IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 March 2024, accepted 14 May 2024, date of publication 21 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3403569

TOPICAL REVIEW

Multi-Label Lifelong Machine Learning: A Scoping Review of Algorithms, Techniques, and Applications

MOHAMMED AWAL KASSIM[1], HERNA VIKTOR[1], AND WOJTEK MICHALOWSKI[2]
[1]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 7K8, Canada
[2]Telfer School of Management, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Herna Viktor (hviktor@uottawa.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant GR000540.

ABSTRACT Lifelong machine learning concerns the development of systems that continuously learn from diverse tasks, incorporating new knowledge without forgetting the knowledge they have previously acquired. Multi-label classification is a supervised learning process in which each instance is assigned multiple non-exclusive labels, with each label denoted as a binary value. One of the main challenges within the lifelong learning paradigm is the stability-plasticity dilemma, which entails balancing a model's adaptability in terms of incorporating new knowledge with its stability in terms of retaining previously acquired knowledge. When faced with multi-label data, the lifelong learning challenge becomes even more pronounced, as it becomes essential to preserve relations between multiple labels across sequential tasks. This scoping review explores the intersection of lifelong learning and multi-label classification, an emerging domain that integrates continual adaptation with intricate multi-label datasets. By analyzing the existing literature, we establish connections, identify gaps in the existing research, and propose new directions for research to improve the efficacy of multi-label lifelong learning algorithms. Our review unearths a growing number of algorithms and underscores the need for specialized evaluation metrics and methodologies for the accurate assessment of their performance. We also highlight the need for strategies that incorporate real-world data from varying contexts into the learning process to fully capture the nuances of real-world environments.

INDEX TERMS Continual learning, lifelong learning, machine learning, multi-label classification.

I. INTRODUCTION

The idea of building systems that reason like humans and are capable of performing intellectual tasks has intrigued researchers since the advent of computing. The seminal work of Samuel introduced the concept of machine learning (ML) and demonstrated its potential by teaching a computer to play checkers at a high level [1]. This marked an early recognition that reliable, effective, and robust machine intelligence can be achieved by developing algorithms that are capable of learning and improving their performance over time, as opposed to handcrafting rules for the performance of specific tasks. This development, in addition to several key milestones, accelerated the growth in ML and established a sub-field within artificial intelligence (AI) that focuses on methods for the development of algorithms that can learn from data without requiring explicit programming.

A major drawback of existing ML algorithms is their inability to learn in a continuous manner rather than in isolation [2]. This means that they learn to perform a particular task and operate under the assumption that the data encountered during deployment have the same characteristics as the training data, and are independently distributed or sampled from a static distribution. This is a very limiting assumption as it does not always hold in the real world, and hence the performance of ML models on new data tends to degrade. To build systems capable of learning in a manner similar to humans requires the construction of algorithms

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

that learn continuously and have the ability to identify new tasks and learn to perform them. Algorithms that learn in this lifelong manner can help overcome the drawbacks of the current learning paradigm.

Lifelong machine learning is an ML paradigm that focuses on the design and development of algorithms and systems that learn continuously, accumulate knowledge, and are capable of identifying new tasks and learning to perform them [2]. This kind of learning from previously accumulated knowledge can potentially eliminate the need for a large number of labeled training instances. One of the key challenges presented by this paradigm is the stability-plasticity dilemma, a trade-off between a model's plasticity in terms of integrating new knowledge and stability in terms of preventing it from forgetting previously-learned knowledge [3].

One of the many problems that ML is used to address is classification, whereby an algorithm learns to assign labels to test instances based on examples it has encountered during training. In multi-label classification, labels are mutually inclusive; hence, any test instance may be associated with multiple labels simultaneously. Lifelong machine learning in a multi-label setting has interesting use-cases in the real world. In healthcare, for example, a patient may suffer from multiple health conditions at the same time. It is not only important to identify all these conditions, but also to identify the presence of complications or adverse interactions that have not previously been seen.

This paper reviews the literature on multi-label lifelong machine learning, a continuous learning approach in which the learner encounters a sequence of learning tasks, with each task involving a dataset where instances are associated with multiple labels simultaneously. The scope of this review is limited to identifying and analyzing the various lifelong learning algorithms that have been proposed to handle multi-label classification tasks. This includes examining the methodologies and architectures employed by these algorithms to manage the complexities of learning multiple labels over time. In addition, we explore the metrics and datasets used to evaluate the performance of these algorithms. This involves reviewing the criteria for assessing the effectiveness of the algorithms in terms of generalization and retention of knowledge across multiple tasks, as well as the datasets that are commonly used in the field to benchmark these algorithms. We begin by providing separate overviews of each learning paradigm, then proceed to consolidate our findings, highlighting the interplay between the two paradigms. This review is intended to help readers develop insights into the various algorithms, techniques, and applications of multi-label lifelong learning and to provide directions for future research. While numerous papers have reviewed lifelong machine learning or multi-label classification individually, to the best of our knowledge a comprehensive review of this emerging domain combining these two learning paradigms has not yet been published.

