Ocean Data Quality Assessment through Outlier Detection-enhanced Active Learning

Na Li*, Yiyang Qi†*, Ruyue Xin*, Zhiming Zhao*‡
*Multiscale Networked System, Informatics Institute, University of Amsterdam, Netherlands
†Computer Science Department, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
‡LifeWatch ERIC Virtual Lab and Innovation Center (VLIC), Amsterdam, Netherlands
n.li@uva.nl, y.qi@student.vu.nl, r.xin@uva.nl, z.zhao@uva.nl

***Abstract*—Ocean and climate research benefits from global ocean observation initiatives such as Argo, GLOSS, and EMSO. The Argo network, dedicated to ocean profiling, generates a vast volume of observatory data. However, data quality issues from sensor malfunctions and transmission errors necessitate stringent quality assessment. Existing methods, including machine learning, fall short due to limited labeled data and imbalanced datasets. To address these challenges, we propose an Outlier Detection-Enhanced Active Learning (ODEAL) framework for ocean data quality assessment, employing Active Learning (AL) to reduce human experts' workload in the quality assessment workflow and leveraging outlier detection algorithms for effective model initialization. We also conduct extensive experiments on five large-scale realistic Argo datasets to gain insights into our proposed method, including the effectiveness of AL query strategies and the initial set construction approach. The results suggest that our framework enhances quality assessment efficiency by up to 465.5% with the uncertainty-based query strategy compared to random sampling and minimizes overall annotation costs by up to 76.9% using the initial set built with outlier detectors.**

***Index Terms*—ocean data quality control, Argo, machine learning, active learning, initial set construction**

I. Introduction

To support ocean and climate research, several international ocean observation programs and projects, such as Array for Real-time Geostrophic Oceanography (Argo) [1], Global Sea Level Observing System (GLOSS) [2], European Multidisciplinary Seafloor Observatory (EMSO) [3], have been launched for observations and measurements at different sea depths [4]. The Argo observation network, dedicated to profiling the global ocean, comprises thousands of profilers that produce enormous observatory data over time. However, the data records usually suffer from quality problems caused by sensor damage, equipment malfunction, and data transmission errors, which may potentially lead to inaccurate scientific conclusions. Thus, it is of great significance to conduct data Quality Control (QC) before it can be used for any downstream applications [5]. Due to the strict requirements on the data credibility, the QC process is mainly done by domain experts or involves a high degree of human engagement, and hence is extremely laborious and time-consuming [6].

Existing studies have proposed automated and semi-automated data quality assessment approaches to assist QC experts. Traditional methods apply a sequence of rule-based statistical tests on the data instances [6]–[8]. These methods rely on pre-defined quality criteria and are subject to specific observation types. As a result, it can only perform basic examination for the validity of data and still depends on manual inspection for accurate outputs. An emerging research direction is to exploit Machine Learning (ML) methods for ocean data quality assessment [9]–[12]. One branch of studies utilizes anomaly detection techniques [13] to identify erroneous measurements from the dataset [9], [10]. Another line of research employs deep neural networks to classify samples into different quality categories [11].

Nonetheless, two main challenges remain in automated QC research. The first challenge is the lack of labeled data for training ML models. Despite the existence of historically labeled data, it is not trivial to transfer knowledge from one dataset to another due to domain shift or concept drift issues. Targeting this challenge, we propose utilizing AL methods [14] to reduce the number of labeled instances required for optimizing ML models. AL query strategies can select the most informative data instances for labeling and thus improve model performances in a data-efficient manner. In the context of data quality assessment, AL can select the most tricky samples that require manual examination and automate the quality assessment for the rest of the samples.

The second challenge is the cold-start problem in AL methods posed by severe data imbalance w.r.t. quality labels, with erroneous measurements occupying less than 1% of the datasets. In a typical AL scenario, an initial set is required to initialize the classifiers, and the iterative query process follows to refine the classifiers [14]. However, a small initial set built from a severely imbalanced dataset likely contains zero erroneous instances, on which classifiers are unable to learn meaningful representations for the erroneous instances and consequentially mislead the query process. This is referred to as the cold-start problem. To address this issue, we propose to leverage the outlier detectors for initial set construction, which can increase the probability for initial sets to include erroneous samples and improve their effectiveness in initializing classifiers.

