Check for updates

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery

WILEY

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

SYSTEMATIC REVIEW OPEN ACCESS

XAI-Guided Continual Learning: Rationale, Methods, and Future Directions

Michela Proietti[1] | Alessio Ragno[2] | Roberto Capobianco[3]

[1]Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Rome, Italy | [2]INSA Lyon, CNRS, LIRIS UMR 5205, Villeurbanne, France | [3]Sony AI, Zurich, Switzerland

Correspondence: Michela Proietti (mproietti@diag.uniroma1.it)

Received: 4 September 2024 | **Revised:** 4 September 2025 | **Accepted:** 2 October 2025

Editor-in-Chief: Witold Pedrycz

Funding: This work was supported by Sapienza University of Rome, as part of the project XAI-based methods for continual learning (protocol number AR12318876F2DF86).

Keywords: continual learning | explainable artificial intelligence | explanation guided learning

ABSTRACT

Providing neural networks with the ability to learn new tasks sequentially represents one of the main challenges in artificial intelligence. Unlike humans, neural networks are prone to losing previously acquired knowledge upon learning new information, a phenomenon known as catastrophic forgetting. Continual learning proposes diverse solutions to mitigate this problem, but only a few leverage explainable artificial intelligence. This work justifies using explainability techniques in continual learning, emphasizing the need for greater transparency and trustworthiness in these systems and grounding our approach in empirical findings from neuroscience that highlight parallels between forgetting in biological and artificial neural networks. Finally, we review existing work applying explainability methods to address catastrophic forgetting and propose potential avenues for future research.

This article is categorized under:

Fundamental Concepts of Data and Knowledge > Explainable AI

Technologies > Artificial Intelligence

Technologies > Cognitive Computing

1 | Introduction

In recent years, deep learning models have matched or surpassed human performance in various applications, including audio, speech, visual data, and natural language processing (Mathur et al. 2022; Sharma et al. 2021; Wolf et al. 2019). However, while humans can continuously adapt and learn, enabling artificial intelligence (AI) systems to learn sequential tasks remains challenging. Continual learning (CL) addresses this gap by proposing methods inspired by the brain's functioning (Wang, Zhang, et al. 2023; Van de Ven et al. 2020). Despite advancements, many methods face scalability issues due to high memory and computational demands or fixed model capacities (Wang et al. 2024).

The biggest challenge in CL is catastrophic forgetting (McCloskey and Cohen 1989), which stems from the stability-plasticity dilemma (Mermillod et al. 2013): higher stability preserves past knowledge, while higher plasticity enhances new learning. Although artificial and biological neural networks (ANNs, BNNs) differ significantly, techniques inspired by neural mechanisms that aim to find a stability-plasticity trade-off, such as synaptic plasticity (Power and Schlaggar 2017) or memory replay (Skaggs and McNaughton 1996), have

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *WIREs Data Mining and Knowledge Discovery* published by Wiley Periodicals LLC.

managed to reduce accuracy drops in sequential task learning (Kirkpatrick et al. 2016; Aljundi et al. 2018; Zenke et al. 2017; Aljundi, Lin, et al. 2019; Lopez-Paz and Ranzato 2017; Shin et al. 2017). Consequently, given the existing similarities between forgetting events in ANNs and BNNs, the use of explainable artificial intelligence (XAI) in CL might provide useful insights on how and when forgetting occurs and offer valuable guidance for future neuroscientific research (Toneva et al. 2019; Isola, Xiao, et al. 2011; Driscoll et al. 2022; Cossu et al. 2024). Furthermore, recent AI regulations mandate explainability for safe real-world use (Laux et al. 2024; Panigutti et al. 2023), which is an even more crucial requirement for constantly evolving CL systems.

These considerations have led to a new research area leveraging explanations to guide CL training procedures for improved performance, and qualitatively evaluating them in terms of explanation drift (Driscoll et al. 2022; Cossu et al. 2024). The developed methods outperform traditional approaches (Mazumder et al. 2023; Ebrahimi et al. 2021; Saha and Roy 2023a; Gu et al. 2022; Dhar et al. 2019; Ede et al. 2022; Rymarczyk et al. 2023) by guaranteeing more robust shared representations between tasks while offering better interpretability. Although research at the intersection between XAI and CL is still underexplored, the growing number of recent publications indicates its rising importance.

Despite the presence of several surveys on current CL literature (Wang et al. 2024), no review focuses on XAI-guided CL approaches. Additionally, comparing existing methods is difficult due to the variability in scenarios, datasets, and setups. This work fills this gap by reviewing papers at the intersection of XAI and CL, analyzing their benefits and limitations. Finally, we explore parallels between BNNs and ANNs that motivate using XAI in CL and propose future research directions. Figure 1 provides a visual summary of the article's content.

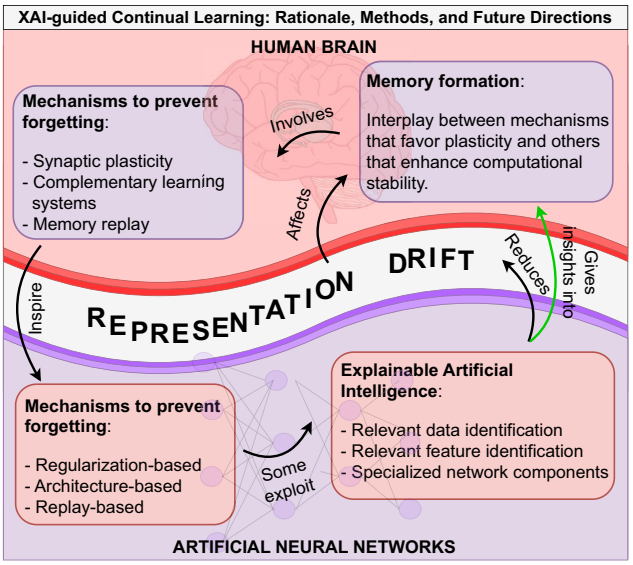

FIGURE 1 | Despite the significant differences existing between artificial and biological neural networks, neuroscience has always been a source of inspiration for artificial intelligence researchers. Most continual learning methods mimic mechanisms in the human brain that support memory formation while preventing forgetting. Among them, several approaches employ explainability techniques to mitigate explanation drift (Cossu et al. 2024), a phenomenon mirroring representation drift in the human brain (Driscoll et al. 2022). In this paper, we survey such work, highlighting how the type of analysis stemming from the use of explainable artificial intelligence in continual learning can open new interesting lines of research at the intersection between artificial intelligence and neuroscience.

2 | Background

This section introduces the necessary background on XAI and CL to frame XAI-guided CL methods in the existing literature and to highlight the importance of this research line.

2.1 | Explainable Artificial Intelligence

Deep learning models achieve impressive performance due to their complexity, enabling them to model intricate input–output relations. However, this complexity limits interpretability. XAI addresses this problem by developing methods to explain model predictions.

XAI literature is wide, and several surveys review existing explainability methods and propose general taxonomies (Barredo Arrieta et al. 2020; Speith 2022; Schwalbe and Finzel 2023). Common categorizations are based on the stage and scope of the explanations:

- **Stage**: Post hoc methods generate explanations for pretrained models (Ribeiro et al. 2016; Bach et al. 2015; Selvaraju et al. 2017), while self-interpretable methods provide explanations alongside predictions (Chen et al. 2019; Rymarczyk et al. 2022).
- **Scope**: Local explanation methods offer understanding at the instance level (Ribeiro et al. 2016; Bach et al. 2015; Selvaraju et al. 2017), while global methods explain the overall model behavior (Chen et al. 2019; Rymarczyk et al. 2022).

As AI systems increasingly impact our daily life, explaining their decisions becomes crucial and is mandated by government regulations (Laux et al. 2024; Panigutti et al. 2023). At the same time, researchers are exploring how explanations can enhance model performance, fairness, transparency, and bias removal, unleashing real-world applications. These new needs have resulted in a new research direction, Explanation Guided Learning (Gao et al. 2024). XAI-guided CL is part of this field, focusing on the use of XAI to balance plasticity and stability for robust shared representations between tasks.

2.2 | Continual Learning

In contrast to traditional machine learning models, which are built to solve individual tasks, each characterized by a static data distribution, the purpose of continual learning (CL) research is to provide AI systems with the ability to learn from a non-stationary data stream.

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Formally, in a CL setting the same model needs to learn a sequence of T tasks sequentially. For each task $t \in [1, T]$, the training dataset is defined as $D_t^{\text{train}} = \{x_t, y_t\}$, where x_t is an input sample and y_t the corresponding label. The test dataset, instead, is independent of the corresponding training set, that is, $D_t^{\text{train}} \cap D_t^{\text{test}} = \emptyset$, but follows the same probability distribution. Then, while training the model on a task t, only samples belonging to D_t^{train} are available. Data belonging to D_i^{train}, with $i = 1, \ldots, t-1$ is unavailable unless we implement some replay strategy (see Table 1).

Depending on the type of input data and on the availability of task identities during training and testing, we can identify different learning scenarios that we describe in detail in Section 2.2.1. Additionally, in Table 1 we summarize the different categories of CL methods identified by Wang et al. (2024). Although we provide references to popular traditional methods, please refer to Wang et al. (2024) for a comprehensive survey.

2.2.1 | Learning Scenarios

There is currently no consensus on the categorization of CL scenarios. Many works identify three classes: task-incremental, class-incremental, and domain-incremental learning (van de Ven et al. 2022), though terms are often used inconsistently. With the aim of presenting XAI-guided methods, we adopt the finer taxonomy by Wang et al. (2024), which takes into account input distributions, label spaces, availability of task identities at both training and testing times, and the way in which data is given as input to the network. Table 2 summarizes this taxonomy, which recognizes the following general scenarios:

- **Instance-incremental learning** (IIL): the training samples belong to the same task and are presented in batches. In this case, we have one single dataset, which is incrementally learned by a network. For instance, Mazumder et al. (2023) perform multiple training sessions on the Google Speech Commands dataset (Warden 2018) in which the network learns to classify samples that were wrongly predicted in the previous sessions.
- **Domain-incremental learning** (DIL): tasks have different data distributions, same label spaces, and there is no need to know task identities. Examples of this scenario are incrementally learning to detect fire on spaceborne, airborne and terrestrial sensor images (Wang, Yu, et al. 2023), or to classify medical images acquired through different hardware (Srivastava et al. 2021).
- **Task-incremental learning** (TIL): tasks have different data distributions, disjoint label spaces, and task identities are known at both training and testing times;
- **Class-incremental learning** (CIL): tasks have different data distributions, disjoint label spaces, and task identities are known just at training time;
- **Task-free continual learning** (TFCL): tasks have different data distributions, disjoint label spaces, and task identities are not known.

An example dataset for the last three scenarios is MNIST-split, in which the original dataset is divided into five disjoint tasks, each consisting of the classification of two digits (Ede et al. 2022). In this work, we consider online continual learning, blurred boundary continual learning, and continual pre-training (Wang et al. 2024) as possible additional characterizations of the

TABLE 1 | Summary of the taxonomy of CL approaches provided by (Wang et al. 2024).

Category	Description	References
Regularization-based	Use a regularization term to achieve a good stability-plasticity trade-off.	(Kirkpatrick et al. 2016; Zenke et al. 2017; Aljundi et al. 2018; Li and Hoiem 2016; Rebuffi et al. 2016; Wang et al. 2022)
Replay-based	Try to approximate or recover the distributions of old tasks by replaying information from old tasks.	(Aljundi, Lin, et al. 2019; Lopez-Paz and Ranzato 2017; Riemer et al. 2019; Shin et al. 2017; Wu et al. 2018; Toldo and Ozay 2022; Kukleva et al. 2021; Liu et al. 2022)
Optimization-based	Act on the optimization process by either projecting gradients in non-interfering directions, using the loss to find flat local minima, or using meta-learning.	(Riemer et al. 2019; Farajtabar et al. 2020; Javed and White 2019; Rajasegaran et al. 2020; Chi et al. 2022)
Representation-based	Produce robust representations by either using self-supervised learning, large-scale pre-training, or representation learning.	(Madaan et al. 2022; Wu et al. 2022; Huo et al. 2023; Asadi et al. 2023; Tao et al. 2020)
Architecture-based	Use task-specific parameters, by either allocating a different parameter subspace for each task, using task-specific components which can be dynamic, or developing modular architectures with parallel sub-networks or sub-modules.	(Ebrahimi et al. 2020; Ostapenko et al. 2021; Veniat et al. 2021; Yang et al. 2023)

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Categorization of CL scenarios provided by Wang et al. (2024), based on input distributions, label spaces, task identity availability, and data provision to the network.

