Neurocomputing 599 (2024) 128073

Contents lists available at ScienceDirect

Neurocomputing

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

Survey Paper

Managing the unknown in machine learning: Definitions, related areas, recent advances, and prospects

Marcos Barcina-Blanco [a,b,*], Jesus L. Lobo [a], Pablo Garcia-Bringas [b], Javier Del Ser [a,c]

[a] TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, 48160, Spain
[b] Faculty of Engineering, University of Deusto, Bilbao, 48007, Spain
[c] University of the Basque Country (UPV/EHU), Bilbao, 48013, Spain

ARTICLE INFO

Communicated by N. Zeng

Keywords:
Open set environments
Open-ended learning
Open-world learning
Open set recognition
Uncertainty
AI alignment

ABSTRACT

In the rapidly evolving domain of machine learning, the ability to adapt to unforeseen circumstances and novel data types is of paramount importance. The deployment of Artificial Intelligence is progressively aimed at more realistic and open scenarios where data, tasks, and conditions are variable and not fully predetermined, and therefore where a *closed set* assumption cannot be hold. In such evolving environments, machine learning is asked to be autonomous, continuous, and adaptive, requiring effective management of uncertainty and the unknown to fulfill expectations. In response, there is a vigorous effort to develop a new generation of models, which are characterized by enhanced autonomy and a broad capacity to generalize, enabling them to perform effectively across a wide range of tasks. The field of machine learning in open set environments poses many challenges and also brings together different paradigms, some traditional but others emerging, where the overlapping and confusion between them makes it difficult to distinguish them or give them the necessary relevance. This work delves into the frontiers of methodologies that thrive in these open set environments, by identifying common practices, limitations, and connections between the paradigms Open-Ended Learning, Open-World Learning, Open Set Recognition, and other related areas such as Continual Learning, Out-of-Distribution detection, Novelty Detection, and Active Learning. We seek to easy the understanding of these fields and their common roots, uncover open problems and suggest several research directions that may motivate and articulate future efforts towards more robust and autonomous systems.

1. Introduction

Up to this point and continuing into the present, the predominant application of Artificial Intelligence (AI) has been centered around models capable of executing particular tasks, often operating under careful guidance, complete supervision and controlled operational settings. Although models used under such circumstances have demonstrated their efficacy in numerous scenarios and remain relevant from a practical perspective, there is an undeniable progressive shift towards emphasizing autonomy and broader applicability in open world scenarios. Consequently, there is a fervent quest for the emergence of a new era of Machine Learning (ML) models characterized by enhanced autonomy and generalization to perform a wide variety of tasks; models capable of handling situations where they encounter data that they were not explicitly trained on [1], particularly in scenarios where they may not even be aware of the existence of certain classes or categories of data.

The *Managing the Unknown in Machine Learning* (MUML) paradigm addresses specific challenges that distinguish them from other traditional ML problems. Unlike conventional ML models that operate subject to fixed conditions (e.g., tasks, classes, features, stable distributions, or learning objectives) – which are known as *closed environment* scenarios [2], MUML models can recognize and handle unknown situations, e.g., new or unseen classes during inference. This is critical in real-world settings, where data can evolve over time or when new categories emerge after the model has been trained. These models are also often designed to update their knowledge base dynamically as they receive new data points. This includes not only recognizing new categories, but also integrating them into the learned knowledge of the model without requiring a complete retraining from scratch. In addition, they must continuously adapt to new data and changing environments; this adaptation can involve learning new features, adjusting to new data distributions, or modifying the decision boundaries. MUML is particularly useful in fields such as security (identifying new

* Corresponding author at: TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, 48160, Spain.
E-mail address: marcos.barcina@tecnalia.com (M. Barcina-Blanco).

https://doi.org/10.1016/j.neucom.2024.128073
Received 5 January 2024; Received in revised form 15 May 2024; Accepted 12 June 2024
Available online 14 June 2024

types of threats), medical diagnostics (new diseases), and robotics (new objects or environments), where the ability to generalize beyond the initial training data is crucial. The major challenges of learning in open scenarios include determining when and how to incorporate new knowledge, managing the potential increase in model complexity, and ensuring that the model remains accurate and robust as it adapts.

Several areas in ML research have focused on addressing the challenging need for MUML, each from a different perspective, some overlapping and others complementing each other. This is the case of Open-Ended Learning (OEL) [3], a very recent paradigm that aims to go one step further and mirror the open-ended nature of human cognition. OEL refers to AI systems designed to engage in an endless learning process, where there is no fixed set of tasks, goals, or data to learn. Rather than optimizing towards a specific goal or mastering a finite set of skills, OEL systems have the ability to generate and pursue new goals, explore new skills, and adapt to emerging challenges autonomously. This approach is inspired by how humans and other animals learn and explore the world in curious, intrinsic goal-driven ways. Despite the inherent utility of generating new modeled knowledge over time, OEL is still a novel research area, and, as such, it still lacks of a global definition, letting alone an extensive catalogue of works. Although we refer to it in this work due to its close relationship to MUML, OEL concentrates on the management of new tasks and goals rather than on the detection and modeling of unknown knowledge, which constitutes the key conundrum that motivates the present study.

In this line of reasoning, one of the most realistic and practical approaches to MUML in classification tasks comprises those formulations that do not assume the so-called *closed set* scenario, e.g., all data samples at inference time belong to at least one of the classes existing in the training data from which the ML model was learned. However, in many real-world circumstances this *closed set* assumption does not necessarily hold. This gives rise to the antagonist concept of *open set* environments where Unknown Classes (UC) can emerge during the inference phase. When this occurs, the model must detect the emergence of UC; otherwise, ML models designed under the *open set* assumption will incorrectly classify samples belonging to UC as any of the Known Classes (KC), often with a high confidence in their predictions. This procedure is indeed what endows the field with the name of *Open Set Recognition* (OSR) [4–6]. OSR models with incremental learning capabilities forge the *Open World Learning* (OWL) paradigm, or also known in the literature as *Open-World Recognition* (OWR).

The detection of UC has been central for the research community working on OSR. However, other tangential areas in ML research have also gravitated on problems that relate closely to the *open set* assumption underneath OSR, such as Novelty Detection (ND) [7], Continual Learning (CL) [8], Out-of-Distribution (OoD) detection [9], Uncertainty Estimation (UE) [10], or Active Learning (AL) [11]. This amalgam of research areas lacks a unified reference work reviewing the literature in each of them, and connecting contributions therein with each other under an unified conceptual framework. This is the overarching goal of this survey: we elucidates the shared methodologies, constraints, and interrelations among them, aiming to dispel the prevalent ambiguity surrounding research domains connected to OSR. In doing so, we reveal unresolved challenges and propose numerous avenues of research that could guide and refine subsequent efforts to develop more resilient ML techniques in open-world scenarios. Through a detailed examination of these interconnected fields, we endeavor to provide clarity, identify existing gaps, and inspire innovative strategies for advancing the robustness of ML systems, an objective of paramount importance in light of the rising importance of AI alignment and technical robustness for the trustworthiness and responsibility of AI-based systems [12,13].

This work is organized as follows. Section 2 elaborates on the motivation and significance of MUML scenarios; Section 3 delves into the OSR concept, the problem definition, and provides a thorough literature review; Section 4 shows a review of the areas related to OSR; Section 5 offers our challenges and prospects for OSR scenarios in years to come; and finally, Section 6 summarizes the main conclusions drawn from our study. A glossary of acronyms used throughout the text is given in Table 1.

2. Motivation and significance of managing the unknown in machine learning

The application of AI increasingly addresses more realistic scenarios where tasks and conditions are changing and not entirely known in advance, where learning must be autonomous, continuous, and adaptive, and where the management of the unknown and its uncertainty must be adequately managed to meet expectations. Under these circumstances, there is an intense pursuit for the development of a new generation of ML models marked by increased self-reliance and the ability to generalize across an extensive array of tasks. However, the majority of these task models continue to operate under a closed-set assumption, under which the information encountered during inference is associated with a finite (*closed*) set of categories identified during the training phase of the models. This assumption often fails to align with the complexities of real-world situations characterized by open sets that cannot be bounded beforehand. In such scenarios, the capability of the model to learn how to solve the task at hand must be augmented with the possibilities to detect, identify, consolidate and efficiently incorporate newly arising categories into their modeled knowledge.

In the literature, we can encounter an mixture of areas related to the paradigm explained above that are often difficult to identify, delineate and distinguish from each other. Despite serious efforts to clarify some of these paradigms [3–6], the community is still lacking of efforts made towards harmonizing all paradigms together (including the emerging ones), and above all, towards rigorously analyzing their overlaps, differential aspects, and challenges. In short words: connecting the dots in MUML research.

The motivation of this work emerges from this noted lack of a reference in the field. Our aim is to provide clarity when presenting each of these existing paradigms of learning in open settings. To this end, we provide a comprehensive overview of MUML, introducing its objective and importance, and then to focus on OSR and its related areas where we have found some confusion and disorder. Specifically, the *raison d'être* of the present work is:

- To identify paradigms that are part of what is known as MUML;
- To focus on those areas that manage UC in the context of classification tasks;
- To establish common grounds on the concepts and definitions managed in OSR, which have been evolving in the last few years;
- To categorize the upsurge of contributions that tackle the OSR problem with different methodological approaches that have been reported in recent times;
- To shed light on the similarities and differences with the aforementioned tangential areas, showing the manifold points of friction with them, such as ND, CL, OoD detection, UE, and AL; and
- To offer a well-informed prospect of problems and research lines of interest for the advance in this field that should drive research in the future.

This will not only benefit researchers and practitioners working in these fields, but also foster cross-disciplinary collaboration and advance the development of robust and reliable ML systems capable of handling realistic open-world scenarios.

3. Open set recognition: Problem statement, literature review and applications

In the context of ML in open environments, the OSR field [4] aims endowing ML models with the capacity to detect (and adapt) their knowledge to the appearance of new classes. From a technical perspective, OSR departs from a limited number of classes available for learning the ML model, whereas samples belonging to UC may appear during the testing phase. Models are required not only to detect and reject unknown samples, but also to correctly classify samples that belong to the KC. This seminal definition of OSR was stated in [14],

Table 1
List of acronyms and their definitions.

Acronym	Definition	Acronym	Definition
AI	Artificial Intelligence	DNN	Deep Neural Networks
ML	Machine Learning	GAN	Generative Adversarial Networks
MUML	Managing the Unknown in ML	NLP	Natural Language Processing
OEL	Open-Ended Learning	FM	Foundational Model
UC	Unknown Class/Classes	LLM	Large Language Model
KC	Known Class/Classes	RL	Reinforcement Learning
OSR	Open Set Recognition	GCRL	Goal Conditioned RL
OWL	Open-World Learning	AD	Anomaly Detection
OWR	Open-World Recognition	IF	Isolation Forest
ND	Novelty Detection	LOF	Local Outlier Factor
CL	Continual Learning	AE	Autoencoder
OoD	Out-of-Distribution	CD	Concept Drift
UE	Uncertainty Estimation	SL	Stream Learning
AL	Active Learning	COD	Continual OD
DL	Deep Learning	CRL	Continual RL
CIL	Class Incremental Learning	ID	In-Distribution
NN	Neural Network	BNN	Bayesian NN
SVM	Support Vector Machine	TTA	Test Time Augmentation
k-NN	k-Nearest Neighbors		

which also introduced the *over-occupied space* problem. This problem refers to the extra feature space that *open set* classifiers assign to each class to create more meaningful boundaries for the KC. While this strategy is effective in improving the generalization ability of the ML model under the *open set* assumption, it leads to incorrect classification of UC in open settings.

Since then, several OSR methodologies have been proposed over the years to cope with the aforementioned problem, from the modification of traditional ML algorithms to the development of Deep Learning (DL) models capable of autonomously detecting UC. More recently, attempts to combine clustering and classification algorithms [15,16] have shown promising results. Before delving into the details of OSR in this section, we first highlight that, although detecting UC is vital for working in open environments, being able to incorporate them into the model's learned knowledge base is equally as important. Both requirements have served as the foundation of a OWR paradigm [17], which combines OSR with Class Incremental Learning (CIL). Since then, OWR has become commonly known as OWL [18,19], which provides models the ability to learn from new data (never seen before) while performing their expected task.

We follow our study by briefly looking at the mathematical statement of the OSR problem and several definitions needed for its proper understanding (Section 3.1), including the *over-occupied space* problem mentioned previously. Then, Section 3.2 provides a literature review of the existing OSR approaches. Section 3.3 reviews some of the real-world applications of OSR. Finally, Section 3.4 provides an overview of the OSR related areas, what they share in common and what makes them differ from each other. We will discuss in depth about the similarities and differences for each area in Section 4.

3.1. Problem definition

Conventional classification tasks formulated under an *open set* assumption assume that all samples belong to one of the KC during training. As such, a classification model M is trained on a dataset comprising KC $\mathcal{Y}_{KC} = \{Y_1, Y_2, \dots, Y_N\}$. Once learned, the model is queried with samples at inference time, which produces a predicted class belonging to $\mathcal{Y}_{KC}$ for each test instance. In this *open set* scenario, given any input $\mathbf{x}$, the classifier M estimates a probability distribution over all training classes $p(y|\mathbf{x})$, with $y \in \mathcal{Y}_{KC}$. This will yield a classification error whenever the classifier is queried with a new instance $\mathbf{x}$ that does not belong to any of the classes in $\mathcal{Y}_{KC}$. Fig. 1(a) schematically represents this problem for a problem with three KC ($|\mathcal{Y}_{KC}| = N = 3$) and 2 features ($|\mathbf{x}| = 2$).

As shown in Fig. 1, traditional *open set* classifiers assign the entire feature space to the N KC to create more meaningful discriminatory boundaries between the KC and, ultimately, achieve a better generalization under the *open set* assumption. However, this over-occupation of the space means that any unknown sample will fall in this space and be incorrectly classified as one of the KC.

OSR tackles this problem by devising a scenario in which knowledge of the full set of classes is not available during training. An *open set* classification model is trained to discriminate between KC in $\mathcal{Y}_{KC}$, but test inputs may also come from UC $\mathcal{Y}_{UC}$. In this situation, the model must be capable of predicting classes from $\mathcal{Y}_{KC} \cup \mathcal{Y}_{UC} = \{Y_1, Y_2, \dots, Y_N, Y_{N+1}, \dots, Y_{N+\Omega}\}$, where $\mathcal{Y}_{UC}$ represent all the UC not seen during training. The model should be able to predict any new instance $\mathbf{x}$ as belonging to one of the KC in $\mathcal{Y}_{KC}$ or classify it as unknown (i.e., as belonging to $\mathcal{Y}_{UC}$).

Fig. 1 also shows all the feature space that is far from the training samples belonging to KC. This is known as the *Open Space*. The work in [14] discusses the risk associated with labeling samples that are far from the training samples in the *Open Space*, giving rise to the *Open Space Risk* $R_O(f)$ defined as:

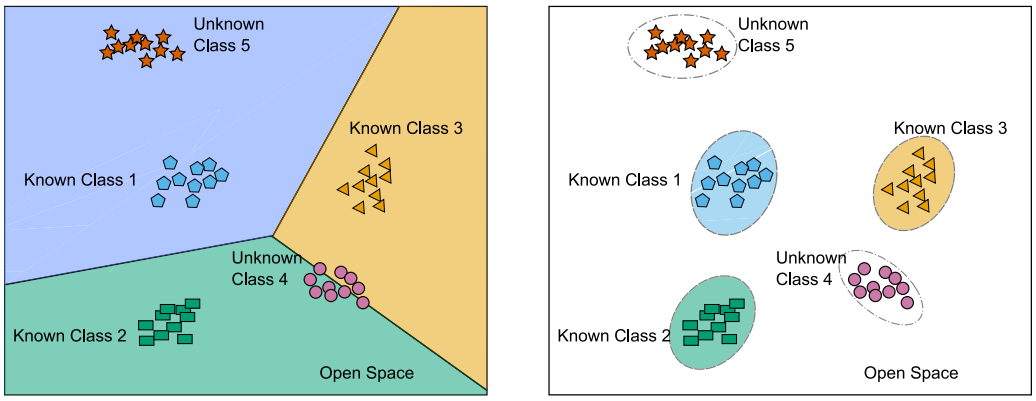
$$R_O(f) = \frac{\int_O f(\mathbf{x}) d\mathbf{x}}{\int_{S_O} f(\mathbf{x}) d\mathbf{x}}, \tag{1}$$

which is a relative measure of positively labeled *Open Space* O to the total measure space S_O, covering the known positive samples and O. The function f is a recognition function that depends on the OSR model/strategy in use, where $f(\mathbf{x}) = 1$ means that a known class is recognized ($f(\mathbf{x}) = 0$ otherwise). The more *Open Space* labeled as positive, the greater R_O becomes, and hence the higher the risk taken by the model when operating in an open set scenario will be.

