IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 March 2024, accepted 12 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377690

RESEARCH ARTICLE

Lifelong Continual Learning for Anomaly Detection: New Challenges, Perspectives, and Insights

KAMIL FABER[1], ROBERTO CORIZZO[1,2], (Member, IEEE), BARTLOMIEJ SNIEZYNSKI[1], AND NATHALIE JAPKOWICZ[2], (Member, IEEE)

[1]Faculty of Computer Science, AGH University of Science and Technology, 30-059 Kraków, Poland
[2]Department of Computer Science, American University, Washington, DC 20016, USA

Corresponding author: Roberto Corizzo (rcorizzo@american.edu)

The paper was supported by funds of the Polish Ministry of Science and Higher Education allocated to the AGH University, and by program "Excellence initiative - research university" for AGH University.

ABSTRACT Anomaly detection is of paramount importance in many real-world domains characterized by evolving behavior, such as monitoring cyber-physical systems, human conditions and network traffic. Current research in anomaly detection leverages offline learning working with static data or online learning focusing on constant adaptation to evolving data. At the same time, lifelong learning represents an emerging trend, answering the need for machine learning models that continuously adapt to new challenges in dynamic environments while retaining past knowledge. Although this aspect could be beneficial to build effective and robust anomaly detection models, lifelong learning research is mainly dedicated to proposing new model update strategies in image classification and reinforcement learning domains. The limited scope addressed by lifelong learning works thus far creates a gap in understanding whether such techniques and capabilities can be fruitfully exploited in anomaly detection contexts, which represents the main motivation of this paper. More specifically, anomaly detection provides unique challenges, such as an evolving normal class and limited availability of anomalies, which significantly differs from the landscape and scenarios of lifelong image classification and reinforcement learning. In this paper, we face this issue by exploring, motivating, and discussing lifelong anomaly detection, as well as providing foundations with regard to scenarios, strategies, and metrics. First, we explain why lifelong anomaly detection is relevant, defining challenges and opportunities to design anomaly detection methods that deal with lifelong learning complexities. Second, we formulate and characterize lifelong learning settings tailored for anomaly detection problems, and design a scenario generation procedure that enables researchers to experiment with lifelong anomaly detection using existing datasets. Third, we perform experiments with popular anomaly detection methods on proposed lifelong scenarios, emphasizing the gap in performance that could be filled with the adoption of lifelong learning. In summary, our efforts are directed at assessing the performance of non-lifelong anomaly detection models in lifelong scenarios and how the adoption of lifelong learning impacts their learning capabilities. Overall, we conclude that the adoption of lifelong anomaly detection is important to design more robust models that provide a comprehensive view of the environment, as well as simultaneous adaptation and knowledge retention.

INDEX TERMS Lifelong anomaly detection, lifelong learning, anomaly detection, continual learning, continual anomaly detection.

I. INTRODUCTION

Anomaly detection is the task of finding anomalous data instances that represent a deviation from the normal

The associate editor coordinating the review of this manuscript and approving it for publication was Xiong Luo.

conditions of a process [1], [2]. The capability to detect anomalous behavior is of paramount importance in many disciplines and real-world applications, such as intrusions in network traffic [3], irregular behavior in cyber-physical systems such as smart grids [4], as well as IoT environments [5], or defects in manufacturing processes [6]. The most widespread approach in machine learning is to model the normal behavior of the system and identify anomalies as data instances that significantly differ from the modeled behavior [7]. This choice reflects the limited availability of anomalies compared to the large availability of normal data, which results in the inability to model the anomaly class accurately. Moreover, it responds to the necessity of detecting anomalies with varying morphology unknown at training time [7], [8].

Most works in anomaly detection deal with the problem in an offline (batch) or online (stream) manner [1]. In evolving environments, models will become outdated and require updates, either as a full retraining stage (batch models), or as an online adaptation stage following concept drift detection (stream models) [9].

Both types of approaches are equally valid depending on the domain characteristics. For instance, offline approaches are commonly used for tasks such as lesion detection in medical images [10] and gravitational waves detection [11]. On the other hand, online approaches are common in domains characterized by a temporal dimension, such as real-time fatigue detection [12] and crowd anomaly detection [13]. Updating models allows them to adapt to the changing conditions of the normal class. However, it is noteworthy that updating the model has the side effect of gradually leading to forgetting past knowledge [14], [15]. Forgetting is a widely known phenomenon in data streams and online learning, and it is considered to be a positive feature in some scenarios as it allows models to focus on the most recent data characteristics [16]. For instance, in crowd anomaly detection [13], forgetting is suitable since it is assumed that only the people present in the current monitored environment (i.e., the most recent data) are the relevant ones to predict anomalies within that particular environment and at that particular time.

On the other hand, lifelong continual machine learning[1] research shows that forgetting is a problem that negatively affects models' performance when previously experienced conditions reoccur in the future [17], [18], [19], [20]. For this reason, lifelong learning seeks to find a balance between adapting to new knowledge while retaining past knowledge, inspired by biology, neuroscience, and computer science [14], [21], [22]. For instance, a popular example in lifelong reinforcement learning is having an agent able to learn how to play new games while not losing the ability to play previously known ones.

[1]Terms "lifelong" and "continual" are used interchangeably in existing literature. From now on, we will use the term lifelong to refer to this learning setting.

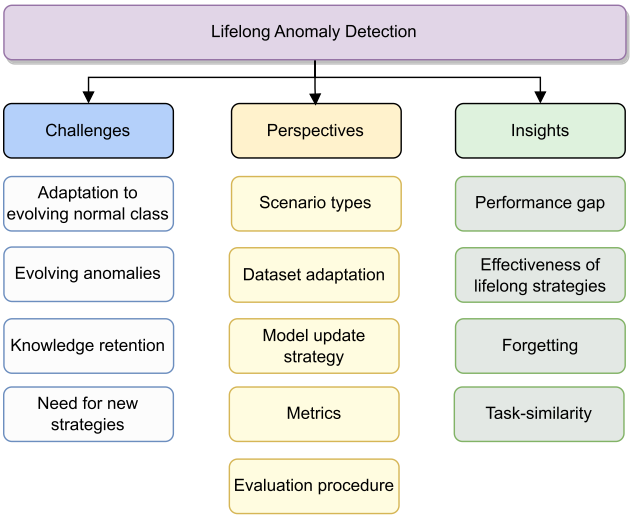

FIGURE 1. General view of lifelong learning in terms of challenges (see Section III), perspectives (see Section IV), and insights (see Section VI) examined in our paper.

Despite the emerging interest in lifelong learning, current research on this topic is mainly focused on computer vision and reinforcement learning [14], [23], [24], [25], with new trends including bridging active learning with open world learning [26] and applications to robotics [27], as well as online image classification [28]. In contrast, anomaly detection problems are still poorly explored. In this paper, we focus on lifelong learning from an anomaly detection perspective, showing that lifelong learning capabilities can bring several advantages in many real-world settings. We argue that considering them will yield more sophisticated models that can detect anomalies while adapting to changing environments and avoiding forgetting knowledge acquired in the past. One of the domains where this capability is crucial is cybersecurity [29]. For example, monitoring network traffic has to deal with dynamic conditions such as changes in the infrastructure, user behaviors, as well as new types of traffic and protocols [30]. Other examples of domains that we further describe in the paper include human condition monitoring and fault detection in industrial settings, although many more can be drawn.

Moreover, we leverage lessons learned in recent lifelong machine learning studies to showcase the current limitations of anomaly detection methods when exposed to lifelong learning scenarios. To this end, we formalize lifelong anomaly detection and devise the desiderata of models and scenarios, building the foundations for future work.

Our analytical scope focuses on two main research questions:

- **RQ1:** Do lifelong scenarios impact the performance of non-lifelong anomaly detection models?
- **RQ2:** Does the adoption of knowledge retention capabilities of lifelong learning provide a valuable improvement in the learning capabilities of existing anomaly detection models in complex lifelong scenarios?

With the first research question, we aim to assess whether current anomaly detection models are effective in lifelong learning scenarios and if there is a gap that needs to be addressed. With the second question, we aim to verify whether adopting lifelong learning strategies can be beneficial in anomaly detection contexts.

The contributions of this study can be summarized as follows:

- Bridging the gap between anomaly detection and lifelong learning, presenting the benefits of lifelong anomaly detection over conventional anomaly detection, which paves the way for new methods that are more robust in real-world domains;
- Devising a characterization of anomaly detection scenarios from a lifelong learning perspective, which sheds light on how to design and build more challenging benchmarks for anomaly detection;
- An open-source[2] implementation for scenario generation method, which can be applied to any anomaly detection dataset, facilitating wider adoption of lifelong learning in anomaly detection settings.
- Evaluating popular anomaly detection methods on our proposed lifelong scenarios, emphasizing challenges and limitations that occur in this setting.

We summarize challenges, perspectives, and insights in Figure 1. Our contributions allow us to highlight the potential of lifelong anomaly detection as a new, promising research direction. The scenarios we designed and the experimental results we obtained in our study help us showcase new challenges, perspectives, and insights brought by lifelong learning research in the context of anomaly detection. By doing so, we aim to increase awareness of the potential of lifelong anomaly detection while rationalizing and providing foundations with regard to scenarios, strategies, and metrics. Our study aims to streamline the adoption of lifelong anomaly detection for researchers and practitioners.

The paper is structured as follows. Section II summarizes related works in lifelong learning and anomaly detection. Section III explores the transition from conventional to lifelong anomaly detection. Section IV describes the proposed lifelong scenario design procedure. Section V discusses the experimental results obtained in our study. Finally, Section VII concludes the paper outlining directions of interest for future work.

II. BACKGROUND

In this section, we first briefly introduce lifelong learning along with the most popular types of methods. Second, we provide an overview of the current landscape of anomaly detection works.

A. LIFELONG LEARNING

Lifelong learning is a continuous process in which a series of different problems, defined as tasks, are presented to a learning method over time [14]. In the most common lifelong learning settings, image classification, a few scenario types built upon task characterization have been proposed. The most popular ones are task-incremental, class-incremental, and domain-incremental [31]. Both class-incremental and task-incremental scenarios provide the model with new, previously unseen classes that need to be incorporated by the model [32], [33]. On the other hand, domain-incremental scenarios provide new distributions of already known classes [34]. Emerging trends involve online lifelong learning [35], [36], simultaneous adaptation in terms of new instances and new classes [37], and repetition of observed tasks [38].

In the lifelong setting, the general goal for the learner is to be able to pick up new skills, adjust to newly presented tasks, and draw on previously learned information to tackle both new obstacles and the recurrence of previously seen tasks [39]. The key difference between lifelong learning and incremental/online learning is that, in lifelong learning, the attention is not solely focused on adaptation but also on knowledge retention and a model's ability to simultaneously handle all tasks, avoiding forgetting [18]. To this end, lifelong learning strategies are inspired by diverse disciplines, including neuroscience and biology [14]. Lifelong learning approaches proposed thus far fall into three main categories.