Scenario	Description
Instance-incremental learning (IIL)	Same input distributions, same label spaces, task identities not required.
Domain-incremental learning (DIL)	Different input distributions, same label spaces, task identities available during training but not required during testing.
Task-incremental learning (TIL)	Different input distributions, disjoint label spaces, task identities available during training and testing.
Class-incremental learning (CIL)	Different input distributions, disjoint label spaces, task identities just available during training.
Task-Free CL (TFCL)	Different input distributions, disjoint label spaces, task identities unavailable during training and optionally available during testing.
Online CL (OCL)	Different input distributions, disjoint label spaces, task identities unavailable during training and optionally available during testing, training saples are provided as a one-pass data stream (i.e., batch_size = 1).
Blurred-boundary CL (BBCL)	Different input distributions, overlapping label spaces, task identities just available during training.
Continual pre-training (CPT)	Several pre-training tasks followed by a downstream task, task identities are available during training but not required during testing.
Few-shot CIL (FSCIL)	First task of b base classes followed by C-way K-shot classification tasks, with C being the number of classes and K being the (limited) number of samples per class. Training is performed in a class-incremental learning setting.

previous general scenarios. For instance, we talk about **online** DIL, TIL, CIL, and TFCL if, during training, samples from sequential tasks are provided as a one-pass data stream. Similarly, in **blurred boundary** TIL, CIL, and TFCL, data label spaces are disjoint but can be partially overlapped. For instance, we might want to first learn to classify the digits 0 and 1, and then the digits 1 and 2. Finally, we have instance–domain–task–class-incremental and task-free continual **pre-training**, when we pre-train a model on sequential tasks to improve the performance on another downstream task.

2.2.1.1 | Few-Shot Class-Incremental Learning. Although most CL approaches rely on the availability of large quantities of annotated data, this is quite an unrealistic assumption, as labeling a continuous stream of data is practically unfeasible. It is, instead, much more realistic that only a very small number of labeled samples is available. For this reason, few-shot class-incremental learning (FSCIL) represents a promising research path in the AI field (Tao et al. 2020; Tian et al. 2024). In FSCIL, the first task is made of b base classes with sufficient training data, while each subsequent task consists of a C-way K-shot classification, with C being the number of classes and K being the (limited) number of samples per class. The objective is to be able to train models on all the tasks sequentially, in a class-incremental learning setting.

2.2.2 | Neuroscience-Inspired Continual Learning Approaches

Neuroscience has long inspired AI research, particularly in CL. Key mechanisms in the brain relevant to CL include synaptic plasticity, hippocampal replay, and complementary learning systems.

2.2.2.1 | Synaptic Plasticity. Synaptic plasticity is crucial for balancing stability and plasticity in the brain (Power and Schlaggar 2017). Individual dendritic branches[1] allow selective updates of important synapses, preventing forgetting. Cichon and Gan (2015) provide an example of this behavior, showing that different motor tasks activate distinct dendrites, reducing memory interference. Regularization-based approaches in AI take inspiration from these findings, identifying key synaptic weights and reducing their plasticity for new tasks by adding a regularization term to the loss function (Kirkpatrick et al. 2016; Aljundi et al. 2018; Zenke et al. 2017).

In the brain, synaptic plasticity regulates the mechanism of long-term potentiation, consisting of an increase in synaptic efficiency, which leads to synaptic consolidation in the local circuits (Lüscher and Malenka 2012). However, memory consolidation also occurs at a system level through a slower process called systems memory consolidation. Frankland and Bontempi (2005) first described this phenomenon, noting that recent memories are more prone to interference than older ones, suggesting increased memory robustness over time. The study also suggests that the hippocampus is involved in short-term memory's formation and retrieval. According to the **complementary learning systems** theory, durable memories form through the interaction between the hippocampus and neocortex (McClelland et al. 1995; Kumaran et al. 2016). Based on these findings, Pham et al. (2021) and Blakeman and Mareschal (2020) propose architectures with fast- and slow-learning components for CL in neural networks, enabling the formation of more robust representations.

2.2.2.2 | Memory Replay. Memory replay is believed to facilitate the transfer of knowledge from short-term to long-term storage by reactivating hippocampal activity patterns during sleep and rest, thus preventing catastrophic forgetting (Skaggs and McNaughton 1996; Rudoy et al. 2009). AI research mirrors this mechanism with episodic memory buffers to reduce forgetting in CL scenarios (Aljundi, Lin, et al. 2019; Lopez-Paz and Ranzato 2017; Van de Ven et al. 2020). However, the data requirements

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of these methods limit their scalability. These observations raise questions about how the brain implements similar strategies with bounded memory requirements. Subsequent rodent studies show hippocampal sequences that predict future goals, shortcuts, or events beyond past experiences, leading to theories that model the hippocampus as a hierarchical generative model with reactivations in the form of generative replay (Stoianov et al. 2022). Consequently, several works in AI use generative replay to mitigate catastrophic forgetting in Shin et al. (2017) and Wu et al. (2018). Despite eliminating the need for large replay buffers, these systems require computationally intensive training of generative models, which may still result in catastrophic forgetting.

2.3 | Forgetting in Artificial and Biological Neural Networks

As described in Section 2.2.2, CL research in neuroscience and AI has a long, intertwined history. This interaction has significantly advanced our understanding of memory protection mechanisms, yet open questions remain that could benefit from joint efforts across these research communities.

Catastrophic forgetting can also occur in the human brain under specific conditions. For example, Pallier et al. (2003) show that Korean-born adults adopted by French families between ages 3 and 8 retain no residual knowledge of Korean, exhibiting brain activations similar to those for entirely foreign languages. Similarly, Mareschal et al. (2007) study sequential category learning in 3- and 4-month-old infants, finding strong task-order dependence: infants retain the category "dog" only when it is learned before "cat," likely due to category similarity. Comparable effects have been observed in ANNs, where intermediate task similarity produces the highest forgetting, while increasing dissimilarity reduces it (Lee et al. 2021; Ramasesh et al. 2021; Lin et al. 2023).

A particularly relevant biological phenomenon is **representation drift**, the gradual change of neural activity patterns encoding the same stimulus over time. While this might appear detrimental to stable memory, Driscoll et al. (2022) argue that drift can also be functional, enabling the brain to flexibly connect or differentiate experiences occurring at different times. CL in ANNs faces an analogous challenge: as new tasks are learned, the internal representations of old ones shift, often leading to catastrophic forgetting. At the level of explanations, this manifests as **explanation drift** (Cossu et al. 2024), where saliency maps or other attribution signals for the same input change substantially after learning new tasks. Explanation drift therefore provides a window into representational changes in ANNs and can serve as an early indicator of when forgetting is likely to occur. Some approaches already exploit this connection: for instance, Ebrahimi et al. (2021) combine replay with regularization to stabilize explanations over time, reducing forgetting compared to non-XAI-guided replay methods.

There are also similarities in individual forgetting events across ANNs and BNNs. Toneva et al. (2019) demonstrate that certain samples are consistently more prone to being forgotten or remembered, regardless of architecture or learning algorithm. This suggests that memorability depends on intrinsic properties of the input stimulus, a finding that aligns with neuroscientific studies on human memorability (Isola, Parikh, et al. 2011; Isola, Xiao, et al. 2011; Isola et al. 2014; Khosla et al. 2015; Bainbridge et al. 2013; Bainbridge et al. 2017; Xie et al. 2020). Such parallels reinforce the idea that explanation signals can reveal which inputs or features are most critical for long-term retention.

Building on these analogies, neuroscientific insights can be translated into concrete design principles for XAI-guided CL. Just as synaptic consolidation stabilizes important connections in the brain, **importance-weighted stability** can be enforced in ANNs by penalizing large changes in explanations for previously seen data (e.g., the regularization used in Ebrahimi et al. (2021)). Likewise, dendritic specialization inspires methods that **freeze or grow network components based on neuron relevance**, as in RNF (Ede et al. 2022), which preserves the most important neurons for past tasks while leaving others free to adapt to new ones. Finally, hippocampal replay motivates **selective replay guided by explanations**, where attribution maps identify the most informative experiences to store or regenerate, thus reducing memory and computational requirements (Saha and Roy 2023b; Shim et al. 2021).

In summary, using XAI in CL not only provides practical benefits such as improved performance and memory efficiency, but also offers theoretical advantages by linking explanation drift in ANNs with representation drift in neuroscience. This connection enables explanation signals to serve not only as interpretability tools, but also as principled drivers of continual learning.

3 | XAI-Guided Continual Learning

Table 3 lists the existing XAI-guided CL approaches with their abbreviations. The number of papers on this topic increases yearly, with many of them being accepted at top conferences, indicating growing interest from the research community. In the following sections, we discuss the datasets and scenarios in XAI-guided CL papers, emphasizing the need for common benchmarks and open-source implementations for easier comparison. We then categorize existing approaches based on the target of explanations, describing them in detail.

3.1 | Paradigm Formalization

While the notion of *XAI–guided continual learning* has been used informally in prior work, it remains unclear what qualifies as a genuine instance of this paradigm. The key idea is to treat explanations not merely as post hoc interpretability tools, but as *training signals* that influence how CL is performed. This requires identifying two elements: (i) an *explanation operator*, which extracts information about how the model produces its predictions, and (ii) a *guidance operator*, which uses such explanations to shape training dynamics, for instance by selecting replay samples, masking salient features, freezing network components, or regularizing explanation stability.

An **explanation operator** is a mapping

$$\mathcal{E}_\phi\colon\ \mathcal{X}\times\Theta\to\mathbb{R}^d,$$

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | List of existing XAI-guided CL approaches, with corresponding abbreviation, reference, venue, and year of publication, and code availability.

Name	Abbreviations	References	Venue	Code
Learning without memorizing	LwM	Dhar et al. (2019)	CVPR 2019	✓
Remembering for the right reasons	RRR	Ebrahimi et al. (2021)	ICLR 2021	✓
Adversarial Shapley value experience replay	ASER	Shim et al. (2021)	AAAI 2021	✓
Semi-quantized activation neural networks	SQANN	Tjoa and Guan (2022)	arXiv 2022	✓
Relevance-based neural freezing	RNF	Ede et al. (2022)	CD-MAKE 2022	
Dual view consistency	DVC	Gu et al. (2022)	CVPR 2022	✓
Experience packing and replay	EPR	Saha and Roy (2023b)	WACV 2023	
XAI-increment	XAI-I	Mazumder et al. (2023)	EUSIPCO 2023	
Interpretable class-InCremental LEarning	ICICLE	Rymarczyk et al. (2023)	ICCV 2023	✓
Shape and semantics-based selective regularization	S3R	Zhang et al. (2023)	IEEE Medical Imaging 2023	✓
Saliency-augmented memory completion	SAMC	Bai et al. (2023)	SDM 2023	✓

ALGORITHM 1 | Schematic Training Loop for XAI-Guided Continual Learning.

Initialize Model Parameters θ, Explainer ϕ, and Memory $\mathcal{M}$
for each task $t = 1, \ldots, T$ **do**
  for each batch $\mathcal{B}_t$ in task t **do**
    Compute explanations $e \leftarrow \mathcal{E}_\phi(\mathcal{B}_t, \theta)$
    Apply guidance operator:
    $$\left(\widetilde{\mathcal{B}}_t, \widetilde{\ell}_t, \widetilde{\mathcal{M}}_t, \mathcal{C}_t\right) \leftarrow \mathcal{G}\left(\mathcal{B}_t, \theta, \mathcal{M}, \mathcal{E}_\phi\right)$$
    Compute Total Loss:
    $$L = \widetilde{\ell}_t + \lambda_{\text{stab}}\mathrm{S} + \lambda_{\text{res}}\Omega + \lambda_{\text{faith}}\mathrm{F}$$
    Update parameters: $\theta \leftarrow \text{Update}\left(\theta, L, \mathcal{C}_t\right)$
    Update memory: $\mathcal{M} \leftarrow \widetilde{\mathcal{M}}_t$
  end for
end for

that returns an explanation $e = \mathcal{E}_\phi(x, \theta)$ in an explanation space $\mathbb{R}^d$ (e.g., saliency maps, prototype similarities, concept activations). $\mathcal{E}_\phi$ can be produced via a post hoc explainability method or by a self-interpretable model itself.