Therefore, there is a direct connection between the *Open Space Risk* and the *over-occupied space* problem as the latter contributes to such risk: when the decision boundaries of a classifier are too expansive or overly aggressive, it increases the probability that the model will encounter and wrongly classify inputs from the *Open Space*. Thus, overly broad decision boundaries can lead to higher *Open Space Risk*. Minimizing *Open Space Risk* involves carefully defining the decision boundaries so that they do not overly encroach on the *Open Space*. This requires a balance between being restrictive enough to avoid large *Open Space Risk* while still being general enough to correctly classify new examples of KC. Techniques in OSR often try to mitigate these issues by regulating how much feature space the model is allowed to classify in a confident fashion.

3.2. Literature review

Open Space Risk (R_O) is at the center of OSR. The methods must be able to correctly classify the KC by minimizing the empirical risk



(a) Closed-set classifier. (b) Open set classifier.

Fig. 1. Differences between *closed set* and *open set* classification models: (a) Decision boundaries of an *closed set* classifier for a problem with $N = 3$ KC and $\Omega = 2$ UC. All the feature space is divided between the KC, so that samples from UC arriving at inference time will be incorrectly classified, in some cases (e.g., unknown class 4) with high class probability. (b) Illustration of the delimited feature space that each KC occupies in *open set* classifiers, allowing for an effective detection of UC.

and recognizing the UC samples that are outside the space assigned to the KC by minimizing the R_O. Since OSR was first presented as such in the community, many works have been done in this area from different modeling perspectives. Previous surveys on OSR [4–6,20] have agreed on the categorization of existing methods in two main categories: *discriminative* and *generative*, although we can find some approximations that combined the two [21]. Discriminative methods model the existing KC data into tighter spaces to reduce R_O, while generative approaches synthesize fake UC samples to train the model and cover the *Open Space*. A taxonomy for representative OSR works and its closest research areas is shown in Table 2. We now examine contributions classified in this taxonomy, including those proposing discriminative approaches (Section 3.2.1) and generative methods (Section 3.2.2).

3.2.1. Discriminative methods

As stated above, discriminative methods aim at reducing the *Open Space Risk* (R_O) by modeling the data into smaller spaces in the feature space. This has been approached mainly by adapting traditional ML methods [14,17,33–43], or by tailoring the Neural Network (NN) training process to perform inference in environments subject to the appearance of new concepts [45–57].

The authors in [14] explained the OSR problem as a balance between reducing R_O and the empirical risk. For this purpose, they proposed a so-called "1-vs-set Machine" model based on Support Vector Machines (SVMs). In [33] a framework that combines a Conditional Random Field and a Probability of Inclusion SVM was proposed for the rejection of UC. The authors in [34] presented the Extreme Value Machine, a model based on statistical Extreme Value Theory, where this approach naturally reduces R_O and decides which new samples are unknown based on a threshold. In [35], the authors have proposed a few-shot learning method that generalizes the maximum likelihood principle to detect samples that do not belong to any KC, and that can be applied to any pre-trained model. Other approaches naturally suited for predicting UC are Gaussian Processes [36], because they assign an uncertainty measure to predictions.

Distance-based models have also been modified in order to work in *open set* scenarios. One of the first works contributing in this direction was the Nearest Non-Outlier algorithm, a modification of the Nearest Class Mean classifier [17]. Furthermore, other traditional distance-based classifiers have been adapted in different ways, such as the Nearest Neighbor Classifier [37,38]. In [37] an extension to the k-nearest neighbors (k-NN) algorithm was proposed, coined as Nearest Neighbor Distance Ratio. More recently, the work in [38] has assumed that samples from KC/UC are likely to be surrounded by other samples belonging to KC/UC. The Otsu's method [230] has been used to select the optimal threshold to classify into KC and UC; after the classification is done, the samples considered to belong to KC are passed to a Random Forest classifier learned from the training samples to perform regular *open set* classification. In [39], the authors presented a weightless artificial NN model to achieve an effective combination of classification with precise identification of extraneous data.

In recent years, prototype learning has elicited promising results in the OSR field. Prototypes are representative samples that reflect other samples from its class. The use of prototypes allows for more compact feature representations of classes, which naturally creates clearer boundaries between KC and UC in *open set* scenarios. In [40] the authors have introduced a Convolutional Prototype Network where prototypes per class are jointly learned during training. The authors in [41] not only used prototypes to represent classes, but also learned discriminative reciprocal points for them. These reciprocal points can be understood as the opposite of the prototype, i.e., what is not the class under target. They help to explore the *Open Space*, naturally reducing the *Open Space Risk*. Similarly, the authors in [42] have proposed that implicitly learned prototypes tend to create undesired prototypes from low-quality samples, and show redundancy in similar prototypes of one category. Unlike previous works, the work in [43] has resorted to Gaussian distributions to represent prototypes instead of sets of feature vectors; they argued that the distribution of KC in the latent feature space can be represented by one or several Gaussian-distributed prototypes.

Deep Neural Networks (DNN) offer powerful representation abilities that can be beneficial for any ML task. DNN used for classification task usually include a last softmax layer to produce a probability distribution over the KC, posing a problem in OSR settings. Furthermore, DNNs may produce high confidence scores for samples belonging to UC [45]. Approaches that rely on thresholding softmax probability scores and assuming that UC inputs produce low probability for all classes were proven to not reach optimal solutions in OSR [46]. Several workarounds were explored to overcome this issue, such as replacing the Softmax function with OpenMax [46]. Other proposals [47,48] replaced the softmax function with a "1-vs-rest" layer based on sigmoid functions, which is able to naturally reject unknown samples. In [49], an OSR algorithm using class conditional auto-encoders was presented, in which the encoder learned *closed set* classification while the decoder learned to identify KC and UC. The authors in [50] have proposed deep compact hypersphere classifiers for OSR. They have regarded hypersphere centers as learnable parameters, updating them based on the features of new samples. The recent work in [51] has proposed to swap the One-hot Encoded Label vector of a NN with a distance profile representation that estimate the distances to a relative neighborhood

Table 2
Overview of recent/relevant works and applications on MUML and related fields.

Research areas	Methods	References
Open-Ended Learning (OEL)		[22–24]
Open-World Learning (OWL)		[17,25–32]
Open Set Recognition (OSR)	Discriminative	[14,15,33–58]
	Generative	[59–72]
	Applications	[73–98]
Continual Learning (CL)	Regularization-based	[99–102]
	Experience replay	[103–108]
	Optimization-based	[109–111]
	Representation-based	[112,113]
	Architecture-based	[114–117]
	Applications	[27,118–134]
Novelty Detection (ND)	Traditional	[135–141]
	Deep Learning	[142–149]
	Applications	[150–165]
Out-of-Distribution (OoD) detection	Energy-based	[166–170]
	Outlier exposure	[171–174]
	Density-based	[175–180]
	Distance-based	[181–187]
	Reconstruction-based	[188–190]
	Applications	[190–198]
Uncertainty Estimation (UE)	Bayesian Neural Networks	[199–204]
	Ensemble-based	[205–207]
	Test Time Augmentation	[208]
	Applications	[204,207,209–216]
Active Learning (AL)	Information based	[217–221]
	Expected change	[222,223]
	Representation-based	[15,224]
	Applications	[225–229]

for any given sample. The density is then computed from this vector and can be interpreted as probability via normalization, which makes it potentially suitable for OSR. In another recent work [52], the authors have suggested forcing a NN to learn mutually different attention maps (the parts of an input that are important for making predictions). They have showed that differences in attention maps leads to diverse representation and the reduction of *Open Space Risk*. Alternatively, the authors in [53] have argued that representing data in the frequency domain instead of the usual spatial domain is better suited for OSR. Other approaches have formulated AL as an OSR problem [54], or used generative information [55]. Other works such as [56,57] have incorporated incremental learning to OSR. They have warned that simply incrementally learning from the UC can be detrimental for the boundaries between the KC and UC features. To overcome this issue, they have proposed a framework designed for OSR and class-incremental learning. Finally, an OSR python module that facilitates the correct experimental evaluation of DNN-based solutions has also been developed very recently [231].

As exposed in this first literature analysis, former discriminative attempts to solve this problem have tried to directly modify the classifiers so that they can deal with inputs belonging to UC. Alternatively, clustering models can be combined with classification to address this problem [4]. Clustering models depart from a measure of similarity among samples, so that those belonging to a certain KC are closer to each other than to those belonging to the rest of KC and UC. They naturally partition the feature space when modeling KC. In order to deal with the *over-occupied space* problem, clustering algorithms can be used to better adjust KC to a confined area, while classification models are used to discriminate between them.

Many efforts for combining clustering and classification exist for *open set* scenarios, which can be implemented either sequentially or simultaneously. Those that work in a sequential fashion first apply clustering techniques over the training data to characterize the feature space corresponding to the KC. This characterization is then exploited to improve the robustness of the subsequent classification model against data belonging to UC. A common approach following this sequential procedure can be found in [232], where both techniques have been combined for malicious attacks classification. Another proposal using clustering to improve classification is [233], which first applied Density-Based Spatial Clustering of Applications with Noise over the training data The obtained centroids were used as the extreme vectors of an Extreme Vectors Machine, hence overriding the need for comparing all training samples and significantly reducing the computational effort of the overall classifier. In [16], images belonging to each class of camera models were first clustered; then a Support Vector Data Description model was trained on each of the clusters to create several hyperspheres. Samples that fall outside of their modeled space are considered to belong to UC. The rationale behind this approach is that several hyperspheres adapt better to the distribution of each class and thus reduce the *over-occupied space*. While these methods result in better performance, they do not take full advantage of the combination, since the clustering techniques do not benefit from the information produced in the classification stage [234].

Techniques simultaneously applying clustering and classification aim to ensure that both algorithms receive feedback from each other during the training phase. Early contributions in this line proposed combining both criteria into a single objective of the model, which was then optimized [234,235]. Although there has been far more research in this line of work [236–239], none of them assume an *open set*

scenario. These works share that clustering and optimization depend on the cluster centers, and use a relation matrix in order to compute the relationship between cluster membership and class membership. Then, each of them applies its own multi-objective optimization framework to find the matrix and centers that best satisfy the clustering and classification indexes. More recent multi-objective optimization approaches [240,241] may be worth exploring to better apply these simultaneous strategies.

Other simultaneous approaches combined ensembles of classifiers and clustering models [242]. The class probability distribution for each test instance (from the classifiers ensemble) and a similarity matrix (from the clusters ensemble) are combined together to produce a posterior class probability assignment for each test instance. In [15] this work was extended to make it more suitable for open world settings. While this approach rendered a good performance in OSR, it still undergoes issues when the space of new UC overlaps with the space of KC, and requires significant processing time to retrain the model after each iteration. In addition, the work in [58] combined a discriminative classifier and a flow-based model to address OSR. Flow models can predict the probability density for each sample: samples that have a large probability density are likely to belong to the training distribution, whereas samples that have a small probability density value usually belong to an UC. Instead of adding a classifier after the flow model, a joint embedding was created for both the classifier and the flow model. This was done in order to ensure more discriminative expressiveness in the embedding space (rather than using the space that the flow model creates on its own).

To the best of our knowledge, beyond the works reviewed above very scarce attempts have been done to combine clustering and classification towards reducing the *over-occupied space* from an OSR perspective. Our prospects offered later in this manuscript will further revolve around this niche.

3.2.2. Generative methods

Unlike their discriminative counterparts, generative methods seek to produce synthetic data samples that represent the UC, so that the model is trained over real and synthetic samples. This results in a more traditional classification task with the addition of one or several extra classes. Generative Adversarial Networks (GAN) were first applied to OSR [59]. These models generate synthetic samples and assign them a new label that represents the UC, so that, when added to the training data from which the classifier is learned, allows predicting samples belonging to UC. Other attempts reported in [60,61] also synthesized samples that are representative of UC, showing that, when used to train the model, make it more robust against samples from actual UC. Similarly, the authors in [62] generated counterfactual images, i.e., those similar to real samples from the KC that are close to the UC. The algorithm proposed in [63] generates synthetic anomalies close to the *open set* boundary of the classifier; the distance at which samples are generated is balanced against a constraint that ensures that the *open set* classification performance is not degraded. Synthetic samples generated in this manner are used to reduce R_O while not increasing empirical risk. The authors in [64] have proposed to use two GANs: generated adversarial samples that have low noise variance are used to increase the density of KC, while samples with high noise variance are dispersed in *Open Space* to decrease R_O. Sometimes the learned model from the known-class domain might be unsuitable for the unknown-class domain. The authors in [65] have addressed this by designing a dual-space consistent sampling approach and a conditional dual-adversarial generative network that generates discriminative features of both KC and UC, which are able to accommodate any inductive OSR method. Few-shot and Zero-shot learning methods have also been adapted for OSR [66,67,123]. Adversarial samples have also been used to enhance other discriminative methods such as prototype learning [68]. In the case of [69], authors have studied the relations between semi-supervised learning and OSR in the context of GANs. Data augmentation techniques have also been applied to OSR [70]. As can be inferred from the reviewed literature, the key to generative methods is to synthesize representative samples of all possible UC, and thus maximize the decrease of R_O resulting from the inclusion of such samples in the training data of the classifier.

3.3. Relevance and applications of open set recognition

OSR plays a crucial role in improving the generalization capabilities and robustness of ML models in dynamic real-world environments. The reason is that real-world applications are prone to encountering data from UC. If models used for medical diagnosis or autonomous driving incorrectly classified UC into existing KC, it would have undesirable consequences for the people interacting with those systems. From a practical standpoint, OSR has been proven beneficial to identify the unknown in practical applications prone to the appearance of unseen classes, including the field of computer vision [73–80] or applications such as cybersecurity [81–84], bio-metrics [85], face recognition [86,87], monitoring systems [88,89], or sport video analysis [90], to mention a few. Natural Language Processing (NLP) is another field where OSR has been applied to improve the reliability of dialog systems [91] or to handle characters out of the training vocabulary [92] and text classification [93]. OSR has also been applied for action recognition tasks [94], which require the model to identify video actions from KC while simultaneously detecting unknown behavioral actions in the scene. The authors have leveraged the semantic information through the creation of a concept relation graph and the use of visual prototypes to preserve the intrinsic semantic structure of the classes.

In recent times, Foundational Models (FM) have gained a lot of popularity. Their versatility and generalization capabilities are very valuable for many fields, such as NLP or computer vision. They have shaped many of the recent contributions and advances in AI research. The OSR field is no exception, and several applications that make use of FMs have already been reported. In [95] the authors proposed EdgeFM, an edge-cloud cooperative system that can achieve OSR capability by querying the FM for selective knowledge on uncertain data samples. This is done to customize the domain-specific knowledge and architecture of the small models that exist on the edge. Similarly, the authors in [96] have investigated how to leverage the rich knowledge of Large Language Models (LLMs) to handle the OSR image classification task. They propose an OSR framework named Large Model Collaboration that is able to leverage and extract the implicit knowledge of several large models in a training-free manner. Other OSR image classification works [97] have also leveraged the rich knowledge of LLMs. Here, the authors consider the tags annotated in the images as insufficient, so they employ the LLMs to generate visual descriptions for each tag category. This added knowledge improves the generalization capability of the model and allows it to deal with a wider range of visual descriptions from open (unseen) categories.

3.4. Research areas related to open set recognition

The appearance of new classes is a common phenomenon in real-life ML applications, receiving increased attention in the community given the progressively higher universal access and exposure of ML models in recent times. As a result, different research areas have embraced this problem as part of their scope, which often overlap. This makes it difficult to establish a clear boundary between them. We now provide clarity in this regard.