In this paper, we present an Outlier Detection-Enhanced Active Learning (ODEAL) framework for ocean data quality assessment, which applies AL to reduce the workload of QC experts in a semi-automated data quality assessment workflow and employs outlier detection algorithms to initialize the

learning models effectively with a minimal number of samples. To validate our method, we conducted extensive experiments on five realistic datasets provided by Argo [15]. The results suggest that the AL method can increase the F1-score by 465.5% compared to random sampling, and the outlier detection-initialized approach reduces overall annotation cost by 76.9%. Our contributions are three-fold:

- presenting a novel AL-based data quality assessment framework to reduce the workload of human analysts;
- proposing using outlier detection to construct the initial set for a highly imbalanced dataset to solve the cold-start problem and minimize the overall annotation cost;
- providing empirical evidence and insights for the effectiveness of the proposed method via extensive experiments.

The related datasets and codes can be found here[1].

II. Related work

Automated and semi-automated QC approaches on large-scale ocean observatory data have been studied to support human experts. Classical automated QC procedures involve constant value check, spike and step check, range check, and stability check [6]–[8] to screen out gross errors from measurements but are subject to human experts for fine-grained quality examination. More advanced methods utilize ML models, such as Multi-layer Perceptron (MLP) and ARIMA, to analyze large-scale data [9]–[11]. A common approach is using anomaly detection for quality assessment, considering the infrequent occurrence of erroneous samples. Castelão [10] characterizes the typical behavior of the data by estimating a probability density function (PDF) for each feature, and the survival function (SF) of the estimated PDF is used to quantify the anomalousness of a certain measurement. This approach reduces the error by at least 50% when applied to 13 years of hydrographic profiles. Zhou et al. [9] focus on time-series data and employ the ARIMA model to detect erroneous measurements. The proposed method was applied to pH and CTD data from the Xiaoqushan Seafloor Observatory and achieved an F1 score of 0.9506 in outlier detection. However, ARIMA is a time series forecasting model that requires clean and well-preprocessed time series data for training, which implicitly demands quality labels for historical data points.

Different from the above methods, Mieruch et al. [11] frame data quality assessment as a binary classification problem, where each sample is categorized as of *good* or *bad* qualities. They present SalaciaML, aiming to mimic the skillful QC experts with a deep learning artificial neural network in identifying potentially erroneous samples. A MLP is trained on more than 2 million temperature measurements over the Mediterranean Seas spanning the last 100 years and detects correctly more than 90% of all good and/or bad data in 11 out of 16 Mediterranean regions. However, it does not consider the correlations between different water properties.

[1]https://github.com/QCDIS/odeal

The high accuracy achieved by the reviewed studies is at the cost of a large amount of labeled data, which is usually the most tricky problem in real-world scenarios. Therefore, our goal is to decrease the demand for labeled data in data quality assessment by leveraging AL methods.

III. Methodology

A. Problem Definition

To formally define the problem, let $x_t^s \in \mathbb{R}$ be an observation record measured at time point t by sensor s. Given a time frame T, data records produced by a set of sensors S are $\{[x_t^{s_i}, x_t^{s_2}, \cdots, x_t^{s_d}]\} \in \mathbb{R}^{|T|\times|S|}, s_i \in S, t \in T$. For simplicity, we will denote $[x_t^{s_i}, x_t^{s_2}, \cdots, x_t^{s_d}]$ as x_t thereafter. The task of quality assessment is to assign a quality flag $\hat{y}_t$ to a data instance x_t. In this work, we only consider binary quality labels, where each instance is classified as *with* OR *without quality issues*, and thus we have $\hat{y}_t \in \{0, 1\}$. The quality flag 1 means there are quality issues in the data instance, while 0 means there is no quality problem.