A **guidance operator** is a policy

$$\mathcal{G}: \ \left(\mathcal{B}_t, \theta_t, \mathcal{M}_t, \mathcal{E}_\phi\right) \ \mapsto \ \left(\widetilde{\mathcal{B}}_t, \widetilde{\ell}_t, \widetilde{\mathcal{M}}_t, \mathcal{C}_t\right),$$

which modifies the batch, the loss, the memory, and/or architectural constraints $\mathcal{C}_t$ based on explanations.

Definition 1. A continual learning procedure is XAI–guided if explanations $\mathcal{E}_\phi$ influence the objective, memory policy, or parameter allocation via $\mathcal{G}$. Classical continual learning is recovered as the special case where $\mathcal{G}$ ignores $\mathcal{E}_\phi$.

To make the role of explanations explicit, we summarize the paradigm in Algorithm 1. The procedure alternates between computing explanations for current and past data, applying a guidance operator that adapts batches, losses, memory contents, or architectural constraints, and updating the model accordingly. The optimization objective combines multiple components: (i) the task loss $\widetilde{\ell}_t$ on the current batch (possibly modified by guidance), (ii) a stability term S weighted by λ_{stab}, which penalizes large drifts in explanations for previously seen data, (iii) a resource term Ω weighted by λ_{res}, which constrains memory usage or architectural growth, and (iv) an optional faithfulness term F weighted by λ_{faith}, encouraging explanations to remain aligned with human-interpretable concepts or ground-truth annotations.

3.2 | Datasets and Scenarios

The heterogeneity of datasets and scenarios complicates comparisons between approaches, as shown in Table 4, with methods being tested on different dataset–scenario combinations. Below, we describe each dataset, providing relevant references:

- Split-MNIST (Zenke et al. 2017) and Split-CIFAR10 (Shim et al. 2021): Ede et al. (2022), Shim et al. (2021), and Gu et al. (2022) split them into five binary classification tasks, respectively.
- Permuted-MNIST (Goodfellow et al. 2013): the first task is the original dataset, followed by nine tasks with random pixel permutations applied consistently within each task but differently across tasks (Ede et al. 2022).
- Split-CIFAR-100 (Rebuffi et al. 2016): split into 10 tasks of 10 classes each by Ebrahimi et al. (2021); Shim et al. (2021); Dhar et al. (2019), 20 tasks of 5 classes each by Saha and Roy (2023b,0); Ebrahimi et al. (2021); Dhar et al. (2019), and 50 tasks of 2 classes each by Dhar et al. (2019)

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Summary of the datasets and scenarios addressed by the existing XAI-based approaches for CL.

Dataset	Tasks	IIL	TIL	CIL	FSCIL
Split-MNIST	5	—	RNF	—	—
Permuted-MNIST	10	—	RNF	—	—
Split-CIFAR-10	5	—	SAMC	ASER, DVC	—
Split-CIFAR-100	10	—	LwM	ASER, DVC, LwM, RRR	—
	20	—	EPR, LwM, SAMC	RRR, EPR	—
	50	—	LwM	LwM	—
CIFAR-10/−100	1/5	—	RNF	—	—
	1/8	—	RNF	—	—
	1/10	—	RNF	—	—
Split-CUB-200	4	—	—	ICICLE	—
	10	—	LwM	LwM, ICICLE, EPR	—
	20	—	EPR	ICICLE	—
	11[1]	—	—	—	RRR
Split-ImageNet	10	—	—	RRR	RRR
Split-ImageNet-small	10	—	LwM	LwM	—
Split-miniImageNet	10	—	—	ASER, DVC	—
	20	—	EPR, SAMC	EPR	—
Split-adience	2	—	RNF	—	—
Adience/ImageNet	—	—	RNF	—	—
Split-caltech-101	10	—	LwM	LwM	—
Google speech commands	—	XAI-I	—	—	—
Split-stanford cars	4,7,14	—	—	ICICLE	—

Note: The second column reports the number of tasks, while the last four columns correspond to the different scenarios, namely instance-incremental learning (IIL), task-incremental learning (TIL), class-incremental learning (CIL), and few-shot class-incremental learning (FSCIL).

- CIFAR-10/CIFAR-100 (Ede et al. 2022): the first task is CIFAR-10, followed by CIFAR-100 split into five tasks of 10 random classes, 10 tasks of 10 random classes, or 8 tasks of semantically similar classes.
- CUB-200 (Dhar et al. 2019): split into 20 tasks of 10 classes each by Saha and Roy (2023b, 0) and Rymarczyk et al. (2023), 10 tasks of 10 classes each by Gu et al. (2022); Dhar et al. (2019), 4 tasks of 50 classes and 10 tasks of 20 classes by Rymarczyk et al. (2023), and 11 tasks (first with 100 classes, others with 10 classes each) by Ebrahimi et al. (2021) in an FSCIL scenario.
- Split-Imagenet (Rebuffi et al. 2016): Ede et al. (2022) split it into 6 tasks of 100 classes each, while Ebrahimi et al. (2021) and Dhar et al. (2019) divide it into 10 tasks of 10 random classes each.
- Split-Imagenet-small (Rebuffi et al. 2016): subset of 100 classes from Imagenet, consisting of 10 tasks of 10 classes each.
- Split-miniImagenet (Shim et al. 2021): split into 20 tasks of 5 classes each for TIL by Saha and Roy (2023b, 0), and 10 tasks of 10 classes each for CIL by Saha and Roy (2023b, 0), Shim et al. (2021), and Gu et al. (2022).
- Split-Adience (Ede et al. 2022): split into two tasks of 4 classes each, representing different age groups in mixed and ordered setups.
- Adience/ImageNet: Ede et al. (2022) train the models on ImageNet, prune them, and then train them on the Adience dataset.
- Caltech-101 (Fei-Fei et al. 2006): Dhar et al. (2019) consider 10 classes per task.
- Google Speech Commands (Warden 2018): Mazumder et al. (2023) convert the audio files into spectrograms and use them in an IIL scenario.
- Stanford cars (Rymarczyk et al. 2023): split into 4, 7, or 14 tasks of 49, 28, and 14 classes, respectively.

3.3 | Papers Categorization

This section provides a taxonomy of XAI-guided methods. Most approaches use feature attribution techniques like GradCAM (Selvaraju et al. 2017), GradCAM++ (Chattopadhay et al. 2018), SmoothGrad (Smilkov et al. 2017), FullGrad (Srinivas and

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Fleuret 2019), LIME (Ribeiro et al. 2016), or LRP (Bach et al. 2015) to prevent explanation drift. Tjoa and Guan (2022) and Rymarczyk et al. (2023), instead, propose self-explainable architectures: the former introduces an expandable network with semi-quantized activations, while the latter uses a prototype-based model. Similarly, XAI-guided methods span all CL categories, but most representatives belong to replay-, regularization-, and architecture-based classes. Consequently, any categorization of XAI-guided CL approaches by explanation type, stage, scope, or CL category would be imbalanced and not adaptable to new explanations. Thus, our categorization focuses on the object of explanation: data, features, or network components.

3.3.1 | Relevant Data Identification

Relevant data identification involves finding representative samples to enhance class discrimination. Toneva et al. (2019) show that removing certain unforgettable samples from the training set does not affect network performance, suggesting some samples are less informative. This insight inspired XAI-guided memory replay approaches.

Shim et al. (2021) introduce adversarial shapley value experience replay (ASER), using Shapley values (Shapley 1953) to select samples for future replay. ASER strategically chooses samples near class boundaries and from different classes, outperforming standard selection strategies like Maximal Interfered Retrieval (Aljundi, Belilovsky, et al. 2019).

Dual view consistency (DVC) (Gu et al. 2022), instead, randomly selects samples for storage but uses a strategy to choose replay samples from the memory bank. For each training batch, DVC performs a virtual parameter update to select memory bank samples with highly interfered gradients, combining them with the training batch for the actual update. Despite not using a proper explanation technique, thanks to a combined use of gradient-based sample selection and representation learning, the authors manage to reduce representation drift by favoring consistency.

Table 5 compares the two approaches in terms of average final accuracy and forgetting (i.e., how much the accuracy of each task has dropped with respect to when it was first trained). Results show DVC's superiority in terms of both metrics, suggesting that storing samples based on inter-task interference is more effective than selecting samples based on within-task discriminativeness. While effective, these approaches have a major drawback: high memory requirements, making them impractical as the number of tasks increases.

3.3.2 | Relevant Feature Identification

As discussed in Section 2.3, several studies in AI and neuroscience show that the memorability of input stimuli is influenced by intrinsic properties (Toneva et al. 2019; Xie et al. 2020; Isola, Xiao, et al. 2011). This section explores XAI-guided CL methods that use explainability to identify relevant features.

TABLE 5 | Accuracy and forgetting of ASER and DVC on Split-miniImageNet, Split-CIFAR-100, and Split-CIFAR-10 in CIL scenario, averaged across 15 seeds.

(a) Mini-ImageNet			
	Buffer size		
	1k	**2k**	**5k**
Accuracy			
ASER	11.5 ± 0.6	13.5 ± 0.8	17.8 ± 1.0
DVC	15.4 ± 0.7	17.2 ± 0.8	19.1 ± 0.9
Forgetting			
ASER	33.8 ± 1.3	30.5 ± 1.3	25.1 ± 0.8
DVC	25.1 ± 0.7	23.1 ± 0.7	21.9 ± 0.8
(b) CIFAR-100			
	Buffer size		
	1k	**2k**	**5k**
Accuracy			
ASER	14.3 ± 0.5	17.8 ± 0.5	22.8 ± 1.0
DVC	19.7 ± 0.7	22.1 ± 0.9	24.1 ± 0.8
Forgetting			
ASER	43.0 ± 0.5	37.9 ± 0.6	29.6 ± 0.9
DVC	30.6 ± 0.7	27.8 ± 1.0	26.1 ± 0.5
(c) CIFAR-10			
	Buffer size		
	0.2k	**0.5k**	**1k**
Accuracy			
ASER	29.6 ± 1.0	38.2 ± 1.0	45.1 ± 2.0
DVC	45.4 ± 1.4	50.6 ± 2.9	52.1 ± 2.5
Forgetting			
ASER	56.4 ± 1.6	47.5 ± 1.3	39.6 ± 2.0
DVC	27.2 ± 2.5	21.3 ± 3.1	19.7 ± 2.9

Note: Experiments are performed using buffer sizes 1, 2, and 5k for Split-miniImageNet and Split-CIFAR-100, and 0.2, 0.5, and 1 k for Split-CIFAR-10.