Regularization-based strategies work by introducing constraints on weight updates during the incremental training of neural networks. One of the initial ideas is to prevent updates for weights learned on previously trained tasks [40].

Another approach is to freeze the first layers in the model architecture to mitigate forgetting previous tasks while leaving the last fully connected layer unfrozen to be updated with new tasks [41]. On the other hand, methods such as EWC [18] and LWF [42] modify the regularization loss using a knowledge distillation to prevent drastic changes in already learned weights, which are important to solve previous tasks.

Dynamic architectures strategies adaptively manipulate the model architecture during the learning process. Methods usually expand the network by adding new neurons or layers as they encounter new tasks [43].

More efficient methods augment dynamic adaptation with pruning capabilities, which keep model capacity under control by removing insignificant weights. Popular examples are PackNet [44] and Winning Subnetworks [45], which splits the model into independent sub-networks, each specialized in addressing a different task. This type of approach is also referred to as forget-free, which holds only under assumptions such as the availability of task labels and unlimited capacity.

Replay-based strategies ensure that the knowledge from previously seen tasks is taken into consideration by the updated model by recurringly incorporating a summarized version of data from previous tasks in model updates. The most standard replay-based techniques focus on preserving knowledge by storing data samples from previously learned tasks in a memory buffer and replaying them during model update [14].

[2]https://github.com/lifelonglab/lifelong-anomaly-detection-scenarios

There are a few strategies devised to select the most relevant samples for every task, keep the replay buffer size compact, and, in turn, limit resource usage [46]. The second category leverages generative models to generate artificial data samples from previous tasks each time the model is updated [47]. By doing so, generative replay eases the burden of storing data samples, reducing the impact of memory occupation that affects conventional replay strategies.

B. ANOMALY DETECTION

We now turn our attention to anomaly detection. As we noted in Section I, anomaly detection has become crucial for decision support in many domains. For instance, the work in [5] emphasizes the importance of anomaly detection in IoT environments, such as transportation systems, health care systems, smart objects, and industrial systems. Similarly, anomaly detection in log sequences is critical for ensuring operational and security integrity in heterogeneous systems [48]. Another interesting example is Industry 4.0, with an example of 3D printing [49], where early identification of malfunctions is fundamental to limit economic losses.

In addition to the online *vs.* offline distinction mentioned in Section I, anomaly detection methods can also be characterized as supervised, semi-supervised, and unsupervised [7], [50], based on data and labels availability.

Supervised methods require labels for both normal and anomaly classes. They also need to be concerned about the class imbalance problem, and they are generally limited by the fact that they can only identify known anomalies rather than discover new ones. Semi-supervised methods are trained using exclusively normal data, and try to identify anomalies in unseen data based on their difference with respect to the learned data distribution of the normal class. Unsupervised methods are different in that they make no assumptions about labels in training data and are simply data-driven, i.e., they fit the model based on all the available unlabelled data. Works in the literature [50] suggest that semi-supervised methods should be preferred if enough labeled normal data is available in order to achieve more robust models.

Due to the peculiarities of the learning settings, semi-supervised and unsupervised methods typically entail one-class learning models. Relevant examples of one-class learning anomaly detection methods are: *i)* Variational Autoencoder (VAE) [51], a neural-network reconstruction-based model with generative capabilities. The model is trained in a one-class manner by minimizing reconstruction error on training data, and the reconstruction error is used as an anomaly score; *ii)* One-Class Support Vector Machine (OCSVM) [52], which provides anomaly scores comparing new data with a hyperplane-based decision boundary learned during the training stage; *iii)* Local Outlier Factor (LOF) [53], which yields an anomaly score based on the ratio between the local density of new data samples with respect to the average local density of its nearest neighbors; *iv)* Isolation Forest (IF) [54], which provides ensembles of trees and considers the length of the path from root to leaf to determine the anomaly score of new samples: a shorter (or longer) path means that a data point is more (or less) likely to be an anomaly; *v)* Copula-based anomaly detection (COPOD) [55], which predicts the degree of ''extremeness'' of data samples based on tail probabilities of an empirical copula, a multivariate cumulative distribution function.

However, while these methods are well-established and perform well in a wide number of scenarios, they do not provide simultaneous knowledge retention and model adaptation, lacking lifelong capabilities. We can observe that recent research works started addressing lifelong anomaly detection. Examples include the adoption of meta-learning to estimate parameters for multiple tasks in one-class image classification [56], transfer learning in video anomaly detection [57], change-point detection coupled with memory organization [58], [59], and leveraging user feedback to improve model performance [60], [61].

Despite the clear advantages that lifelong learning could provide in anomaly detection methods, the number of published works is still rather limited. We attribute this scarcity to the novel and emerging nature of the subject and to the lack of established practices, protocols, and guidelines. Our study attempts to fill this gap by increasing the awareness of the potential of lifelong anomaly detection, while rationalizing and providing foundations with regard to scenarios, strategies, and metrics, which foster a simplified adoption of the lifelong learning framework for researchers and practitioners.

III. FROM ANOMALY DETECTION TO LIFELONG ANOMALY DETECTION

In this section, we start by analyzing the challenges arising in dynamic real-world scenarios and emphasizing the limitations of currently adopted anomaly detection approaches. Second, we discuss the advantages lifelong anomaly detection could bring to the anomaly detection landscapes.

A. WHEN IS NON-LIFELONG ANOMALY DETECTION NOT ENOUGH?

Offline anomaly detection has shown to be useful in different applications where it is possible to gather and process background data, such as post-incident analysis [50], breast cancer detection [10] and gravitational waves detection [11]. However, offline models are not sufficient for many real-world dynamic applications as they do not consider any change in the normal class.

Online anomaly detection methods partially address this limitation, providing learning systems with continuous updating capabilities. However, the underlying assumption is that only the most recent information is required to maintain satisfactory performance on the anomaly detection task. In this context, forgetting is a desirable property that allows the model to prevent obsolescence [62]. This behavior is considered to be sufficient to deal with many dynamic learning settings such as crowd anomaly detection [13]

and fatigue detection [12]. It is also possible to detect whether concept drift occurred by monitoring the statistical properties or the model's error rate, and updating anomaly detection models to reflect the most recent conditions of the environment [62]. However, methods coupled with concept drift detection also follow the assumption that only the most recent data is relevant for the anomaly detection task. As a result, these methods are prone to forgetting past knowledge, which is a shortcoming in domains with recurring tasks.

Many real-world domains may greatly benefit from the adoption of lifelong anomaly detection, as they are inherently characterized by dynamic and quickly evolving conditions, as well as recurring conditions. These challenges require model capabilities that foster simultaneous adaptation and knowledge retention. In the following, we describe three out of many possible real-world domains where such model capabilities are required.

First, monitoring human conditions to detect harmful states must be able to deal with many human activities, each presenting a unique definition of the normal class. In this setting, new life habits bring new activities that can be assimilated as tasks to be learned by the model (e.g., jogging, characterized by a high heart rate that should not be considered anomalous behavior). Forgetting activities carried out in the past is not acceptable in this setting and brings a number of practical disadvantages, which are systematically described in Section III-B.

Another example is the detection of intrusions in a cloud environment, which requires the ability to deal with a dynamic environment where multiple virtual servers (cloud instances) are added or removed over time. Such instances have different characteristics of normal behavior that depend on active services and user interactions. The system must be able to adjust and detect anomalies in traffic patterns in new cloud instances, but, at the same time, it should not decrease its performance when analyzing traffic from already monitored cloud instances.

Looking at a different domain, the identification of faults and malfunctioning in cyber-physical systems, such as water treatment plants or smart grids, is characterized by a very dynamic environment with multiple operating conditions and different uncontrollable inputs (e.g., geophysical factors in nature) that change over time. Moreover, components also age over time or are replaced with other components with different specifications. In this context, the model should be able to deal with tasks corresponding to different operating conditions.

All these domains are characterized by challenges such as a number of evolving emerging conditions that require prompt model adaptation, as well as recurring conditions that require the ability to preserve the knowledge of previously observed conditions. This duality creates the ideal conditions for the adoption of lifelong anomaly detection. However, to verify this assumption, it is important to assess whether lifelong scenarios impact the performance of non-lifelong anomaly detection models **(RQ1)**, and whether adopting knowledge retention capabilities actually results in an improvement for such models in lifelong scenarios **(RQ2)**.

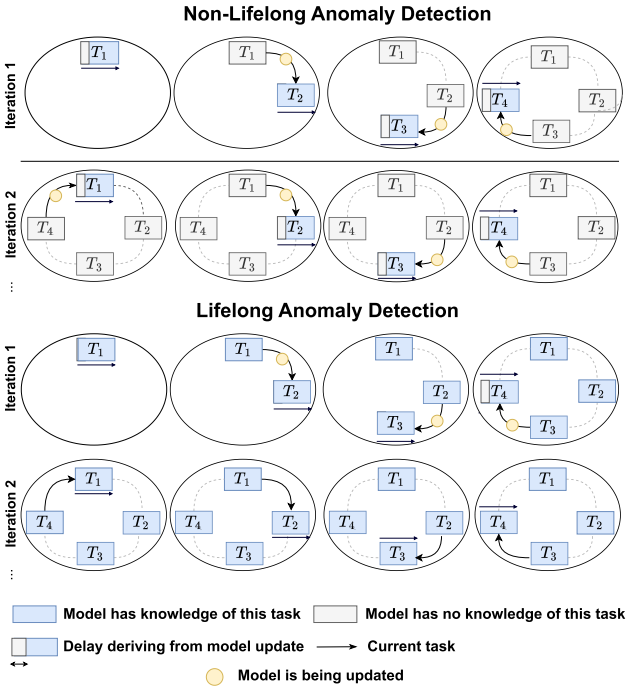

FIGURE 2. A scenario with four recurring tasks (T_1, T_2, T_3, T_4). Conventional anomaly detection requires constant model updates and results in detection delays. Lifelong learning mitigates this burden by retaining knowledge of tasks.

B. LIFELONG ANOMALY DETECTION TO THE RESCUE

Figure 2 shows a representative scenario that compares conventional anomaly detection with model updates to lifelong anomaly detection. In the second iteration, lifelong anomaly detection does not require model updates after a recurrence of each task. In contrast, conventional anomaly detection keeps updating the model, resulting in detection delays, i.e., false predictions, until the model has incorporated the new task. Moreover, a scenario with 100 iterations would require just 4 model updates for lifelong anomaly detection *vs.* 400 model updates for conventional anomaly detection, during which detection delays will occur. Many real-world scenarios with recurrence could be mapped to it, including sequences of human activities, geophysical phenomena such as weather patterns, and operating conditions of cyber-physical systems.