B. Outlier Detection-Enhanced Active Learning Framework

We propose an AL-based data quality assessment framework called Outlier Detection-Enhanced Active Learning (ODEAL), which aims to label all the data instances with high accuracy while minimizing the annotation cost, i.e., the number of samples that need to be manually labeled. It operates under the pool-based AL paradigm. Figure 1 depicts the high-level structure of the proposed framework. It consists of the *Initial set construction* and the *Active learning cycles* phases. In the first phase, we build an initial set D_I with anomaly detectors, which select the top-N_I anomalous samples from the target dataset D, where N_I is the size of D_I. And the remaining samples are put into the unlabeled set D_U. D_I is sent to human experts for labels and then used for classifier initialization.

During the AL cycles, the labeled set D_L stores all the labeled instances generated by the human annotators. The target models are quality assessors that assign quality labels to data instances. A human expert is involved as the oracle providing legitimate quality labels for queried samples. The AL workflow is described below:

1) The quality assessment models are trained on the initial set (Step 1).
2) The quality assessment models predict the quality labels on all instances in the unlabeled set (Step 2).
3) If the budget is exhausted or the models are confident enough for the predicted results, output the predictions (Step 3);
4) Otherwise, the models produce intermediate results for the query strategy to select the K most informative instances (Step 4).
5) The selected instances are presented to human annotators for labeling. Meanwhile, they are moved from the unlabeled set to the labeled set (Step 5).
6) The quality assessment models are updated using the labeled set (Step 6).

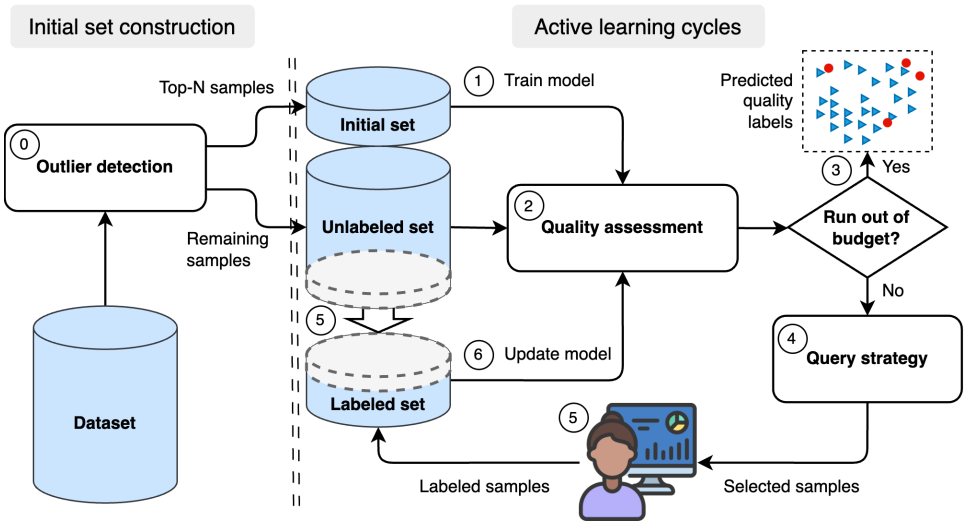


Fig. 1: High-level structure of the proposed ODEAL framework

7) Repeat Steps 2, 4, 5, and 6 until the annotation budget exhausts or the confidence threshold is reached.

For implementation, the annotation budget and the confidence threshold should be defined under the operational condition depending on the requirements. It is also application-dependent in terms of quality assessment model selection and query strategy development. We illustrate our choices of quality assessment models and query strategy in the ocean observation use case context. However, it is at the readers' discretion to make their decisions.

C. Initial Set Construction

The importance of the initial set is often overlooked by previous studies. Most of them create an initial set using a pre-defined number of randomly selected samples and exclude the labeling of the initial set from the annotation budget. There are two problems with it. First, in real applications, the labels of the initial set are also provided by human experts and should be included in the overall budget. Second, the initial set can significantly affect the following AL process. Insufficient initial instances may cause cold-start issues, while redundant initial instances will result in a wastage of annotation costs. Moreover, some classifiers, e.g., CatBoost, can only learn from different classes of data, which will require a large initial set when the abnormal samples are elusive in the dataset, i.e., a highly imbalanced dataset.