Some approaches implement light-memory replay by storing only relevant features, avoiding the memory overload of storing raw samples. Experience packing and replay (EPR) (Saha and Roy 2023b,0) and Saliency-augmented memory completion (SAMC) (Bai et al. 2023) leverage saliency maps to store the most relevant parts of input images. Figure 2 illustrates the differences between these methods in computing and replaying relevant features. Specifically, given a saliency map $S_I \in \mathbb{R}^{H \times W}$, EPR computes the most salient patch by performing an average pooling with kernel size equal to the desired patch size $W_p \times W_p$. The maximum average-pooled value corresponds to the top-left coordinates (x_p, y_p) of the most salient patch P, extracted from the input image I as:

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

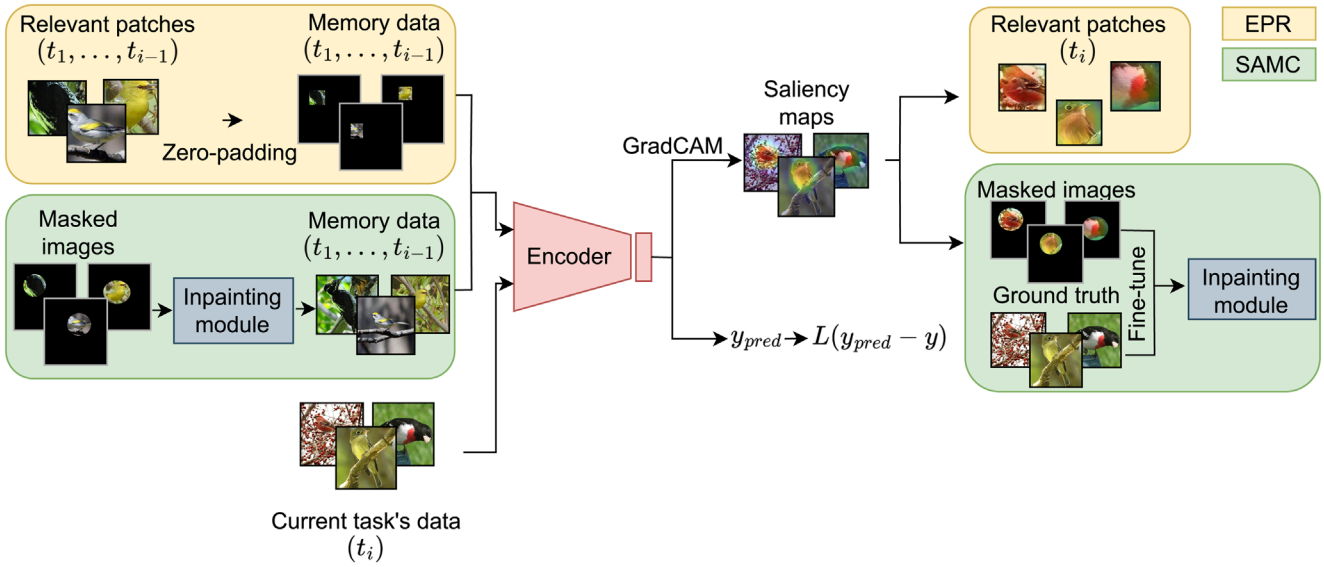

FIGURE 2 | A comparative overview of EPR (Saha and Roy 2023b,0) and SAMC (Bai et al. 2023). In both approaches, during $task_i$'s training, the network receives $task_i$'s data and data from the replay buffer for tasks $t_1, \ldots, t_{i-1}$. In the case of EPR (yellow), replay data consists of relevant patches, which are zero-padded before passing them to the network. Instead, in the case of SAMC (green), replay data consists of masked images stored in coordinate format (Bell and Garland 2009), that are first converted to sparse matrices, then passed to an inpainting module, which reconstructs the original images, which are then passed to the network. In both cases, the network outputs predictions and saliency maps are computed using GradCAM. These are then used to extract the most salient patches from a subset of the training images in EPR and to obtain masked images in SAMC. Both patches and masked images are stored for replay while training on new tasks. In SAMC, before moving to the next task's training, the masked images, together with the corresponding original images, are used to fine-tune the inpainting module.

$$P = I(x_p : x_p + W_p, y_p : y_p + W_p)$$

During replay, the authors perform zero-padding, ensuring that the patch is in the same position as the original image through the stored (x_p, y_p) coordinates. To avoid accuracy loss, EPR stores more patches than needed and prioritizes those that result in correct predictions after zero-padding. Differently, SAMC computes sparse masks by thresholding saliency values:

$$I' = \mathbb{1}\{S_I > \mu\}$$

and store them as sparse matrices in coordinate format (Bell and Garland 2009). Then, an inpainting module, trained with the classification network via bilevel optimization, reconstructs images for replay. The lower level addresses the main task objective, while the upper level minimizes reconstruction loss. EPR and SAMC achieve performance comparable to state-of-the-art replay methods like GEM (Lopez-Paz and Ranzato 2017), MER (Riemer et al. 2019), and iCaRL (Rebuffi et al. 2016), using less memory.

The results in Table 6a compare results used by SAMC and EPR on Split-CIFAR-100 in a TIL scenario. Both approaches achieve significantly higher performance than the finetune baseline (which simply performs the training without applying any CL strategy). However, EPR consistently outperforms SAMC at comparable memory budgets, achieving higher accuracy despite using a smaller memory buffer (e.g., 58.5% with only 85 samples against 56.2% for SAMC with 200). This suggests that patch-based replay guided by saliency is more memory-efficient than SAMC's inpainting-based reconstruction, and also does not require training a separate inpainting model.

Despite effectively using explanations to improve performance and reduce memory requirements, the authors of EPR (Saha and Roy 2023b) and SAMC (Bai et al. 2023) do not analyze their impact on explanation drift. In contrast, as shown in Figure 3, Dhar et al. (2019), Ebrahimi et al. (2021), and Rymarczyk et al. (2023) address this issue by adding regularization or penalty terms to the loss function, discouraging changes in saliency maps, or learning task-specific prototypes. Remembering for the right reasons (RRR) Ebrahimi et al. (2021) uses a memory buffer $\mathcal{M}_{rep}$ for raw samples and another one $\mathcal{M}_{RRR}$ for reference saliency maps. After training on task t_i, RRR replays stored samples from tasks $t_1, \ldots, t_{i-1}$ and computes new saliency maps. It then calculates the L1 loss between the new and reference saliencies:

$$L_{RRR} = \mathbb{E}_{(I,\hat{S}_I)\sim(\mathcal{M}_{rep},\mathcal{M}_{RRR})} \left\| S_I - \hat{S}_I \right\|_1$$

where I is the input image, and S_I and $\hat{S}_I$ are the corresponding new and reference saliencies. Adding RRR to methods like EWC (Kirkpatrick et al. 2016) and iTAML (Rajasegaran et al. 2020) improves overall accuracy and prevents explanation drift, but it requires double the memory of traditional replay-based methods. However, RRR also performs well in FSCIL by storing only a few images per task along with their saliency maps. *Learning without Memorizing* (LwM) Dhar et al. (2019), instead, features a XAI-informed distillation loss, preventing explanation drift without storing replay data. At each incremental step t, a student model M_t learns new classes based on a teacher model M_{t-1},

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | (a) Average accuracy of SAMC and EPR on Split-CIFAR-100 (20 tasks) in the TIL scenario.

(a) Split-CIFAR-100 (20 tasks, TIL)			
	Buffer size	**# Runs**	**ACC (↑)**
Finetune	—	5	42.9 ± 2.1
SAMC	200	5	56.2 ± 1.0
EPR	85	5	58.5 ± 1.2
	170	5	60.8 ± 0.3
(b) Split-CUB-200 (10 tasks, CIL)			
	Buffer size	**# Runs**	**ACC (↑)**
Finetune	—	3	12.9 ± 1.7
LwM	—	3	29.4 ± 3.2
ICICLE	—	3	60.2 ± 3.5
RRR	85	5	62.9 ± 1.3
	170	5	67.1 ± 1.3
EPR	85	5	72.1 ± 0.9
	170	5	73.5 ± 1.3

Note: Results for SAMC are taken from Bai et al. (2023), while results for EPR and the finetune baseline are taken from Saha and Roy (2023b). (b) Average accuracy of LwM, ICICLE, RRR, and EPR on Split-CUB-200 (10 tasks, CIL). Results for LwM, ICICLE, and the finetune baseline are taken from Rymarczyk et al. (2023), while results for EPR and RRR are taken from Saha and Roy (2023b), to guarantee similar experimental setups when possible.

trained on N base classes. This is done using a combination of a classification loss L_C, to allow M_t to learn the new classes, a distillation loss L_D, to account for shared visual semantics between the new and old classes, and a XAI-informed distillation loss L_{LwM}, to more explicitly discourage drastic representation changes. Given an input image belonging to one of the new classes I_n, $Q_{t-1}^{I_n,c}$ and $Q_t^{I_n,c}$ are the vectorized attention maps of length l generated by M_{t-1} and M_t, respectively. Let b be the top base class predicted by M_t for I_n, the attention distillation is computed as:

$$L_{LwM} = \sum_{j=1}^{l} \left\| \frac{Q_{t-1,j}^{I_n,b}}{\| Q_{t-1}^{I_n,b} \|_2} - \frac{Q_{t,j}^{I_n,b}}{\| Q_t^{I_n,b} \|_2} \right\|_1$$

that is, as the sum of element wise $L1$ difference of the normalized, vectorized attention maps. This method requires storing parameters of both teacher and student models and is computationally expensive due to the need to compute saliency maps for all training images.

While the presented works use post hoc explainability methods, interpretable class-incremental learning (ICICLE) Rymarczyk et al. (2023) adapts ProtoPNet (Chen et al. 2019), a self-interpretable network, to the CIL scenario. To prevent forgetting, ICICLE uses an XAI-based similarity distillation loss to minimize changes in prototype similarities:

$$L_{ICICLE} = \sum_{i=0}^{H} \sum_{j=0}^{W} \left| sim\left(p^{t-1}, z_{i,j}^t\right) - sim\left(p^t, z_{i,j}^t\right) \right| \cdot S_{i,j}$$

where $sim\left(p^{t-1}, z_{i,j}^t\right)$ is computed for the model stored before training task t and S is a binary mask representing the γ quantile of the pixels with the highest similarity. Similarity is computed as $sim(p_i, z_j) = log\frac{|z_j - p_i|_2 + 1}{|z_j - p_i|_2 + \eta}$, with $\eta \ll 1$. Additionally, ICICLE equalizes logits of different tasks to reduce task recency bias. This approach outperforms LwM and other standard methods, like EWC (Kirkpatrick et al. 2016) and LwF (Li and Hoiem 2016), while reducing the explanation drift. Variants like ProtoPool (Rymarczyk et al. 2022) could further reduce memory needs by sharing prototypes among classes.

Table 6b reports results for LwM, ICICLE, RRR, and EPR on Split-CUB-200 in the CIL scenario. LwM provides the lowest improvement over the finetune baseline, while ICICLE nearly doubles performance by leveraging self-explainable prototypes and no replay buffer. Replay-based approaches further improve results: RRR outperforms ICICLE, but at the cost of doubling memory requirements, since it must store both raw images and saliency maps. EPR achieves the highest accuracy overall (73.5 % with buffer size 170), confirming that storing only salient patches offers an effective balance between interpretability and efficiency. Importantly, these differences reveal trade-offs across methods: regularization-only approaches like LwM scale poorly, prototype-based methods like ICICLE improve stability but are still limited, and replay-based approaches that integrate explanations (EPR, RRR) currently dominate in both accuracy and memory efficiency. An interesting future research direction would be to combine ICICLE with patch-based experience replay, similarly to EPR, to leverage the advantages of both approaches.

To avoid storing information about old tasks, the authors of Mazumder et al. (2023) propose a fully regularization-based method, XAI-Increment (XAI-I), for IIL. XAI-I performs multiple training passes, computes LIME (Ribeiro et al. 2016) explanations for the true and predicted classes of wrongly predicted samples, and augments the training set with these samples. Then, XAI-I uses a weighted loss, where weights are the distances between LIME explanations for the true and predicted classes, focusing the model on features relevant to the correct class and correcting wrong predictions.