In the OSR sphere we find the OWL paradigm [18]. This concept was first mentioned in [17], where the authors expanded the OSR definition to OWR by adding the CIL task to it. By jointly considering the OSR and CIL tasks, an OWR model should perform the following tasks: (i) to detect samples belonging to UC classes (while still classifying samples from KC); (ii) to choose which samples to label and add to the model;

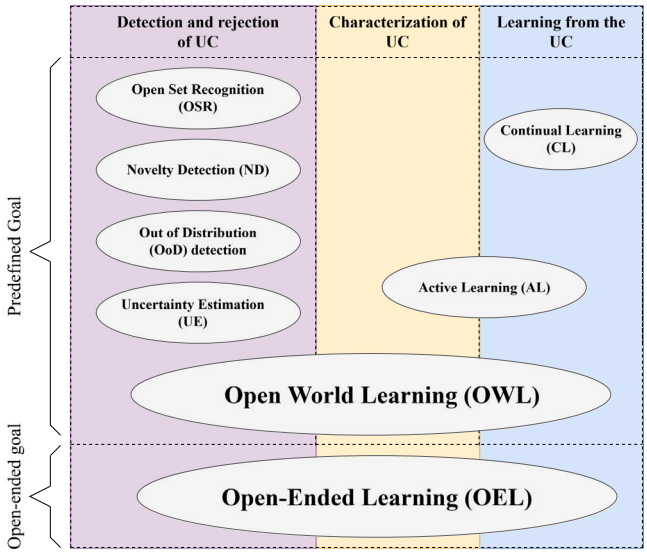


Fig. 2. Relationship between OEL, OWL, and OSR and their tangential areas. Each colored area represents one of the three main tasks of OWL. OSR, ND, OoD detection, and UE are primarily concerned with detecting UC. The main task of CL is learning from the UC. AL is concerned with both characterizing and learning from the UC. Although the previous areas are restricted to a predefined goal, OEL is capable of defining its own goals.

(iii) to label those samples and to update the model with them. Nowadays, OWR is commonly referred to as OWL [18,19]. Although branded with a different name, the learning process in OWL is still defined as per the aforementioned three steps that a model must be able to perform in an autonomous way. The first is to detect and reject UC samples, the second involves characterizing new classes from the rejected samples, and the third is to update the model with the knowledge of the new discovered classes. Once again, this combination of tasks is intertwined with other research areas such as CL, leading to potential confusion due to their overlapping goals and subtle nuances. The emergence of OWL is caused by the growing complexity and diversity of real-world applications, and techniques falling within this area have been applied to problems arising in DL [25], 3D point classification tasks [26], object detection in computer vision [27–29] and NLP [30,31], among others.

While significantly more robust in non-stationary environments, OWL models still lack the autonomy to face new tasks and learn how to solve them. An OWL model trained to detect images of dogs and cats may be able to learn to detect images from cars, but it will still be limited to the classification of images. Although the details of OEL are out of the scope of this work, it is worth mentioning that addresses this challenge by allowing models to participate in an endless process of learning, where they are able to choose, learn, and adapt to new tasks. Although a global definition of OEL does not exist to date, some contributions have considered it from a Goal Conditioned Reinforcement Learning (GCRL) perspective [3]. They define the problem of OEL as a sequence of Reinforcement Learning (RL) problems with goals that come from an open-ended generation process. They then combine this definition with CL and define lifelong open-ended GCRL, which adds to GCRL agents the need for not forgetting how to solve previous RL problems and achieve goals while learning how to solve a new RL problem or achieve other goals. This is why most of the current works related to this problem are from the RL perspective [22–24].

Apart from OSR and OWL, the problem of new classes has also been prominently investigated in ND, CL, OoD, UE, and AL. Although these areas share very similar goals, they differ in the way they achieve them or their benchmarks, which causes confusion for newcomers to any of the fields.

Fig. 2 shows the three main tasks of OWL and how each area relates to them. The main goal of OSR, ND, OoD and UE is to detect the samples from an UC (or any abnormal samples). AL asks an oracle for information on unlabeled samples and then updates the model with that knowledge, so it is concerned with both characterization and learning from the UC. Finally, the main task of CL is to update the model without forgetting old knowledge, so they mainly focus on the third task. OWL encompasses all three main tasks but, similarly to the other areas, it is still limited to a predefined goal. However, OEL is capable of defining its own goals and evolve them continuously over time. We note that, in OEL, not every goal that a model chooses to learn has to be related with the classification scenarios that OWL, OSR and the other areas usually assume. The figure signifies that, in the event that a specific goal includes classification, the model should be able to perform the same tasks as OWL.

4. Tangential areas in OSR research

Once all fields of study have been located and have already deepened in OSR, in the following subsection we will focus on those fields (tangential areas) related to OSR and UC management: ND, OoD, UE, and AL. We leave here OWR/OWL because it is a variant of OSR with CL, and OEL because it is a paradigm more focused on the management of new tasks and goals than on the appearance of new classes. Therefore, in this section, we review each of the related areas, explain the similarities and differences with OSR, and provide a brief look at their applications and challenges. Fig. 3 depicts a visual summary of the tangential areas to be discussed in this section, alongside the general challenges and research directions later offered in Section 5.

4.1. Novelty detection

This area focuses on the detection of new test samples that have not been observed by the model during training. It should not be confused with Anomaly Detection (AD). While AD assumes that abnormal samples may already exist in the training data and focuses on finding them, ND assumes clean training data without any anomalies and expects a semantic shift (namely, appearance of novel classes) to occur during the testing phase [7,243,244]. In ND the normal pattern is represented by a dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ where each $\mathbf{x}_i$ represents a sample. A model that captures the distribution of normal or typical data points in the dataset is created, and it is then used to assign a novelty score $\xi(\mathbf{x})$ to each testing sample. Larger novelty scores correlate with more abnormal samples. Based on this, a ND threshold γ_{ND} is defined so that a sample is considered normal when $\xi(\mathbf{x}) \leq \gamma_{ND}$ and abnormal when $\xi(\mathbf{x}) > \gamma_{ND}$. Although this is a general definition of ND, many different methods exist in the literature to set the parameters of the model producing the score, as well as the threshold γ_{ND}.

Clearly, ND shares many similarities with OSR. However, there are differences. To begin with, ND is generally considered an unsupervised task, while in OSR the model has access to the labels of the KC. This is consequent with their own respective goals. Unlike OSR, ND focuses only on detecting novel samples, hence classification of samples belonging to the KC is not required. Although ND tasks may also be formulated in multi-class settings, their ultimate goal is the detection of novel samples, i.e., a binary classification problem to discriminate between KC and UC.

4.1.1. Literature review

ND has been an active research area for several decades, which has led to the existence of a wide variety of approaches. One-class models have become prevalent in this area [135] due to the inherently unsupervised nature of ND. Isolation Forests (IF) [136] split the data based on random attributes until the data is separated, which causes anomalies to be split from the rest of the data earlier in the process, i.e., anomalies exhibit shorter path lengths. The authors in [137] have

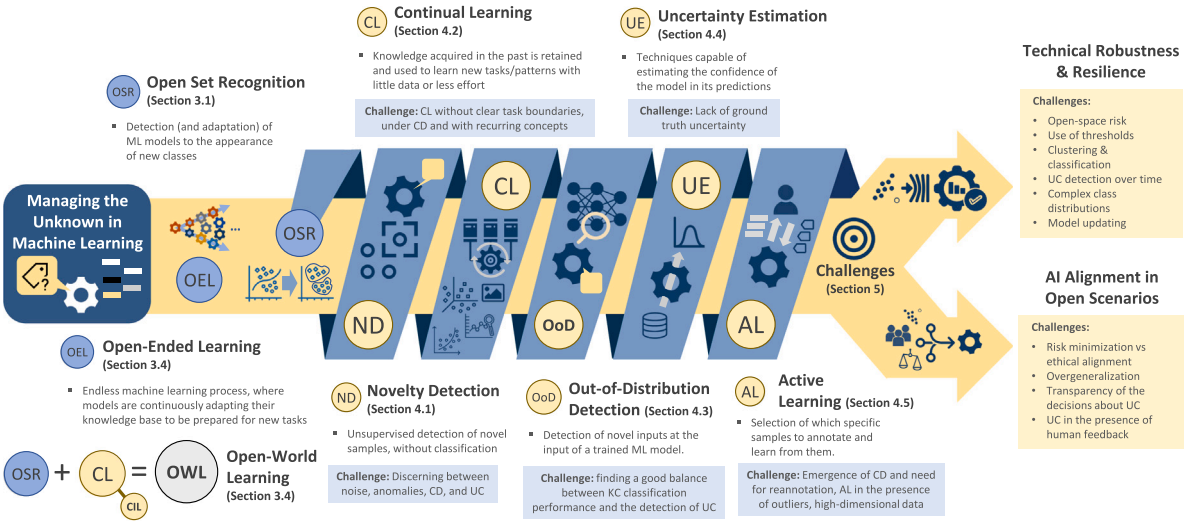


Fig. 3. Visual summary of the MUML paradigm tackled in this survey, together with the tangential areas to OSR analyzed in Sections 4 and 3.4, and the general challenges in Section 5.

enhanced IF by adding a probability density function that guides the splits of the dataset. They have argued that splits through gaps (which separate clusters and outliers) are more effective at isolating outlier samples earlier. Density-based methods such as Local Outlier Factor (LOF) are another popular branch of traditional methods for AD and ND. These methods compare the local density of a sample with that of its neighbors [138] to find anomalous samples. Despite effective, calculating the local density of all the neighbors in the designated distance is often time-consuming. To lessen this computational cost, the authors in [139] have proposed a growing gas NN and applying a LOF algorithm to the observed data. Other traditional methods such as three-way clustering [140] and margin distribution [141] have also been explored for ND.

DL has spurred the appearance of several ND methods [142] due to their ability to learn expressive representations from complex data. Autoencoders (AE) NN architectures are forced to learn important regularities of the data to minimize reconstruction error; since anomalies are difficult to reconstruct from the generated representations, they are expected to have greater reconstruction errors. This is a straightforward approach for detecting anomalies through the adoption of the reconstruction error as the novelty score γ_{ND}. Consequently, AE have been thoroughly investigated for AD and ND [143–145]. This idea, that normal samples can be generated better than abnormal samples, has also been harnessed by GAN architectures [146] and diffusion models [147]. Distance-based approaches in DL [148] project the data into a low dimensional space before calculating any distance, which avoids the problems that distance metrics face when working with high dimensional data. Similarly, deep clustering methods better adapt to the structure of the data, which improves performance in detecting abnormal samples for large and complex data [149].

4.1.2. Applications and challenges

The ability to detect abnormal patterns or behaviors by only learning from normal experiences has proven to be very useful for many real-world applications such as credit card fraud detection [150], malware detection [151], and even the identification of anomalies in the activities of daily living of the elderly in the domestic context [152]. ND has also seen application in computer vision tasks such as image AD for industrial inspection [153], or ND for surveillance videos [154]. With the growing number of scientific publications, some NLP applications that measure the novelty of scientific publications based on their title [155] or based on triplets [156] have emerged. The emergence of IoT systems has also brought forward the need for systems capable of handling large amounts of time-series data while detecting unusual behavior in real-time [157].

One of the most interesting areas of application for ND is data stream mining, where the distribution of the data is prone to change over time. This scenario poses additional challenges to ND, namely Concept Drift (CD), the appearance of noise and outliers, and feature shift [245]. CD refers to sudden, gradual or recurrent changes in the data distribution, which causes the feature space occupied by the KC to shift. This phenomenon is relevant to ND because it can be challenging to differentiate between a sample affected by CD and a sample from an UC [158]. Approaches employing micro-clusters, which are a tuple with information that represents clusters, have been prevalent in the field [159–161]. The way they have tackled this problem is by detecting anomalous samples and storing them in a buffer. When enough samples arrive at the buffer, clustering and density based techniques are applied to identify new UCs. Samples that are not considered to belong to an UC, are considered as being affected by CD. Other solutions have proposed the use of k-NN ensembles [162], prototypes [163], or the application noise reduction techniques [164]. Nevertheless, the core concept of discerning between noise, anomalies, CD, and UC remains an open challenge. We note that in some of the works in this field, ND is referred to as *concept evolution*, which is the same task, but it also involves adapting the model to the UC that have appeared.

4.2. Continual learning

Also referred to as lifelong (machine) learning, CL focuses on a knowledge-driven paradigm, where the knowledge acquired in the past is retained and used to learn new tasks/patterns with little data or less effort [8,246]. Under this paradigm, we usually find several learning scenarios such as domain incremental learning, Task-Incremental Learning, CIL, online learning, or Stream Learning (SL); in essence, all of them are associated with the gradual provision of data throughout their lifetime. Despite in many cases the distinction between these terms is often not strictly defined, nuances aside, they face the challenge of learning from data that are continuously generated over time (data stream). As mentioned before, the OWL paradigm usually considers a CIL scenario.

In CIL the model learns to recognize new classes over time while retaining the ability to recognize previously learned classes. A sequence of B training tasks is defined as $\mathcal{T} = \{T_1, T_2, \ldots, T_B\}$, where $T_b = \{(\mathbf{x}_i^b, y_i^b)\}_{i=1}^{N_b}$ is the $b-ıh$ training step with N_b training samples. $\mathbf{x}_i^b$ is the instance ith that belongs to class $y_i^b \in \mathcal{Y}_b$. The label space of each

task b does not contain any overlapping classes ($\mathcal{Y}_b \cap \mathcal{Y}'_b = \emptyset$) and only the data from T_b is accessible during task b. The objective of the model is to acquire knowledge from the current task T_b while preserving the knowledge from previous tasks. Typical CIL scenarios assume no class overlap between tasks, whenever classes overlap $\mathcal{Y}_b \cap \mathcal{Y}'_b \neq \emptyset$ it is referred to as *blurry* CIL [247]. Some CL and OWL works in the field do not make this distinction and assume that overlapping classes between tasks is a possibility because it is a more common scenario in real-world data streams. With respect to OSR, there are a few works that explicitly refer to the problem of OSR in CL [248]. This is because, as explained in Section 3.4, the combination of CIL and OSR is referred to as OWL [249,250].

4.2.1. Literature review

While the detection of UC is a complementary task, updating the model with new knowledge while retaining previous knowledge remains the main objective of most studies dealing with CL. Therefore, the community working in this area aims to tackle the well-known catastrophic forgetting issue, a phenomenon by which previously learned information is degraded or lost when new information is added incrementally to an already learned model [251]. In order to address this challenge, many different CL methods have been developed over the years.

Regularization-based methods aim to preserve important knowledge by adding regularization terms to balance old and new tasks. This was achieved in [99] via elastic weight consolidation, where changes to the most important parameters of previous tasks of a NN were penalized. Instead, the work in [100] estimated the importance of the parameters based on their contributions to the loss function during training. Since then, more sophisticated methods have been developed such as adding an evaluation function that prioritizes past tasks based on their difficulty [101] or implementing an additional auxiliary network [102], that imprints plasticity to a mainly stable model. Experience replay is another popular branch of CL methods. As the name suggests, these methods store old representative samples and use them to recover previous data distributions. Works in this direction have focused on efficient sample selection and exploitation [103,104]. Newer strategies such as AdaER [105] have enhanced the performance of the memory buffer through an entropy-balanced reservoir sampling strategy to maximize the information entropy. Instead of storing actual samples, generative replay methods opt to synthesize their own samples from a previous task to improve the training of the classifier and they can be of several types such as GANs [106], auto-encoders [107], and diffusion-based [108]. Despite these and other efforts reported in the CL literature, and the existence of conferences that are thematically centered in this area (see e.g. Conference on Lifelong Learning Agents, CoLLAs[1]), the plasticity-stability trade-off is still far from being properly understood and autonomously balanced in general CL settings.

Optimization-based methods explicitly design the learning process to accommodate new tasks [109]. In [110] each task has been learned by updating the network weights only in the direction orthogonal to the subspace spanned by all previous task inputs, which ensured no interference with tasks that have already been learned. The so-called Trust Region Gradient Projection method proposed in [111] has introduced a 'trust region' to select the most related old tasks to the new task by using the norm of gradient projection onto the subspace of previous samples. Representation based methods create informative representations of data during the learning process. These approaches are more suited whenever there are no labels available for any task [112,113], which is appealing for OWL methods since they usually need to learn from UC, which lack annotation.

Instead of sharing a set of parameters, architecture-based methods construct specific parameters for each task. Usually these methods isolate several parameter subsets in a growing network architecture, as done in the work in [114], where each task has a corresponding mask that adapts a pretrained vision transformer to the new task. Another related instance is [115], which differentiates between helpful and harmful information for old tasks. While expandable network architectures pose obvious advantages, they may become overly complex over time. This is why recent works seek minimal and efficient expansion policies for the CL model [116]. Alternatively, the authors in [117] have proposed an ensemble of subnetworks that learn all incremental tasks in parallel, arguing that it helps reduce the discrepancy between task distributions. Updating the model with new knowledge is a vital part of OWL, and many ideas can be adapted from CL methods.