To address the above challenges, we propose exploiting outlier detectors (or anomaly detectors) for forming the initial set D_I. The objective here is to identify anomalies for inclusion in D_I, while maintaining its compact size. Because there is a scarcity of labeled data, the outlier detection algorithms must function in an unsupervised manner or be fine-tuned with a minimal amount of labeled samples. While it is possible to utilize the outlier detectors alone for implementing AL methods, updating an unsupervised outlier detector with labeled data is complex. As a result, our research places its emphasis on employing outlier detectors to only build the initial set. Let ψ be an outlier detector, and then we have an outlier score $\mu_t = \psi(x_t)$ for data instance x_t. D_I is composed of the top-N_I samples from the dataset D ranked by outlier scores, i.e.,

$$D_I = \{x_t \mid \mu_t \in \text{top-}N_I \text{ outlier scores}\}, \tag{1}$$

where N_I is the size of D_I. This method increases the possibility of including abnormal samples in a small initial set to warm up the classifiers and is expected to be effective in severely imbalanced datasets. In this work, we consider three common outlier detection algorithms, i.e., Isolation Forest (iForest) [16], One-Class Support Vector Machine (OCSVM) [17], and Local Outlier Factor (LOF) [18].

D. Quality Assessment Model Selection

The classifiers (also called target models or learners in the AL paradigm) being investigated in our study include k-nearest Neighbor (KNN) [19], Extreme Gradient Boosting (XGBoost) [20], Categorical Boosting (CatBoost) [21] and Light Gradient Boosting Machine (LightGBM) [22]. KNN is an instance-based algorithm that classifies data points by considering the class of their nearest neighbors. XGBoost is an ensemble algorithm utilizing gradient boosting, excelling in structured data tasks, and featuring sequential model correction. CatBoost is a gradient boosting approach tailored for categorical data, automatically handling categorical features during training. LightGBM is a high-efficiency gradient boosting framework using histogram-based techniques for faster training, suitable for large datasets. XGBoost, CatBoost, and LightGBM are effective in dealing with imbalanced data, as they assign higher weights to misclassified samples, which give more attention to the minority class.

E. Query Strategy

We adopt the Uncertainty-based Sampling (US) strategies for classifiers to select instances. These query strategies are

based on the hypothesis that the more uncertain one classifier is about the predictive result, the more informative the data sample is. Thus, it will be more useful for optimizing the classifier. There are various ways to compute the *uncertainty* for predictions, such as *confidence*, *margin*, and *entropy*. In binary classification scenarios, the corresponding query strategies, i.e., *least confident*, *smallest margin*, and *maximum entropy*, become equivalent [14]. Here we describe the prediction entropy query strategy. Supposing that a classifier ϕ can output the class probability $P_\phi(\hat{y}_t|x_t), \hat{y}_t \in \{0,1\}$ for a sample x_t, the entropy-based uncertainty of the classifier is defined as:

$$H_{EN}(x_t) = -\sum_{\hat{y}_t \in \{0,1\}} P_\phi(\hat{y}_t|x_t) \log P_\phi(\hat{y}_t|x_t). \quad (2)$$

The sample with maximum entropy will be requested for labeling:

$$x^*_{EN} = \arg\max_t H_{EN}(x_t) \quad (3)$$

IV. Experiments

A. Experiment Setup

1) Dataset: We use ocean observatory data provided by Argo [15], an international program that collects subsurface ocean water properties such as temperature, salinity, and currents across the global earth using a fleet of robotic instruments. The instruments called *floats* or *profilers*, drift with the ocean currents and move up and down between the surface and a mid-water level. We build five datasets using records produced by five similar floats in order to reduce the influence of environmental and operational factors. The five floats were equipped with the same sensors and deployed in the same month (March 2019) within the same area (Atlantic Ocean). We consider totally six features, including *datetime*, *latitude*, *longitude*, *pressure*, *temperature* and *salinity*. The datasets have extensive quality labels for all the data samples provided by human analysts. To mimic the real usage scenario, we only use labels of queried samples, treating them as acquired from human experts. When processing the quality flags, we treat samples labeled as *good data* as error-free, denoted as 0, and others as erroneous data, denoted as 1. Originally, the quality flags were assigned for each feature. To get a global quality label for the data, we treat the data instance as 0 only if all the features have the label 0.