In Figure 3, we show a comparison between the conventional anomaly detection approaches and a counterpart for anomaly detection that entails lifelong learning. In this example, normal class data evolves over time in a task-sequential manner. As new tasks are presented, the model is updated – either in an online manner, possibly following concept drift detection, or in a lifelong learning manner. It can be observed that the non-lifelong anomaly detection with a model update approach entails forgetting as a way

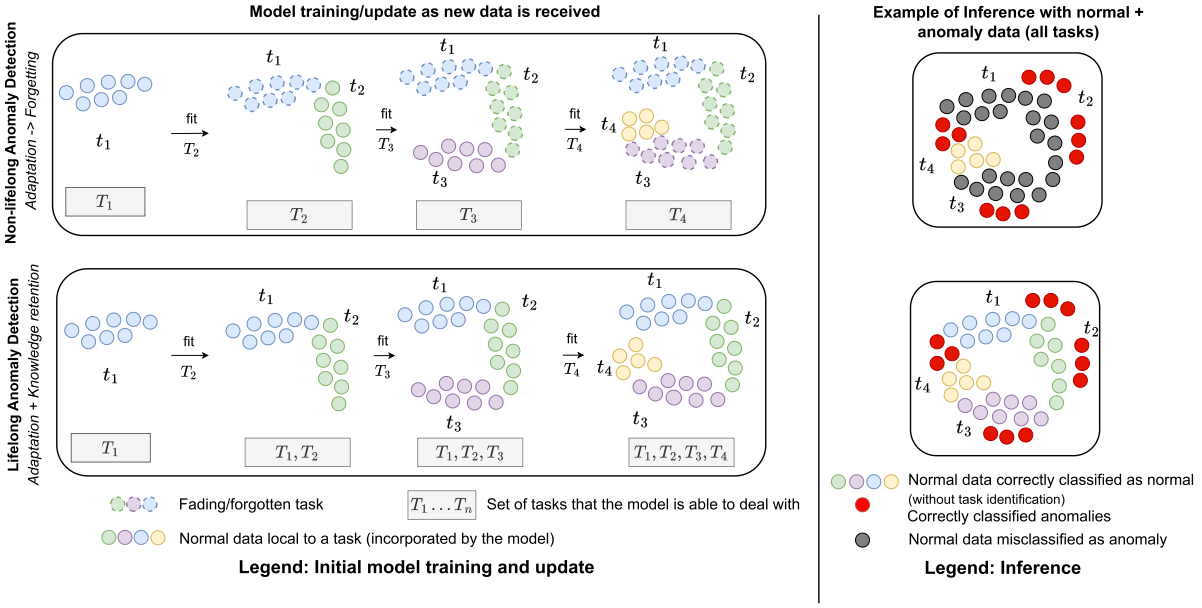

FIGURE 3. **Comparison of training/update and inference for non-lifelong and lifelong anomaly detection in the scenario with four tasks (T_1, T_2, T_3, T_4). In non-lifelong anomaly detection, the model forgets the previous tasks as soon as a new task is learned (left – top). In contrast, the lifelong anomaly detection model aims to retain knowledge of all tasks (left – bottom). This characteristic has a serious impact on the model's behavior during inference (right). In non-lifelong anomaly detection, after learning task T_4, the model misclassifies data from previous tasks as anomalous since it considers only data from the current task as normal behavior (right – top). On the other hand, the ideal lifelong anomaly detection model retains the knowledge of all tasks, preventing the misclassification of normal data from previous tasks as anomalous (right – bottom). This difference in behavior between non-lifelong and lifelong anomaly detection may lead to a discrepancy in their performance scores. .**

to focus on the most recent data, leading to mistakenly classifying normal data from previous tasks as anomalous. On the other hand, the lifelong learning approach for model update aims at adapting the model to incorporate new tasks without forgetting previous tasks. The advantage is to exploit knowledge from the combination of the different tasks to provide a more comprehensive and robust anomaly detection model, which correctly identifies normal data from all tasks as normal behavior.

Overall, even though the conventional anomaly detection with model update approach would be, in principle, able to re-learn previously forgotten tasks as they reoccur, this has several drawbacks, as indicated above. From a practical viewpoint, frequent model updates require additional computational resources in the presence of highly recurring tasks. Moreover, delays in the responsiveness of the model caused by the time difference between the appearance of a recurring task and updating the model once concept drift is detected can be costly in terms of false or missed detections. We aim to verify this assumption throughout an experimental analysis involving non-lifelong anomaly detection models exposed to lifelong learning scenarios **(RQ1)**.

By analyzing the challenges pertaining to the lifelong anomaly detection learning setting and by generalizing the examples presented in Figure 2 and Figure 3, we identify the following drawbacks of conventional anomaly detection with constant model updates:

- By forgetting past knowledge and only adapting to the new normal class distribution, the system may trigger a large number of false positives when recurring patterns are presented, leading to a dramatic decrease in performance until a retraining phase is undertaken;
- Inability for the models to leverage skills learned in the past in combination with recent skills to solve tasks in a more compelling way;
- Experimental settings that are too simplistic to reflect the complexity of many real-world challenges, which usually involve the appearance of new conditions and the recurrence of old conditions, requiring a more comprehensive evaluation across all tasks;
- A consistent use of computational resources for data processing and model training to deal with recurring tasks;
- From a theoretical viewpoint, the model will not provide a comprehensive view of the environment without the possibility to leverage task similarity and knowledge transfer across a combination of tasks to solve every single task.

These drawbacks make current anomaly detection models rather simplistic in comparison to sophisticated human-level intelligence when faced with complexities and challenges brought by lifelong anomaly detection scenarios **(RQ1)**. Intuitively, humans are exposed to different aspects of reality, building up skills incrementally throughout their lifespan and improving their general knowledge base. Introducing similar

capabilities in models should allow them to provide a more sophisticated behavior that translates into more reliable and accurate predictions [63] (**RQ2**).

In practical terms, by leveraging discoveries in biology, neuroscience, and computer science, lifelong learning enables the possibility of incorporating comprehensive knowledge in models. Examples include the adoption of task similarity, curriculum learning, yielding positive forward and backward transfer across different tasks, as already demonstrated in image classification [64], object detection [14] and reinforcement learning problems [23].

Overall, we argue that the adoption of lifelong learning in anomaly detection would yield the ability to consider the combination of all these aspects, providing more complex learning strategies that lead to more informed decisions. Instead of constantly forgetting and learning each individual condition, an ideal model could leverage past knowledge to retain performance across all conditions. We aim to verify this assumption by designing experiments that uncover whether the adoption of lifelong learning knowledge retention strategies can be beneficial for non-lifelong anomaly detection models (**RQ2**).

In summary, the minimal set of advantages for the adoption of a lifelong anomaly detection approach includes capabilities such as: *i)* simultaneous adaptation and knowledge preservation; *ii)* inference that exploits a more comprehensive knowledge of the domain or environment at hand; *iii)* resource-savvy model updates compared to conventional anomaly detection methods with constant model updates; *iv)* more realistic experimental settings and evaluation schemes that consider all tasks in combination.

IV. LIFELONG ANOMALY DETECTION: SCENARIOS AND EVALUATION PROTOCOLS

Given the novelty of lifelong anomaly detection, as shown by the limited availability of research works on the subject, it is important to devise procedures and guidelines to standardize its adoption. This section aims to provide new perspectives on how to address lifelong learning challenges in anomaly detection. To this end, we devise a categorization of lifelong learning scenarios, provide a scenario creation algorithm, and evaluation protocols that can guide researchers interested in the problem.

A. LIFELONG LEARNING SCENARIOS: AN ANOMALY DETECTION PERSPECTIVE

Current lifelong learning approaches are focused on classification tasks, where *tasks* are defined as sets of classes (e.g., in the MNIST dataset, any combination of two classes among its 10 classes), and the learning workflow encompasses a sequence of n tasks $T = t_1, t_2, \ldots, t_n$ where the model is challenged to learn new tasks without forgetting previous tasks.

Another important element of comparison is the type of *learning scenarios*. We recall that lifelong image classification usually describes three types of scenarios: task-incremental, class-incremental, and domain-incremental [31]. These scenarios differ based on the availability of task labels and task boundaries. For all scenarios except domain-incremental, in the simple classification example mentioned above (MNIST), task labels identify a specific subset of 10 classes currently being presented to the model, whereas task boundaries identify the beginning/end of such task. On the other hand, in a domain-incremental scenario, tasks represent a new distribution of already known classes, where task labels identify specific distributions and task boundaries indicate the moment when distribution changes. The availability of the task labels and boundaries depends on the degree of available domain knowledge [65]. A more detailed description of learning scenarios in lifelong classification may be found in [31]. It is worth noting that emerging scenarios are being proposed to account for the limitations of previously existing ones. One example is the consideration of the temporal dimension in online lifelong learning scenarios [66].

The notion of a task in anomaly detection clearly differs from the conventionally adopted definition in image classification since we usually deal with two classes: normal and anomaly. Tasks in this context represent various aspects of the normal class, which is expected to evolve over time. Moreover, the normal class may also change its role depending on the context, i.e., what is normal in one context can be anomalous in another, further increasing the complexity of the problem. To differentiate lifelong anomaly detection setting, we define a self-consistent behavior[3] of the normal class, alongside the specific anomalies occurring with it, as a *concept*. For instance, in monitoring human conditions to detect harmful states, the entire normal class can be thought of as a set of concepts: resting, jogging, and eating, all presenting different characteristics. In lifelong anomaly detection, multiple consistent behaviors of the normal class are presented over time instead of new classes as in class and task-incremental scenarios, and we are focused on the evolution of a single normal class instead of the evolution of all classes as in domain-incremental scenarios. For example, a high heart rate can be considered an anomaly in resting conditions but is expected during jogging, and therefore it does not represent an anomaly in this context. Therefore, to deal with the inadequacy of lifelong image classification scenarios in the context of anomaly detection, we define distinctive scenarios for this setting. Following the example of anomaly detection in human conditions, *concept identifiers* define a consistent behavior of the normal class (a specific activity), whereas *concept boundaries* represent explicit information on whether the currently analyzed concept (a specific activity) has changed. Concept identifiers and concept boundaries may correspond to task labels and task boundaries, respectively, in lifelong image classification.

[3] A behavior could correspond to a new distribution, change of a performed activity, or a new state of the environment, depending on the specific analytical context considered.

Based on this consideration, we identify the following learning scenarios in increasing order of complexity:

- *Concept-aware*: Known concept identifier and concept boundaries.
- *Concept-incremental*: Unknown concept identifier but known concept boundaries.
- *Concept-agnostic*: Unknown concept identifier and concept boundaries

In reference to our example of anomaly detection in human conditions, a concept-aware scenario implies that the model is aware of the currently processed activity and its lifespan (at both training and inference time). On the other hand, a concept-incremental scenario only provides an indication that a change of activity has occurred without any identifying information about the specific activities. Finally, a concept-agnostic scenario is the most challenging, as it does not provide any supporting information about the current activity being performed and its lifespan. These notions are general and can be adopted in any domain.

B. SCENARIOS DESIGN

In the following, we propose a scenario design procedure that applies to most datasets and enables researchers and practitioners to transition their current scenarios and evaluation setup toward lifelong anomaly detection.