Finally, among XAI-guided approaches, only one is applied to the medical domain and focuses on image segmentation. Shape and Semantics-based Selective Regularization (S3R) (Zhang et al. 2023) uses shape-based and semantic-based importance weights for selective regularization to avoid forgetting shape and semantic knowledge. XAI allows the visualization of the relevant features memorized in the identified weights. The proposed Importance Activation Mapping (IAM) highlights regions in the input image activated by weights with high importance scores. First, we need to transform parameter importance into kernel importance as:

$$\hat{\Omega}_{ij} = \frac{1}{N_j} \sum_{k=1}^{N_j} \Omega_{ijk}$$

where Ω_{ijk} is the importance of parameter k in kernel j of layer i. Then, we can compute a heatmap as the weighted sum of all the feature channels in layer i and applying ReLU to clip the negative values:

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

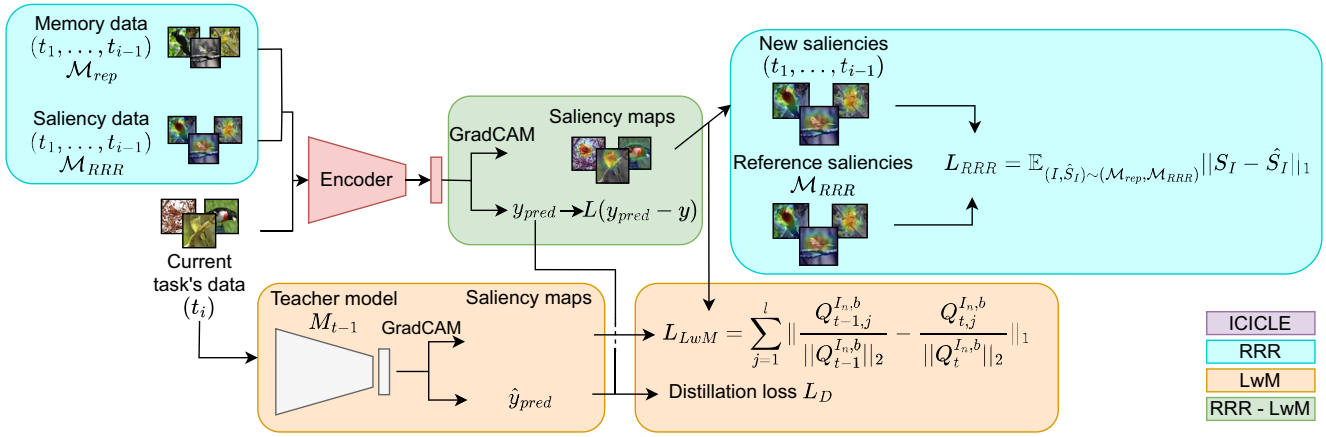

FIGURE 3 | A comparative overview of LwM (Dhar et al. 2019), RRR (Ebrahimi et al. 2021), and ICICLE (Rymarczyk et al. 2023). RRR is the only method among the three that implements memory replay, by storing both raw images ($\mathscr{M}_{rep}$) and the corresponding saliency maps ($\mathscr{M}_{RRR}$). RRR and LwM give current task's data as input to the classification model, which outputs predictions used to compute GradCAM explanations. RRR computes the regularization term L_{RRR}, by comparing the produced explanations with those stored in $\mathscr{M}_{RRR}$. In a similar way, LwM computes an attention distillation loss L_{LwM}, by comparing the saliency maps for current task's data to the ones produced by a teacher model, trained on tasks $1, 2, \ldots, t-1$, after receiving the same inputs as the student model. In the case of ICICLE, instead, current task's data is passed to the encoder and then projected into a prototypical part space by f_A, followed by a sigmoid activation function. Subsequently, each image part is compared to the prototypical parts of all tasks up to t by separate prototypical layers, which output similarity scores. These scores are used to compute the regularization term L_{ICICLE}, where $\text{sim}\left(p^{t-1}, z_{i,j}^{t}\right)$ is computed for the model stored before training task t. The similarity scores are also passed to separate fully-connected tasks for each task to obtain a prediction that is used to compute the classification loss.

$$S_i = \text{ReLU}\left(\sum_{j=1}^{N_i} \hat{\Omega}_{ij} \cdot A_{ij}\right)$$

with A_{ij} being the $j-$th feature channel in layer i. While S3R does not directly use heatmaps to guide training, it analyzes explanation drift and draws inspiration from cognitive neuroscience theories, such as the dissociated shape-semantic perception mechanism (Bracci and de Beeck 2015) and the cascade neurobiological model (Fusi et al. 2005), enhancing the understanding of human brain mechanisms.

3.3.3 | Specialized Network Components

Instead of focusing on relevant data or features, the approaches in this section identify critical network components to prevent catastrophic forgetting, as suggested by Räuker et al. (2023). Architecture-based methods, especially those relying on parameter specialization, benefit greatly from explainability. Tjoa and Guan (2022) introduce the semi-quantized activation neural network (SQANN), a self-interpretable model that adds a new neuron when an out-of-distribution sample is encountered. This is achieved using a selective activation function with three activation levels to guide neuron addition. Relevance-based neural freezing (RNF), on the other hand, (Ede et al. 2022) uses LRP (Bach et al. 2015) to compute importance scores for the network's units. Neurons are then ordered by importance, and an iterative pruning procedure removes the least relevant units. Once a threshold is reached (e.g., based on accuracy or available units), the remaining neurons are frozen for subsequent tasks. As evident from the visual summary in Figure 4, despite being effective, this approach suffers from limited scalability, since the network saturates early as the number of tasks grows. Similarly, it is unfeasible to keep adding neurons to SQANN as more out-of-distribution samples are encountered.

3.4 | Comparative Insights and Trade-Offs

When analyzed comparatively, the presented approaches reveal consistent trade-offs in memory footprint, scalability, reliance on attribution quality, and explanation stability. In this section, we analyze these trade-offs, together with failure modes and limitations of the presented approaches.

3.4.1 | Memory Footprint

Replay-based approaches such as EPR (Saha and Roy 2023b), SAMC (Bai et al. 2023), and ASER (Shim et al. 2021) require explicit storage of data or features. Saliency-guided replay can reduce the size of stored inputs (patches or masks), but memory usage still grows with the number of tasks. RRR (Ebrahimi et al. 2021), on the other hand, stores saliency maps in addition to raw samples, thus doubling the memory demand. Regularization-based methods like LwM (Dhar et al. 2019) and architecture-based approaches, such as RNF (Ede et al. 2022) and SQANN (Tjoa and Guan 2022), avoid storing data from old tasks, but generally perform worse than replay-based approaches or suffer from limited scalability due to either model saturation or capacity growth.

3.4.2 | Scalability

Replay methods are effective in small- to medium-scale benchmarks (Split-MNIST, Split-CIFAR), but scale poorly with many tasks or high-dimensional data, unless combined with strong

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

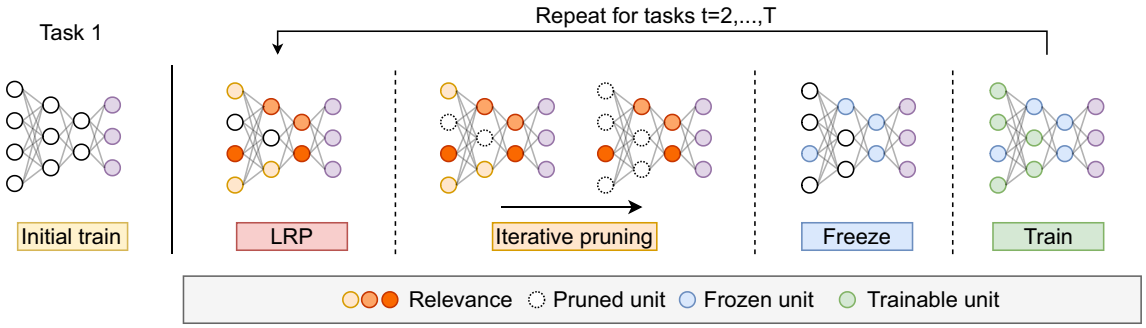

FIGURE 4 | Visual summary of RNF (Ede et al. 2022). After each task training, LRP is used to compute relevance scores for the network's neurons. Based on such scores, neurons are iteratively pruned starting from the least important ones until a certain threshold condition is reached. The remaining neurons are then frozen and are not used for subsequent tasks to avoid impacting the performance of the old ones.

compression, like in EPR (Saha and Roy 2023b) or SAMC (Bai et al. 2023). Regularization approaches scale better in terms of memory but often require computing explanations at each training step, which is expensive in large-scale or online CL. Architecture-based methods do not scale well with the number of tasks: RNF (Ede et al. 2022) saturates as neurons are progressively frozen, and SQANN (Tjoa and Guan 2022) grows unbounded with task diversity.

3.4.3 | Reliance on Explanation Quality

All surveyed approaches depend critically on the quality of explanations. Saliency-guided methods, such as EPR (Saha and Roy 2023b), SAMC (Bai et al. 2023), and LwM (Dhar et al. 2019), assume that attribution maps highlight semantically relevant features, but are vulnerable to explanation instability under distributional shift. Prototype-based methods like ICICLE (Rymarczyk et al. 2023), instead, embed explanations in their prediction process, reducing the risk of post hoc instability but at the expense of model flexibility and computational efficiency, since new prototypes must be added or updated as tasks accumulate. Replay-based strategies such as RRR (Ebrahimi et al. 2021) offer a compromise by enforcing explanation consistency over time, but they require substantial memory overhead to store both data and saliency maps.

3.4.4 | Summary

Overall, replay-based methods are intuitive and effective but face memory and scalability limits; regularization methods are memory-efficient but computationally costly due to the need for continuous explanation computation; and architecture-based methods, despite not using external memory buffers, risk saturation or unbounded growth. The choice of method thus reflects a trade-off between resource availability (memory vs. compute vs. capacity) and the desired degree of explanation stability. This comparative perspective shows both the promise and the limitations of current XAI-guided CL and motivates research into hybrid approaches that combine selective replay, stability regularization, and lightweight self-interpretable components while also addressing robustness and causality.

3.5 | Limitations

Beyond these trade-offs, XAI-guided CL faces several foundational challenges that limit its practical reliability. First, explanations are often *unstable under distributional shift*: saliency maps or prototypes for the same input can vary drastically when the model is fine-tuned on new tasks, making it difficult to determine whether drift reflects meaningful adaptation or harmful forgetting. Second, most post hoc attribution methods are *non-causal*: they highlight correlations in the learned features rather than causal drivers of the prediction. As a result, relying on such explanations to guide memory or parameter selection risks amplifying spurious correlations present in the data. Third, explanation signals themselves can be *volatile*: gradients and relevance maps may change significantly across training runs due to stochastic optimization, introducing additional noise into the guidance process. The use of self-interpretable models partially addresses these issues by providing causal explanations that directly affect predictions, which represent a more trustworthy driver of continual learning.

4 | Future Directions

The application of XAI to CL is a relatively new research direction with few papers, leaving much room for advancement. Here, we outline some promising research lines.

4.1 | Task-Free Continual Learning and Few-Shot Class-Incremental Learning

The goal of CL research is to design AI systems that can incrementally learn new skills by adapting to the surrounding environment.

This requires considering task-free CL, where task identities are unknown during training and testing. Prototype-based networks could be used to automatically infer task identities by leveraging similarities to the learned prototypes. Some works have already attempted to do so by learning prototypical latent spaces through representation learning techniques. For instance, Adel (2024) uses an auxiliary prototypical

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

network (Snell et al. 2017) to ensure learning a latent space in which classes belonging to all seen tasks are clearly distinguishable. If the prototypical network identifies the same task as the most similar to the previous and current data streams, the authors assume that there is no task change. In this way, they manage to perform automatic identification of task boundaries. However, prototypical networks do not directly provide insights into why they make certain predictions, which are crucial in critical applications, such as healthcare. Contrarily, prototype-based models (Chen et al. 2019; Rymarczyk et al. 2022; Nauta et al. 2023), by learning interpretable prototypes, allow a double benefit: generalization and explainability. In particular, PIP-Net (Nauta et al. 2023) can already recognize out-of-distribution data and, if applied to CL scenarios, it would allow automatic task identification without requiring additional assumptions.

Another big challenge in real-world applications, such as robotics and healthcare, is limited data availability due to constraints such as real-time requirements or privacy concerns. FSCIL research addresses this by developing models that can learn new tasks from a few samples. Ebrahimi et al. (2021) already demonstrated that the use of XAI in this context allows leveraging the knowledge about previous tasks to effectively and efficiently learn new ones from a few samples. The use of self-interpretable models could further enhance FSCIL by building semantically meaningful latent spaces that allow faster learning of new tasks based on previously acquired knowledge. We provide more details in the next section on self-interpretable models.

4.1.1 | Self-Interpretable Models

Most XAI-guided CL approaches leverage post hoc explainability methods, thus needing an intermediate step between task trainings to compute explanations. This is time- and computationally expensive and requires knowing when a task's training ends. For this reason, post hoc explanations are unsuitable for complex scenarios like OCL, BBCL, and FSCIL.

For instance, using a self-interpretable model for FSCIL would allow us to exploit the learned prototypes, logic rules, or concepts to learn a new task from a few samples by leveraging similarities with previous tasks. In Figure 5, we consider a medical image classification system trained to recognize various skin conditions (e.g., melanoma, eczema, psoriasis). Initially, it learns to classify a set of common conditions using a self-interpretable model that builds prototypes (representative feature vectors) or logic rules based on image features (color, texture, shape) and metadata (e.g., patient age, location of the lesion). Then, when a new and rare skin condition is discovered, with only a handful of annotated images available, a self-interpretable model allows comparing the new condition's samples with existing prototypes (e.g., lesions with similar color patterns or shapes) or applying existing logic rules (e.g., "If the lesion is asymmetrical and has irregular borders, then it might be malignant") to make preliminary classifications. This similarity-based reasoning enables faster and more reliable generalization from very few examples, while still being interpretable to dermatologists. Such an approach enhances trust and adoption in high-stakes fields like healthcare, where interpretability and data scarcity are both critical. Continual learning is especially important because medical data is often distributed over time, with new diseases, imaging modalities, or diagnostic protocols continuously emerging. Unlike traditional XAI, which helps interpret a model's decisions at a fixed point in time, XAI-guided CL supports ongoing model adaptation while maintaining transparency. For example, a model trained to detect common lung conditions from chest X-rays may later need to incorporate knowledge about rare or newly discovered diseases without retraining from scratch or losing previous capabilities. Here, XAI can help identify which visual concepts, such as texture anomalies or shape patterns, are reused or modified during task transitions, enabling human experts to validate the learning process.