The statistics of the datasets are listed in Table I. The dataset DS_{high} exhibits a substantial error rate of 33.72%, whereas the remaining four datasets, namely DS_{low}1-4, demonstrate exceedingly minimal error rates, all below 1%. Each dataset is randomly split into 60% training, 20% validation, and 20% test subsets while maintaining the same error rate for each subset. All the features are first normalized using the Z-score method before feeding into the classifiers.

2) Evaluation metrics: We use *F1-score* to evaluate the performance of the classifiers.

B. Effectiveness of AL Query Strategies

We examine the effectiveness of the Uncertainty-based Sampling (US) method by comparing it to Random Sampling (RS). The initial set is formed by randomly selected instances. Considering the error rates of the datasets, we set the initial set size to 1000 for DS_{high} and 100 for the rest datasets. Figure 2 shows the results for four classifiers on five datasets.

TABLE I: Statistics of the datasets.

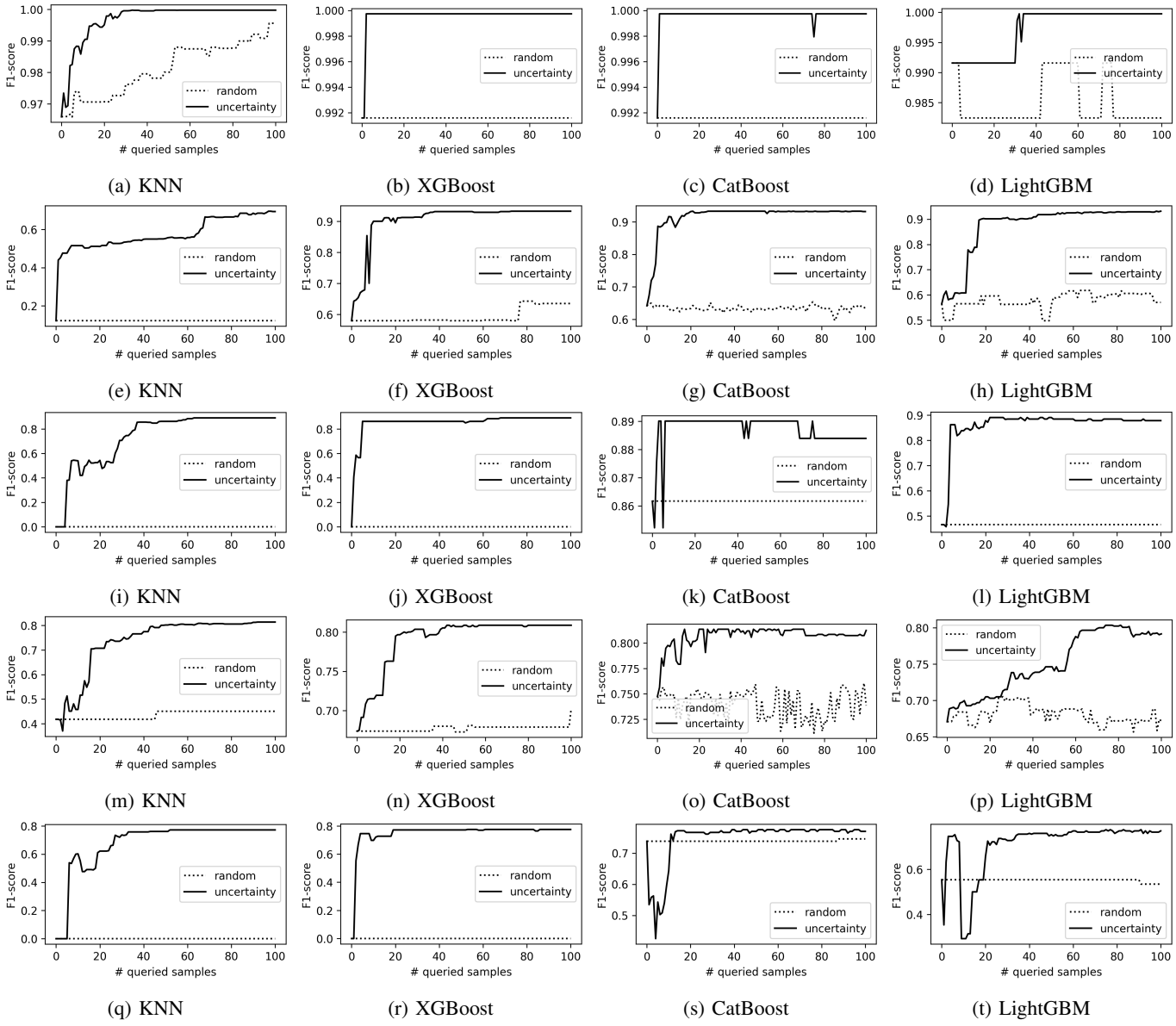
Dataset	Float code	Launch date	Error rate	Training samples	Test samples
DS_{high}	4903217	21/03/2019	33.72%	179,539	59,847
$DS_{low}1$	4903218	10/03/2019	0.84%	175,583	58,528
$DS_{low}2$	4903220	07/03/2019	0.16%	181,009	60,337
$DS_{low}3$	4903052	20/03/2019	0.69%	179,000	59,667
$DS_{low}4$	4903054	23/03/2019	0.23%	177,455	59,152