4.2.2. Applications and challenges

In real-world applications, providing task identities is not a realistic assumption. Therefore, CIL scenarios have become increasingly popular as the 'default' setting for CL in several areas [251]. One of the most widely studied applications is Incremental Object Detection or Continual Object Detection (COD). In this situation, the model needs to deal with the appearance of new objects that were previously unknown and had appeared as background. This causes issues to the model's understanding of what an object is [118–120]. In [121] the authors have evaluated techniques commonly used in COD such as knowledge distillation [252,253] and exemplar replay, and have proposed a novel method to solve their issues. The authors in [122] have applied COD to streaming video, through two feed-forward NNs that act as slow and fast learners. Additionally, in order to emulate a more realistic setting, they also have assumed an online setting, where training data is only allowed to undergo one training pass, and models have been trained on the entire dataset for only one epoch. COD applications are closely related to OWL applications in object detection [27,123], where the model needs to classify all the detected objects in a scene as either KC or UC, while also incrementally learning the UC whenever their labels become available. Similarly to COD, works dealing with semantic segmentation tasks formulated under CL settings also consider that samples from previous and new classes can appear at the same time [124–126].

RL treats the learning process as reaching the best policy, rather than an endless adaptation. Continual Reinforcement Learning (CRL) addresses this, and proposes a setting in which the agent never stops learning [24]. The work in [127] has evaluated current CRL methods and proposed a ReLU based mitigation strategy that facilitates CL in a changing environment. Common NLP tasks such as text classification [128] and dialog systems [129] have received much attention. More recently, CL has become an important part of LLMs due to their necessity of reflecting the evolving human knowledge [130]. Some applications include continual pretraining [131] to improve end-task performance, continual instruction tuning that teaches LLMs to follow instructions and learn downstream tasks [132,133], and continual alignment [134], which seeks to adapt the model to newer societal values, human preferences and ethical guidelines. Some CL methods rely on task identities and boundaries, but data from real-world applications often come without clear task boundaries. Due to their nature, task-agnostic models are more suitable for OWL and CIL scenarios. Managing CD also presents a challenge for CL methods because it can cause old tasks to re-emerge further down the line. If not correctly detected, the model may learn a 'duplicate' of the old task which wastes its resources. This is a similar challenge to recurring CD in SL [254].

4.3. Out-of-distribution detection

This area is very closely related to OSR because it also challenges the *closed set* assumption. It states that the test distribution $P_{test}(\mathbf{x}, y)$ is different from the training distribution $P_{train}(\mathbf{x}, y)$. Therefore, the OoD

[1] Conference on Lifelong Learning Agents (CoLLAs), https://lifelong-ml.cc/, accessed on May 13th, 2024.

detection problem can be defined as a instance of the traditional supervised learning problem where $P_{test}(\mathbf{x}, y) \neq P_{train}(\mathbf{x}, y)$, and $P_test(\mathbf{x}, y)$ is not known during the training phase [9]. The difference between both distributions can be caused by many factors such as a change in the data distribution overtime or a training set that does not represent the actual data correctly. OoD recognizes two types of distribution shifts: *covariate shift* $P(\mathbf{x}) \neq P'(\mathbf{x})$ and *semantic shift* $P(y) \neq P'(y)$. The first one refers to a shift in the distribution of the input features, but the relationship between the input features and the labels remains the same [244]. The second one is caused by a shift in the meaning of the data itself, e.g, OoD samples coming from classes that have not been seen previously by the model. In the context of a semantic shift, clear parallels between OoD detection and OSR can be drawn. Both train the model to detect UC during inference, while also correctly classifying the KC.

Even if their goals are aligned, OoD detection features several small differences when compared to OSR. Firstly, as stated before, OSR exclusively deals with a semantic shift $P(y) \neq P'(y)$, while OoD detection can also cover covariate shift $P(\mathbf{x}) \neq P'(\mathbf{x})$. Secondly, some OoD detection methods involve exposing the model to a number of OoD samples during the training process to help models learn to discriminate between In-Distribution (ID) and OoD data. The use of these training OoD data clashes with the assumptions made in OSR. Another difference arises in the benchmark protocol followed by studies related to these two research areas: while OSR models usually split a single dataset into partitions with samples belonging to known (ID) and unknown (OoD) classes, OoD models consider an entire dataset as the ID data, whereas several different datasets are utilized as OoD data. All in all, OoD detection techniques can be used to perform OSR in multi-class settings.

4.3.1. Literature review

The detection of OoD samples has been approached from different means in the literature [244], even to the point of comparing it to OSR [255]. A straightforward approach is to directly deconstruct the problem into several one-versus-all binary classification tasks [181]. Some methods make use of the model output such as the maximum softmax probability in order to determine whether a sample is ID or OoD. The authors in [256] proposed ODIN, a post-hoc method that detects OoD images by using temperature scaling and adding small perturbations to inputs, which enlarges the difference between the softmax score of ID and OoD samples. However, NNs are prone to giving overconfident predictions on samples far from the training data. As a way to overcome this issue, the authors in [166] argued that energy-based models are more suitable for OoD detection. The reason is that energy scores align with the probability density of the samples, and thus are less susceptible to overconfident predictions. These methods consider samples with higher energy as OoD and samples with lower energy as ID [167,168]. Some efforts to enhance energy-based models have been made by generating high-sparsity representations of ID data [169] or by adding a metric learning based distance function to the energy score function [167]. Energy-based methods have also been applied to graph-structured data for OoD detection [170]. Other classification based methods expose the model to outliers (e.g., OoD samples) to improve their performance. Although these methods achieved comparable results [171], having access to outliers during training may not always be feasible. Instead, other methods synthesize their own outliers through the use of GANs [172] to create samples close to the decision boundary. Diffusion models have also been applied to OoD detection [173,174], arguing that they are able to synthesize virtual OoD features that are close to the classification boundary between ID and OoD objects.

Density based methods estimate the density of the training data and consider the data located in low density regions as OoD [175]. Flow-based models are a popular choice in the field due to their ability to represent complex data distributions and providing an exact likelihood function [176–179]. Instead of performing clustering on the input data, the authors in [180] have performed the clustering over the space of internal activations of a NN classifier to detect the OoD samples. Distance based methods work on the assumption that the OoD samples are far away from the training data in the input space, which is the same assumption made for the UC samples in OSR [182,183]. While Mahalanobis distance approaches [184,185] have been prevalent in the field, recent works seek to maximize inter-class dispersion and intra-class compactness [186]. In [187] the authors have created expressive class-specific descriptor vectors for all ID classes and then applied a cosine similarity calculation between descriptor vectors consistently to identify OoD samples.

Finally, reconstruction-based methods follow the intuition that OoD samples would have a larger reconstruction error than ID samples due to the lack of knowledge of the model [188,189]. However, effectively reconstructing ID inputs and not OoD inputs requires careful tuning [190], and so the authors have proposed using Diffusion denoising probabilistic models, which allows better controlling the amount of injected noise. Many of the methods reviewed apply the same principles as OSR to detect OoD samples, so they can be used to detect UC in OSR.

4.3.2. Applications and challenges

Being one of the most similar areas to OSR, OoD has also been applied to computer vision tasks such as medical classification of images [190–192], where OoD is useful for evaluating diseases or object detection [193], which is essential for autonomous driving tasks. NLP is another popular area of application for OoD methods [194]. Recently they also have been applied to LLMs for fake text detection [195]. Other applications that find OoD methods useful include intrusion detection systems [196,197] to prevent unknown or zero-day attacks and industrial AD [198] to identify defects, incorrect parts, and damages in industrial components.

Although it has already been shown to excel in several practical applications, OoD detection also faces challenges still to be answered. The most closely related one to this article is how to reach a good balance between KC classification performance and the detection of UC, dealing with overlapping class distributions, and managing changes in the data distribution over time [244]. Other prominent challenge in both OSR and OoD is the lack of theoretical characterization of the OoD or OSR problems [9]. Although sometimes unpredictable, taking into consideration the types of distributional shifts the model will encounter at inference (which can be sometimes known a priori) is critical for the design of effective and practical OoD and OSR methods.

4.4. Uncertainty estimation

This area aims to provide ML models with the ability to estimate and quantify the uncertainty associated their individual predictions [10]. There are two main types of uncertainty: aleatory uncertainty (data uncertainty) and epistemic uncertainty (model uncertainty). Aleatory uncertainty is attributed to the inherent random processes in nature which are reflected in the data. Epistemic uncertainty is due to model problems that can be caused by suboptimal architectures, errors during training, or lack of knowledge about UC. In theory, epistemic uncertainty could be reduced by improving the learning process, architecture, or the training data of a network. However, achieving zero model uncertainty is not feasible; thus, quantifying model uncertainty is very important.

A Bayesian framework allow model uncertainty to be formalized as a probability distribution over the model parameters θ given the observed data $\mathcal{D}$. The data uncertainty is given by as a probability distribution over the model outputs y assuming a fixed set of model parameters θ. Epistemic uncertainty can be then represented as a distribution of the prediction y given the input data $\mathbf{x}$ and the dataset $\mathcal{D}$, defined as:

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \theta) p(\theta|\mathcal{D})\, d\theta. \tag{2}$$

Additionally, based on the input $\mathbf{x}$, the uncertainty can be classified into three main categories: (i) *in-domain* uncertainty, which is caused by an input drawn from the same distribution as the training data; (ii) *domain-shift* uncertainty, which is caused by an input from a shifted training distribution; and (iii) *out-of-domain* uncertainty, which is caused by inputs from an unknown data distribution. Data uncertainty is only able to capture in-domain uncertainty because it is caused by the noise and/or the quality of the samples the model has been learned from. However, model uncertainty is especially interesting for tasks such as OSR, OoD and ND, since it is also able to capture domain-shift and out-of-domain uncertainty. Although the main objective of UE is not the detection and rejection of UC inputs, uncertainty quantification methods can and have already been used to achieve this, as shown in the next subsection.

4.4.1. Literature review

Since no clear way of representing uncertainty exists, many methods for estimating uncertainty have been developed throughout the years. Single deterministic methods are explicitly modeled to quantify uncertainties with a single model and a single pass of the input data [257]. Bayesian Neural Networks (BNNs) estimate a posterior distribution over the parameters of a network, and thus have been extensively used to compute the uncertainty in networks through various methods such as dropout [199,200], deep ensembles [201], Markov Chain Monte Carlo methods [202]. Although BNNs have shown great performance, they are computationally expensive to train and scale. Some works have attempted to circumvent this issue by constructing an efficient Dirichlet distribution [203] or by using variational AE [204]. Ensembles are another popular branch of UE methods. They seek to improve the performance of the model by combining multiple models. Although ensembles were originally conceived to improve the accuracy of models, the variety among the predictions of an ensemble can be used as a way to measure the uncertainty of a prediction. In order to increase variety among members, ensemble methods applied boosting [205] and bagging strategies [206]. Ensemble methods have also shown their robustness against OoD data [207]. Test Time Augmentation (TTA) has also been used to measure uncertainty. Similarly to ensemble methods, by creating several variations of a test sample, results on multiple predictions of the same sample, which are then used to measure the uncertainty [208].

UE has also shown its ability to detect unknown samples in fields such as OSR [71,72] and OoD [258]. Although not explicitly mentioned in their titles, UE methods are generally also tested for OoD data [204, 207].

4.4.2. Applications and challenges

In the real world, there are many factors that add uncertainty to the output of a model. Removing every single one of them is a challenging task for any practical application. Furthermore, the distribution of real-world data may not necessarily be completely represented by the available training data, and is prone to shift over time. UE has found its major use in medical data analysis [209–211], because quantifying and understanding the uncertainty associated with medical predictions, diagnosis and treatments is vital to make informed decisions that ultimately affect humans. As a matter of fact, UE methods for medical analysis have usually been paired with TTA strategies due to the lack of large amounts of high-quality, thoroughly, reliably annotated medical data [204,207]. Since in modern times NLP heavily relies on black-box methods, estimating the uncertainty of their predictions has become a crucial research direction [212]. With the appearance of LLMs, some works have focused on measuring the uncertainty of the outputs of these models to know whether their predictions can be trusted [213, 214]. Another example of applications of UE has been robotics [215], where such techniques are used to know when a robotic agent should ask for help. Chemistry has also leveraged uncertainty quantification to diagnose the chemical component that introduces uncertainty in molecular property predictions [216]. A major challenge for many UE methods is the lack of ground truth uncertainty, which means a shift in the data distribution may render the uncertainty values given by a model invalid.

4.5. Active learning

AL is a strategy by which the algorithm selects the most informative data points to label from a pool of unlabeled samples, aiming to maximize learning efficiency with minimal labeling effort [11]. In this paradigm, the model or an external system queries an oracle (usually a human annotator) for the labels of specific samples. The newly annotated points are then used to update the model. This process is repeated until a termination condition is met. AL techniques usually make use of a small set of labeled data $\mathcal{D}_\ell$, where the model is initially trained on the KC, and a large set of unlabeled data $\mathcal{D}_\circ$. After the initial training, a query strategy is applied to select the most uncertain samples $\mathbf{x} \in \mathcal{D}_\circ$ until a stopping criterion is met. Each time a sample $\mathbf{x}_i$ is selected by the query strategy, a label y_i is requested from an oracle and added to $\mathcal{D}_\ell \cup (\mathbf{x}_i, y_i)$. After this process is completed, the model is updated according to $\mathcal{D}_\ell$ augmented with new samples.

It is worth noting that $\mathcal{D}_\circ$ may or may not contain samples from UC (classes not present in the labeled set $\mathcal{D}_\ell$). Although both OSR and AL deal with the limitations of the knowledge of a model, when faced with a sample from an UC their approach is different. OSR strategies directly detect a sample from UC, whereas AL strategies would initially incorrectly classify it and then query the oracle for its label. Depending on the query strategy under consideration, samples from UC may not be selected for querying, which would leave the model unable to correctly classify them in the next iteration. Alternatively, if selected, learning about new classes this way naturally reduces the *over-occupied space*. We note that the process of characterizing and learning from UC is also part of the OWL paradigm. The key difference between these two is that in AL the model queries an external oracle for the labels, while in OWL this process should ideally be autonomously performed by the model itself [4].

4.5.1. Literature review

AL algorithms differ mainly in their query strategy. Each query strategy has an utility function that generates an utility score for each instance, which is used to select the samples to query.

Information-based strategies search for the samples with the highest uncertainty that are expected to be close to the decision boundary. This has been approached in many ways, for multi-class settings the entropy method can be used [217], where all classes are taken into account and samples with the largest entropy are queried. In [218] a set of models was trained on different subsets of samples, and then the unlabeled samples where their predictions disagree were queried. Recent works such as [219] have enhanced this strategy by exploiting the meta characteristics of data streams, which are used to dynamically tune the uncertainty threshold required to select a sample for querying.

Other approaches rely on computing the expected change in the model or expected error [222,223], which query the points that would reduce the expected future error of the model. Existing AL strategies in an *open set* scenario, where unknown samples exist among the unlabeled known samples, have focused on distinguishing and selecting the samples from the KC while discarding the ones from UC [220]. However, the work in [221] has leveraged both known and unknown distributions to select better informative samples in each query.

Representation-based strategies make use of the structure of the unlabeled data to find the points that better represent the input space. The utility function measures the representativeness of the samples and queries the most representative ones. In [224] the authors performed clustering on the entire input space and then selected the nearest samples to the cluster centers for querying. A similar strategy was

followed in [15], where it allowed the model to detect and learn new concepts.

The utility functions in AL are designed to evaluate the uncertainty or representativeness of each sample. Since some OSR methods have made use of the same concepts for detecting samples from the UC such as entropy [259,260], it is reasonable to suggest that each field could adapt solutions from the other. For example, following the assumption that the UC are far from the KC strategies, the representation-based category would be better suited to find samples from UC. This is because they encourage the exploration of the *open space*. On the other hand, information-based strategies would be more suitable when the UC are close to the decision boundaries of the KC.

