Unsupervised Unknown Unknown Detection in Active Learning

Prajit T Rajendran, Huascar Espinoza, Agnes Delaborde, Chokri Mraidha

Unsupervised Unknown Unknown Detection in Active Learning

Prajit T. Rajendran[1,*], Huascar Espinoza[2], Agnes Delaborde[3] and Chokri Mraidha[1]

[1]CEA, List, F-91120, Palaiseau, France
[2]KDT JU, Avenue de la Toison d'Or 56-60, 1060 Brussels, Belgium
[3]Laboratoire National de Metrologie et d'Essais, Trappes, France

Abstract

Unknown unknowns in machine learning signify data points outside the distribution of known data and constitute blindspots of traditional machine learning models. As these data points typically involve rare and unexpected scenarios, the models may make wrong predictions, potentially leading to catastrophic situations. Detecting "unknown unknowns" is essential to ensure machine learning systems' reliability and robustness and avoid unexpected failures in real-world safety-critical applications. This paper proposes an Unsupervised Unknown Unknown Detection in Active Learning (U3DAL) to detect "unknown unknowns" in a stream-based data setting using active learning data selection mechanisms that rely on uncertainty and diversity. The effectiveness of the proposed approach is validated on the Imagenet-A dataset and across different metrics, demonstrating that it outperforms existing methods for detecting "unknown unknowns".

Keywords

Active learning, safety, unknown unknowns

1. Introduction

1.1. Motivation

Thanks to its ability to make accurate predictions based on patterns and trends in data, machine learning has become a popular tool across various industries and use cases. However, regarding the use of such models in safety critical applications, there are some potential downsides such as distribution shift, adversarial examples, lack of explainability, out of distribution examples, anomalies, unknown unknowns and more. Unknown unknowns refer to data points that are outside the distribution of known data and, therefore, represent blind spots of traditional machine learning models[1]. These data points typically involve rare and unexpected scenarios, and if a model is not able to detect them, it may make wrong predictions, potentially leading to catastrophic situations. Model monitoring mechanisms such as purely uncertainty based techniques fail in this regard, because the model is highly confident about its misprediction.

Detecting unknown unknowns in machine learning can be challenging because these are unanticipated issues that have not been previously encountered or accounted for in the design phase[2, 3, 4]. Some of the simpler, yet not fully sufficient ways to deal with them are discussed below:

- **Anomaly detection:** Anomalies maybe present in the data, which may confuse the model to make confident mispredictions[5]. During testing or deployment, anomaly detectors could be deployed to identify potentially anomalous inputs or states. In training time, it is possible to analyze the data thoroughly to determine biases and irregularities so that these anomalies are not passed on to the model. This is harder when we have no access to what the true data is and what the anomaly is, which is typical in stream-based data settings.
- **Out-of-distribution detection:** Machine learning models perform poorly when shown data points which are very different from previously seen data points[6]. Detecting potential out-of-distribution samples that may not belong to any known classes or categories could also help in identifying potential unknown unknowns. Note that out-of-distribution samples is a subset of unknown unknowns, which includes all data points which are high-confidence mispredictions by the model.
- **Adversarial Attack Detection:** Adversarial inputs may confuse the model to make highly confident mispredictions lead to unknown unknowns [7]. There are various techniques to tackle adversarial exampes, which could also help in mitigating some unknown unknowns.
- **Human-in-the-Loop:** Humans are equipped with conceptual knowledge and hence can iden-

The IJCAI-2023 AISafety and SafeRL Joint Workshop, August 21, 2023, Macau, SAR, China
*Corresponding author.
Prajit.THAZHURAZHIKATH@cea.fr (P. T. Rajendran); Huascar.Espinoza@kdt-ju.europa.eu (H. Espinoza); agnes.delaborde@lne.fr (A. Delaborde); Chokri.MRAIDHA@cea.fr (C. Mraidha)