Algorithm 1 presents a general pseudo-code for scenario design. It requires users to define a few parameters, which determine the creation of diverse scenarios: the number of desired concepts **c**, Normal (**N**), and Anomaly (**A**) data from a given dataset, and three functions: ϕ, γ, and λ. The algorithm leverages concept creation functions ϕ and γ to create normal and anomaly concepts based on normal and anomaly data, respectively. Their goal is to transform the original dataset into self-consistent sets of data points having common characteristics. An example implementation for ϕ and γ is a clustering algorithm of choice, based on user preferences. The assignment function λ matches each normal concept with an anomaly concept, leading to a combined concept containing one normal and one anomaly concept, which allows for model training and evaluation. The sequence of these combined concepts defines the complete scenario. An example implementation for λ is mapping an anomaly cluster to its closest normal cluster.

Focusing on Algorithm 1, first, we create concepts for the normal class through a concept creation function ϕ (Line 1). The concept creation function can leverage any aspect or feature value that allows us to delineate the boundaries of one concept. Second, we create concepts for the anomaly class through anomalous concepts creation function γ (Line 2). Third, for each normal concept C_{N_i}, we select a corresponding anomaly concept C_{A_j} using a function λ (Line 3-5). The combination of C_{N_i} and C_{A_j} is a concept added to the lifelong scenario (Line 6). Each time a concept is built, the selected anomaly concept C_{A_j} is removed from the set of available anomaly concepts C_A so that it appears only once throughout the scenario (Line 7). The algorithm returns the resulting scenario as a sequence of concepts (Line 9), each of which may need to be separated into training and evaluation data depending on the learning settings, e.g., unsupervised or semi-supervised.

The proposed method allows us to create diverse scenarios depending on the user selection of ϕ, γ, and λ. For example, it is possible to leverage k-Means to cluster normal and anomaly concepts (leveraging ϕ, and γ functions) and assign each anomaly concept to the closest normal concept (leveraging λ function). Alternative scenarios can be designed by customizing ϕ, γ, and λ, paving the way for a wide range of possible scenarios. To simplify this task, our code is publicly available and can be used to easily generate scenarios for any dataset chosen

Algorithm 1 Scenario Design Protocol

Input: c – Number of desired concepts
Input: **N**, **A** – Normal/Anomaly data
Input: ϕ – Concepts creation function for normal data
Input: γ – Concepts creation function for anomalies
Input: λ – Assignment function
1 $C_N \leftarrow \phi(N, c)$ // Create concepts $\{C_{N_0}, C_{N_1}, \ldots, C_{N_c}\}$
2 $C_A \leftarrow \gamma(A, c)$ // Create concepts $\{C_{A_0}, C_{A_1}, \ldots, C_{A_c}\}$
3 $T \leftarrow \emptyset$ // Result scenario
4 **for** $C_{N_i} \in C_N$ **do**
5 $\quad j \leftarrow \lambda(C_A, C_{N_i})$ // Match anomaly-normal concepts
6 $\quad T \leftarrow T \cup (C_{N_i}, C_{A_j})$ // Add concepts to scenario
7 $\quad C_A \leftarrow C_A - C_{A_j}$ // Remove used anomaly concept
8 **end**
9 **return** T

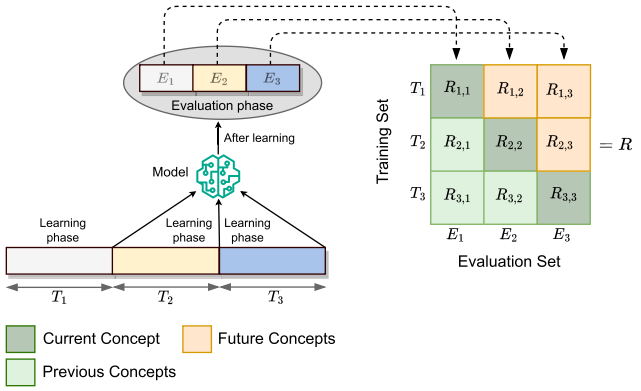

FIGURE 4. Lifelong evaluation protocol. The model handles a sequence of concepts $i = 1, 2, 3$. For each concept i, the model is trained on training set T_i (learning phase). After each learning phase, the evaluation phase is triggered, where the model anomaly detection performance (in terms of ROC-AUC) is evaluated on all testing sets E_j from all concepts (previous, current, and future). The evaluation protocol creates a matrix R, in which the entry $R_{i,j}$ represents model performance in terms of ROC-AUC on concept j after learning concept i. This matrix is used to compute final metric values, such as Lifelong ROC-AUC, BWT, and FWT.

Algorithm 2 Pseudo-Code of the Evaluation Protocol

Input: **T** – Sequence of N training sets
Input: **E** – Sequence of N testing sets
Input: L – Learning model
Input: ρ – Evaluation function computing ROC–AUC
1 $R_{N\times N} = \{\}$ // Initialize results matrix NxN
2 **for** $T_i \in$ **T do**
3 $L \leftarrow$ update L with T_i // Train/update model
4 **for** $E_j \in E$ **do**
5 $R_{i,j} \leftarrow \rho(L, E_j)$ // Evaluate L on E_j
6 **end**
7 **end**
8 **return** R

by users: https://github.com/lifelonglab/lifelong-anomaly-detection-scenarios. We note that different choices of ϕ and γ may lead to concept imbalance, i.e., some concepts may present significantly fewer samples than others. This situation may be problematic in some settings, or exacerbate the learning complexity for some anomaly detection models. In these cases, consideration of strategies for imbalanced learning such as resampling [67], cost-sensitive learning, and special-purpose algorithms [68]. It is worth noting that, in our experiments, we did not experience high imbalance ratios for concepts generated with our protocol.

C. MODEL EVALUATION

Lifelong learning scenarios require a continuous evaluation across all concepts. To realize this goal, we adopt a lifelong learning evaluation protocol that considers the performance of all concepts across a learning scenario.

Algorithm 2 provides a general overview applicable to a vast number of use cases. Without loss of generality, the protocol can be modified to accommodate specific requirements, e.g., recurring concepts, time-based concepts, unsupervised learning, etc. Moreover, our protocol can support any base model of choice, and any data preprocessing step, such as data augmentation and missing data treatment, to deal with specific data challenges.

First, the evaluation protocol initializes a matrix R to accommodate anomaly detection results for specific tasks (Line 1). Second, the protocol iterates over training sets for all concepts (Line 2). For each concept, the model is trained/updated (Line 3) and evaluated on all testing sets for all concepts (Lines 4-5), i.e., previous, current, and future concepts. Our protocol yields a matrix R, where entries $R_{i,j}$ define the ROC-AUC metric of the model evaluated on concept j after learning concept i. A graphical representation of this protocol is shown in Figure 4.

The matrix R can be used to directly compute lifelong learning metrics, such as backward and forward transfer.

These metrics allow us to assess model behavior more extensively than standard performance metrics by taking into account the model's performance on different concepts (previous, current, and future).

Inspired by [69], we propose a **Lifelong ROC-AUC** – a lifelong variant of ROC-AUC that can adequately assess models' performance on all concepts after learning every new concept, instead of models' performance on just a single concept. It is defined as:

$$\text{Lifelong ROC-AUC} = \frac{\sum_{i\geq j}^{N} R_{i,j}}{\frac{N(N+1)}{2}} \tag{1}$$

The metric is computed considering previously learned concepts, including the current concept, which corresponds to averaging over $\frac{N(N+1)}{2}$ entries from lower triangular. We favor ROC-AUC over threshold-dependent metrics such as Precision, Recall, and F–Score, since it allows us to evaluate the model's performance more comprehensively. ROC–AUC may be swapped with other metrics of choice without impacting the validity of the protocol.

Backward Transfer for ROC–AUC (BWT) measures the impact of learning new concepts on the performance of all previously learned concepts. Negative backward transfer suggests that the model is prone to forgetting. A strongly negative value is also sometimes regarded as catastrophic forgetting. On the other hand, positive backward transfer suggests that learning new concepts benefits models' performance on previously learned concepts. Backward transfer is computed over all concepts as:

$$BWT = \frac{\sum_{i=2}^{N}\sum_{j=1}^{i-1} R_{i,j} - R_{j,j}}{\frac{N(N-1)}{2}} \tag{2}$$

The impact of learning each concept on the model's performance on future concepts is measured by **Forward Transfer for ROC–AUC** (FWT). Forward transfer can also be thought of as the zero-shot model performance on future concepts since it assesses model performance on unseen concepts. It partially depends on concept similarity (task similarity) and the model's knowledge transfer ability. It is computed as:

$$FWT = \frac{\sum_{i<j}^{N} R_{i,j}}{\frac{N(N-1)}{2}} \tag{3}$$

It is noteworthy that the protocol slightly differs based on the learning setting. Specifically, in concept-aware and concept-incremental scenarios, batches T_i (training) and E_i (evaluation) correspond to the single $i-$th concept. As for concept-agnostic settings, a batch does not necessarily correspond to a single concept since the setting assumes that no explicit concept boundaries are provided to the lifelong algorithm. As a result, the evaluation may require considering multiple batches as belonging to the same concept or a single batch including data for more than one concept.

V. EXPERIMENTS

Our experiments are directed at answering two main research questions: *i)* Do lifelong scenarios impact the performance of

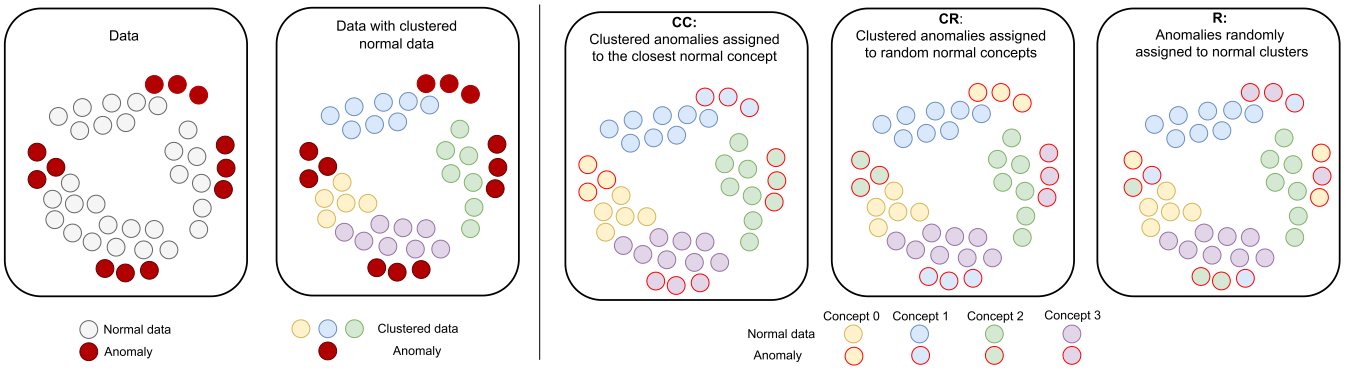

FIGURE 5. **Lifelong scenarios variants based on different choices of concept creation functions γ and λ: *i)* clustered anomaly concepts assigned to the closest normal concept (CC), *ii)* clustered anomaly concepts assigned randomly to normal concepts (CR), and *iii)* anomalies randomly assigned to normal concepts (R).**

non-lifelong anomaly detection models?; ii) Does adopting lifelong learning provide a valuable improvement in the learning capabilities of anomaly detection models in lifelong scenarios? We empirically address these two questions in the following two subsections and provide insights on anomaly detection performance, forgetting, and task similarity. To answer the above question, we design three lifelong scenario variants based on different choices:

- **CC**: clustered anomaly concepts assigned to the closest normal concept, where γ and ϕ leverage a clustering function, and λ maps a given normal concept to the closest anomaly concept.
- **CR**: clustered anomaly concepts assigned randomly to normal concepts, where γ and ϕ leverage a clustering function, and λ assigns a random anomaly concept to a given normal concept.
- **R**: anomalies randomly assigned to normal concepts, where γ leverage a clustering function, ϕ creates anomaly concepts using random sampling, and λ assigns a random anomaly concept to a given normal concept.