The results suggest that the US method is effective for all datasets, either with high error rates or low error rates, but is remarkably successful on severely imbalanced datasets. First, we observe that on DS_{high} XGBoost and CatBoost rapidly reach their best F1-score with less than 10 queried samples using US, while the RS method makes no difference in the models' performance throughout the AL cycles. LightGBM shows similar behavior though it needs more samples to optimize. An exception is KNN, where the RS method still increases its accuracy but takes much longer to achieve optimal performance compared to the US method. Compared with RS, US improve the F1-score for KNN, XGBoost, CatBoost and LightGBM on DS_{high} by 0.4%, 0.8%, 0.8%, and 1.8%, respectively. The subtle progress is due to the fair performance achieved by classifiers before the AL query process. Second, US completely beats RS on four low-error datasets, i.e., DS_{low}1-4, regardless of classifiers. On $DS_{low}1$, compared with RS, the US approach increases the F1-score of KNN, XGBoost, CatBoost and LightGBM by 465.5%, 46.8%, 47.1%, and 63.2%, respectively with merely 100 queried samples. And on $DS_{low}3$, the improvement of F1-score for KNN, XGBoost, CatBoost and LightGBM are 80.5%, 15.1%, 10.0%, and 17.8%, respectively. On $DS_{low}2$ and $DS_{low}4$, KNN and XGBoost report F1-score of 0 using RS but rapidly reach the highest F1-score with US. CatBoost and LightGBM see performance increase of 2.6% and 88.3% respectively on $DS_{low}2$, 3.2% and 44.1% respectively on $DS_{low}4$, when comparing US to RS.

C. Effectiveness of using outlier detectors for initial set construction

We compare iForest [16], OCSVM [17], and LOF [18] to detect outliers on $DS_{low}4$ and report the results in Table II. We use OCSVM with non-linear kernels (RBF). $N_I/2$ randomly selected samples are used to train OCSVM, together with the top-$N_I/2$ predicted anomalous samples to form the initial set. The number of estimators in iForest is set to 100. The number of neighbors in LOF is set to 10. The contamination value for iForest and LOF are defined as the dataset's error rate.