4.1.2 | Application Domains

All existing works propose methods to be used in image classification tasks, with the exception of SQANN (Tjoa and Guan 2022), which works on tabular data. However, CF is common to all types of architectures and learning mechanisms, thus creating the need for adapting the existing methods and developing new ones for other application domains, such as reinforcement learning or graph data.

For example, Graph Concept Whitening (GCW) (Proietti et al. 2024) is a technique used to turn a black-box graph neural network into a self-interpretable model, and it was used in drug discovery to disentangle and align latent representations with human-interpretable chemical concepts, such as aromaticity or polarity. In a continual learning setting, where new drug-related tasks (e.g., predicting toxicity, solubility, or binding affinity) are introduced over time, GCW might allow the model to preserve and reuse the previously learned latent space where such concepts are disentangled. For instance, if a model has already learned a concept associated with non-polar groups influencing solubility, it can leverage this to more quickly adapt to a new task involving drug permeability, where similar features are relevant. This interpretability-aware approach enhances both transfer learning and the explainability and reliability of predictions, which are crucial in high-stakes domains like drug development.

Future research should explore how concept-level alignment across tasks can be visualized and tracked over time, how to quantify the transferability of interpretable features, and how to detect when new concepts are needed. Developing benchmarks and evaluation protocols tailored to such dynamic, interpretable settings would be key to advancing the field.

Another promising application domain is continual reinforcement learning (CRL). RL represents indeed the most natural application for CL, as agents typically interact with dynamic environments and must continuously adapt by learning new skills (Khetarpal et al. 2022). Integrating self-interpretable models into continual RL could enable agents to explicitly reason about their decision-making processes, thereby improving their adaptability and facilitating human oversight and validation, for instance via human-understandable concepts (Dai et al. 2022). Nonetheless, this approach faces open challenges, mostly related to the definition of stable and transferable concepts. Prototype-based

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

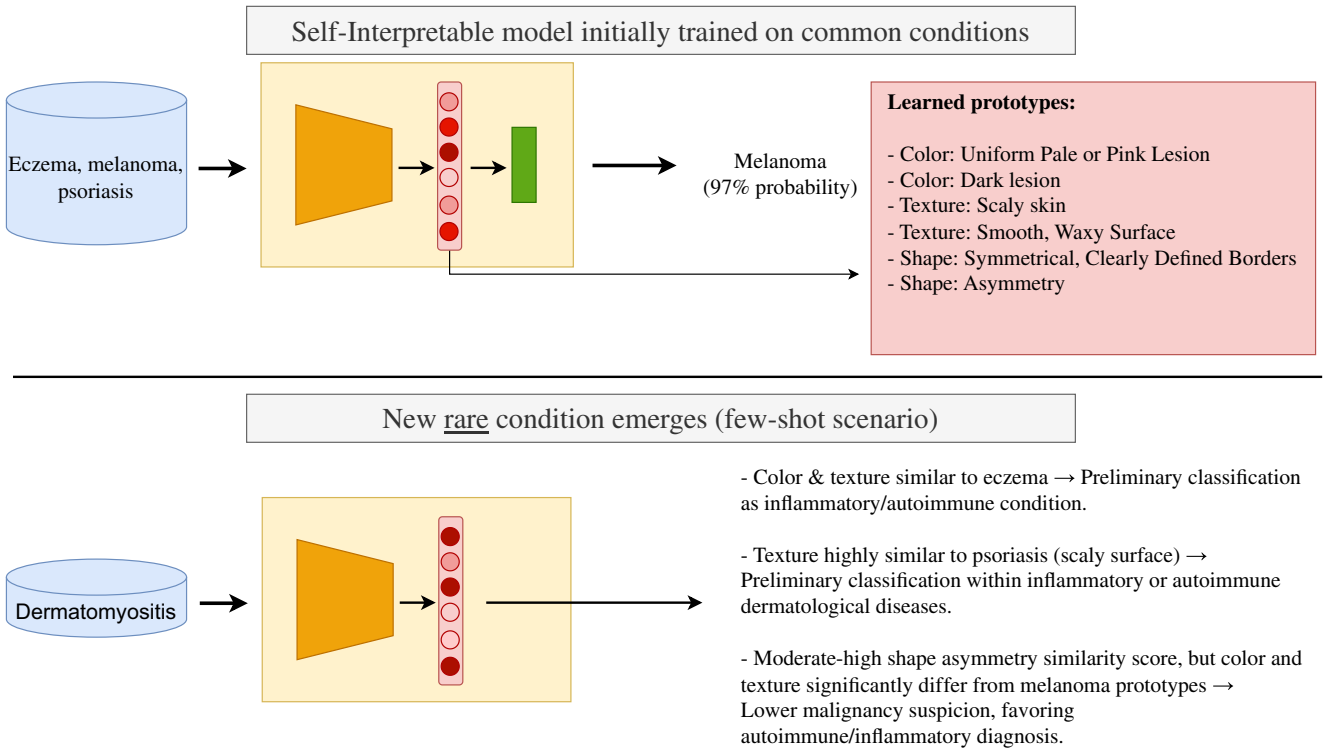

FIGURE 5 | Illustrative example of a self-interpretable prototype-based model applied to incremental medical diagnosis. The model builds human-understandable prototypes from skin lesion images (e.g., color, texture, shape) and uses them to classify conditions such as melanoma or eczema. When new and rare conditions appear, the model can relate them to previously learned prototypes, enabling faster generalization from few samples while maintaining interpretability for clinicians. This example highlights how self-interpretable approaches can support CL in high-stakes domains, where data are scarce, new tasks emerge over time, and transparency is critical for trust and adoption.

approaches might therefore be more suited in situations in which concept selection is non-trivial. However, most of the current prototype-based approaches are post hoc, as they are interpretable versions of pre-trained black-box agents (Ragodos et al. 2022; Borzillo et al. 2023; Kenny et al. 2023).

Future work should try to develop methods that integrate prototype learning directly into the training process, thus achieving intrinsic interpretability and reducing reliance on post hoc explanations. Additionally, research should address how interpretable RL agents can efficiently identify, refine, and merge concepts across multiple tasks, particularly in complex or partially observable environments. A seminal work in this area is the one by Proietti et al. (2025), which leverages prototype-based reasoning for automatic event-guided replay, shows promise in the application of self-interpretable architectures in CRL. Finally, designing metrics for assessing the stability and adaptability of learned prototypes or concepts, along with developing standardized evaluation benchmarks for interpretable CRL, would greatly facilitate progress in this domain.

4.1.3 | Open-Source Libraries

A recurring limitation in this area is the lack of standardized implementations and benchmarks, which makes it difficult to compare methods fairly. To facilitate progress, future work should focus on unifying experimental setups and providing open, reproducible frameworks. As a first step in this direction, we have released *XAI4CL*, an open-source repository built on top of Avalanche (Carta et al. 2023), which integrates several XAI-guided CL methods within a common interface.[2]

Avalanche is a modular and extensible PyTorch-based library tailored for CL, developed with a focus on reproducibility, scalability, and ease of experimentation. It is structured around five core modules: *benchmarks*, *training*, *models*, *evaluation*, and *logging*, each supporting different stages of a CL pipeline. The *benchmark* module facilitates the definition and manipulation of CL scenarios. It provides both standard and custom benchmarks by organizing data into streams and experiences, representing sequential learning tasks. This enables the flexible simulation of scenarios like TIL, CIL, and DIL learning. The *training* module offers a suite of predefined strategies and supports the construction of hybrid approaches by combining multiple techniques. Central to this design is a plugin-based architecture, which allows researchers to inject additional behavior into training loops without modifying core strategy implementations. In this context, we integrate XAI-guided CL methods as *strategy plugins* implementing attribution-based regularizers, replay buffers with saliency selection, or other explanation-driven components which can be used with the *Naive* baseline (i.e., standard sequential training), in combination with other compatible strategies (e.g., EWC). The *model* module introduces support for dynamic architectures, allowing the network structure to evolve over time, a key requirement in lifelong learning. Multi-head classifiers, progressive networks, and other adaptive architectures are readily supported. For performance tracking, the *evaluation* module includes an extensive set of metrics, ranging from accuracy

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to memory usage, and supports both standalone and plugin-based usage. These metrics can be visualized or stored using the *logging* module.

Finally, Avalanche's design philosophy emphasizes composability. Strategies are built atop reusable templates, and the plugin interface allows XAI components to seamlessly interact with internal states of the learning process, such as modifying the loss function before each update or altering the data stream based on explanation scores. This flexibility has been key in enabling our unified implementation of XAI-guided CL methods within the XAI4CL repository. Although still a work in progress, XAI4CL encourages contributions and lays the foundation for more systematic empirical studies on XAI-guided CL.

4.1.4 | Study of the Causes of Forgetting

Toneva et al. (2019) study forgetting events in neural networks trained on single image classification tasks. They show that the memorability of input samples is independent of the architecture and learning algorithm, and that (un)forgettability depends on intrinsic properties of the samples.

Li et al. (2023) extend this analysis to CL settings using a visual analytics framework. Using filter-wise input attributions, they identify features causing sample misclassification. This work represents an early example of how XAI can facilitate this type of analysis. However, how new tasks interfere with previously learned representations remains unclear. Are concepts represented within the network overwritten by new knowledge? If learned concepts relevant to old tasks are preserved, are they still correctly accessed or do the patterns of activations for old tasks change? These are some of the intriguing questions XAI could provide an answer for.

Another important question that remains open is whether explanation drift is only detrimental to the model's performance on old tasks. Driscoll et al. (2022) hypothesize that representational drift is needed in the brain to correctly learn new information and that forgetting is avoided by achieving a balance with opposite mechanisms, such as memory replay, that allow preserving old knowledge. It would be interesting to find a way to quantify the correlation between performance drop and explanation drift, and to understand whether explanation drift could be used as an early indicator of catastrophic forgetting.

Understanding when and how forgetting occurs will indeed help build robust shared representations, maintaining high performance on old tasks while learning new ones efficiently. Additionally, combining XAI techniques with metrics that identify the most informative samples early in training (Paul et al. 2021) can make the process faster and computationally cheaper.

4.1.5 | Bridging the Gap With Neuroscience

The brain has always been a source of inspiration for CL approaches because of its ability to learn continuously over a long time span without ever forgetting catastrophically, in the absence of pathological conditions. Zhuang et al. (2022) present a life-long visual learning benchmark containing continuous visual experiences from children over several years of their development, captured from a naturalistic, first-person perspective. Using such data, the authors evaluate the performance of several ANNs trained sequentially on these video segments, reflecting a human-like visual learning trajectory across developmental timescales. Specifically, models are assessed on their ability to continuously learn visual features without catastrophic forgetting, which is tested periodically through classification tasks on MiniImageNet. The results showed that algorithms that performed best on popular CL benchmarks were actually outperformed by earlier approaches on this new benchmark. This discrepancy was mainly due to newer algorithms' inability to effectively leverage sparse and low-diversity visual experiences, implying that mechanisms such as negative sampling are essential to emulate human-like continual learning.

Beyond highlighting performance differences, this benchmark opens intriguing questions about representational alignment between BNNs and ANNs. Indeed, there is currently an active area of research studying the alignment between biological and artificial representations, aiming to enhance both the interpretability and effectiveness of ANNs by drawing inspiration from BNNs, and vice versa (Caucheteux and King 2022; Goldstein et al. 2022; Cadieu et al. 2014; Toneva and Wehbe 2019). To the best of the authors' knowledge, no existing work has explicitly explored this alignment within continual learning scenarios. Employing XAI methods in this context would enable researchers to identify representational differences and commonalities between biological and artificial systems, thus providing insights into the cognitive mechanisms underlying continual learning in both cases.