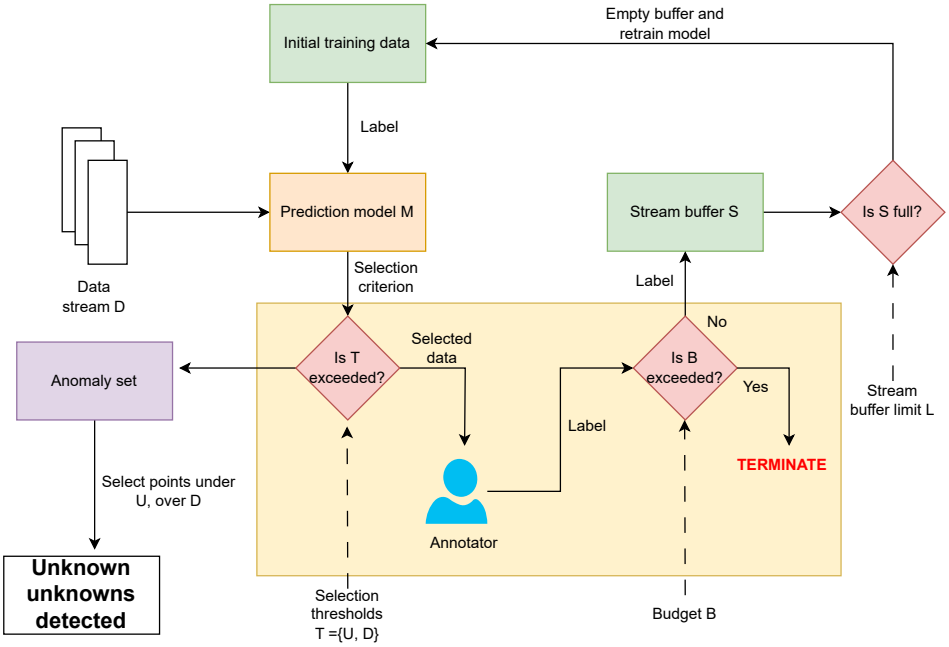

Figure 1: Block diagram of proposed approach

tify potentially dangerous situations with their expert knowledge. If a human is present in the loop, they can assist the model by covering its blindspots, hence mitigating some of the unknown unknowns [8].

- **Robustness testing:** If it is possible to test the model under different scenarios, data distributions and perturbations, some dangers of unknown unknowns could be mitigated. However, in a stream-based active learning setting wherein the data arrives one by one, it is not possible to mitigate the danger of unknown unknowns in advance; it is necessary to detect these unknown unknowns in real time.

The above methods can assisting in identifying safety issues to some extent, but it is not possible to detect or account for all unknown unknowns.

To ensure the reliability and robustness of machine learning systems, it is crucial to detect unknown unknowns. In this paper, we propose a new approach called Unsupervised Unknown Unknown Detection in Active Learning (U3DAL) to detect unknown unknowns in a stream-based data setting using active learning data selection mechanisms that rely on uncertainty and diversity thresholds.

In active learning, a model is trained with a subset of initial labeled data. Based on a predefined function called the acquisition function, the remaining data points are analyzed to determine which of them are complex or interesting enough to be labelled by the human [9]. Some of the common functions include uncertainty, which is a measure of how confident the model is in its predictions [10] and diversity, measuring how the distance of the instances in the stream from those already in the training set[11]. The acquisition function is designed to select the most informative or diverse data points to be labeled, within the constraint of the budget, without compromising on performance[8].

Stream-based active learning is a type of active learning wherein the data arrives in a continuous stream [12]. Learning in real-time is essential in applications where the data distribution is time variant. A challenging aspect of the stream-based learning approach is that it is not possible to access future data points, and therefore the decision of whether or not to choose a data point for querying to the human oracle has to be made as the data arrives.

In this paper, we aim to solve the problem of detecting unknown unknowns in stream-based active learning setting in an unsupervised manner without access to what constitutes a "good" data point or "bad" data point beforehand. As the model has no access to future data points and needs to make a decision to query the current data point one by one, it is interesting to determine which points could be potentially unsafe as they arrive. Since stream-based active learning methods have thresholds

for data selection by design, we hypothesize that these thresholds can help us determine unknown unknown data points. Moreover, through our empirical experiments, we aim to explore the link between the unknown unknown detection capability and the threshold levels.

Contributions: This paper proposes an unknown unknown detection mechanism in a stream-based active learning application, making use of the thresholds for uncertainty and diversity. The contributions of this paper are listed as follows:

- Defined a novel unknown unknown detection algorithm which uses the thresholds for uncertainty and diversity to determine low entropy and high diversity points.
- Conducted an empirical study with the datasets Mini Imagenet and Imagenet-A, comparing with state-of-the-art approaches in anomaly detection.
- Studied the impact of the uncertainty and diversity thresholds over several acquisition functions in terms of the unknown unknown detection capability.