For all scenario variants, we use k-Means as the clustering function. These scenarios are conceptually represented as two-dimensional plots in Figure 5. In our experiments, we extract between 5 and 20 concepts depending on the complexity and the size of each dataset.

In our experiments, we employ a diverse set of datasets encompassing cybersecurity and smart grids.

- **NSL-KDD** [70]: Widely adopted dataset containing records of network traffic gathered during the DARPA Intrusion Detection Systems evaluation program;
- **UNSW-NB15** [71]: The dataset containing records of network traffic describing hybrid real modern normal and contemporary synthesized attack activities;
- **Energy** [72]: Sensor-based anomaly detection in photovoltaic/solar power plants. The data was collected from power plants located in Italy (17 plants, 2.5 years);
- **Wind** [4]: Wind power production dataset, containing anomalous patterns occurred in eolic/wind parks (Wind). The data was modeled using the Weather Research & Forecasting (WRF) model (5 plants, 2 years).

All datasets present fairly different feature representations and represent data collected by heterogeneous systems. They also present different types of anomalies and complexity levels.

We use five popular anomaly detection models described in Section II: One-Class Support Vector Machines (OC-SVM), Local Outlier Factor (LOF), Isolation Forests (IF), Variational Auto-Encoders (VAE), and Copula-based Outlier Detection (COPOD).

We leverage the following learning strategies:

- **Naive lifelong**: models are updated as new data becomes available, without any smart lifelong learning strategy to tune adaptation and knowledge retention. By updating the model only based on the new data, a reasonable expectation is that the model will gradually or catastrophically forget knowledge of previously presented data. It can be considered as a lower-bound non-lifelong baseline learning strategy.
- **Multiple Single-Task Experts (MSTE)**: a way to simulate upper-bound model performance in a non-lifelong scenario. In this strategy, a pool or ensemble of models, each of which is an expert for a single concept, is adopted. Whenever a new concept is presented, a new model is trained on the new data and added to the pool. We note that this is an unrealistic setting since, in real-world scenarios, it requires extremely high computational resources to deal with a potentially infinite number of models, and the availability of concept identifiers, which are not available for concept-incremental and concept-agnostic scenarios.
- **Replay**: a replay-based method that preserves selected data samples from previous concepts in a memory buffer, which is limited in size by a parameter known as a budget. When the model faces a new task (concept), the replay buffer is updated to include the data from the new concept. As a result, the replay buffer contains knowledge of all concepts presented so far. The replay

buffer is then used while updating the model to mitigate forgetting [46]. The expectation is that, by providing a summarized representation of all concepts, the model should, in principle, be able to preserve a satisfactory performance on all concepts, without a significant degree of forgetting for any of the concepts. In our experiments, we use a simple balanced replay buffer with a very constrained budget of 3,000 data samples.

To this end, we adopt the learning evaluation workflow in Algorithm 2, which provides the model with the challenge of adapting to new data while retaining knowledge of previously observed data. Technically, we create a matrix where the performance of the model is measured on all concepts after learning each single concept. Metric computation takes place according to Equations 1-3.

We address **RQ1** by devising experiments to assess whether non-lifelong anomaly detection methods are impacted by the challenges brought by lifelong scenarios. To this aim, leverage the two mentioned strategies (Naive lifelong and MSTE) to showcase if there is a gap between models' performance in non-lifelong scenarios vs. lifelong scenarios, which would suggest the need for adopting lifelong learning strategies. Particularly, we are interested in observing whether MSTE achieves higher performance values in terms of ROC-AUC, BWT, and FWT compared to Naive. This expectation is motivated by the observation that Naive only adapts to new data without any knowledge retention mechanism, while MSTE does not experience forgetting due to the creation of a single expert model for each task.

We address **RQ2** by devising experiments to assess whether the adoption of lifelong learning strategies has the potential to increase the performance of non-lifelong anomaly detection models in lifelong scenarios. To this aim, we analyze the impact of the adoption of a lifelong Replay strategy on the performance of all base models to show that the adoption of lifelong learning strategies may be beneficial to improve model performance in a lifelong anomaly detection scenario. Experimental results[4] are shown in Tables 1, 2.

VI. RESULTS DISCUSSION

In this section, we discuss results alongside two main perspectives: the impact of lifelong scenarios on non-lifelong anomaly detection problems, and the impact of the adoption of lifelong learning strategies providing knowledge retention capabilities.

A. IMPACT OF LIFELONG SCENARIOS ON NON-LIFELONG ANOMALY DETECTION

In this subsection, we focus on providing insights on how non-lifelong anomaly detection methods are impacted by the challenges brought by lifelong scenarios (**RQ1**).

[4]We note that our results are obtained by averaging multiple executions with different hyperparameter values (see Appendix) to showcase the reliability of the results, similarly to a cross-validation evaluation [73], [74].

We present the experimental results in Table 1. The general trend that can be observed across all methods and datasets is that there is a gap between the anomaly detection performance in terms of ROC-AUC achieved by widely adopted anomaly detection methods and the hypothetical upper-bound defined by multiple single-task experts (MSTE). Notably, with the Energy dataset, base models with a Naive learning strategy are significantly outperformed by the MSTE approach (for example, for Isolation Forest, CC: 0.64 vs. 0.88 — CR: 0.65 vs. 0.97 — R: 0.59 vs. 0.97). This is not an isolated case but applies to the other datasets as well.

It is noteworthy that MSTE achieves very high performance in most cases, highlighting that the scenario generation procedure decomposes each dataset's complexities into sub-complexities (concepts), which are much more manageable to learn in isolation but much more challenging when provided as a lifelong scenario. This phenomenon has been observed in [65] in the context of non-lifelong one-class classification. Overall, it looks clear that non-lifelong anomaly detection methods are penalized in lifelong scenarios, leading to sub-optimal performance scores, as evident by the low ROC-AUC values. We present a graphical illustration of this phenomenon in Figure 8, which shows a clear performance gap between non-lifelong and lifelong strategies.

Another lifelong metric worth analyzing is the backward transfer (BWT) as it allows us to analyze how learning new tasks affects model performance on previous tasks. We recall that negative values of BWT indicate that learning new tasks introduces forgetting in previously learned tasks, whereas positive values are indicative of effective knowledge transfer capabilities across tasks (see Section IV-C). We observe that all base models with Naive strategy present negative values of BWT, e.g., with Wind (R), models showcase values from -0.11 to -0.52, which shows that they are affected by a degree of forgetting (from mild to catastrophic). This phenomenon is more clearly visible in Figure 6 (VAE) and 7 (LOF), where the performance on C_0 drops from 1 to 0.69 (VAE) and from 0.99 to 0.41 (LOF) after the last concept C_4 is learned. Negative backward transfer is also observed for C_2, where model performance gradually drops from 0.99 to 0.5 (VAE) and from 0.99 to 0.34 (LOF) after learning C_3. Subsequently, learning C_4 leads to increased forgetting of C_2, and the performance on C_2 drops to 0.078 (VAE) and 0.098 (LOF). Another exciting aspect that can be observed is a positive backward transfer, which indicates that learning a new concept improves the performance of the model on a previous concept. This is emphasized by the performance on C_0 before and after learning C_2. Results show that learning C_2 increases the performance on C_0 from 0.51 to 0.96 (VAE) and from 0.42 to 0.55 (LOF). It means that the model can leverage the knowledge acquired while learning C_2 to present better anomaly detection capabilities on concept C_0 by leveraging similarity between concepts (task-similarity). By analyzing BWT, we uncover concept similarity relationships and measure models' ability to retain and reuse

TABLE 1. Experimental results for Naive lifelong strategy with all methods (IF, LOF, COPOD, OC-SVM, and VAE) and datasets (Energy, NSL-KDD, UNSW, Wind) in the three concept-incremental scenarios (CC, CR, R), according to ROC-AUC, BWT, FWT metrics. For ROC-AUC, we also report results obtained with multiple single-task experts (MSTE) in parenthesis (upper bound).

	IF			LOF			COPOD			OC-SVM			VAE		
Dataset	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT
Energy (CC)	0.64 (0.88)	-0.29	0.61	0.71 (0.96)	-0.31	0.81	0.85 (0.91)	-0.07	0.90	0.76 (0.94)	-0.22	0.84	0.75 (0.92)	-0.21	0.81
Energy (CR)	0.65 (0.97)	-0.40	0.53	0.70 (0.99)	-0.35	0.73	0.91 (0.98)	-0.09	0.86	0.75 (0.99)	-0.29	0.71	0.74 (0.99)	-0.31	0.68
Energy (R)	0.59 (0.97)	-0.46	0.56	0.66 (1.00)	-0.41	0.70	0.89 (0.98)	-0.11	0.87	0.69 (0.99)	-0.37	0.70	0.68 (0.99)	-0.38	0.69
NSL-KDD (CC)	0.68 (0.97)	-0.32	0.77	0.62 (0.93)	-0.33	0.56	0.54 (0.66)	-0.14	0.43	0.67 (0.99)	-0.36	0.85	0.67 (0.98)	-0.35	0.85
NSL-KDD (CR)	0.70 (0.96)	-0.28	0.72	0.60 (0.94)	-0.37	0.54	0.52 (0.62)	-0.13	0.44	0.75 (0.96)	-0.23	0.73	0.73 (0.96)	-0.25	0.73
NSL-KDD (R)	0.85 (0.99)	-0.16	0.82	0.63 (0.95)	-0.35	0.52	0.74 (0.81)	-0.08	0.67	0.88 (1.00)	-0.13	0.85	0.87 (1.00)	-0.14	0.86
UNSW (CC)	0.49 (0.73)	-0.29	0.45	0.57 (0.88)	-0.38	0.45	0.32 (0.43)	-0.11	0.29	0.59 (0.78)	-0.22	0.53	0.55 (0.81)	-0.32	0.50
UNSW (CR)	0.60 (0.94)	-0.41	0.50	0.67 (0.98)	-0.38	0.49	0.48 (0.67)	-0.24	0.24	0.73 (0.97)	-0.31	0.56	0.72 (0.98)	-0.32	0.51
UNSW (R)	0.53 (0.90)	-0.45	0.45	0.60 (0.96)	-0.42	0.46	0.60 (0.76)	-0.18	0.44	0.55 (0.91)	-0.44	0.46	0.56 (0.94)	-0.46	0.44
Wind (CC)	0.83 (0.90)	-0.10	0.74	0.65 (0.98)	-0.49	0.58	0.89 (0.93)	-0.06	0.89	0.77 (0.96)	-0.28	0.77	0.76 (0.96)	-0.30	0.76
Wind (CR)	0.74 (0.95)	-0.32	0.71	0.62 (0.99)	-0.57	0.53	0.89 (0.96)	-0.12	0.90	0.72 (1.00)	-0.42	0.70	0.68 (0.98)	-0.46	0.73
Wind (R)	0.74 (0.95)	-0.31	0.69	0.65 (0.99)	-0.52	0.50	0.90 (0.97)	-0.11	0.91	0.74 (0.99)	-0.38	0.66	0.71 (0.99)	-0.42	0.64

TABLE 2. Experimental results for Replay strategy with all methods (IF, LOF, COPOD, OC-SVM, and VAE) and datasets (Energy, NSL-KDD, UNSW, Wind) in the three concept-incremental scenarios (CC, CR, R), according to ROC-AUC, BWT, FWT metrics.