5 | Conclusions

CL is an open problem in AI research. Although the literature on the topic is wide, just a few methods exploit XAI. This work surveys existing XAI-guided CL approaches, organizing them in a taxonomy based on what is explained (data, features, network's components). We also provide open-source implementations of some of the presented approaches using Avalanche (Carta et al. 2023), encouraging external contributions. Finally, we present possible future lines of research to encourage further research on the topic.

Author Contributions

Michela Proietti: conceptualization (equal), data curation (lead), formal analysis (lead), methodology (equal), software (lead), writing – original draft (lead), writing – review and editing (equal). **Alessio Ragno:** conceptualization (equal), methodology (equal), writing – original draft (equal), writing – review and editing (equal). **Roberto Capobianco:** conceptualization (equal), methodology (equal), writing – original draft (equal), writing – review and editing (equal).

Acknowledgments

This work has been carried out while Michela Proietti and Alessio Ragno were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. Open access publishing facilitated by Universita degli Studi di Roma La Sapienza, as part of the Wiley - CRUI-CARE agreement.

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Related WIREs Articles

Explainable artificial intelligence and machine learning: A reality rooted perspective

Explainable artificial intelligence: an analytical review

The role of lifelong machine learning in bridging the gap between human and machine learning: A scientometric analysis

Endnotes

[1] Extensions from a neuron's cell body that receive signals from other neurons.

[2] All the code is available at https://github.com/KRLGroup/XAI4CL.

References

Adel, T. 2024. "Similarity-Based Adaptation for Task-Aware and Task-Free Continual Learning." *Journal of Artificial Intelligence Research* 80: 377–417.

Aljundi, R., F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. 2018. "Memory Aware Synapses: Learning What (not) to Forget." In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, 144–161. Springer-Verlag. https://doi.org/10.1007/978-3-030-01219-9_9.

Aljundi, R., E. Belilovsky, T. Tuytelaars, et al. 2019. "Online Continual Learning With Maximal Interfered Retrieval." In *Advances in Neural Information Processing Systems, Volume 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates Inc.

Aljundi, R., M. Lin, B. Goujaud, and Y. Bengio. 2019. "Gradient Based Sample Selection for Online Continual Learning." In *Advances in Neural Information Processing Systems, Volume 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates Inc.

Asadi, N., M. Davar, S. P. Mudur, R. Aljundi, and E. Belilovsky. 2023. "Prototype-Sample Relation Distillation: Towards Replay-Free Continual Learning." In *International Conference on Machine Learning*, 1093–1106. PMLR.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. 2015. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PLoS One* 10: e0130140.

Bai, G., C. Ling, Y. Gao, and L. Zhao. 2023. *Saliency-Augmented Memory Completion for Continual Learning*, 244–252. Society for Industrial and Applied Mathematics.

Bainbridge, W. A., D. D. Dilks, and A. Oliva. 2017. "Memorability: A Stimulus-Driven Perceptual Neural Signature Distinctive From Memory." *NeuroImage* 149: 141–152.

Bainbridge, W. A., P. Isola, and A. Oliva. 2013. "The Intrinsic Memorability of Face Photographs." *Journal of Experimental Psychology. General* 142, no. 4: 1323–1334.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (Xai): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Bell, N., and M. Garland. 2009. "Implementing Sparse Matrix-Vector Multiplication on Throughput-Oriented Processors." In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, 1–11.

Blakeman, S., and D. Mareschal. 2020. "A Complementary Learning Systems Approach to Temporal Difference Learning." *Neural Networks* 122: 218–230.

Borzillo, C., A. Ragno, and R. Capobianco. 2023. "Understanding Deep Rl Agent Decisions: A Novel Interpretable Approach With Trainable Prototypes." In *XAI.it@AI*IA*. CEUR Workshop Proceedings.

Bracci, S., and H. P. O. de Beeck. 2015. "Dissociations and Associations Between Shape and Category Representations in the Two Visual Pathways." *Journal of Neuroscience* 36: 432–444.

Cadieu, C. F., H. Hong, D. L. Yamins, et al. 2014. "Deep Neural Networks Rival the Representation of Primate It Cortex for Core Visual Object Recognition." *PLoS Computational Biology* 10, no. 12: e1003963.

Carta, A., L. Pellegrini, A. Cossu, H. Hemati, and V. Lomonaco. 2023. "Avalanche: A Pytorch Library for Deep Continual Learning." *Journal of Machine Learning Research* 24, no. 363: 1–6.

Caucheteux, C., and J.-R. King. 2022. "Brains and Algorithms Partially Converge in Natural Language Processing." *Communications Biology* 5, no. 1: 134.

Chattopadhay, A., A. Sarkar, P. Howlader, and V. N. Balasubramanian. 2018. "Grad-Cam++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847. IEEE.

Chen, C., O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. 2019. "This Looks Like That: Deep Learning for Interpretable Image Recognition." *Advances in Neural Information Processing Systems* 32.

Chi, Z., L. Gu, H. Liu, Y. Wang, Y. Yu, and J. Tang. 2022. "Metafscil: A Meta-Learning Approach for Few-Shot Class Incremental Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14166–14175. CVPR conference proceedings.

Cichon, J., and W.-B. Gan. 2015. "Branch-Specific Dendritic Ca2+ Spikes Cause Persistent Synaptic Plasticity." *Nature* 520, no. 7546: 180–185.

Cossu, A., F. Spinnato, R. Guidotti, and D. Bacciu. 2024. "Drifting Explanations in Continual Learning." *Neurocomputing* 597: 127960.

Dai, Y., H. Ouyang, H. Zheng, H. Long, and X. Duan. 2022. "Interpreting a Deep Reinforcement Learning Model With Conceptual Embedding and Performance Analysis." *Applied Intelligence* 53, no. 6: 6936–6952.

Dhar, P., R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa. 2019. "Learning Without Memorizing." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR conference proceedings.

Driscoll, L. N., L. Duncker, and C. D. Harvey. 2022. "Representational Drift: Emerging Theories for Continual Learning and Experimental Future Directions." *Current Opinion in Neurobiology* 76: 102609.

Ebrahimi, S., F. Meier, R. Calandra, T. Darrell, and M. Rohrbach. 2020. "Adversarial Continual Learning." In *Computer Vision – ECCV 2020*, edited by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, 386–402. Springer International Publishing.

Ebrahimi, S., S. Petryk, A. Gokul, et al. 2021. "Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting." *Applied AI Letters* 2, no. 4: e44.

Ede, S., S. Baghdadlian, L. Weber, et al. 2022. "Explain to Not Forget: Defending Against Catastrophic Forgetting With Xai." In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 1–18. Springer International Publishing.

Farajtabar, M., N. Azizan, A. Mott, and A. Li. 2020. "Orthogonal Gradient Descent for Continual Learning." In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics,*

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Volume 108 of Proceedings of Machine Learning Research, edited by S. Chiappa and R. Calandra, 3762–3773. PMLR.

Fei-Fei, L., R. Fergus, and P. Perona. 2006. "One-Shot Learning of Object Categories." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, no. 4: 594–611.

Frankland, P. W., and B. Bontempi. 2005. "The Organization of Recent and Remote Memories." *Nature Reviews Neuroscience* 6, no. 2: 119–130.

Fusi, S., P. J. Drew, and L. F. Abbott. 2005. "Cascade Models of Synaptically Stored Memories." *Neuron* 45: 599–611.

Gao, Y., S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao. 2024. "Going Beyond Xai: A Systematic Survey for Explanation-Guided Learning." *ACM Computing Surveys* 56, no. 7: 1–39.

Goldstein, A., Z. Zada, E. Buchnik, et al. 2022. "Shared Computational Principles for Language Processing in Humans and Deep Language Models." *Nature Neuroscience* 25, no. 3: 369–380.

Goodfellow, I. J., M. Mirza, X. Da, A. C. Courville, and Y. Bengio. 2013. "An Empirical Investigation of Catastrophic Forgeting in Gradient-Based Neural Networks." *CoRR*, abs/1312.6211.

Gu, Y., X. Yang, K. Wei, and C. Deng. 2022. "Not Just Selection, but Exploration: Online Class-Incremental Continual Learning via Dual View Consistency." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7442–7451. CVPR conference proceedings.

Huo, F., W. Xu, J. Guo, H. Wang, and Y. Fan. 2023. "Non-Exemplar Online Class-Incremental Continual Learning via Dual-Prototype Self-Augment and Refinement." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 11: 12698–12707.

Isola, P., D. Parikh, A. Torralba, and A. Oliva. 2011. "Understanding the Intrinsic Memorability of Images." In *Advances in Neural Information Processing Systems, Volume 24*, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Curran Associates Inc.

Isola, P., J. Xiao, D. Parikh, A. Torralba, and A. Oliva. 2014. "What Makes a Photograph Memorable?" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, no. 7: 1469–1482.

Isola, P., J. Xiao, A. Torralba, and A. Oliva. 2011. "What Makes an Image Memorable?" In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 145–152. CVPR conference proceedings.

Javed, K., and M. White. 2019. "Meta-Learning Representations for Continual Learning." In *Advances in Neural Information Processing Systems, 32*. NeurIPS conference proceedings.

Kenny, E. M., M. Tucker, and J. Shah. 2023. "Towards Interpretable Deep Reinforcement Learning With Human-Friendly Prototypes." In *The Eleventh International Conference on Learning Representations*. ICLR conference proceedings.

Khetarpal, K., M. Riemer, I. Rish, and D. Precup. 2022. "Towards Continual Reinforcement Learning: A Review and Perspectives." *Journal of Artificial Intelligence Research* 75: 1401–1476.

Khosla, A., A. S. Raju, A. Torralba, and A. Oliva. 2015. "Understanding and Predicting Image Memorability at a Large Scale." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. ICCV conference proceedings.

Kirkpatrick, J., R. Pascanu, N. C. Rabinowitz, et al. 2016. "Overcoming Catastrophic Forgetting in Neural Networks." *Proceedings of the National Academy of Sciences of the United States of America* 114: 3521–3526.

Kukleva, A., H. Kuehne, and B. Schiele. 2021. "Generalized and Incremental Few-Shot Learning by Explicit Learning and Calibration Without Forgetting." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9020–9029. ICCV conference proceedings.

Kumaran, D., D. Hassabis, and J. L. McClelland. 2016. "What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated." *Trends in Cognitive Sciences* 20, no. 7: 512–534.

Laux, J., S. Wachter, and B. Mittelstadt. 2024. "Trustworthy Artificial Intelligence and the European Union Ai Act: On the Conflation of Trustworthiness and Acceptability of Risk." *Regulation & Governance* 18, no. 1: 3–32.

Lee, S., S. Goldt, and A. Saxe. 2021. "Continual Learning in the Teacher-Student Setup: Impact of Task Similarity." In *Proceedings of the 38th International Conference on Machine Learning, Volume 139 of Proceedings of Machine Learning Research*, edited by M. Meila and T. Zhang, 6109–6119. PMLR.

Li, Z., and D. Hoiem. 2016. "Learning Without Forgetting." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40: 2935–2947.

Li, Z., J. Xu, W.-L. Chao, and H.-W. Shen. 2023. "Visual Analytics on Network Forgetting for Task-Incremental Learning." *Computer Graphics Forum* 42, no. 3: 437–448.

Lin, S., P. Ju, Y. Liang, and N. Shroff. 2023. "Theory on Forgetting and Generalization of Continual Learning." In *Proceedings of the 40th International Conference on Machine Learning, Volume 202 of Proceedings of Machine Learning Research*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, 21078–21100. PMLR.

Liu, H., L. Gu, Z. Chi, et al. 2022. "Few-Shot Class-Incremental Learning via Entropy-Regularized Data-Free Replay." In *European Conference on Computer Vision*, 146–162. Springer Nature Switzerland.

Lopez-Paz, D., and M. Ranzato. 2017. "Gradient Episodic Memory for Continual Learning." In *Advances in Neural Information Processing Systems, 30*. NeurIPS conference proceedings.

Lüscher, C., and R. C. Malenka. 2012. "Nmda Receptor-Dependent Long-Term Potentiation and Long-Term Depression (Ltp/Ltd)." *Cold Spring Harbor Perspectives in Biology* 4, no. 6: a005710.