4.5.2. Applications and challenges

AL has been extensively applied in many modern applications, including LLM [225]. Here AL has been used to tackle the problem of finding the most informative demonstrations for in-context learning. AL is also useful for massively reducing labeling costs in tasks that are formulated over image/video data; this includes semantic segmentation [227] or object detection [226], where AL can reduce annotation costs via entropy querying. Beyond general application-agnostic ML tasks, AL methods have shown promise in guiding and significantly speeding the exploration of biomass fuel solutions [228] and in molecular modeling and drug discovery [229]. Regardless of the area of application, some of the challenges that the field is currently facing are the need for a large number of queries before the model achieves good results, or learning an effective stopping condition for the AL algorithms. Other challenges are shared with the previous fields: (i) the emergence of CD in data streams, which causes the data distribution to change and old labeled samples to become irrelevant or harmful to the model; (ii) the presence of outliers in the data, which causes the model to select and waste queries on them; and (iii) high dimensional data such as images and videos, which makes the process of querying more difficult and costly.

5. OSR challenges and beyond

In what follows, we offer our view on the challenges that this area faces for the future, and outline research directions that can be pursued to address them efficiently:

Open space risk. Despite the activity noted in the area (exposed by the high scientific productivity noted in the last couple of years), our literature review has revealed that reducing the *Open Space Risk* remains an open problem. Decision boundaries delineated by off-the-shelf ML classifiers tend to over-occupy the feature space for the KC. Managing the space out of the feature space regions corresponding to the KC is still a hard task that increases its difficulty with the openness of the problem [14]. Similar to OoD, OSR methods usually work on assumptions such as the separation of KC and UC in the representation space, but rarely provide a strong theoretical foundation behind them [9]. Works that analyze how learning the representation of KC alters the representation of UC [44,261] may prove helpful for managing *Open Space Risk*. An important albeit often overlooked benefit of reducing the *over-occupied space* is that it avoids false positives. This is critical for some real-world applications such as medical image analysis [258], where classifying an unknown pathology as another one entails significant consequences. The same holds for other high-risk areas such as security [81]. Although the *over-occupied space* is a common issue for all perspectives that deal with UC, it has never been technically approached from the perspective of ND and OoD detection. We envision that there are profitable synergies between these two close research areas and OSR, that can endow ML models with an improved robustness against UC.

Use of thresholds. Many OSR, ND, OoD detection, UE, and AL models rely on a threshold for distinguishing between KC or UC. Computing an optimal threshold is not straightforward and can be time consuming [38]. Furthermore, these thresholds can vary greatly depending on the openness of the problem and the a priori information about the characteristics of the UC that the model will encounter when running inference in the open world. Even a threshold for the same set of classes can yield a high *Open Space Risk* if the data distribution changes and the threshold is not updated accordingly. While an optimal threshold can be dynamically set for each task or set of data, it can still become a problem in evolving data streams, where new UC may arrive regularly and the model continuously produces predictions for the instances arriving in the stream. Calibrating a suitable threshold for each situation is very valuable for applications in streaming environments [122,219]. In addition, thresholds can be easily interpreted by non-expert users and stakeholders, which helps them understand the decisions and the performance of the application. Self-adaptive parameter tuning methods [262], including dynamic evolutionary algorithms or reinforcement learning agents [32], can be interesting research paths to follow in order to achieve OSR techniques capable of operating in highly-varying open settings.

Combining clustering and classification. As evidenced by our literature analysis, most OSR approaches relying on clustering and classification combine both techniques in a sequential way [16,233]. This practice has two inherent limitations: on one hand, the clustering algorithm does not benefit from the information provided by the classifier, even though feedback such as the uncertainty about its prediction can be a valuable input for determining whether the test instance is unknown. Combining clustering and classification is beneficial in applications that deal with class imbalance, such as fault diagnosis [98], e.g., by creating homogeneous clusters to train the classifier on. Very scarce works ensure that both processes benefit from each other in an *open set* scenario. New strategies to hybridize clustering and classification should be targeted by the community, such as clustering over the space of internal activations when dealing with NN classifiers [180] or employing multi-objective optimization frameworks [241].

Detecting unknown samples and identifying new classes over time. Performing classification of each new sample in isolation is an standard evaluation protocol in OSR [6]. Therefore, instances that arrive before are not taken into account in the decision. However, it is often the case that UC arise gradually over time, imprinting a correlation between successive instances that can be modeled for the sake of a more precise detection of its membership to an KC or UC. Considering this time correlation is of utmost importance when dealing with scenarios with intermittently appearing UC: in this case, allowing the OSR technique to memorize long-term relationships between instances detected as UC can be crucial to expedite the UC identification when one of such UC reappears, and/or to support a preemptive reservation of the *over-occupied space* as per the emergence/disappearance dynamics of UC. These scenarios are common in IoT environments [165] and data mining [263], where concepts can reappear regularly and the resources of the device on which the model is deployed are limited. Data distillation and neural embeddings can provide a low-dimensional representation of samples detected to belong to UC, so that the relationship between unknown concepts over time can be learned and eventually exploited for identification and/or consolidation in the knowledge base of the model. Additionally, the existence of CD in data streams poses additional difficulties not only for OSR, but also for every area reviewed in this survey. Current methods rely on the assumption that samples from UC possess a higher degree of abnormality than samples affected by CD [161], so that they can be distinguished from each other (usually by means of a threshold). Relying on distance and density measures is often not enough for addressing more complex and shifting distributions. Employing additional meta information [219] may prove beneficial for discriminating between CD and UC.

Disentangling complex class distributions. A major fraction of the contributions on OSR apply clustering to data and identify new classes from the obtained clusters [249,250]. This naive approach does not perform well when KC and UC overlap, or when having underpopulated KC that do not provide enough statistical support to decide whether a new instance belongs to them. This is often the case of OSR in ML tasks formulated over real-world image data subject to a strong content variability (i.e., low domain specificity). Although some approaches make use of prototypes [43] or metric learning [167,259] to create a more meaningful feature space, we note a lack of reliable strategies for identifying novel classes from unknown samples that perform well over these challenging KC distributions. In this regard, we foresee that any means to augment the knowledge of the model about newly emerging classes, including domain-specific meta-information, can help the model identify such classes more reliably.

Updating the model. After UC have been discovered, the model needs to be updated with the new knowledge. Completely retraining the model [249,250] is not a feasible adaptation strategy in environments with tight computation constraints. Similarly, the human annotation of new classes is not affordable in most practical problems. Incremental learning strategies such as the update and maintenance of micro-clusters [160] have recently shown to be promising in SL, but have not been extensively researched in OSR and OWL. Likewise, such model updating techniques can be combined with imbalance learning or – when interaction with the data source is possible – AL strategies to enrich the learning process with diverse instances of the newly discovered UC. Finally, well-known issues in CL settings, including catastrophic forgetting, also hold in OWL applications [25]. As such, updating the model with new classes while discriminating known concepts should be also given attention in the future, especially under periodic occurrence patterns and extreme verification latency. A similar problem has already appeared in SL under the name of Recurring CD [254], which can serve to inspire new developments in OWL.

5.1. AI alignment in open scenarios

Before concluding the challenges, we draw the attention of the community to the emerging issue of *AI alignment* in MUML and, in particular, OSR. We strongly believe that it is already necessary to include alignment as a new point of view in all data modeling problems, especially in those where autonomy, uncertainty, and little or no human intervention are baseline conditions to be met.

AI alignment focuses on making AI behave in line with human intentions and values [12], more than its capabilities or performance. This field is gaining relevance due to the risk of suffering undesirable behaviors in AI systems. Some of these risks are sycophancy [264,265], unreliable answers [266], or deception [267], among others. The alignment problem is receiving increasing awareness in recent times, not only from the scientific point of view [12,268], but also from institutions and agencies like the United Nations Educational, Scientific and Cultural Organization (UNESCO), which has recently elaborated a comprehensive international framework to shape the ethical development and use of AI technologies[2].

The fact that an AI-based system operates in an open scenario can exacerbate misalignment problems [269–271]. The alignment problem in open world scenarios and OSR emerges when AI systems must make decisions or perform actions based on inputs that were not present during training. This problem unchains several new challenges:

- When an AI encounters an unknown input, it must decide how to respond in a way that minimizes risk and aligns with ethical guidelines. This is challenging because the AI lacks specific training on these samples, and therefore must rely on general principles or fallback strategies that reflect human values.
- AI systems might overgeneralize their training data and incorrectly classify unknown inputs as belonging to one of the known categories, leading to errors. The alignment problem here involves ensuring that AI remains cautious and seeks further information or guidance when faced with uncertainty.
- Maintaining user trust requires transparent communication about the confidence levels of AI and the reasoning behind its classification of inputs as unknown. This transparency helps align the operations of AI with user expectations, and builds trust in its decision making process.
- Ideally, an AI system should not only recognize when it encounters an unknown, but also learn from these samples to improve its performance. Aligning this learning process with human values involves careful consideration of privacy, consent, and relevance of new data to the intended purposes of AI.

Implementing mechanisms for human-in-the-loop feedback allows the AI to learn from its handling of unknown samples in a way that reflects human values and judgments. But at the same time, human feedback is not without its problems. The feedback mechanism designed to improve or guide the behavior of AI inadvertently leads to behaviors that diverge from the intended goals or values. This misalignment often stems from the complexity of interactions between the AI system, its environment, and the feedback it receives. In AI systems, feedback is crucial for learning and adaptation. It can come in many forms, such as reinforcement signals, corrections from users, or data reflecting the consequences of the AI actions. Ideally, this feedback helps the AI to align its actions more closely with human values, objectives, and expectations. In open scenarios, the misalignment becomes particularly salient and challenging due to the nature of such environments, where the goals are not strictly defined, and the learning process is expected to continue indefinitely, exploring new and unforeseen areas of the problem space. For example, in OEL environments, systems are designed to encourage AI to generate novel behaviors, solutions, or ideas beyond initial programming or human expectations. However, this exploratory freedom, combined with feedback mechanisms meant to guide learning, can lead to unexpected forms of misalignment between the AI behaviors and the broader objectives or values intended by its creators or users (e.g., the feedback-induced misalignment [12]). We definitely envision a future where the need for adaptation in OWL is justified not only for performance reasons, but also for aligning the behavior of the AI-based system to human values represented by corrective feedback signals.

6. Concluding remarks

Open set environments call for more autonomous ML models capable of dealing with unknown situations. Consequently, the OSR and OWL fields have garnered significant attention in recent times. However, their resemblance to established fields like ND or OoD detection can present challenges for researchers in understanding where such fields currently stand, as well as in identifying substantive research niches towards safe and autonomous ML models.

This review has addressed this issue by examining the current state of OSR and the different strategies followed in the literature to tackle the *Open Space Risk* problem. In doing so, and in response to the aforementioned entanglement of different ML research areas around the open set learning paradigm, we have highlighted and clarified how such areas link to OSR, including ND, CL, OoD detection, UE and AL. Based on our definitions and the literature analysis, we have discussed several challenges, both in each of such tangential areas and the OSR field as a whole. Among them, we highlight the following open

[2] Global AI Ethics and Governance Observatory, *Recommendation on the Ethics of AI*, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics, accessed on May 13th, 2024.

problems, which should focalize efforts in the forthcoming years: (i) scaling the current approaches up to consider complex known class distributions, (ii) improving the incremental consolidation of newly discovered classes by the model in scarce data regimes; and (iii) the exploitation of the temporal correlation between test samples when detecting and characterizing new concepts over time. On a closing note to our prospects, we have emphasized the growing concern with the alignment of AI-based systems with human values and objectives, which stems from the widespread adoption and ubiquity of the access and use of these systems in open settings. Research in open set learning is imperative for ensuring the technical robustness of AI-based models in the open world, but can also spawn new problems around the provision of technical performance goals under the guidance of human-induced alignment signals.

We hope that this survey establishes itself as a reference for new researchers willing to have a clear overview of the field and its manifold related areas, and as a motivating landmark to join forces and delve deeper into the OSR/OWL research areas for a safer and better aligned AI.

CRediT authorship contribution statement

Marcos Barcina-Blanco: Writing – review & editing, Writing – original draft, Visualization, Investigation, Conceptualization. **Jesus L. Lobo:** Writing – review & editing, Visualization, Supervision, Resources, Methodology, Conceptualization. **Pablo Garcia-Bringas:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Javier Del Ser:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Data availability

No data was used for the research described in the article.

Acknowledgments

All authors have read and approved the version of the manuscript to be published. This research has received funding from the Basque Government, Spain (BEREZ-IA project with grant number KK-2023/00012), and the research groups MATHMODE (IT1456-22) and D4K-Deusto for Knowledge (IT1528-22).

References

[1] I. Triguero, D. Molina, J. Poyatos, J. Del Ser, F. Herrera, General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance, Inf. Fusion 103 (2024) 102135.
[2] Z.-H. Zhou, Open-environment machine learning, Natl. Sci. Rev. 9 (8) (2022) nwac123.
[3] O. Sigaud, G. Baldassarre, C. Colas, S. Doncieux, R. Duro, N. Perrin-Gilbert, V.-G. Santucci, A definition of open-ended learning problems for goal-conditioned agents, 2023, arXiv preprint arXiv:2311.00344.
[4] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3614–3631.
[5] A. Mahdavi, M. Carvalho, A Survey on Open Set Recognition, in: 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering, AIKE, 2021, pp. 37–44, arXiv:2109.00893 [cs].
[6] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-set recognition: A good closed-set classifier is all you need? 2022, arXiv:2110.06207 [cs].
[7] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M.H. Rohban, M. Sabokrou, A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges, 2021, arXiv preprint arXiv:2110.14051.
[8] L. Wang, X. Zhang, H. Su, J. Zhu, A comprehensive survey of continual learning: Theory, method and application, 2023, arXiv preprint arXiv:2302.00487.
[9] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, P. Cui, Towards out-of-distribution generalization: A survey, 2021, arXiv preprint arXiv:2108.13624.
[10] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl 1) (2023) 1513–1589.
[11] A. Tharwat, W. Schenck, A survey on active learning: State-of-the-art, practical challenges and research directions, Mathematics 11 (4) (2023) 820.
[12] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, et al., AI alignment: A comprehensive survey, 2023, arXiv preprint arXiv:2310.19852.
[13] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M.L. de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation, Inf. Fusion 99 (2023) 101896.
[14] W.J. Scheirer, A. de Rezende Rocha, A. Sapkota, T.E. Boult, Toward open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2012) 1757–1772.
[15] L.F. Coletta, M. Ponti, E.R. Hruschka, A. Acharya, J. Ghosh, Combining clustering and active learning for the detection and learning of new image classes, Neurocomputing 358 (2019) 150–165.
[16] B. Wang, Y. Wang, J. Hou, Y. Li, Y. Guo, Open-Set source camera identification based on envelope of data clustering optimization (EDCO), Comput. Secur. 113 (2022) 102571.
[17] A. Bendale, T. Boult, Towards open world recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1893–1902.
[18] J. Parmar, S.S. Chouhan, V. Raychoudhury, S.S. Rathore, Open-world machine learning: Applications, challenges, and opportunities, 2022, arXiv:2105.13448 [cs].
[19] F. Zhu, S. Ma, Z. Cheng, X.-Y. Zhang, Z. Zhang, C.-L. Liu, Open-world machine learning: A review and new outlooks, 2024, arXiv preprint arXiv:2403.01759.
[20] W.J. Scheirer, L.P. Jain, T.E. Boult, Probability models for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2317–2324.
[21] P. Perera, V.I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, V.M. Patel, Generative-discriminative feature representations for open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11814–11823.
[22] N. Castanet, S. Lamprier, O. Sigaud, Stein variational goal generation for reinforcement learning in hard exploration problems, 2022, arXiv preprint arXiv:2206.
[23] A.A. Team, J. Bauer, K. Baumli, S. Baveja, F. Behbahani, A. Bhoopchand, N. Bradley-Schmieg, M. Chang, N. Clay, A. Collister, et al., Human-timescale adaptation in an open-ended task space, 2023, arXiv preprint arXiv:2301.07608.
[24] D. Abel, A. Barreto, B. Van Roy, D. Precup, H.P. van Hasselt, S. Singh, A definition of continual reinforcement learning, Adv. Neural Inf. Process. Syst. 36 (2024).
[25] M. Mundt, Y. Hong, I. Pliushch, V. Ramesh, A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning, Neural Netw. 160 (2023) 306–336.
[26] X. Zhu, R. Zhang, B. He, Z. Guo, Z. Zeng, Z. Qin, S. Zhang, P. Gao, Pointclip v2: Prompting clip and gpt for powerful 3D open-world learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2639–2650.
[27] K. Joseph, S. Khan, F.S. Khan, V.N. Balasubramanian, Towards open world object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5830–5840.
[28] A. Gupta, S. Narayan, K. Joseph, S. Khan, F.S. Khan, M. Shah, Ow-detr: Open-world detection transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9235–9244.
[29] F. Gao, W. Zhong, Z. Cao, X. Peng, Z. Li, OpenGCD: Assisting open world recognition with generalized category discovery, 2023, arXiv preprint arXiv:2308.06926.
[30] Y. Wang, W. Xiao, Z. Tan, X. Zhao, Caps-OWKG: A capsule network model for open-world knowledge graph, Int. J. Mach. Learn. Cybern. 12 (2021) 1627–1637.
[31] S. Mazumder, B. Liu, Open-world continual learning: A framework, in: Lifelong and Continual Learning Dialogue Systems, Springer, 2024, pp. 21–47.
[32] J. Balloch, Z. Lin, R. Wright, X. Peng, M. Hussain, A. Srinivas, J. Kim, M.O. Riedl, Neuro-symbolic world models for adapting to open world novelty, 2023, arXiv preprint arXiv:2301.06294.
[33] H. Ma, R. Xiong, Y. Wang, S. Kodagoda, L. Shi, Towards open-set semantic labeling in 3D point clouds: Analysis on the unknown class, Neurocomputing 275 (2018) 1282–1294.
[34] E.M. Rudd, L.P. Jain, W.J. Scheirer, T.E. Boult, The extreme value machine, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2017) 762–768.
[35] M. Boudiaf, E. Bennequin, M. Tami, A. Toubhans, P. Piantanida, C. Hudelot, I. Ben Ayed, Open-set likelihood maximization for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24007–24016.