	IF			LOF			COPOD			OC-SVM			VAE		
	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT	ROC-AUC	BWT	FWT
Energy (CC)	0.82	-0.06	0.64	0.95	-0.01	0.78	0.89	-0.01	0.89	0.83	-0.04	0.85	0.77	-0.16	0.83
Energy (CR)	0.78	-0.13	0.51	0.96	-0.01	0.64	0.89	-0.04	0.83	0.85	-0.06	0.77	0.81	-0.15	0.66
Energy (R)	0.75	-0.18	0.58	0.97	-0.02	0.61	0.87	-0.06	0.83	0.84	-0.07	0.76	0.75	-0.25	0.68
NSL-KDD (CC)	0.85	-0.11	0.92	0.89	-0.03	0.33	0.79	0.00	0.78	0.71	-0.21	0.89	0.82	-0.09	0.93
NSL-KDD (CR)	0.81	-0.07	0.82	0.88	-0.02	0.39	0.72	-0.02	0.62	0.79	-0.11	0.74	0.84	-0.04	0.76
NSL-KDD (R)	0.95	-0.03	0.90	0.90	-0.01	0.34	0.92	0.04	0.84	0.93	-0.03	0.86	0.93	-0.01	0.85
UNSW (CC)	0.51	-0.03	0.35	0.83	-0.00	0.32	0.43	0.00	0.29	0.56	-0.02	0.54	0.65	-0.12	0.38
UNSW (CR)	0.68	-0.02	0.45	0.89	-0.01	0.39	0.67	0.00	0.30	0.82	0.01	0.57	0.73	-0.05	0.33
UNSW (R)	0.48	-0.08	0.35	0.82	-0.05	0.40	0.52	-0.05	0.39	0.50	-0.08	0.46	0.66	-0.08	0.39
Wind (CC)	0.91	-0.01	0.74	0.97	-0.01	0.60	0.92	-0.01	0.90	0.91	-0.04	0.83	0.86	-0.13	0.72
Wind (CR)	0.88	-0.08	0.70	0.97	-0.03	0.57	0.89	-0.06	0.89	0.88	-0.08	0.75	0.84	-0.18	0.74
Wind (R)	0.90	-0.06	0.67	0.98	-0.02	0.52	0.91	-0.06	0.89	0.91	-0.06	0.72	0.86	-0.17	0.72

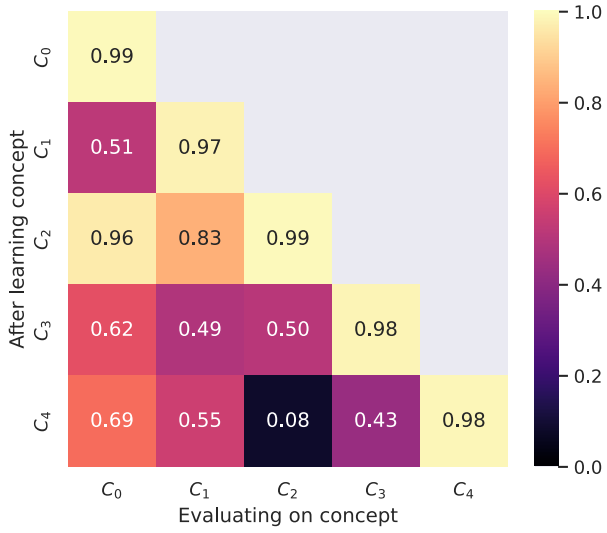

FIGURE 6. Concept-level ROC-AUC performance for the lifelong learning scenario. Each row $i = 0, 1, \ldots$ represents the performance on all concepts observed so far after learning the concept C_i (WIND (R) dataset; Naive lifelong strategy; VAE base model).

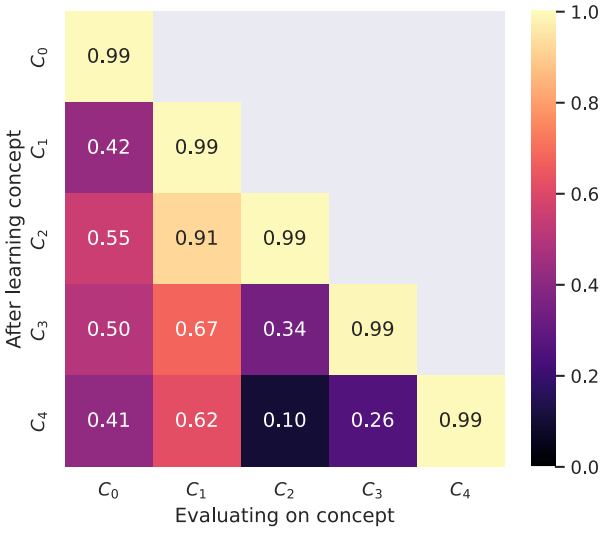

FIGURE 7. Concept-level ROC-AUC performance for the lifelong learning scenario. Each row $i = 0, 1, \ldots$ represents the performance on all concepts observed so far after learning the concept C_i (WIND (R) dataset; Naive lifelong strategy; LOF base model).

knowledge for improving its overall performance across all concepts.

Finally, moving our focus to Forward Transfer (*FWT*), we can observe values from 0.24 – UNSW (CR) with

COPOD to 0.91 – Wind (R) with COPOD. This result suggests a moderate-to-high concept similarity that could be leveraged by models. Interestingly, these values are higher than those commonly seen in image classification since the one-class learning setting is inherently different from multi-class classification. Specifically, since we compute FWT using ROC-AUC as a base metric, the random reference value is 0.5, whereas, in image classification, it is the ratio between 1 and the number of classes in the single task. This difference exacerbates the complexity of a comparison of FWT in these two settings. However, we argue that interpreting *FWT* in this context requires additional in-depth research focused on leveraging concepts similarity for one-class anomaly detection.

In summary, we observed a gap in ROC-AUC performance (Naive lifelong *vs*. MSTE) alongside with the presence of forgetting (emphasized by negative values of BWT) and degrees of concept similarity (observed via BWT and *FWT*). These phenomena show that lifelong learning scenarios generated with our approach allow us to adapt non-lifelong datasets to lifelong anomaly detection scenarios, introducing lifelong challenges which yield more demanding conditions for models (**RQ1**).

Additionally, our results show that tackling anomaly detection problems from a lifelong learning perspective can enable a more comprehensive evaluation of models, including measuring the impact of forgetting and task transferability on models. Concept-level granularity in results also allows us to uncover relationships between concepts in the learning scenario, as well as highlight specific concepts for which models are challenged more than others, enabling a more comprehensive evaluation and in-depth model analysis.

B. IMPACT OF THE ADOPTION OF A REPLAY-BASED LIFELONG LEARNING STRATEGY

In this subsection, we analyze the impact of the adoption of a lifelong knowledge retention strategy (Replay) on the performance of anomaly detection models. Our effort is aimed at verifying whether the adoption of lifelong learning strategies may be beneficial to improve model performance in a lifelong anomaly detection scenario (**RQ2**).

In order to assess the impact of introducing Replay strategy, we compare anomaly detection performance (ROC-AUC) and backward transfer (BWT) for Replay (Table 2) and the Naive lifelong strategy (Table 1). We can observe that the Replay strategy brings various levels of improvement, as shown by higher values of ROC-AUC and BWT. In some cases, the improvement margin is particularly high. For example, in Energy (CC), the Replay strategy can improve ROC-AUC by 0.18 (IF). There are also many cases in which the improvement margin is moderate. For example, in Wind (CC), Replay yields a ROC-AUC improvement of 0.08 (IF). Finally, the improvement margin is quite limited in cases such as UNSW (CC), in which Replay improves the performance of Naive by just 0.02 (IF). Similar patterns can be observed across all base models (IF, LOF, COPOD, OC-SVM, VAE). A summarized visual perspective of these results is shown in Figure 8, which emphasizes the differences between Naive, Replay, and MSTE.

Results in Table 2 show that in the majority of cases (54 out of 60), the simple Replay strategy achieves better results in terms of ROC-AUC than the Naive lifelong approach. Most of the exceptions regard the UNSW (R) scenario. We attribute this result to the concept complexity in UNSW and the simplicity of the Replay strategy. In this scenario, it is evident that more complex lifelong strategies are required to outperform the Naive baseline.

The improvement in performance is also supported by the observation of a decrease in forgetting, as shown by improvements in Backward Transfer (BWT) results when comparing Naive lifelong and Replay. Improvements are considerable, for example, in UNSW (R), where BWT for VAE goes from -0.46 to -0.08 for VAE, as well as in Wind (CR), where BWT goes from -0.46 to -0.18. Improvements with the Replay strategy compared to Naive lifelong can also be observed in Figure 9 in comparison to Figure 6, where we observe a significant increase in performance and a decrease in forgetting in all cases except two (performance on C_0 after learning C_2, and performance on C_1 after learning C_1). These two cases highlight that, although the Replay strategy is expected to improve the average model performance due to consideration of all concepts, it also provides additional challenges for the model, which is tasked to learn multiple concepts. As a result, minor decreases in performance for specific concepts should be expected.

We note that there is still a gap between Replay results and simulated upper-bound MSTE results. We can observe that MSTE presents better results than Replay in most cases (56 out of 60). There are 4 cases in which the Replay strategy is better than MSTE. The reason behind this phenomenon can be attributed to the fact that while MSTE can build specialized models for each concept, it cannot leverage knowledge from multiple tasks, thus exploiting task similarity, which, in contrast, is supported in a basic form by Replay.