Madaan, D., J. Yoon, Y. Li, Y. Liu, and S. J. Hwang. 2022. "Representational Continuity for Unsupervised Continual Learning." Preprint, arXiv:2110.06976, October 13.

Mareschal, D., M. H. Johnson, S. Sirois, M. Spratling, M. S. Thomas, and G. Westermann. 2007. *Neuroconstructivism-I: How the Brain Constructs Cognition*. Oxford University Press.

Mathur, R., T. Chintala, and D. Rajeswari. 2022. "Identification of Illicit Activities & Scream Detection Using Computer Vision & Deep Learning." In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1243–1250. IEEE.

Mazumder, A. N., N. Lyons, A. Pandey, A. Santra, and T. Mohsenin. 2023. "Harnessing the Power of Explanations for Incremental Training: A Lime-Based Approach." In *2023 31st European Signal Processing Conference (EUSIPCO)*, 1–5. IEEE.

McClelland, J. L., B. L. McNaughton, and R. C. O'Reilly. 1995. "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory." *Psychological Review* 102, no. 3: 419–457.

McCloskey, M., and N. Cohen. 1989. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem." *Psychology of Learning and Motivation - Advances in Research and Theory* 24: 109–165.

Mermillod, M., A. Bugaiska, and P. Bonin. 2013. "The Stability-Plasticity Dilemma: Investigating the Continuum From Catastrophic Forgetting to Age-Limited Learning Effects." *Frontiers in Psychology* 4: 504.

Nauta, M., J. Schlötterer, M. van Keulen, and C. Seifert. 2023. "Pip-Net: Patch-Based Intuitive Prototypes for Interpretable Image

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Classification." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. CVPR conference proceedings.

Ostapenko, O., P. Rodriguez, M. Caccia, and L. Charlin. 2021. "Continual Learning via Local Module Composition." In *Advances in Neural Information Processing Systems, Volume 34*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, 30298–30312. Curran Associates Inc.

Pallier, C., S. Dehaene, J.-B. Poline, et al. 2003. "Brain Imaging of Language Plasticity in Adopted Adults: Can a Second Language Replace the First?" *Cerebral Cortex* 13, no. 2: 155–161.

Panigutti, C., R. Hamon, I. Hupont, et al. 2023. "The Role of Explainable Ai in the Context of the AI Act." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1139–1150. ACM conference proceedings.

Paul, M., S. Ganguli, and G. K. Dziugaite. 2021. "Deep Learning on a Data Diet: Finding Important Examples Early in Training." In *Advances in Neural Information Processing Systems, Volume 34*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, 20596–20607. Curran Associates Inc.

Pham, Q., C. Liu, and S. Hoi. 2021. "Dualnet: Continual Learning, Fast and Slow." In *Advances in Neural Information Processing Systems*, edited by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. NeurIPS conference proceedings.

Power, J. D., and B. L. Schlaggar. 2017. "Neural Plasticity Across the Lifespan." In *Wiley Interdisciplinary Reviews: Developmental Biology, 6*. Wiley Interdisciplinary Reviews.

Proietti, M., A. Ragno, B. L. Rosa, R. Ragno, and R. Capobianco. 2024. "Explainable Ai in Drug Discovery: Self-Interpretable Graph Neural Network for Molecular Property Prediction Using Concept Whitening." *Machine Learning* 113, no. 4: 2013–2044.

Proietti, M., P. R. Wurman, P. Stone, and R. Capobianco. 2025. "Protocrl: Prototype-Based Network for Continual Reinforcement Learning." In *Reinforcement Learning Conference*. RLC conference proceedings.

Ragodos, R., T. Wang, Q. Lin, and X. Zhou. 2022. "Explaining a Reinforcement Learning Agent via Prototyping." In *Advances in Neural Information Processing Systems*, edited by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. NeurIPS conference proceedings.

Rajasegaran, J., S. H. Khan, M. Hayat, F. S. Khan, and M. Shah. 2020. "Itaml: An Incremental Task-Agnostic Meta-Learning Approach." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13585–13594. CVPR conference proceedings.

Ramasesh, V. V., E. Dyer, and M. Raghu. 2021. "Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics." Preprint, arXiv:2007.07400.

Räuker, T., A. Ho, S. Casper, and D. Hadfield-Menell. 2023. "Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SATML)*, 464–483. IEEE.

Rebuffi, S.-A., A. Kolesnikov, G. Sperl, and C. H. Lampert. 2016. "ICARL: Incremental Classifier and Representation Learning." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5533–5542. CVPR conference proceedings.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD conference proceedings.

Riemer, M., I. Cases, R. Ajemian, et al. 2019. "Learning to Learn Without Forgetting by Maximizing Transfer and Minimizing Interference." Preprint, arXiv:1810.11910.

Rudoy, J. D., J. L. Voss, C. E. Westerberg, and K. A. Paller. 2009. "Strengthening Individual Memories by Reactivating Them During Sleep." *Science* 326, no. 5956: 1079.

Rymarczyk, D., Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński. 2022. "Interpretable Image Classification With Differentiable Prototypes Assignment." In *European Conference on Computer Vision*, 351–368. Springer Nature Switzerland.

Rymarczyk, D., J. van de Weijer, B. Zieliński, and B. Twardowski. 2023. "ICICLE: Interpretable Class Incremental Continual Learning." In *European Conference on Computer Vision*, 351–368. Springer Nature Switzerland.

Saha, G., and K. Roy. 2023a. "Online Continual Learning With Saliency-Guided Experience Replay Using Tiny Episodic Memory." *Machine Vision and Applications* 34, no. 4: 65.

Saha, G., and K. Roy. 2023b. "Saliency Guided Experience Packing for Replay in Continual Learning." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5273–5283. IEEE/CVF WCACV conference proceedings.

Schwalbe, G., and B. Finzel. 2023. "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts." In *Data Mining and Knowledge Discovery*, 1–59. Springer Nature.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Gradcam: Visual Explanations From Deep Networks via Gradient-Based Localization." In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. ICCV conference proceedings.

Shapley, L. S. 1953. "A Value for N-Person Games."

Sharma, V., M. Gupta, A. Kumar, and D. Mishra. 2021. "Video Processing Using Deep Learning Techniques: A Systematic Literature Review." *IEEE Access* 9: 139489–139507.

Shim, D., Z. Mai, J. Jeong, S. Sanner, H. Kim, and J. Jang. 2021. "Online Class-Incremental Continual Learning With Adversarial Shapley Value." *Proceedings of the AAAI Conference on Artificial Intelligence* 35, no. 11: 9630–9638.

Shin, H., J. K. Lee, J. Kim, and J. Kim. 2017. "Continual Learning With Deep Generative Replay." In *Advances in Neural Information Processing Systems*. NeurIPS conference proceedings.

Skaggs, W. E., and B. L. McNaughton. 1996. "Replay of Neuronal Firing Sequences in Rat Hippocampus During Sleep Following Spatial Experience." *Science* 271, no. 5257: 1870–1873.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. 2017. "Smoothgrad: Removing Noise by Adding Noise." Preprint, arXiv:1706.03825.

Snell, J., K. Swersky, and R. Zemel. 2017. "Prototypical Networks for Few-Shot Learning." In *Advances in Neural Information Processing Systems, Volume 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, et al. Curran Associates Inc.

Speith, T. 2022. "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. ACM conference proceedings.

Srinivas, S., and F. Fleuret. 2019. "Full-Gradient Representation for Neural Network Visualization." In *Advances in Neural Information Processing Systems, Volume 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates Inc.

Srivastava, S., M. Yaqub, K. Nandakumar, Z. Ge, and D. Mahapatra. 2021. "Continual Domain Incremental Learning for Chest X-Ray Classification in Low-Resource Clinical Settings." In Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3, pages 226–238. Springer.

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Stoianov, I., D. Maisto, and G. Pezzulo. 2022. "The Hippocampal Formation as a Hierarchical Generative Model Supporting Generative Replay and Continual Learning." *Progress in Neurobiology* 217: 102329.

Tao, X., X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong. 2020. "Few-Shot Class-Incremental Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR conference proceedings.

Tian, S., L. Li, W. Li, H. Ran, X. Ning, and P. Tiwari. 2024. "A Survey on Few-Shot Class-Incremental Learning." *Neural Networks* 169: 307–324.

Tjoa, E., and C. Guan. 2022. "Two Instances of Interpretable Neural Network for Universal Approximations." Preprint, arXiv:2112.15026.

Toldo, M., and M. Ozay. 2022. "Bring Evanescent Representations to Life in Lifelong Class Incremental Learning." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16711–16720. CVPR conference proceedings.

Toneva, M., A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon. 2019. "An Empirical Study of Example Forgetting During Deep Neural Network Learning." Preprint, arXiv:1812.05159.

Toneva, M., and L. Wehbe. 2019. "Interpreting and Improving Natural-Language Processing (in Machines) With Natural Language-Processing (in the Brain)." In *Advances in Neural Information Processing Systems, 32*. NeurIPS conference proceedings.

Van de Ven, G. M., H. T. Siegelmann, and A. S. Tolias. 2020. "Brain-Inspired Replay for Continual Learning With Artificial Neural Networks." *Nature Communications* 11, no. 1: 4069.

van de Ven, G. M., T. Tuytelaars, and A. S. Tolias. 2022. "Three Types of Incremental Learning." *Nature Machine Intelligence* 4: 1185–1197.

Veniat, T., L. Denoyer, and M. Ranzato. 2021. "Efficient Continual Learning With Modular Networks and Task-Driven Priors." Preprint, arXiv:2012.12631. 2020 December 23.

Wang, L., X. Zhang, Q. Li, et al. 2023. "Incorporating Neuro-Inspired Adaptability for Continual Learning in Artificial Intelligence." *Nature Machine Intelligence* 5, no. 12: 1356–1368.

Wang, L., X. Zhang, H. Su, and J. Zhu. 2024. "A Comprehensive Survey of Continual Learning: Theory, Method and Application." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 8: 5362–5383.

Wang, M., D. Yu, W. He, P. Yue, and Z. Liang. 2023. "Domain-Incremental Learning for Fire Detection in Space-Air-Ground Integrated Observation Network." *International Journal of Applied Earth Observation and Geoinformation* 118: 103279.

Wang, Z., L. Liu, Y. Duan, and D. Tao. 2022. "Continual Learning Through Retrieval and Imagination." *Proceedings of the AAAI Conference on Artificial Intelligence* 36: 8594–8602.

Warden, P. 2018. "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition." Preprint, ArXiv, abs/1804.03209.

Wolf, T., L. Debut, V. Sanh, et al. 2019. "Transformers: State-of-the-Art Natural Language Processing." Preprint, arXiv:1910.03771.

Wu, C., L. Herranz, X. Liu, Y. Wang, J. van de Weijer, and B. Raducanu. 2018. "Memory Replay Gans: Learning to Generate New Categories Without Forgetting." In *Advances in Neural Information Processing Systems, 31*. NeurIPS conference proceedings.

Wu, T.-Y., G. Swaminathan, Z. Li, et al. 2022. "Class-Incremental Learning With Strong Pre-Trained Models." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9591–9600. CVPR conference proceedings.

Xie, W., W. A. Bainbridge, S. K. Inati, C. I. Baker, and K. A. Zaghloul. 2020. "Memorability of Words in Arbitrary Verbal Associations Modulates Memory Retrieval in the Anterior Temporal Lobe." *Nature Human Behaviour* 4: 937–948.

Yang, B., M. Lin, Y. Zhang, et al. 2023. "Dynamic Support Network for Few-Shot Class Incremental Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 3: 2945–2951.

Zenke, F., B. Poole, and S. Ganguli. 2017. "Continual Learning Through Synaptic Intelligence." In *International Conference on Machine Learning*, 3987–3995. PMLR.

Zhang, J., R. Gu, P. Xue, et al. 2023. "S3r: Shape and Semantics-Based Selective Regularization for Explainable Continual Segmentation Across Multiple Sites." *IEEE Transactions on Medical Imaging* 42, no. 9: 2539–2551.

Zhuang, C., Z. Xiang, Y. Bai, et al. 2022. "How Well Do Unsupervised Learning Algorithms Model Human Real-Time and Life-Long Learning?" In *Advances in Neural Information Processing Systems, Volume 35*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 22628–22642. Curran Associates Inc.

19424795, 2025, 4, Downloaded from https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70046 by University Di Roma La Sapienza, Wiley Online Library on [29/04/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License