[36] T. Chen, G. Feng, P.M. Djurić, Improving open-set recognition with Bayesian metric learning, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2024, pp. 6185–6189.

[37] P.R. Mendes Júnior, R.M. De Souza, R.d.O. Werneck, B.V. Stein, D.V. Pazinato, W.R. de Almeida, O.A. Penatti, R.d.S. Torres, A. Rocha, Nearest neighbors distance ratio open-set classifier, Mach. Learn. 106 (3) (2017) 359–386.

[38] X.-m. Hui, Z.-g. Liu, A new k-NN based open-set recognition method, in: 2022 17th International Conference on Control, Automation, Robotics and Vision, ICARCV, IEEE, 2022, pp. 481–486.

[39] D.O. Cardoso, J. Gama, F.M. França, Weightless neural networks for open set recognition, Mach. Learn. 106 (9–10) (2017) 1547–1567.

[40] H.-M. Yang, X.-Y. Zhang, F. Yin, Q. Yang, C.-L. Liu, Convolutional prototype network for open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2358–2370.

[41] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, Y. Tian, Learning open set network with discriminative reciprocal points, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 507–522.

[42] J. Lu, Y. Xu, H. Li, Z. Cheng, Y. Niu, Pmal: Open set recognition via robust prototype mining, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 1872–1880.

[43] J. Liu, J. Tian, W. Han, Z. Qin, Y. Fan, J. Shao, Learning multiple Gaussian prototypes for open-set recognition, Inform. Sci. 626 (2023) 738–753.

[44] Z. Xia, P. Wang, G. Dong, H. Liu, Spatial location constraint prototype loss for open set recognition, Comput. Vis. Image Underst. 229 (2023) 103651.

[45] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436.

[46] A. Bendale, T.E. Boult, Towards open set deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1563–1572.

[47] L. Shu, H. Xu, B. Liu, Doc: Deep open classification of text documents, 2017, arXiv preprint arXiv:1709.08716.

[48] L. Shu, H. Xu, B. Liu, Unseen class discovery in open-world classification, 2018, arXiv preprint arXiv:1801.05609.

[49] P. Oza, V.M. Patel, C2ae: Class conditioned auto-encoder for open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2307–2316.

[50] H. Cevikalp, B. Uzun, Y. Salk, H. Saribas, O. Köpüklü, From anomaly detection to open set recognition: Bridging the gap, Pattern Recognit. 138 (2023) 109385.

[51] J. Komorniczak, P. Ksieniewicz, Distance profile layer for binary classification and density estimation, Neurocomputing (2024) 127436.

[52] Y. Wang, J. Mu, P. Zhu, Q. Hu, Exploring diverse representations for open set recognition, 2024, arXiv preprint arXiv:2401.06521.

[53] L. Liu, R. Wang, Y. Wang, L. Jing, C. Wang, Frequency shuffling and enhancement for open set recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 3675–3683.

[54] J.K. Mandivarapu, B. Camp, R. Estrada, Deep active learning via open-set recognition, Front. Artif. Intell. 5 (2022) 737363.

[55] M. Vendramini, H. Oliveira, A. Machado, J.A. dos Santos, Opening deep neural networks with generative models, in: 2021 IEEE International Conference on Image Processing, ICIP, IEEE, 2021, pp. 1314–1318.

[56] J. Xu, C. Grohnfeldt, O. Kao, OpenIncrement: A unified framework for open set recognition and deep class-incremental learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3303–3311.

[57] B. Ma, Y. Cong, Y. Ren, IOSL: Incremental open set learning, IEEE Trans. Circuits Syst. Video Technol. 34 (4) (2024) 2235–2248.

[58] H. Zhang, A. Li, J. Guo, Y. Guo, Hybrid models for open set recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 102–117.

[59] Z. Ge, S. Demyanov, Z. Chen, R. Garnavi, Generative openmax for multi-class open set classification, 2017, arXiv preprint arXiv:1707.07418.

[60] I. Jo, J. Kim, H. Kang, Y.-D. Kim, S. Choi, Open set recognition by regularising classifier with fake data generated by generative adversarial networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 2686–2690, ISSN: 2379-190X.

[61] S. Kong, D. Ramanan, OpenGAN: Open-set recognition via open data generation, 2021, arXiv:2104.02939 [cs].

[62] L. Neal, M. Olson, X. Fern, W.-K. Wong, F. Li, Open set learning with counterfactual images, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 613–628.

[63] J. Goodman, S. Sarkani, T. Mazzuchi, A generative approach to open set recognition using distance-based probabilistic anomaly augmentation, IEEE Access 10 (2022) 42232–42242, Conference Name: IEEE Access.

[64] D. Pal, S. Bose, B. Banerjee, Y. Jeppu, Morgan: Meta-learning-based few-shot open-set recognition via generative adversarial network, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6295–6304.

[65] J. Sun, Q. Dong, Conditional feature generation for transductive open-set recognition via dual-space consistent sampling, Pattern Recognit. 146 (2024) 110046.

[66] D. Pal, D. More, S. Bhargav, D. Tamboli, V. Aggarwal, B. Banerjee, Domain adaptive few-shot open-set learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18831–18840.

[67] H. Wang, G. Pang, P. Wang, L. Zhang, W. Wei, Y. Zhang, Glocal energy-based learning for few-shot open-set recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7507–7516.

[68] Z. Xia, P. Wang, G. Dong, H. Liu, Adversarial kinetic prototype framework for open set recognition, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14, Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[69] E.-R. Engelbrecht, J.A. du Preez, On the link between generative semi-supervised learning and generative open-set recognition, Sci. Afr. 22 (2023) e01903.

[70] G. Jiang, P. Zhu, Y. Wang, Q. Hu, Openmix+: Revisiting data augmentation for open set recognition, IEEE Trans. Circuits Syst. Video Technol. (2023).

[71] M. Mundt, I. Pliushch, S. Majumder, V. Ramesh, Open set recognition through deep neural network uncertainty: Does out-of-distribution detection require generative classifiers? in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[72] C. Pires, M. Barandas, L. Fernandes, D. Folgado, H. Gamboa, Towards knowledge uncertainty estimation for open set recognition, Mach. Learn. Knowl. Extract. 2 (4) (2020) 505–532.

[73] K. Mazur, E. Sucar, A.J. Davison, Feature-realistic neural fusion for real-time, open set scene understanding, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, 2023, pp. 8201–8207.

[74] S. Sisti, W. Bennette, Open-set recognition for automatic target recognition: practical considerations for obtaining out of distribution examples, in: Automatic Target Recognition XXXII, vol. 12096, SPIE, 2022, pp. 145–152.

[75] C. Zhao, D. Du, A. Hoogs, C. Funk, Open set action recognition via multi-label evidential learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22982–22991.

[76] P. Saranrittichai, C.K. Mummadi, C. Blaiotta, M. Munoz, V. Fischer, Multi-attribute open set recognition, in: DAGM German Conference on Pattern Recognition, Springer, 2022, pp. 101–115.

[77] J.-J. Shao, X.-W. Yang, L.-Z. Guo, Open-set learning under covariate shift, Mach. Learn. (2022) 1–17.

[78] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023, arXiv preprint arXiv:2303.05499.

[79] K. Peng, C. Yin, J. Zheng, R. Liu, D. Schneider, J. Zhang, K. Yang, M.S. Sarfraz, R. Stiefelhagen, A. Roitberg, Navigating open set scenarios for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 4487–4496.

[80] M. Chen, J.-Y. Xia, T. Liu, L. Liu, Y. Liu, Open set recognition and category discovery framework for SAR target classification based on K-contrast loss and deep clustering, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (2024).

[81] M. Soltani, B. Ousat, M. Jafari Siavoshani, A.H. Jahangir, An adaptable deep learning-based intrusion detection system to zero-day attacks, J. Inform. Secur. Appl. 76 (2023) 103516.

[82] J. Guo, Y. Xu, W. Xu, Y. Zhan, Y. Sun, S. Guo, Mdenet: multi-modal dual-embedding networks for malware open-set recognition, 2023, arXiv preprint arXiv:2305.01245.

[83] J. Guo, H. Wang, Y. Xu, W. Xu, Y. Zhan, Y. Sun, S. Guo, Multimodal dual-embedding networks for malware open-set recognition, IEEE Trans. Neural Netw. Learn. Syst. (2024).

[84] L. Du, Z. Gu, Y. Wang, C. Gao, Open world intrusion detection: An open set recognition method for can bus in intelligent connected vehicles, IEEE Netw. (2024).

[85] H. Shao, D. Zhong, Towards open-set touchless palmprint recognition via weight-based meta metric learning, Pattern Recognit. 121 (2022) 108247.

[86] W. Liu, Y. Wen, B. Raj, R. Singh, A. Weller, SphereFace revived: Unifying hyperspherical face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2023) 2458–2474, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[87] R.H. Vareto, M. Günther, W.R. Schwartz, Open-set face recognition with neural ensemble, maximal entropy loss and feature augmentation, in: 2023 36th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, IEEE, 2023, pp. 55–60.

[88] T. Li, Z. Wen, Y. Long, Z. Hong, S. Zheng, L. Yu, B. Chen, X. Yang, L. Shao, The importance of expert knowledge for automatic modulation open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[89] C.-S. Shieh, F.-A. Ho, M.-F. Horng, T.-T. Nguyen, P. Chakrabarti, Open-set recognition in unknown ddos attacks detection with reciprocal points learning, IEEE Access (2024).

[90] R. Torres, Computer Science in Sport, Springer, 2024, pp. 217–222, http://dx.doi.org/10.1007/978-3-662-68313-2-26, Ch. Open-Set Recognition.

[91] Y. Zheng, G. Chen, M. Huang, Out-of-domain detection for natural language understanding in dialog systems, 2022, arXiv:1909.03862 [cs].

[92] C. Liu, C. Yang, H.-B. Qin, X. Zhu, C.-L. Liu, X.-C. Yin, Towards open-set text recognition via label-to-prototype learning, Pattern Recognit. 134 (2023) 109109.

[93] A. Fedotova, A. Kurtukova, A. Romanov, A. Shelupanov, Semantic clustering and transfer learning in social media texts authorship attribution, IEEE Access (2024).

[94] Y. Hu, J. Gao, J. Dong, B. Fan, H. Liu, Exploring rich semantics for open-set action recognition, IEEE Trans. Multimed. (2023).

[95] B. Yang, L. He, N. Ling, Z. Yan, G. Xing, X. Shuai, X. Ren, X. Jiang, EdgeFM: Leveraging foundation model for open-set learning on the edge, 2023, arXiv preprint arXiv:2311.10986.

[96] H. Qu, X. Hui, Y. Cai, J. Liu, Lmc: Large model collaboration with cross-assessment for training-free open-set object recognition, Adv. Neural Inf. Process. Syst. 36 (2024).

[97] X. Huang, Y.-J. Huang, Y. Zhang, W. Tian, R. Feng, Y. Zhang, Y. Xie, Y. Li, L. Zhang, Inject semantic concepts into image tagging for open-set recognition, 2023, arXiv preprint arXiv:2310.15200.

[98] C. Wang, C. Xin, Z. Xu, A novel deep metric learning model for imbalanced fault diagnosis and toward open-set classification, Knowl.-Based Syst. 220 (2021) 106925.

[99] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A.A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proc. Natl. Acad. Sci. 114 (13) (2017) 3521–3526.

[100] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: International Conference on Machine Learning, PMLR, 2017, pp. 3987–3995.

[101] W. Cong, Y. Cong, G. Sun, Y. Liu, J. Dong, Self-paced weight consolidation for continual learning, IEEE Trans. Circuits Syst. Video Technol. (2023).

[102] S. Kim, L. Noci, A. Orvieto, T. Hofmann, Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11930–11939.

[103] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P.K. Dokania, P.H. Torr, M. Ranzato, On tiny episodic memories in continual learning, 2019, arXiv preprint arXiv:1902.10486.

[104] S. Ho, M. Liu, L. Du, L. Gao, Y. Xiang, Prototype-guided memory replay for continual learning, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[105] X. Li, B. Tang, H. Li, AdaER: An adaptive experience replay approach for continual lifelong learning, Neurocomputing 572 (2024) 127204.

[106] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, G. Mori, Lifelong GAN: Continual learning for conditional image generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2759–2768.

[107] A. Ayub, A.R. Wagner, Eec: Learning to encode and regenerate images for continual learning, 2021, arXiv preprint arXiv:2101.04904.

[108] R. Gao, W. Liu, Ddgr: Continual learning with deep diffusion-based generative replay, in: International Conference on Machine Learning, PMLR, 2023, pp. 10744–10763.

[109] Y. Kong, L. Liu, Z. Wang, D. Tao, Balancing stability and plasticity through advanced null space in continual learning, in: European Conference on Computer Vision, Springer, 2022, pp. 219–236.

[110] Y. Guo, W. Hu, D. Zhao, B. Liu, Adaptive orthogonal projection for batch and online continual learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 6783–6791.

[111] S. Lin, L. Yang, D. Fan, J. Zhang, Trgp: Trust region gradient projection for continual learning, 2022, arXiv preprint arXiv:2202.02931.

[112] D. Rao, F. Visin, A. Rusu, R. Pascanu, Y.W. Teh, R. Hadsell, Continual unsupervised representation learning, in: Advances in Neural Information Processing Systems, vol. 32, 2019.

[113] M. Davari, N. Asadi, S. Mudur, R. Aljundi, E. Belilovsky, Probing representation forgetting in supervised and unsupervised continual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16712–16721.

[114] M. Xue, H. Zhang, J. Song, M. Song, Meta-attention for vit-backed continual learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 150–159.

[115] H. Jin, E. Kim, Helpful or harmful: Inter-task association in continual learning, in: European Conference on Computer Vision, Springer, 2022, pp. 519–535.

[116] Q. Gao, Z. Luo, D. Klabjan, F. Zhang, Efficient architecture search for continual learning, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[117] L. Wang, X. Zhang, Q. Li, J. Zhu, Y. Zhong, Coscl: Cooperation of small continual learners is stronger than a big one, in: European Conference on Computer Vision, Springer, 2022, pp. 254–271.

[118] L. Wang, X. Zhang, K. Yang, L. Yu, C. Li, L. Hong, S. Zhang, Z. Li, Y. Zhong, J. Zhu, Memory replay with data compression for continual learning, 2022, arXiv preprint arXiv:2202.06592.

[119] A.G. Menezes, G. de Moura, C. Alves, A.C. de Carvalho, Continual object detection: A review of definitions, strategies, and challenges, Neural Netw. 161 (2023) 476–493.

[120] J. Kim, H. Cho, J. Kim, Y.Y. Tiruneh, S. Baek, SDDGR: Stable diffusion-based deep generative replay for class incremental object detection, 2024, arXiv preprint arXiv:2402.17323.