The improvements in ROC-AUC and BWT observed with Replay when compared with Naive suggest that adopting lifelong learning techniques and learning strategies referenced in Section II, as well as designing new ones, will be beneficial for the advancement of anomaly detection in lifelong learning scenarios (**RQ2**). At the same time, the discussed gap in performance between Replay and MSTE suggests that more robust lifelong strategies may be devised to further improve model performance in complex lifelong anomaly detection scenarios.

C. SUMMARY OF GAINED INSIGHTS

In summary, our experimental analysis discussed above led us to learn the following insights:

- A performance gap exists between non-lifelong and lifelong learning strategies in lifelong anomaly detection scenarios (comparing MSTE vs. Naive) (**RQ1**). This

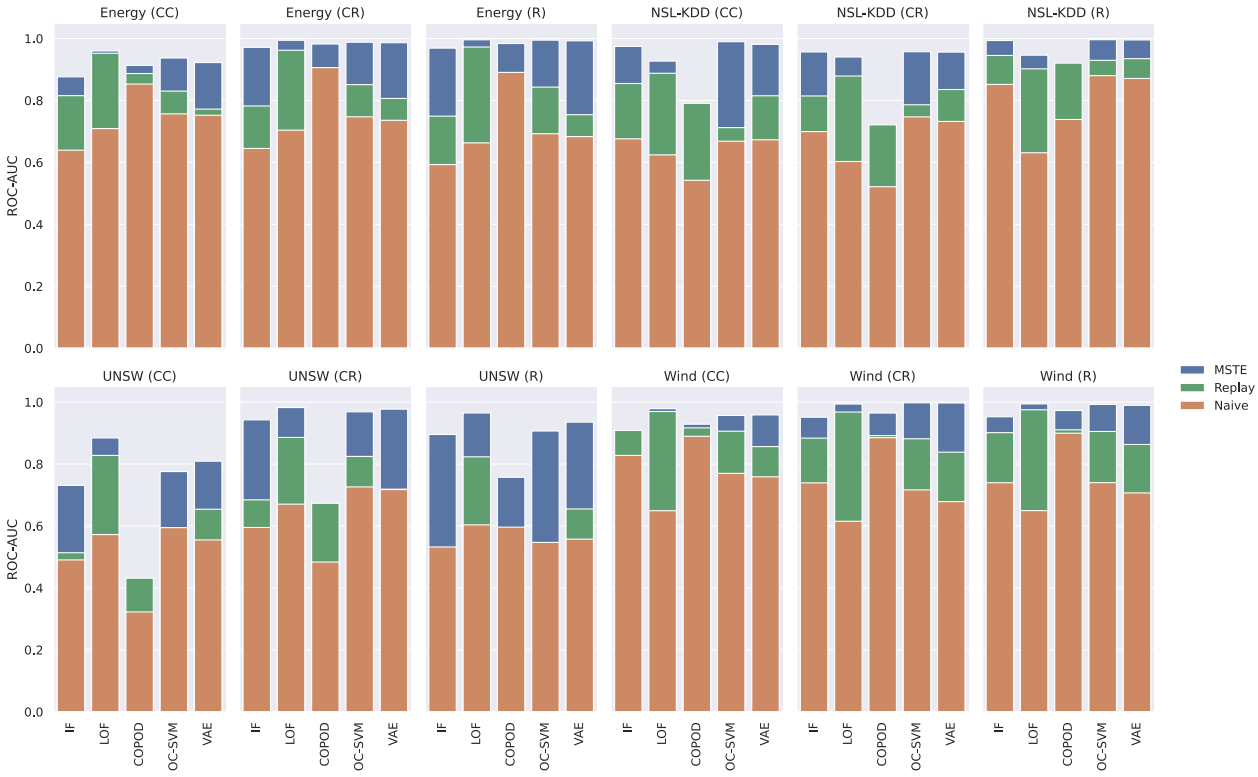

FIGURE 8. Summary of experimental results for all datasets (Energy, UNSW, Wind) in three scenario types (CC, CR, R) comparing non-lifelong (Naive), lifelong (Replay), and upper bound (MSTE) learning strategies. This figure illustrates the performance gap between non-lifelong and lifelong strategies in lifelong anomaly detection scenarios.

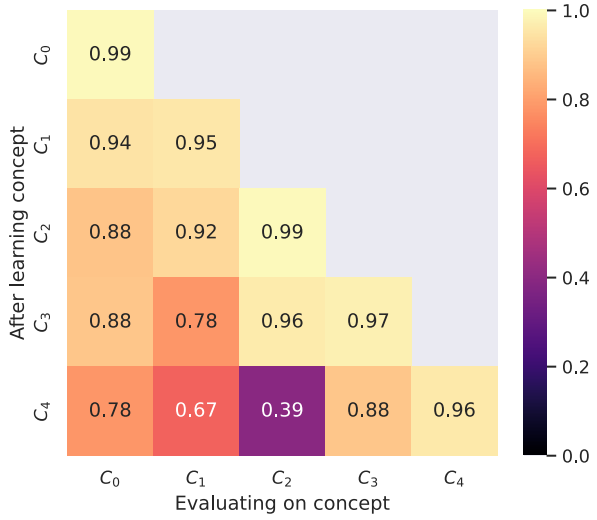

FIGURE 9. Concept-level ROC-AUC performance for the lifelong learning scenario. Each row $i = 0, 1, \ldots$ represents the performance on all concepts observed so far after learning the concept C_i (WIND (R) dataset; Replay strategy; LOF base model).

gap was systematically observed across all scenarios and datasets, and needs to be addressed to increase the reliability of anomaly detection models.

- In such scenarios, the adoption of (even simple) lifelong learning strategies such as Replay enables an improvement in anomaly detection performance when compared to non-lifelong models (comparing Replay vs Naive) (**RQ2**).
- Forgetting was broadly observed for many anomaly detection base models when exposed to the complexity of lifelong scenarios, as emphasized by negative values of BWT. This phenomenon suggests that there is an opportunity to build more robust models that simultaneously deal with adaptation and knowledge retention (**RQ1, RQ2**).
- Our in-depth lifelong evaluation of model performance revealed cases of positive concept similarity (task similarity). This aspect translates in high values of the forward transfer metric (FWT), leading to above-random anomaly detection performance when models are faced with data from not yet learned concepts. Concept similarity is also leveraged by models to improve performance on previously learned concepts after learning a similar concept, resulting in improved values of backward transfer (BWT).

VII. CONCLUSION

In this paper, we addressed lifelong learning in the context of anomaly detection. In general, our contribution stands in the definition of a common ground for research on this topic and showcasing challenges, perspectives and insights brought by lifelong learning for anomaly detection. Specifically, we devised a learning setting characterization that could be

TABLE 3. Hyperparameters for all the anomaly detection models considered in our experiments. All models are trained and evaluated five times using all hyperparameter values in the sets shown in the table, and the final results are averaged.

VAE	hidden layers size $\in \{(16, 4, 16); (12, 8, 12); (19, 9, 19); (32, 16, 32); (32, 8, 32)\}$
LOF	n-neighbors $\in \{5; 10; 15; 20; 25\}$
IF	n-estimators $\in \{25; 35; 50; 75; 100\}$
OC-SVM	$\nu \in \{0.01; 0.02; 0.03; 0.05; 0.1\}$
	$\gamma \in \{0.01; 0.02; 0.03; 0.05; 0.1\}$
COPOD	parameterless

useful to adopt lifelong learning in the context of anomaly detection. Moreover, we designed a procedure for scenario generation that can be used to create lifelong learning scenarios adopting any standard anomaly detection dataset. Insights from our experiments revealed that anomaly detection in lifelong learning scenarios is a challenging problem, and that there is a performance gap between non-lifelong and lifelong learning strategies, indicating that lifelong scenarios are challenging for commonly adopted non-lifelong anomaly detection methods. Moreover, we observed that lifelong learning strategies such as Replay have the potential to tackle these challenges. Overall, we advocate that lifelong learning is essential in anomaly detection to further bring real-life complexity to the experimental setting, providing advantages compared to static and online scenarios currently adopted in the literature. We identified a number of domains, such as cybersecurity, human activity, and industrial processes, where such capabilities can be fruitful due to their dynamic characteristics. We showed that lifelong learning metrics and concept-level performance observed in the learning scenario enable a more detailed model evaluation that uncovers the impact of forgetting and task transferability.

As we provided an overview of lifelong learning challenges from an anomaly detection perspective, we believe that our work will enable other researchers to take action by working on open problems that are relevant in lifelong anomaly detection. To start with, anomaly detection researchers may adopt the scenarios (e.g., concept-aware, concept-incremental) in their benchmark datasets to expose anomaly detection models to new complexities. Moreover, they can adopt lifelong metrics (Lifelong ROC-AUC, BWT, FWT), as well as the evaluation protocol proposed in this paper. In the future, efforts should be directed at designing or improving lifelong scenarios and metrics, so that they reflect the real-world anomaly detection complexities even better. Avenues for future research also include the design of different types of lifelong learning strategies beyond replay-based, such as regularization and architectural, which are popular in image classification, and could be tailored to lifelong anomaly detection.

APPENDIX. HYPERPARAMETERS OF THE DIFFERENT BASE MODELS

For reproducibility, Table 3 shows the five hyperparameter configurations for all models considered in our experiments.