2. Related works

Detection of unknown unknowns and anomalies in machine learning is of paramount importance in the case of deployment in safety critical applications. Several studies have researched about effective techniques to tackle these problems. Isolation Forest, proposed by Liu et al. [13], is a powerful anomaly detection algorithm capable of efficiently handling high-dimensional data, and is a popular choice in the industry. It utilizes the principle of isolating anomalies, making it potentially suitable for detecting unknown unknowns efficiently. The Isolation Forest algorithm constructs a random forest of isolation trees, where anomalies are expected to have shorter average path lengths. Studies such as Liu et al. [13] have demonstrated the effectiveness of Isolation Forest in identifying anomalies in diverse applications, including network intrusion detection and fraud detection. Isolation forest is marked by its ability to handle high-dimensional data and its resistance to outliers and this makes it a popular choice in anomaly detection tasks.

Local Outlier Factor (LOF), introduced by Breunig et al. [14] is another widely studied anomaly detection technique. LOF measures the degree of local deviation of a data point with respect to its neighboring points, enabling it to identify anomalies based on the concept of differing densities. Various studies have focussed the application of LOF in anomaly detection tasks such as Papadimitriou et al. [15], where LOF was applied for outlier detection in sensor networks. Schubert et al. [16] used LOF for detecting anomalies in spatial databases. LOF has been shown to be effective in various domains, including cybersecurity and finance, where the detection of unknown unknowns is crucial for identifying emerging threats or fraud.

Apart from Isolation Forest and LOF, several other techniques have also been introduced for anomaly detection. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is one such algorithm proposed by Ester et al. [17]. DBSCAN groups together densely connected data points and identifies outliers as points that do not belong to any cluster. Studies such as Tang et al. [18] have applied DBSCAN for anomaly detection in computer networks, reporting its ability to detect unknown unknowns. In some works such as Hodge and Austin [19] proposed an ensemble approach that combines Isolation Forest, LOF, and other techniques to enhance the overall detection performance, ensemble-based methods such as the combination of multiple anomaly detection algorithms, have been explored.

By making use of intrinsic data characteristics, unsupervised anomaly detection methods has the potential to be applied in various domains such as fraud detection, cyber security and safety critical applications, where the identification of unknown unknowns is of paramount importance to improve safety. In this work, U3DAL, a novel unsupervised anomaly detection method is proposed.

3. U3DAL Method

Figure 2 demonstrates the quadrant of knowledge in machine learning [8]. In the top left are the known knowns. These are the data points which the model is confident about, and makes correct predictions. Therefore these data points have a low predictive entropy and are familiar, hence are not too distant from what is already seen by the model. Here, we trust the model to make the correct decisions. Known unknowns are data points which the model is underconfident about and makes wrong predictions. The dangerous situations arising from these data points can be captured easily using uncertainty based monitors. Here, we know that the model should not be trusted. Unknown knowns are human blindspots, such as latent features, but they are rich features in the model's perspective and facilitates better prediction capabilities. The last category in the quadrant consists of the unknown unknowns. These are the data points which the model makes mispredictions with a high confidence. Therefore they are categorized by a low predictive entropy (high confidence) and high diversity (different from data seen previously) score.

Figure 1 shows the block diagram of the proposed approach. As in a typical stream-based learning setting, there is a prediction model M trained on the initially available labelled data. The data stream is passed into the

