizing maps application in a remote water quality monitoring system. IEEE Trans. Instrum. Meas. 54 (1), 322–329.
Ramanathan, N., Balzano, L., Burt, M., Estrin, D., Harmon, T., Harvey, C., Jay, J., Kohler, E., Rothenberg, S., Srivastava, M., 2006. Rapid Deployment with Confidence: Calibration and Fault Detection in Environmental Sensor Networks. Technical Reports. Center for Embedded Network Sensing.
Rosén, C., Lennox, J., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Water Res. 35 (14), 3402–3410.
Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M., 2018. Deep one-class classification. International Conference on Machine Learning, pp. 4393–4402.
Russo, S., Disch, A., Blumensaat, F., Villez, K., 2019. Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data. Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2019).
Russo, S., Lürig, M., Hao, W., Matthews, B., Villez, K., 2020. Active learning for anomaly detection in environmental data. Environ. Model. Softw. 134, 104869.
Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E., 2018. Adversarially learned one-class classifier for novelty detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3379–3388.
Sillito, R.R., Fisher, R.B., 2008. Semi-supervised learning for anomalous trajectory detection. BMVC, vol. 1, pp. 1–035.
Trilles, S., Belmonte, Ò., Schade, S., Huerta, J., 2017. A domain-independent methodology to analyze IoT data streams in real-time. A proof of concept implementation for anomaly detection from environmental data. Int. J. Digital Earth 10 (1), 103–120.
Villez, K., Ruiz, M., Sin, G., Colomer, J., Rosén, C., Vanrolleghem, P.A., 2008. Combining multiway principal component analysis and clustering for efficient data mining of historical data sets of SBR processes. Water Sci. Technol. 57 (10), 1659–1666.
Yu, J., 2012. A nonlinear kernel gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. Chem. Eng. Sci. 68 (1), 506–519.
Zheng, A., Casari, A., 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. "O'Reilly Media, Inc.".
Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synth. Lect. Artif. Intell.Mach. Learn. 3 (1), 1–130.