REFERENCES

[1] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier Analysis*. Cham, Switzerland: Springer, 2013, pp. 1–40.
[2] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," *Proc. VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, May 2022.
[3] K. Faber, L. Faber, and B. Sniezynski, "Autoencoder-based IDS for cloud and mobile devices," in *Proc. IEEE/ACM 21st Int. Symp. Cluster, Cloud Internet Comput. (CCGrid)*, May 2021, pp. 728–736.
[4] R. Corizzo, M. Ceci, G. Pio, P. Mignone, and N. Japkowicz, "Spatially-aware autoencoders for detecting contextual anomalies in geo-distributed data," in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2021, pp. 461–471.
[5] M. Fahim and A. Sillitti, "Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review," *IEEE Access*, vol. 7, pp. 81664–81681, 2019.
[6] A. L. Alfeo, M. G. C. A. Cimino, G. Manco, E. Ritacco, and G. Vaglini, "Using an autoencoder in the design of an anomaly detector for smart manufacturing," *Pattern Recognit. Lett.*, vol. 136, pp. 272–278, Aug. 2020.
[7] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.
[8] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.
[9] G. Fenza, M. Gallo, and V. Loia, "Drift-aware methodology for anomaly detection in smart grid," *IEEE Access*, vol. 7, pp. 9645–9657, 2019.
[10] R. Hou, Y. Peng, L. J. Grimm, Y. Ren, M. A. Mazurowski, J. R. Marks, L. M. King, C. C. Maley, E. S. Hwang, and J. Y. Lo, "Anomaly detection of calcifications in mammography based on 11,000 negative cases," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1639–1650, May 2022.
[11] R. Corizzo, M. Ceci, E. Zdravevski, and N. Japkowicz, "Scalable auto-encoders for gravitational waves detection from time series data," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113378.
[12] R. Laxhammar and G. Falkman, "Online learning and sequential anomaly detection in trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1158–1173, Jun. 2014.
[13] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, Jan. 2017.
[14] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.
[15] A. Cossu, A. Carta, V. Lomonaco, and D. Bacciu, "Continual learning for recurrent neural networks: An empirical evaluation," *Neural Netw.*, vol. 143, pp. 607–627, Nov. 2021.
[16] J. Gama, *Knowledge Discovery From Data Streams*. Boca Raton, FL, USA: CRC Press, 2010.
[17] G. I. Parisi, X. Ji, and S. Wermter, "On the role of neurogenesis in overcoming catastrophic forgetting," 2018, *arXiv:1811.02113*.
[18] K. James, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.
[19] B. Wickramasinghe, G. Saha, and K. Roy, "Continual learning: A review of techniques, challenges and future directions," *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 1–21, Dec. 2023.
[20] H. Liu, Y. Yang, and X. Wang, "Overcoming catastrophic forgetting in graph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 8653–8661.
[21] A. Chaudhry, A. Gordo, P. Dokania, P. Torr, and D. Lopez-Paz, "Using hindsight to anchor past knowledge in continual learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 8, pp. 6993–7001.
[22] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.
[23] D. Abel, Y. Jinnai, S. Y. Guo, G. Konidaris, and M. Littman, "Policy and value transfer in lifelong reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 20–29.
[24] J. Parmar, S. Chouhan, V. Raychoudhury, and S. Rathore, "Open-world machine learning: Applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 10, pp. 1–37, Oct. 2023.
[25] M. M. Baker, "A domain-agnostic approach for characterization of lifelong learning systems," *Neural Netw.*, vol. 160, pp. 274–296, Mar. 2023.

[26] M. Mundt, Y. Hong, I. Pliushch, and 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.*, vol. 160, pp. 306–336, Mar. 2023.

[27] X. Nie, Z. Deng, M. He, M. Fan, and Z. Tang, "Online active continual learning for robotic lifelong object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, 2024.

[28] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, Jan. 2022.

[29] K. Faber, B. Sniezynski, and R. Corizzo, "Distributed continual intrusion detection: A collaborative replay framework," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Los Alamitos, CA, USA, Dec. 2023, pp. 3255–3263.

[30] D. Javaheri, S. Gorgin, J.-A. Lee, and M. Masdari, "Fuzzy logic-based DDoS attacks and network traffic anomaly detection methods: Classification, overview, and future perspectives," *Inf. Sci.*, vol. 626, pp. 315–338, May 2023.

[31] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," 2019, *arXiv:1904.07734*.

[32] E. Belouadah, A. Popescu, and I. Kanellos, "A comprehensive study of class incremental learning algorithms for visual tasks," *Neural Netw.*, vol. 135, pp. 38–54, Mar. 2021.

[33] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[34] N. Gunasekara, H. Gomes, A. Bifet, and B. Pfahringer, "Adaptive neural networks for online domain incremental continual learning," in *Discovery Science*, P. Pascal and D. Lenco, Eds. Cham, Switzerland: Springer, 2022, pp. 89–103.

[35] M. De Lange and T. Tuytelaars, "Continual prototype evolution: Learning online from non-stationary data streams," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8230–8239.

[36] A. Soutif–Cormerais, A. Carta, A. Cossu, J. Hurtado, H. Hemati, V. Lomonaco, and J. Van de Weijer, "A comprehensive empirical evaluation on online continual learning," 2023, *arXiv:2308.10328*.

[37] V. Lomonaco, D. Maltoni, and L. Pellegrini, "Rehearsal-free continual learning over small non-I.I.d. batches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 989–998.

[38] A. Cossu, G. Graffieti, L. Pellegrini, D. Maltoni, D. Bacciu, A. Carta, and V. Lomonaco, "Is class-incremental enough for continual learning?" *Frontiers Artif. Intell.*, vol. 5, pp. 1–6, Mar. 2022.

[39] Y.-H. Wang, C.-Y. Lin, T. Thaipisutikul, and T. K. Shih, "Single-head lifelong learning based on distilling knowledge," *IEEE Access*, vol. 10, pp. 35469–35478, 2022.

[40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.

[41] V. Lomonaco and D. Maltoni, "CORe50: A new dataset and benchmark for continuous object recognition," in *Proc. Conf. Robot Learn.*, 2017, pp. 17–26.

[42] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[43] T. Diethe, T. Borchert, E. Thereska, B. Balle, and N. Lawrence, "Continual learning in practice," in *Proc. NeurIPS Continual Learn. Workshop*, 2018, pp. 1–9.

[44] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," 2017, *arXiv:1711.05769*.

[45] H. Kang, R. J. L. Mina, S. Rizky, H. Madjid, J. Yoon, M. Hasegawa-Johnson, S. Ju-Hwang, and C. D. Yoo, "Forget-free continual learning with winning subnetworks," in *Proc. ICML*, 2022, pp. 10734–10750.

[46] P. Buzzega, M. Boschini, A. Porrello, and S. Calderara, "Rethinking experience replay: A bag of tricks for continual learning," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2180–2187.

[47] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. NeurIPS*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–10.

[48] P. K. Mvula, P. Branco, G.-V. Jourdan, and H. L. Viktor, "HEART: Heterogeneous log anomaly detection using robust transformers," in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2023, pp. 673–687.

[49] J. Sendorek, T. Szydlo, M. Windak, and R. Brzoza-Woch, "Dataset for anomalies detection in 3D printing," in *Proc. Int. Conf. Comput. Sci.*, Jun. 2021, pp. 647–653.

[50] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.

[51] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[52] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 582–588.

[53] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.

[54] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, Dec. 2008, pp. 413–422.

[55] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: Copula-based outlier detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 1118–1123.

[56] A. Frikha, D. Krompaß, and V. Tresp, "ARCADe: A rapid continual anomaly detector," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10449–10456.

[57] K. Doshi and Y. Yilmaz, "Continual learning for anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1025–1034.

[58] R. Corizzo, M. Baron, and N. Japkowicz, "CPDGA: Change point driven growing auto-encoder for lifelong anomaly detection," *Knowl.-Based Syst.*, vol. 247, Jul. 2022, Art. no. 108756.

[59] K. Faber, R. Corizzo, B. Sniezynski, and N. Japkowicz, "VLAD: Task-agnostic VAE-based lifelong anomaly detection," *Neural Netw.*, vol. 165, pp. 248–273, Aug. 2023.

[60] K. Faber, R. Corizzo, B. Sniezynski, and N. Japkowicz, "Active lifelong anomaly detection with experience replay," in *Proc. IEEE 9th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2022, pp. 1–10.

[61] M. Du, Z. Chen, C. Liu, R. Oak, and D. Song, "Lifelong anomaly detection through unlearning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 1283–1297.

[62] B. Krawczyk and M. Woźniak, "One-class classifiers with incremental learning and forgetting for data streams with concept drift," *Soft Comput.*, vol. 19, no. 12, pp. 3387–3400, Dec. 2015.

[63] D. Kudithipudi et al., "Biological underpinnings for lifelong learning machines," *Nature Mach. Intell.*, vol. 4, no. 3, pp. 196–210, 2022.

[64] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2018, *arXiv:1812.00420*.

[65] S. Sharma, A. Somayaji, and N. Japkowicz, "Learning over subconcepts: Strategies for 1-class classification," *Comput. Intell.*, vol. 34, no. 2, pp. 440–467, May 2018.

[66] Y. Ghunaim, A. Bibi, K. Alhamoud, M. Alfarra, H. A. A. K. Hammoud, A. Prabhu, P. H. S. Torr, and B. Ghanem, "Real-time evaluation in online continual learning: A new hope," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11888–11897.

[67] K. Fujiwara, M. Shigeno, and U. Sumita, "A new approach for developing segmentation algorithms for strongly imbalanced data," *IEEE Access*, vol. 7, pp. 82970–82977, 2019.

[68] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 111, pp. 1–57, Dec. 2022.

[69] N. Díaz-Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni, "Don't forget, there is more than forgetting: New metrics for continual learning," 2018, *arXiv:1810.13166*.

[70] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.

[71] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6.

[72] R. Corizzo, M. Ceci, and N. Japkowicz, "Anomaly detection and repair for accurate predictions in geo-distributed big data," *Big Data Res.*, vol. 16, pp. 18–35, Jul. 2019.

[73] J. Khan, E. Lee, and K. Kim, "A higher prediction accuracy-based alpha–beta filter algorithm using the feedforward artificial neural network," *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1124–1139, Dec. 2023.

[74] J. Khan, M. Fayaz, A. Hussain, S. Khalid, W. K. Mashwani, and J. Gwak, "An improved alpha beta filter using a deep extreme learning machine," *IEEE Access*, vol. 9, pp. 61548–61564, 2021.

KAMIL FABER received the B.S. and M.S. degrees from the AGH University of Krakow, where he is currently pursuing the Ph.D. degree in computer science. He was a Research Intern and later a Research Assistant with the DARPA-funded project with American University, from 2021 to 2022. His research interests include lifelong learning, anomaly detection, and machine learning applications in cybersecurity.

ROBERTO CORIZZO (Member, IEEE) received the Ph.D. degree. He is an Assistant Professor with the Department of Computer Science, American University. He has coauthored over 40 articles, including 13 publications in journals, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, *Neural Networks*, and *Machine Learning*. He was involved in the scientific committee of international conferences and served as a reviewer for several international journals.

BARTLOMIEJ SNIEZYNSKI received the Ph.D. degree in computer science from the AGH University of Science and Technology, Kraków, Poland, in 2004. In 2004, he was a Postdoctoral Fellow with the Machine Learning and Inference Laboratory, George Mason University, Fairfax, VA, USA, under the supervision of Prof. R. S. Michalski. Currently, he is an Associate Professor with the Institute of Computer Science, AGH University of Science and Technology. His research interests include machine learning, multi-agent systems, and knowledge engineering. He is a member of the Polish Information Processing Society (PTI) and the Polish Artificial Intelligence Society (PSSI).

NATHALIE JAPKOWICZ (Member, IEEE) is a Professor of computer science with American University. Previously, she was with the School of Electrical Engineering and Computer Science, University of Ottawa, where she leads the Laboratory for Research on Machine Learning for Defense and Security. Over the years, she has supervised over 30 graduate students; received funding from Canadian federal and provincial institutions (NSERC, DRDC, Health Canada, OCE, and MITACS CITO); and worked with private companies (Girih, Larus Technologies, Weather Telematics, TechInsights, and Ciena). She has published over 100 articles, papers, and books, including *Evaluating Learning Algorithms: A Classification Perspective*, with Mohak Shah (Cambridge University Press, 2011); and *Big Data Analysis: New Algorithms for a New Society*, with Jerzy Stefanowski (Springer, 2016).