prediction model, and the acquisition function decides whether the data point should be selected to be labelled by the human annotator or not. This selection is based on a preset criterion such as uncertainty or threshold, and requires thresholds for each of the criteria. Data points exceeding the threshold are passed on to the annotator to provide labels. The human oracle can only provide labels until the budget B is exhausted. The label and data point are passed to a stream buffer. When the buffer is full, the data is appended to the previously used training data and the prediction model M is re-trained. This process continues till either the budget B runs out, the data stream D stops or the prediction model reaches a sufficient level of performance. After each training of the model, it is possible for the model to be used as an unknown unknown detector as well, apart from its original functionality of classification, regression etc. This is the core idea of U3DAL- making use of the thresholds U, D and the prediction model M to determine whether a given data point is an unknown unknown data point or not. If the normalized (min-max, for instance) predictive entropy of a given data point is lower than the threshold U, and if its distance score is greater than the threshold D, U3DAL classifies that point as an unknown unknown. To evaluate the efficacy of this approach, U3DAL is compared with other state of the art approaches such as Isolation forest and LOF on the same anomaly set (all data points of which are curated to be very complex anti-examples) to compare how many of the unknown unknowns are detected accurately. Note that the approach is unsupervised because the model is not provided any prior information regarding which samples constitute unknown unknowns. The labelling that takes place in this pipeline refers to the human oracle providing class labels to the corresponding data points, which only influences the classification performance of the model on the trained task of classification and not on unknown unknown detection. The unknown unknown detection model is based on the uncertainty and diversity thresholds of selection and are not dependent on the class labels provided by the human oracle.

In U3DAL, the measurement of uncertainty is entropy, which is a well-established measure in the active learning domain. Predictive entropy is a measure of the spread of the probability distribution over all the possible classes. High entropy indicates increased randomness, which means that the model is unsure about the true class, whereas low entropy indicates that the model is confident in its prediction, regardless of its accuracy. High entropy data points are usually close to the decision boundary and therefore can be categorized as the known unknowns of the model. Identifying these data points which are close to the boundary and labelling them selectively results in an improved performance without the need to label all instances.

Figure 2: Image of a ladybug from the Mini Imagenet dataset

Figure 3: Confusing image of a ladybug from the Imagenet-A dataset

The diversity measurement is performed using the Z-score, which is a distance metric which considers the data distribution. The Z-score distance measures the distance of a data point from the centroid of the instances from the training set. A high Z-score indicates that the datapoint is not similar to the training data points seen by the model, whereas a low Z-score distance indicates that the data point is similar to those seen before[20]. Every time the model is re-trained, the mean and standard deviation vectors of the distribution of all data points of the training set encountered thus far are calculated. These vectors are then used to calculate the distance of novel data points from the distribution of the data previously seen. The Z-score of a value x with mean μ and standard deviation σ is defined as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Table 1
Classification accuracy over the train set and anomaly set for different acquisition functions (15-class problem)

No. of data points	Random		Uncertainty		Diversity	
used for training	Validation set	Anomaly set	Validation set	Anomaly set	Validation set	Anomaly set
1000	0.261	0.052	0.261	0.052	0.261	0.052
2000	0.387	0.085	0.417	0.076	0.385	0.088
3000	0.449	0.105	0.432	0.081	0.404	0.096
4000	0.516	0.118	0.506	0.098	0.428	0.113

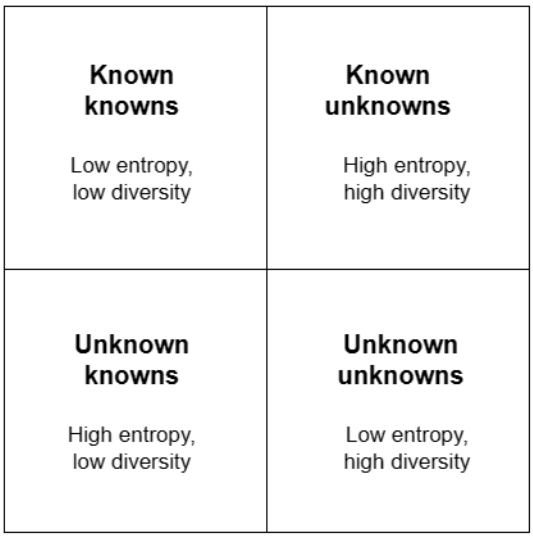

Figure 4: Link between unknown unknowns, predictive entropy and feature diversity