[121] Y. Liu, B. Schiele, A. Vedaldi, C. Rupprecht, Continual detection transformer for incremental object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23799–23808.

[122] J.Z. Wu, D.J. Zhang, W. Hsu, M. Zhang, M.Z. Shou, Label-efficient online continual object detection in streaming video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19246–19255.

[123] Y. Li, X. Yang, H. Wang, X. Wang, T. Li, Learning to prompt knowledge transfer for open-world continual learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 13700–13708.

[124] S.Y. Gadre, K. Ehsani, S. Song, R. Mottaghi, Continuous scene representations for embodied AI, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14849–14859.

[125] E. Camuffo, S. Milani, Continual learning for LiDAR semantic segmentation: Class-incremental and coarse-to-fine strategies on sparse data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2447–2456.

[126] T.-D. Truong, H.-Q. Nguyen, B. Raj, K. Luu, Fairness continual learning approach to semantic scene understanding in open-world environments, Adv. Neural Inf. Process. Syst. 36 (2024).

[127] Z. Abbas, R. Zhao, J. Modayil, A. White, M.C. Machado, Loss of plasticity in continual deep reinforcement learning, in: Conference on Lifelong Learning Agents, PMLR, 2023, pp. 620–636.

[128] Y. Huang, Y. Zhang, J. Chen, X. Wang, D. Yang, Continual learning for text classification with information disentanglement based regularization, 2021, arXiv preprint arXiv:2104.05489.

[129] B. Liu, S. Mazumder, Lifelong and continual learning dialogue systems: Learning during conversation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 15058–15063.

[130] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, G. Haffari, Continual learning for large language models: A survey, 2024, arXiv preprint arXiv:2402.01364.

[131] Z. Ke, Y. Shao, H. Lin, T. Konishi, G. Kim, B. Liu, Continual pre-training of language models, 2023, arXiv preprint arXiv:2302.03241.

[132] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, A. Almahairi, Progressive prompts: Continual learning for language models, 2023, arXiv preprint arXiv:2301.12314.

[133] Y. Wang, Y. Liu, C. Shi, H. Li, C. Chen, H. Lu, Y. Yang, Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions, 2024, arXiv preprint arXiv:2403.11435.

[134] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, N. Zhang, Editing large language models: Problems, methods, and opportunities, 2023, arXiv preprint arXiv:2305.13172.

[135] S.Y. Yerima, A. Bashar, Semi-supervised novelty detection with one class SVM for SMS spam detection, in: 2022 29th International Conference on Systems, Signals and Image Processing, IWSSIP, vol. CFP2255E-ART, 2022, pp. 1–4, ISSN: 2157-8702.

[136] A. Mensi, M. Bicego, A novel anomaly score for isolation forests, in: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20, Springer, 2019, pp. 152–163.

[137] M. Tokovarov, P. Karczmarek, A probabilistic generalization of isolation forest, Inform. Sci. 584 (2022) 433–449.

[138] Z. Xu, D. Kakde, A. Chaudhuri, Automatic hyperparameter tuning method for local outlier factor, with applications to anomaly detection, in: 2019 IEEE International Conference on Big Data, Big Data, IEEE, 2019, pp. 4201–4207.

[139] X. Yang, K. Peng, L. An, P. Huang, P. Feng, GMBLOF: A machine learning algorithm of novelty detection based on local outlier factor, in: 2022 5th International Conference on Pattern Recognition and Artificial Intelligence, PRAI, IEEE, 2022, pp. 20–25.

[140] A. Shah, N. Azam, B. Ali, M.T. Khan, J. Yao, A three-way clustering approach for novelty detection, Inform. Sci. 569 (2021) 650–668.

[141] F. Zhu, W. Zhang, X. Chen, X. Gao, N. Ye, Large margin distribution multi-class supervised novelty detection, Expert Syst. Appl. 224 (2023) 119937.

[142] G. Pang, C. Shen, L. Cao, A.V.D. Hengel, Deep learning for anomaly detection: A review, ACM Comput. Surv. (CSUR) 54 (2) (2021) 1–38.

[143] M. Salehi, A. Arya, B. Pajoum, M. Otoofi, A. Shaeiri, M.H. Rohban, H.R. Rabiee, ARAE: Adversarially robust training of autoencoders improves novelty detection, 2020, arXiv:2003.05669 [cs].

[144] S.-Y. Lo, P. Oza, V.M. Patel, Adversarially robust one-class novelty detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2023) 4167–4179, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[145] Y. Huang, Y. Li, G. Jourjon, S. Seneviratne, K. Thilakarathna, A. Cheng, D. Webb, R.Y.D. Xu, Calibrated reconstruction based adversarial autoencoder model for novelty detection, Pattern Recognit. Lett. 169 (2023) 50–57.

[146] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, N. Ding, GAN-based anomaly detection: A review, Neurocomputing 493 (2022) 497–535.

[147] H. Mirzaei, M. Salehi, S. Shahabi, E. Gavves, C.G.M. Snoek, M. Sabokrou, M.H. Rohban, Fake it till you make it: Towards accurate near-distribution novelty detection, 2022, arXiv:2205.14297 [cs].

[148] G. Pang, L. Cao, L. Chen, H. Liu, Learning representations of ultrahigh-dimensional data for random distance-based outlier detection, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2041–2050.

[149] A.G. Roselin, P. Nanda, S. Nepal, X. He, Intelligent anomaly detection for large network traffic with optimized deep clustering (ODC) algorithm, IEEE Access 9 (2021) 47243–47251.
[150] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection, Inform. Sci. 557 (2021) 317–331.
[151] N. Shang, A. Wang, Y. Ding, K. Gai, L. Zhu, G. Zhang, A machine learning based golden-free detection method for command-activated hardware Trojan, Inform. Sci. 540 (2020) 292–307.
[152] M. Freitas, V. de Aquino Piai, R. Dazzi, R. Teive, W. Parreira, A. Fernandes, I.M. Pires, V.R.Q. Leithardt, Identification of abnormal behavior in activities of daily life using novelty detection, in: International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer, 2022, pp. 559–570.
[153] Z. Liu, Y. Zhou, Y. Xu, Z. Wang, Simplenet: A simple network for image anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20402–20411.
[154] S. Tang, Z. Wang, C. Yu, C. Sun, Y. Li, J. Xiao, Fast and accurate novelty detection for large surveillance video, CCF Trans. High Perform. Comput. (2024) 1–20.
[155] D. Jeon, J. Lee, J.M. Ahn, C. Lee, Measuring the novelty of scientific publications: A fasttext and local outlier factor approach, J. Informetrics 17 (4) (2023) 101450.
[156] K. Gupta, A. Ahmad, T. Ghosal, A. Ekbal, SciND: A new triplet-based dataset for scientific novelty detection via knowledge graphs, Int. J. Digit. Libraries (2024) 1–21.
[157] M.J. Hossen, J.M.Z. Hoque, T.T. Ramanathan, J.E. Raja, et al., Unsupervised novelty detection for time series using a deep learning approach, Heliyon 10 (3) (2024).
[158] S.U. Din, J. Shao, J. Kumar, C.B. Mawuli, S.H. Mahmud, W. Zhang, Q. Yang, Data stream classification with novel class detection: A review, comparison and challenges, Knowl. Inf. Syst. 63 (2021) 2231–2276.
[159] S.U. Din, J. Shao, Exploiting evolving micro-clusters for data stream classification with emerging class detection, Inform. Sci. 507 (2020) 404–420.
[160] G. Liao, P. Zhang, H. Yin, X. Deng, Y. Li, H. Zhou, D. Zhao, A novel semi-supervised classification approach for evolving data streams, Expert Syst. Appl. 215 (2023) 119273.
[161] L. Xu, X. Ding, H. Peng, D. Zhao, X. Li, ADTCD: An adaptive anomaly detection approach towards concept-drift in IoT, IEEE Internet Things J. (2023).
[162] J. Zhang, T. Wang, W.W.Y. Ng, W. Pedrycz, KNNENS: A k-nearest neighbor ensemble-based method for incremental learning under data stream with emerging new classes, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–8, Conference Name: IEEE Transactions on Neural Networks and Learning Systems.
[163] S.U. Din, A. Ullah, C.B. Mawuli, Q. Yang, J. Shao, A reliable adaptive prototype-based learning for evolving data streams with limited labels, Inf. Process. Manage. 61 (1) (2024) 103532.
[164] H. Guo, H. Xia, H. Li, W. Wang, Concept evolution detection based on noise reduction soft boundary, Inform. Sci. 628 (2023) 391–408.
[165] V. Agate, S. Drago, P. Ferraro, G.L. Re, Anomaly detection for reoccurring concept drift in smart environments, in: 2022 18th International Conference on Mobility, Sensing and Networking, MSN, IEEE, 2022, pp. 113–120.
[166] W. Liu, X. Wang, J.D. Owens, Y. Li, Energy-based Out-of-distribution Detection, 2021, arXiv:2010.03759 [cs].
[167] A. Joshi, S. Chalasani, K.N. Iyer, Semantic driven energy based out-of-distribution detection, in: 2022 International Joint Conference on Neural Networks, IJCNN, IEEE, 2022, pp. 01–08.
[168] S. Elflein, Master's Thesis: Out-of-distribution Detection with Energy-based Models, 2023, arXiv:2302.12002 [cs].
[169] Q. Chen, W. Jiang, K. Li, Y. Wang, Improving energy-based out-of-distribution detection by sparsity regularization, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2022, pp. 539–551.
[170] Q. Wu, Y. Chen, C. Yang, J. Yan, Energy-based out-of-distribution detection for graph neural networks, 2023, arXiv:2302.02914 [cs].
[171] D. Hendrycks, M. Mazeika, T. Dietterich, Deep anomaly detection with outlier exposure, 2018, arXiv preprint arXiv:1812.04606.
[172] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, K. Czarnecki, Out-of-distribution detection in classifiers via generation, 2019, arXiv preprint arXiv:1910.04241.
[173] A. Wu, D. Chen, C. Deng, Deep feature deblurring diffusion for detecting out-of-distribution objects, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13381–13391.
[174] X. Du, Y. Sun, J. Zhu, Y. Li, Dream the impossible: Outlier imagination with diffusion models, Adv. Neural Inf. Process. Syst. 36 (2024).
[175] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: International Conference on Machine Learning, PMLR, 2022, pp. 20827–20840.
[176] E. Zisselman, A. Tamar, Deep residual flow for out of distribution detection, 2020, arXiv:2001.05419 [cs, stat].
[177] P. Kirichenko, P. Izmailov, A.G. Wilson, Why normalizing flows fail to detect out-of-distribution data, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 20578–20589.
[178] C. Horvat, J.-P. Pfister, Density estimation on low-dimensional manifolds: An inflation-deflation approach, J. Mach. Learn. Res. 24 (61) (2023) 1–37.
[179] E.D. Cook, M.-A. Lavoie, S.L. Waslander, Feature density estimation for out-of-distribution detection via normalizing flows, 2024, arXiv preprint arXiv:2402.06537.
[180] A. Martinez-Seras, J. Del Ser, J. L. Lobo, P. Garcia-Bringas, N. Kasabov, A novel out-of-distribution detection approach for spiking neural networks: Design, fusion, performance evaluation and explainability, Inf. Fusion 100 (2023) 101943.
[181] Z. Cheng, X.-Y. Zhang, C.-L. Liu, Unified classification and rejection: A one-versus-all framework, 2023, arXiv preprint arXiv:2311.13355.
[182] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022, pp. 20827–20840, ISSN: 2640-3498.
[183] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, Y. Li, Delving into out-of-distribution detection with vision-language representations, in: Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 35087–35102.
[184] J. Ren, S. Fort, J. Liu, A.G. Roy, S. Padhy, B. Lakshminarayanan, A simple fix to Mahalanobis distance for improving near-ood detection, 2021, arXiv preprint arXiv:2106.09022.
[185] H. Anthony, K. Kamnitsas, On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging, in: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer, 2023, pp. 136–146.
[186] Y. Ming, Y. Sun, O. Dia, Y. Li, How to exploit hyperspherical embeddings for out-of-distribution detection? 2022, arXiv preprint arXiv:2203.04450.
[187] S. Wilson, T. Fischer, N. Sünderhauf, F. Dayoub, Hyperdimensional feature fusion for out-of-distribution detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2644–2654.
[188] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, S. Vernekar, Improving reconstruction autoencoder out-of-distribution detection with Mahalanobis distance, 2018, arXiv preprint arXiv:1812.02765.
[189] Y. Yang, R. Gao, Q. Xu, Out-of-distribution detection with semantic mismatch under masking, in: European Conference on Computer Vision, Springer, 2022, pp. 373–390.
[190] M.S. Graham, W.H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, J. Cardoso, Denoising diffusion models for out-of-distribution detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2947–2956.
[191] A.A. Dovganich, A.V. Khvostikov, Y.A. Pchelintsev, A.A. Krylov, Y. Ding, M.C. Farias, Automatic out-of-distribution detection methods for improving the deep learning classification of pulmonary X-ray images, J. Image Graph. 10 (2) (2022) 56–63.
[192] C. González, K. Gotkowski, M. Fuchs, A. Bucher, A. Dadras, R. Fischbach, I.J. Kaltenborn, A. Mukhopadhyay, Distance-based detection of out-of-distribution silent failures for COVID-19 lung lesion segmentation, Med. Image Anal. 82 (2022) 102596.
[193] Z. Song, G. Zhang, L. Liu, L. Yang, S. Xu, C. Jia, F. Jia, L. Wang, Robofusion: Towards robust multi-modal 3D obiect detection via SAM, 2024, arXiv preprint arXiv:2401.03907.
[194] Q. Wu, H. Jiang, H. Yin, B.F. Karlsson, C.-Y. Lin, Multi-level knowledge distillation for out-of-distribution detection in text, 2022, arXiv preprint arXiv:2211.11300.
[195] Z. Lai, X. Zhang, S. Chen, Adaptive ensembles of fine-tuned transformers for llm-generated text detection, 2024, arXiv preprint arXiv:2403.13335.
[196] A. Corsini, S.J. Yang, Are existing out-of-distribution techniques suitable for network intrusion detection? in: 2023 IEEE Conference on Communications and Network Security, CNS, IEEE, 2023, pp. 1–9.
[197] Y.A. Farrukh, S. Wali, I. Khan, N.D. Bastian, Detecting unknown attacks in iot environments: An open set classifier for enhanced network intrusion detection, in: MILCOM 2023-2023 IEEE Military Communications Conference, MILCOM, IEEE, 2023, pp. 121–126.
[198] J. Hyun, S. Kim, G. Jeon, S.H. Kim, K. Bae, B.J. Kang, ReConPatch: Contrastive patch representation learning for industrial anomaly detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2052–2061.
[199] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
[200] A. Mobiny, P. Yuan, S.K. Moulik, N. Garg, C.C. Wu, H. Van Nguyen, Dropconnect is effective in modeling uncertainty of Bayesian deep networks, Sci. Rep. 11 (1) (2021) 5458.
[201] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
[202] G. Luo, M. Blumenthal, M. Heide, M. Uecker, Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models, Magn. Reson. Med. 90 (1) (2023) 295–311.
[203] M. Hobbhahn, A. Kristiadi, P. Hennig, Fast predictive uncertainty for classification with Bayesian deep networks, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 822–832.