Compared with using random selection to build the initial set, LOF can largely reduce the size of the initial set D_I to contain erroneous instances. According to Table II, LOF successfully recognizes 4, 6, 14, and 21 true anomalies within the



(a) KNN (b) XGBoost (c) CatBoost (d) LightGBM

(e) KNN (f) XGBoost (g) CatBoost (h) LightGBM

(i) KNN (j) XGBoost (k) CatBoost (l) LightGBM

(m) KNN (n) XGBoost (o) CatBoost (p) LightGBM

(q) KNN (r) XGBoost (s) CatBoost (t) LightGBM

Fig. 2: Comparisons of query strategies on different datasets. Legends denote the query strategies: "random" stands for random sampling, and "uncertainty" uncertainty-based sampling. (a)(b)(c)(d): DS_{high}; (e)(f)(g)(h): $DS_{low}1$; (i)(j)(k)(l): $DS_{low}2$; (m)(m)(o)(p): $DS_{low}3$; (q)(r)(s)(t): $DS_{low}4$. K=1 for all datasets.

TABLE II: Number of anomalies identified by outlier detection methods on $DS_{low}4$. N_I stands for the initial set size.

Method	**# anomalies@N_I**			
	N_I=100	**N_I=200**	**N_I=300**	**N_I=400**
iForest	0	0	0	0
OCSVM	0	0	0	0
LOF	4	6	14	21

top-100, 200, 300, and 400 predicted anomalies, respectively, while iForest and OCSVM fail to identify any anomalies in the same settings. LOF is $4/100/0.0023 \approx 17$ times efficient in building the initial set containing erroneous samples in comparison with random selection.

Therefore, we exploit LOF to construct the initial set D_I, denoted as D_I^{LOF}, differing from the one formed by randomly selected instances, denoted as D_I^{RD}. To quantify the efficacy of the LOF-initialized AL method, we employ the reduced annotation cost as the evaluation metric. Let N_I and N_L be the numbers of instances in the initial set and in the labeled set, respectively, and then we have the total annotation cost of $N_I + N_L$. The reduced cost is computed as:

$$Cost_{reduced} = 1 - \frac{N_I^{LOF} + N_L^{LOF}}{N_I^{RD} + N_L^{RD}}. \tag{4}$$

It ranges in $(-\infty, 1]$; the higher, the better. Table III summa-

rizes the comparison results on dataset $DS_{low}4$.

TABLE III: Comparing the effectiveness of LOF-based to randomly-built initial set on dataset $DS_{low}4$. K:1.

Classifier	Random initialization			LOF initialization			Reduced cost
	N_I	N_L	**F1**	N_I	N_L	**F1**	
KNN	740	60	0.7719	400	212	0.7719	23.5%
XGBoost	740	68	0.7753	100	251	0.7753	56.5%
CatBoost	740	9	0.7719	100	73	0.7719	76.9%
LightGBM	740	258	0.7699	400	261	0.7706	33.8%

The results suggest that our outlier detection-initialized AL method significantly reduces the overall annotation cost to achieve comparable performances. CatBoost and XGBoost diminish the initial set size from 740 to 100 and yield 76.9% and 56.5% cost reductions, respectively, while achieving the same F1-scores compared with randomly initialized counterparts. Compared with CatBoost and XGBoost, KNN and LightGBM require more instances for initialization (400 rather than 100), possibly due to their low efficiency in learning imbalanced data. Yet, the LOF initialization method manages to reduce their costs by 23.5% and 33.8%, respectively.

V. Conclusion and Discussion

This paper presents an Outlier Detection-Enhanced Active Learning (ODEAL) framework to reduce the workload of QC experts on ocean data quality assessment tasks. By exploiting the ability of AL query strategies in selecting the most informative samples, our method achieved increases in F1-score of KNN, XGBoost, CatBoost, and LightGBM by 465.5%, 46.8%, 47.1%, and 63.2%, respectively with merely 100 queried samples. The LOF-based initial set construction approach successfully identified 3 erroneous instances within top-100 ranked samples from a highly imbalanced dataset and accomplished a great decrease of annotation cost by 76.9% for CatBoost. To our knowledge, this is the first study dedicated to applying AL to ocean data quality assessment. The promising experimental results on real Argo observatory data provide strong evidence of the effectiveness of this methodology.