Algorithm 1 U3DAL algorithm

Require: Data stream: $d = x_0, x_1, ...x_n$,
Budget: B, Uncertainty selection threshold: $U \in [0, 1]$, Diversity selection threshold: $D \in [0, 1]$, Stream buffer: S, Current uncertainty range: E_{min}, E_{max}, Current diversity range: D_{min}, D_{max}, Training vector mean: μ, Training vector standard deviation: σ, Classification model: M, trained on initially labeled data
Initialize $labeled = 0$
while $labeled < B$ **do**
  $E_{norm} \leftarrow \frac{(E - E_{min})}{(E_{max} - E_{min})}$
  $D_{norm} \leftarrow \frac{(D - D_{min})}{(D_{max} - D_{min})}$
  if selected by acquisition function **then**
    append d_{index} to stream buffer S
    $labeled \leftarrow labeled + 1$
  end if
  if stream buffer S full **then**
    Re-train model, empty buffer
    Initialize $anomalyCount = 0$
    for data point in anomaly set **do**
      if $E_{norm} < U$ and $D_{norm} > D$ **then**
        $anomalyCount += 1$
      end if
    end for
  end if
end while
Output: $anomalyCount$: Number of detected unknown unknown data samples in anomaly set

4. Evaluation

This section introduces the test methodology used in this work and presents the experimental results.

4.1. Test methodology

The proposed algorithm was tested on the classification problem on the Imagenet-A dataset, which is a challenging dataset that causes machine learning model performance to degrade substantially. The authors of [21] report that on this dataset, well known CNN models exhibit an accuracy drop of approximately 90%. The data points are chosen to be those with limited spurious cues, collected with a simple adversarial filtration technique. The Imagenet-A dataset contains images belonging to fifteen classes; 759 of them constitute the anomaly set in this work. Each image is resized to the dimension 224 x 224 x 3, which is a standard input shape for most well known CNN models used in transfer learning. Note that the anomaly set shall not be seen by the model at any stage of training, and will only be used as the dataset to assess the number of anomalies detected after each round of training. The training set for the classification task consists of images from the same fifteen classes, taken from the Mini Imagenet dataset [22]. These images are also resized to the dimension 224 x 224 x 3 and are 9000 in number, 600 from each class. 1000 images are selected to be the initial labelled points in the active learning pipeline, and a further 1000 points are set aside to be the validation set. Data points from Mini Imagenet dataset, fed in a stream to simulate a stream-based active learning setting, are used to train the model to perform image classification. Unknown unknown detection is not the trained task of the model, and is instead accom-

plished using the selection thresholds for entropy and diversity achieved during the training process. Note that the Imagenet-A data samples are the anomalous samples used solely for testing the performance of the unknown unknown detection model and are not seen by the model during test time. Figures 2 and 3 demonstrate how the Mini Imagenet dataset consists of normal data points use for training whereas Imagenet-A consists of more complex and confusing data points.

Since the goal of the work is to evaluate unknown unknown detection in a stream-based setting, the remaining data points are fed into the active learning pipeline one by one. The order of the data points to be fed to the pipeline is shuffled, but the random seed is fixed in order to facilitate comparison between different settings. In the following experiments, the maximum budget B is set to be 4000 data points, meaning that at most 4000 data points out of the dataset are fed to the human oracle for labeling.

Transfer learning based on the Mobilenet backbone [23] is used as the prediction model. As the onus of the paper is on unknown unknown detection, multiple architectures were not tested for the prediction model. However, as the proposed approach is model agnostic, there are no limitations to apply the same for other model architectures. The architecture used is: Mobilenet backbone + GlobalAveragePooling2D + Dense(1024) + Dense(512) + Dense(100) + Dense(15). The penultimate fully connected (Dense) layer acts as the base to extract the intermediate features, in order to compute the diversity score, as well as input to the baselines of Local Outlier Factor (LOF)[14] and Isolation forest[13]. The other parameters are as follows: $B = 4000$, Buffer size = 1000. Since the 15-class classification problem in Mini Imagenet included a total of 9000 images, the total budget B was set to be 4000 (<50% of all images) to simulate a realistic active learning setting with limited time and resources. The buffer size was selected to be 1000 to ensure that the model is not re-trained too often, to follow time constraints of training. The Mobilenet backbone was selected because it is a very popular CNN model used for image classification tasks. Ablation studies are possible with different architectures, budget values, buffer sizes and thresholds and this is deferred to future work.

The algorithms used for unknown unknown detection in this work are as follows:

- **Local Outlier Factor (Baseline)**: Identifies anomalies with the concept that outliers have different densities compared to their neighboring data points.
- **Isolation forest (Baseline)**: Measures the anomaly score based on the average path length required to isolate instances.
- **U3DAL (Our approach)**: Detects anomalous points as those having a low entropy and high diversity, in a stream-based setting.