[204] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, I. Bloch, Encoding the latent posterior of Bayesian neural networks for uncertainty quantification, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[205] A. Malinin, L. Prokhorenkova, A. Ustimenko, Uncertainty in gradient boosting via ensembles, 2020, arXiv preprint arXiv:2006.10562.
[206] S. Jain, G. Liu, J. Mueller, D. Gifford, Maximizing overall diversity for improved uncertainty estimates in deep ensembles, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 4264–4271.
[207] T. Tohme, K. Vanslette, K. Youcef-Toumi, Reliable neural networks for regression uncertainty estimation, Reliab. Eng. Syst. Saf. 229 (2023) 108811.
[208] G. Deng, K. Zou, K. Ren, M. Wang, X. Yuan, S. Ying, H. Fu, SAM-U: Multi-box prompts triggered uncertainty estimation for reliable SAM in medical image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 368–377.
[209] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023), Comput. Biol. Med. (2023) 107441.
[210] L. Lu, M. Yin, L. Fu, F. Yang, Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation, Biomed. Signal Process. Control 79 (2023) 104203.
[211] M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian, et al., Uncertainty-inspired open set learning for retinal anomaly identification, Nature Commun. 14 (1) (2023) 6757.
[212] M. Hu, Z. Zhang, S. Zhao, M. Huang, B. Wu, Uncertainty in natural language processing: Sources, quantification, and applications, 2023, arXiv preprint arXiv:2306.04459.
[213] L. Kuhn, Y. Gal, S. Farquhar, Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023, arXiv preprint arXiv:2302.09664.
[214] Y. Huang, J. Song, Z. Wang, H. Chen, L. Ma, Look before you leap: An exploratory study of uncertainty measurement for large language models, 2023, arXiv preprint arXiv:2307.10236.
[215] S. Singi, Z. He, A. Pan, S. Patel, G.A. Sigurdsson, R. Piramuthu, S. Song, M. Ciocarlie, Decision making for human-in-the-loop robotic agents via uncertainty-aware reinforcement learning, 2023, arXiv preprint arXiv:2303.06710.
[216] C.-I. Yang, Y.-P. Li, Explainable uncertainty quantifications for deep learning-based molecular property prediction, J. Cheminform. 15 (1) (2023) 13.
[217] J. Wu, J. Chen, D. Huang, Entropy-based active learning for object detection with progressive diversity constraint, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9397–9406.
[218] W.H. Beluch, T. Genewein, A. Nürnberger, J.M. Köhler, The power of ensembles for active learning in image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9368–9377.
[219] V.E. Martins, A. Cano, S.B. Junior, Meta-learning for dynamic tuning of active learning on stream classification, Pattern Recognit. 138 (2023) 109359.
[220] K.-P. Ning, X. Zhao, Y. Li, S.-J. Huang, Active learning for open-set annotation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 41–49.
[221] B. Safaei, V. Vibashan, C.M. de Melo, V.M. Patel, Entropic open-set active learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 4686–4694.
[222] G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander, X. Qian, Efficient active learning for Gaussian process classification by error reduction, Adv. Neural Inf. Process. Syst. 34 (2021) 9734–9746.
[223] P. Wei, Y. Zheng, J. Fu, Y. Xu, W. Gao, An expected integrated error reduction function for accelerating Bayesian active learning of failure probability, Reliab. Eng. Syst. Saf. 231 (2023) 108971.
[224] M. Wang, F. Min, Z.-H. Zhang, Y.-X. Wu, Active learning through density clustering, Expert Syst. Appl. 85 (2017) 305–317.
[225] K. Margatina, T. Schick, N. Aletras, J. Dwivedi-Yu, Active learning principles for in-context learning with large language models, 2023, arXiv preprint arXiv:2305.14264.
[226] R. Greer, B. Antoniussen, M.V. Andersen, A. Møgelmose, M.M. Trivedi, The why, when, and how to use active learning in large-data-driven 3D object detection for safe autonomous driving: An empirical exploration, 2024, arXiv preprint arXiv:2401.16634.
[227] F. Wu, P. Marquez-Neila, M. Zheng, H. Rafii-Tari, R. Sznitman, Correlation-aware active learning for surgery video segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 2010–2020.
[228] X. Zheng, G. Jia, Active learning based reverse design of hydrogen production from biomass fuel, Fuel 357 (2024) 129948.
[229] F. Gusev, E. Gutkin, M.G. Kurnikova, O. Isayev, Active learning guided drug design lead optimization based on relative binding free energy modeling, J. Chem. Inform. Model. 63 (2) (2023) 583–594.
[230] D. Liu, J. Yu, Otsu method and K-means, in: 2009 Ninth International Conference on Hybrid Intelligent Systems, vol. 1, 2009, pp. 344–349.
[231] J. Komorniczak, P. Ksieniewicz, Torchosr—A PyTorch extension package for open set recognition models evaluation in Python, Neurocomputing 566 (2024) 127047.
[232] K. Samunnisa, G.S.V. Kumar, K. Madhavi, Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods, Measurement: Sensors 25 (2023) 100612.
[233] J. Henrydoss, S. Cruz, C. Li, M. Günther, T.E. Boult, Enhancing open-set recognition using clustering-based extreme value machine (C-EVM), in: 2020 IEEE International Conference on Big Data, Big Data, 2020, pp. 441–448.
[234] Q. Qian, S. Chen, W. Cai, Simultaneous clustering and classification over cluster structure representation, Pattern Recognit. 45 (6) (2012) 2227–2236.
[235] W. Cai, S. Chen, D. Zhang, A simultaneous learning framework for clustering and classification, Pattern Recognit. 42 (2009) 1248–1259.
[236] N. Bharill, A. Tiwari, An improved multiobjective simultaneous learning framework for designing a classifier, in: 2011 International Conference on Recent Trends in Information Technology, ICRTIT, 2011, pp. 737–742.
[237] R. Liu, Y. Chen, L. Jiao, Y. Li, A particle swarm optimization based simultaneous learning framework for clustering and classification, Pattern Recognit. 47 (6) (2014) 2143–2152.
[238] J. Luo, L. Jiao, R. Shang, F. Liu, Learning simultaneous adaptive clustering and classification via MOEA, Pattern Recognit. 60 (2016) 37–50.
[239] H. Li, F. He, Y. Chen, Learning dynamic simultaneous clustering and classification via automatic differential evolution and firework algorithm, Appl. Soft Comput. 96 (2020) 106593.
[240] H. Li, Z. Wang, C. Lan, P. Wu, N. Zeng, A novel dynamic multiobjective optimization algorithm with non-inductive transfer learning based on multi-strategy adaptive selection, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[241] H. Li, Z. Wang, C. Lan, P. Wu, N. Zeng, A novel dynamic multiobjective optimization algorithm with hierarchical response system, IEEE Trans. Comput. Soc. Syst. (2023).
[242] A. Acharya, E.R. Hruschka, J. Ghosh, S. Acharyya, C 3 e: A framework for combining ensembles of classifiers and clusterers, in: Multiple Classifier Systems: 10th International Workshop, MCS 2011, Naples, Italy, June 15-17, 2011. Proceedings 10, Springer, 2011, pp. 269–278.
[243] M.A. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249.
[244] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, 2021, arXiv preprint arXiv:2110.11334.
[245] S. Agrahari, S. Srivastava, A.K. Singh, Review on novelty detection in the non-stationary environment, Knowl. Inf. Syst. 66 (3) (2024) 1549–1574.
[246] Z. Chen, B. Liu, Lifelong Machine Learning, vol. 1, Springer, 2018.
[247] D.-W. Zhou, Q.-W. Wang, Z.-H. Qi, H.-J. Ye, D.-C. Zhan, Z. Liu, Deep class-incremental learning: A survey, 2023, arXiv preprint arXiv:2302.03648.
[248] M. Mundt, I. Pliushch, S. Majumder, Y. Hong, V. Ramesh, Unified probabilistic deep continual learning through generative replay and open set recognition, J. Imaging 8 (4) (2022) 93.
[249] Y. Gao, Y.-F. Li, B. Dong, Y. Lin, L. Khan, Sim: Open-world multi-task stream classifier with integral similarity metrics, in: 2019 IEEE International Conference on Big Data, Big Data, IEEE, 2019, pp. 751–760.
[250] J. Leo, J. Kalita, Moving towards open set incremental learning: readily discovering new authors, in: Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2, Springer, 2020, pp. 739–751.
[251] L. Wang, X. Zhang, H. Su, J. Zhu, A comprehensive survey of continual learning: Theory, method and application, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[252] P. Wu, Z. Wang, H. Li, N. Zeng, KD-PAR: A knowledge distillation-based pedestrian attribute recognition model with multi-label mixed feature learning network, Expert Syst. Appl. 237 (2024) 121305.
[253] N. Zeng, X. Li, P. Wu, H. Li, X. Luo, A novel tensor decomposition-based efficient detector for low-altitude aerial objects with knowledge distillation scheme, IEEE/CAA J. Autom. Sin. 11 (2) (2024) 487–501.
[254] A.L. Suárez-Cetrulo, D. Quintana, A. Cervantes, A survey on machine learning for recurring concept drifting data streams, Expert Syst. Appl. 213 (2023) 118934.
[255] A. Gillert, U.F. von Lukas, Towards combined open set recognition and out-of-distribution detection for fine-grained classification, in: VISIGRAPP (5: VISAPP), 2021, pp. 225–233.
[256] S. Liang, Y. Li, R. Srikant, Enhancing the reliability of out-of-distribution image detection in neural networks, 2017, arXiv preprint arXiv:1706.02690.
[257] T. Ramalho, M. Miranda, Density estimation in representation space to predict model uncertainty, in: Engineering Dependable and Secure Machine Learning Systems: Third International Workshop, EDSMLS 2020, New York City, NY, USA, February 7, 2020, Revised Selected Papers 3, Springer, 2020, pp. 84–96.
[258] J. Linmans, S. Elfwing, J. van der Laak, G. Litjens, Predictive uncertainty estimation for out-of-distribution detection in digital pathology, Med. Image Anal. 83 (2023) 102655.
[259] B.J. Meyer, T. Drummond, The importance of metric learning for robotic vision: Open set recognition and active learning, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 2924–2931.

[260] R.H. Vareto, Y. Linghu, T.E. Boult, W.R. Schwartz, M. Günther, Open-set face recognition with maximal entropy and objectosphere loss, Image Vis. Comput. 141 (2024) 104862.
[261] J. Park, H. Park, E. Jeong, A.B.J. Teoh, Understanding open-set recognition by Jacobian norm and inter-class separation, Pattern Recognit. 145 (2024) 109942.
[262] J.L. Lobo, J. Del Ser, E. Osaba, Lightweight alternatives for hyper-parameter tuning in drifting data streams, in: 2021 International Conference on Data Mining Workshops, ICDMW, 2021, pp. 304–311.
[263] D.M.V. Sato, S.C. de Freitas, J.P. Barddal, E.E. Scalabrin, A survey on concept drift in process mining, ACM Comput. Surv. 54 (9) (2022) 1–38, arXiv:2112.02000 [cs].
[264] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al., Discovering language model behaviors with model-written evaluations, 2022, arXiv preprint arXiv:2212.09251.
[265] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S.R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S.R. Johnston, et al., Towards understanding sycophancy in language models, 2023, arXiv preprint arXiv:2310.13548.
[266] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023, arXiv preprint arXiv:2302.04023.
[267] P.S. Park, S. Goldstein, A. O'Gara, M. Chen, D. Hendrycks, AI deception: A survey of examples, risks, and potential solutions, 2023, arXiv preprint arXiv:2308.14752.
[268] J. Kirchner, L. Smith, J. Thibodeau, J. McDonell, L. Reynolds, Understanding AI alignment research: A systematic analysis, 2022, arXiv preprint arXiv:2206.02841.
[269] R. Paulus, C. Xiong, R. Socher, A deep reinforced model for abstractive summarization, 2017, arXiv preprint arXiv:1705.04304.
[270] W.B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, P. Stone, Reward (mis) design for autonomous driving, Artificial Intelligence 316 (2023) 103829.
[271] L. Gao, J. Schulman, J. Hilton, Scaling laws for reward model overoptimization, in: International Conference on Machine Learning, PMLR, 2023, pp. 10835–10866.

Marcos Barcina works as a researcher in Artificial Intelligence at TECNALIA. He is currently pursuing the Ph.D. program "Engineering for the Information Society and Sustainable Development" at the University of Deusto in conjunction with TECNALIA. The research performed in this program focuses on AI, specifically on the design, development and experimentation with AI systems capable of adapting and performing their functions in non-stationary environments. He also focuses on writing and publishing his research in journals and conferences with high scientific impact. He graduated in "Computer Engineering + Digital Transformation of the Enterprise" in 2021 at the University of Deusto. He then completed an online master's degree in "Cybersecurity and Privacy" at the Universitat Oberta de Catalunya in 2022. At the same time, he worked as a researcher at the University of Deusto in projects focused on technological development in the industrial sector such as "REMEDY" and "PILAR" where he was able to work with AI technologies. After this, he started his Ph.D. at the University of Deusto and TECNALIA until today.

Dr. Jesús López lives in Bilbao (Spain) and works at TECNALIA as Principal Investigator (PI) in Artificial Intelligence, after a cycle as Software Engineer and Project Manager that started in 2003. Jesús got his "Computer Science Engineering" at the *University of Deusto* in 2003, to then get started in Artificial Intelligence through the "Master in Advanced Artificial Intelligence" at the *National University of Distance Education* (UNED), which he finished in 2014. He then embarked on the Ph.D. program "Information and Communication Technologies in Mobile Networks (TICRM)" at the *University of the Basque Country (UPV/EHU)*, where he specialized in Artificial Intelligence, and in which he obtained in 2018 the Cum Laude grade, the International Mention, and the Extraordinary Ph.D. Award. During his thesis he spent 6 months at the KEDRI group at the *Auckland University of Technology (AUT)*, New Zealand. He has published in high impact journals (Q1) and conferences (e.g. ECML, ICDM, WCCI). He also collaborates with the *Universitat Oberta de Catalunya (UOC)* as Affiliate Professor. His lines of research and interests are: Adaptive AI (Stream Learning, Concept Drift) and AI Alignment, among others.

Dr. Pablo García Currently works as University Associate Professor (Profesor Titular de Universidad), he has been dedicated to Research, Technology and Innovation for 20 years, from positions as Head Researcher of DeustoTech Computing - S3Lab, and also as Director of DeustoTech - Deusto Institute of Technology. He has been also Director of Research at the Faculty of Engineering. He has combined teaching (in several B.Sc.,M.Sc., and Ph.D. courses) with research since 1998, mainly focused in the fields of Operating Systems, Cyber-Security and Artificial Intelligence. Since 2003, he directed the Official Master on Information Security for 11 editions. Also in 2003, he put the first building-blocks for the foundation of the research unit DeustoTech Computing (also named, in the past, S3Lab - Laboratory for Smartness, Semantics and Security), where he coordinated the research activities of around 30 researchers for more than ten years. This research unit -one of the most fruitful at DeustoTech- is devoted to Applied Artificial Intelligence, with application to different domains. In particular, his research background is focused on Machine Learning applied to the fields of (a) Information Security, (b) Industrial Processes, and (c) Genomics and Proteomics. He has over 20 years of experience in R&D management, with tens of projects and technology transfer actions led, for more than fifteen million euro, more than 200 international scientific publications (in the top 9% of spanish researchers and global researchers in spanish institutions, according to Google Scholar and Webometrics ranking), and 20 directed Ph.D. dissertations. He has co-chaired world-class scientific events such as DEXA, CISIS, SOCO, ICEUTE, HAIS, INFOSEC, BIGDAT or DEEP LEARNING BILBAO!

Dr. Javier Del Ser began his career as a Senior Telecommunications Engineer at the *University of the Basque Country (UPV/EHU)*, obtaining his degree in May 2003. Subsequently, in November of that same year he was awarded a doctoral scholarship for pursuing his doctoral thesis at the *Centro de Estudios e Investigaciones Técnicas de Gipuzkoa* (CEIT, Spain). He defended his doctoral thesis (Cum Laude) in Telecommunications from the *University of Navarra* in 2006, and a second doctoral thesis in "Information and Communication Technologies" (also Cum Laude and Extraordinary Doctorate Award) from the *University of Alcala* (Spain) in 2013. He combined his research work with teaching duties as an assistant professor (2003–2005), interim professor (2006) and associate professor at TECNUN (*University of Navarra*). From August to December 2007, he did a postdoctoral stay at the *University of Delaware* (Newark, DE, USA). In 2008, he joined the *Robotiker Foundation* as a Senior Research Scientist and Project Leader at the TELECOM Unit. Javier is currently a Research Professor and Principal Investigator (PI) at TECNALIA RESEARCH & INNOVATION, leader of the Joint Research Lab in Artificial Intelligence together with researchers from the *Basque Center of Applied Mathematics* (BCAM) and the UPV/EHU. He is also the director of the TECNALIA Chair in Artificial Intelligence implemented at the *University of Granada* (Spain), considered one of the leading universities in the world in this discipline of knowledge. He is also an Associate Professor in the Communications Engineering Department of the EHU/UPV, and an external scientific member of the BCAM. His research activity focuses on Artificial Intelligence, machine learning, Deep Learning and, in general, descriptive, prescriptive and predictive analytics. In these fields, he has published to date more than 400 scientific publications, edited 7 books, directed 30 master's theses and 11 doctoral theses, as well as participated/directed more of 50 research projects. Javier is a Senior member of the IEEE.