Nonetheless, there are some limitations associated with this work. First, classifier performances remain sub-optimal on highly imbalanced datasets, which calls for data-cleaning operations or more proficient learning models. Conflicts may exist in data distribution resulting from the inconsistency of data labeling over a long time period. Second, we exploit pool-based AL methods, which require access to the entire dataset and might be unsuitable for real-time applications. Stream-based AL that decides on the current instance for labeling could be more appropriate in such scenarios.

Acknowledgment

We thank EuroArgo (Mr Thierry Carval and Mr Jean-Marie Baudet) and MARIS (Mr Peter Thijsse) for discussing the selected data sets, quality control processes, and data labels. This work has been partially funded by the European Union's Horizon research and innovation program by the CLARIFY (860627), BLUECLOUD 2026 (101094227), ENVRI-FAIR (824068) and ARTICONF (825134), by the LifeWatch ERIC, and by the NWO LTER-LIFE project.

References

[1] D. Roemmich, O. Boebel, Y. Desaubies, H. Freeland, K. Kim, B. King, P.-Y. Le Traon, R. Molinari, B. W. Owens, S. Riser, U. Send, K. Takeuchi, and S. Wijffels, "Argo : The global array of profiling floats," *Observing the Oceans in the 21st Century*, 2001. [Online]. Available: https://archimer.ifremer.fr/doc/00090/20097/

[2] M. Merrifield, T. Aarup, A. Allen, A. Aman, P. Caldwell, E. Bradshaw, R. Fernandes, H. Hayashibara, F. Hernandez, B. Kilonsky *et al.*, "The global sea level observing system (gloss)," *Proceedings of the OceanObs*, vol. 9, 2009.

[3] P. Favali and L. Beranzoli, "Emso: European multidisciplinary seafloor observatory," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 602, no. 1, pp. 21–27, 2009.

[4] M. Lin and C. Yang, "Ocean observation technologies: A review," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 1, pp. 1–18, 2020.

[5] J. A. Cummings, "Ocean data quality control," *Operational oceanography in the 21st Century*, pp. 91–121, 2011.

[6] D. Abeysirigunawardena, M. Jeffries, M. G. Morley, A. O. Bui, and M. Hoeberechts, "Data quality control and quality assurance practices for ocean networks canada observatories," in *OCEANS 2015-MTS/IEEE Washington*. IEEE, 2015, pp. 1–8.

[7] R. Diamant, I. Shachar, Y. Makovsky, B. M. Ferreira, and N. A. Cruz, "Cross-sensor quality assurance for marine observatories," *Remote Sensing*, vol. 12, no. 21, p. 3470, 2020.

[8] A. M. Skålvik, C. Saetre, K.-E. Frøysa, R. N. Bjørk, and A. Tengberg, "Challenges, limitations, and measurement strategies to ensure data quality in deep-sea sensors," *Frontiers in Marine Science*, vol. 10, p. 1152236, 2023.

[9] Y. Zhou, R. Qin, H. Xu, S. Sadiq, and Y. Yu, "A data quality control method for seafloor observatories: The application of observed time series data in the east china sea," *Sensors*, vol. 18, no. 8, p. 2628, 2018.

[10] G. P. Castelão, "A machine learning approach to quality control oceanographic data," *Computers & Geosciences*, vol. 155, p. 104803, 2021.

[11] S. Mieruch, S. Demirel, S. Simoncelli, R. Schlitzer, and S. Seitz, "Salaciaml: A deep learning approach for supporting ocean data quality control," *Frontiers in Marine Science*, vol. 8, p. 611742, 2021.

[12] S. Demirel, S. Mieruch, and R. Schlitzer, "Deep learning for supporting ocean data quality control," *Bollettino di Geofisica*, vol. 12, p. 121, 2021.

[13] R. Xin, H. Liu, P. Chen, and Z. Zhao, "Robust and accurate performance anomaly detection and prediction for cloud applications: a novel ensemble learning-based framework," *Journal of Cloud Computing*, vol. 12, no. 1, pp. 1–16, 2023.

[14] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[15] G. Argo, "Argo float data and metadata from global data assembly centre (argo gdac)," *Seanoe*, 2000.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[17] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.

[18] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[19] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.