LOF is a flexible algorithm, and it can handle different types of data and adapt to various data distributions. It is particularly useful in situations where the normal data points exhibit complex patterns. Isolation forest is efficient and is capable of dealing with high-dimensional data, and is thus useful for detecting anomalies in various applications. The above algorithms are extremely popular in the world of anomaly detection, and they form a good baseline to evaluate the efficacy of the proposed method because of their extensive use in the industry.

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	91	69	56
U=0.6	116	85	68
U=0.7	122	88	70

Table 2
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Random

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	84	62	46
U=0.6	96	69	52
U=0.7	108	77	58

Table 3
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Uncertainty

Threshold	D=0.5	D=0.6	D=0.7
U=0.5	90	69	55
U=0.6	100	76	57
U=0.7	104	78	59

Table 4
Variation of number of unknown unknown data points detected as a function of the uncertainty threshold (U) and diversity threshold (D), acquisition function = Diversity

The proposed method is evaluated with the following acquisition functions:

- **Random selection**: Data points from the stream are selected at random to be queried to the annotator.
- **Entropy/uncertainty-based selection**: Data points are selected to be labeled if they have a predictive entropy higher than a preset threshold.
- **Distance/diversity-based selection**: Data points are selected to be labeled if they have a Z-score higher than a preset threshold.

Table 5
Comparison of the number of unknown unknown data points detected by LOF, Isolation forest, U3DAL

No. of data points	Random			Uncertainty			Diversity		
used for training	IF	LOF	U3DAL	IF	LOF	U3DAL	IF	LOF	U3DAL
1000	4	17	35	5	18	55	15	17	48
2000	9	29	59	22	24	58	30	26	66
3000	16	30	82	27	29	93	37	31	82
4000	23	33	122	38	35	108	44	44	104

Uncertainty and diversity based methods are popular acquisition functions in active learning applications. In uncertainty-based techniques, the focus is on selecting instances that the model is unsure about- dealing with the model blindspots, whereas diversity-based techniques aim to maximize the diversity of the data points in the training set- dealing with the data blindspots. Both approaches possess different advantages, and are popular choices because they improve the robustness and generalization of the model. Random selection on the other hand is a common baseline acquisition function in active learning.

4.2. Experimental results

To demonstrate that the anomaly set is difficult for the prediction model, we evaluate the classification accuracy of the prediction model on the anomaly set over multiple rounds of active learning. Table 1 shows the classification accuracy on the anomaly set for each acquisition function. Note that the initial 1000 data points are the same in each of the acquisition functions. Subsequently, due to the differing data selection mechanism, the prediction performance differs for each acquisition function. It can be seen that the classification accuracy over the anomaly set is significantly lower than that for the validation set. This illustrates that the samples from the anomaly set are vastly more challenging than the ones used for training and validation. It is an expected result because Imagenet-A was curated to be a challenging dataset. Since the model confidently mispredicts the data points, as expected Imagenet-A consists of unknown unknown data points. Thus, in the following experiments, the goal is to evaluate which algorithm can determine the unknown unknown data points contained in the anomaly set with a higher accuracy score.

In the first experiment, we compare the variation of the anomaly detection capability of U3DAL for various uncertainty and diversity thresholds. Note that in U3DAL unknown unknown data points are defined to be the low uncertainty-high diversity data points. This means that the data points with an entropy lower than the current threshold and with a diversity score higher than the current threshold are predicted to be the unknown unknown data points. Tables 2,3, and 4 illustrate how the uncertainty and diversity thresholds influence the number of unknown unknowns correctly detected for each acquisition function. Combinations of 0.5, 0.6 and 0.7 were tested for both the uncertainty and diversity thresholds. It can be observed that the best configuration for this anomaly set is U=0.7 and D=0.5. This implies that only data points with a normalized prediction entropy of lower than 0.7 and those with a normalized Z-score greater than 0.5 are classified as unknown unknown data points. This configuration is shown to detect the highest number of unknown unknown data points. The variation amongst the acquisition functions seem to be insignificant for the most part.

In the second experiment, we stack up the baseline outlier detection methods of LOF and Isolation forest against U3DAL in this use case. We observed that in the stream-based setting with a challenging anomaly set, U3DAL outperformed both LOF and Isolation forest in detecting the unknown unknown data points contained in the anomaly set. Table 5 reports the number of unknown unknown data points detected by the methods LOF, Isolation forest and the proposed method U3DAL. It can be seen that U3DAL, making use of the uncertainty and diversity thresholds, is able to detect more number of unknown unknowns than the baseline methods. As the active learning cycle proceeds and more data points are labelled by the oracle, we can see an improvement in the unknown unknown detection in all of the algorithms. This is expected because as the model is trained further, the predictive performance (influencing the uncertainty score) and the richness of the features (influencing the diversity score) improves drastically. As the model also comes across more data points, it learns the distribution of the data better and when the normalized entropy scores and diversity scores are computed, the thresholds become a better filter for detecting unknown unknowns. In an adaptive threshold setting wherein the threshold changes to adapt for data distribution shift, the performance could be expected to be even better, although it is out of the scope of this work.

5. Conclusion

In this paper, we proposed a novel method titled U3DAL to detect unknown unknowns in an unsupervised manner in a stream-based active learning setting. In order to evaluate the effectiveness of our approach, we conducted experiments on the Imagenet-A dataset, and compared the performance of our approach with existing methods for detecting unknown unknowns. Our results demonstrate that U3DAL outperforms existing methods across different metrics. By detecting unknown unknowns in real-time, our approach can help prevent unexpected failures and ensure the safety and reliability of machine learning systems in real-world safety-critical applications.

Acknowledgments

This work is partially funded by TAILOR, an ICT-48 Network of AI Research Excellence Centers funded by EU Horizon 2020 research and innovation programme under grant agreement No 952215.

References

[1] J. Attenberg, P. Ipeirotis, F. Provost, Beat the machine: Challenging humans to find a predictive model's "unknown unknowns", Journal of Data and Information Quality (JDIQ) 6 (2015) 1–17.

[2] H. Lakkaraju, E. Kamar, R. Caruana, E. Horvitz, Identifying unknown unknowns in the open world: Representations and policies for guided exploration, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[3] P. Zhao, Y.-J. Zhang, Z.-H. Zhou, Exploratory machine learning with unknown unknowns, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 10999–11006.

[4] S. Sharifi Noorian, S. Qiu, U. Gadiraju, J. Yang, A. Bozzon, What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 882–892.

[5] A. Patcha, J.-M. Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer networks 51 (2007) 3448–3470.

[6] I. Ben-Gal, Outlier detection, Data mining and knowledge discovery handbook (2005) 131–146.

[7] S. Thomas, N. Tabrizi, Adversarial machine learning: A literature review, in: Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I 14, Springer, 2018, pp. 324–334.

[8] R. Munro, R. Monarch, Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI, Simon and Schuster, 2021.

[9] B. Settles, Active learning literature survey (2009).

[10] Y. Yang, M. Loog, Active learning using uncertainty information, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2646–2651.

[11] Y. Yang, Z. Ma, F. Nie, X. Chang, A. G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, International Journal of Computer Vision 113 (2015) 113–127.

[12] C. C. Loy, T. M. Hospedales, T. Xiang, S. Gong, Stream-based joint exploration-exploitation active learning, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1560–1567.

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

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, 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.

[15] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, C. Faloutsos, Loci: Fast outlier detection using the local correlation integral, in: Proceedings 19th international conference on data engineering (Cat. No. 03CH37405), IEEE, 2003, pp. 315–326.

[16] E. Schubert, A. Zimek, H.-P. Kriegel, Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection, Data mining and knowledge discovery 28 (2014) 190–237.

[17] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: kdd, volume 96, 1996, pp. 226–231.

[18] J. Tang, Z. Chen, A. W.-C. Fu, D. W. Cheung, Enhancing effectiveness of outlier detections for low density patterns, in: Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 Proceedings 6, Springer, 2002, pp. 535–548.

[19] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artificial intelligence review 22 (2004) 85–126.

[20] P. T. Rajendran, G. Ollier, H. Espinoza, M. Adedjouma, A. Delaborde, C. Mraidha, Safety-aware active learning with perceptual ambiguity and severity assessment (2022).

[21] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, D. Song, Natural adversarial examples, in: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, 2021, pp. 15262–15271.

[22] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Advances in neural information processing systems 29 (2016).

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).