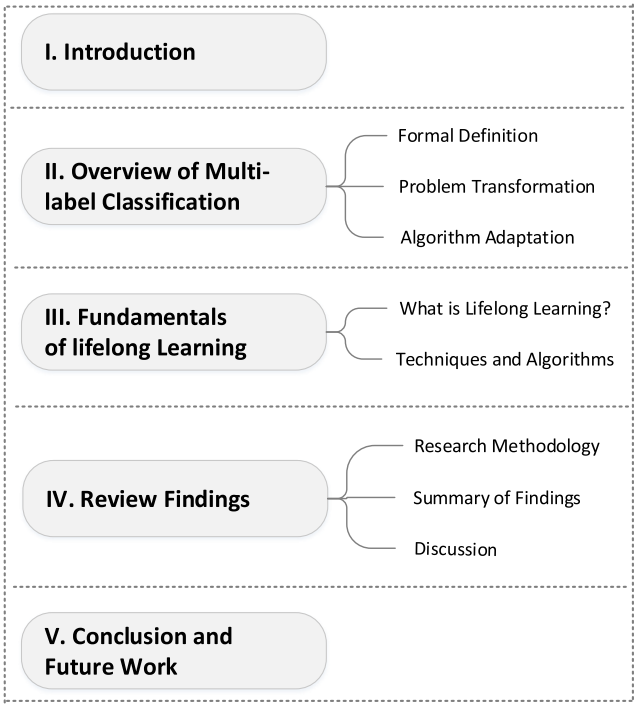


FIGURE 1. Structure of review.

The key contributions of this work are:

1) We provide a detailed analysis of the algorithms that have been proposed for multi-label lifelong machine learning. This includes a discussion of their underlying mechanisms, such as how they handle multiple labels simultaneously, manage knowledge retention, and adapt to evolving data distributions over time
2) Our review systematically presents the metrics and datasets used to evaluate the performance of multi-label lifelong learning algorithms. We highlight the importance of these metrics in assessing various aspects of model performance, such as accuracy, label ranking, and the ability to balance between learning new information and retaining previous knowledge
3) We showcase a range of practical applications where multi-label lifelong learning is making a significant impact, discuss the potential impact of advancements in the field, and identify key research directions.

The rest of the paper is organized as follows. Sections II and III present overviews of multi-label classification and lifelong learning, respectively. We discuss the key techniques used in each domain and highlight the real-world applications of lifelong learning. In Section IV, the research criteria used in this review are explained, the research questions and knowledge gaps this work addresses are stated, and our findings are presented. Section V outlines challenges in multi-label lifelong learning, potential directions for future research, and finally draws conclusions from this work.

II. OVERVIEW OF MULTI-LABEL CLASSIFICATION

In ML, classification is a fundamental task which involves categorizing input data into pre-defined classes or labels [4].

It plays an important role in various domains, including image recognition and text classification. Accurately predicting the class labels of new, previously unseen examples is the main goal of classification. Classification tasks can be broadly categorized into two types: single label and multi-label classification. In single label classification tasks, each instance in a dataset is assigned to only one label. However, in many real-world applications instances may have multiple labels at the same time. Multi-label classification addresses this situation by allowing instances to be simultaneously associated with multiple labels [5], thus expanding the scope of classification and providing a realistic representation of real-world problems.

A. FORMAL DEFINITION

Mathematically, a multi-label data instance is represented by $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i = (x_1, \ldots.., x_d)$ is a d-dimensional vector of features and $y_i = (y_1, \ldots.., y_L)$ is the associated set of labels, with each $y_l \in \{0, 1\}$. As formulated by Zhang and Zhou, the multi-label training set $D = \{(\mathbf{x}_i, y_i)|1 \leq i \leq m, \}$, where m is the total number of training examples; the task of multi-label classification is to create a classifier that learns from D a function f that maps the input space $\mathbf{X}$ to the binary exponential label space [6]. Formally:

$$f : \mathbf{X} \rightarrow 2^Y \tag{1}$$

For any unseen example, the classifier predicts the set of labels based on the learned relationship between the input space $\mathbf{X}$ and the label space 2^Y. Fig. 2 provides some examples of multi-label data across different modalities. The image in this example contains multiple objects (a bowl, bread, broccoli, and an orange). In such a recognition task, the goal is to accurately identify and label each of the distinct objects in the image. To accomplish this, the algorithm must learn to predict multiple labels for a single image. The text represents a single instance from a text classification dataset where a set of predefined categories are assigned to the given text. For the sensor data, each instance is represented by a single row in the table. A combination of features (temperature, humidity, moisture, etc.) could yield different weather conditions (rainy, cloudy) that are not mutually exclusive. It is worth noting that the major difference between multi-class and multi-label classification is that each instance is allowed one— and only one — label in multi-class problems.

As is evident in (1), the size of the label sets increases exponentially as the number of labels increases. This is a major challenge associated with learning from multi-label data. For example, a dataset with five unique labels would have about 32 possible label sets. Increasing the number of labels to 10 would significantly increases the possible label sets to over a thousand (1024 to be precise). It is therefore important to take advantage of any label relations in order to address this challenge. Zhang and Zhang categorized the existing strategies for exploiting label relations into three families, based on the degree of associations considered by each strategy [7].

The first-order strategy addresses multi-label learning by treating each label independently, neglecting the co-existence of other labels. This approach decomposes the multi-label learning problem into separate binary classification tasks for each label [8], [9], [10], resulting in simplicity and efficiency but potentially also sub-optimal performance due to the disregard of label relations. The second-order strategy considers pairwise relations between labels, involving ranking labels as relevant or irrelevant [11], [12], [13] or interactions between pairs of labels [5], [14], [15], [16]. This strategy partially exploits label relations, leading to good generalization performance. Finally, the high-order strategy deals with high-order relations among labels. Some approaches assume the association of all other labels with each label [17], [18], [19], [20], while others address connections among a random subset of labels [21], [22], [23].

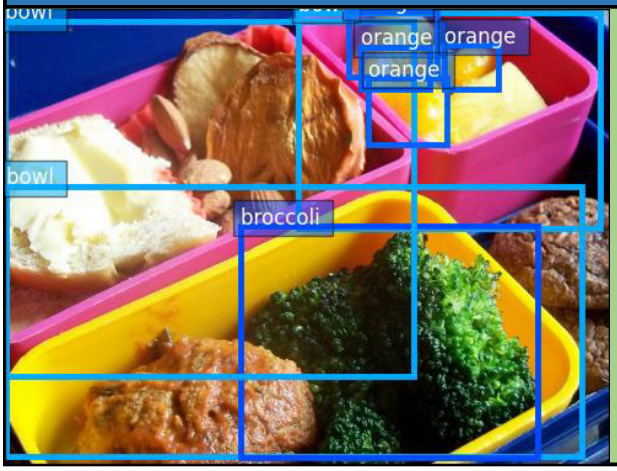
Due to the degree of relation modeled in high order strategies, the problem of learning from a vast output space is alleviated. However, these strategies are more computationally expensive and less scaleable than first and second order strategies. The aforementioned categorization is solely based on the degree of association between labels in any given dataset. Throughout the literature, two main techniques have been developed to handle multi-label classification: problem transformation and algorithm adaptation. This taxonomy is shown in Fig. 3. Each of these techniques adopt one or more of the strategies introduced above. In problem transformation techniques, the original multi-label classification problem is transformed into one or more single label problems, while algorithm adaptation techniques modify existing algorithms to cater to the constraints of multi-label data.

B. PROBLEM TRANSFORMATION

Problem transformation, as previously stated, involves converting the original multi-label problem into one or more binary or multi-class classification problems that can be solved using existing single-label classification algorithms. The fundamental idea is to divide the multi-label problem into a number of easier sub-problems, each of which focuses on determining whether a single label will be present or absent [24]. Binary Relevance (BR) is a first order strategy that is frequently employed in the problem transformation process. BR strategies approach the multi-label problem as a collection of independent binary classification problems, with a separate binary classifier trained for each label. Each binary classifier is trained utilizing the original feature space and the related binary labels for the specific label being predicted. At test time, each classifier independently predicts the presence or absence of its associated label, and the sum of these binary predictions yields the final multi-label prediction.

Label Powerset (LP) strategies transform the multi-label task into a multi-class classification problem, whereby a

Image



Pre-defined Labels

[Bowl, Bread, Broccoli, Orange]

Text

In a recent development in the political landscape, the government has announced a new economic stimulus package aimed at boosting the economy amid ongoing challenges. The package focuses on technology-driven initiatives to promote innovation and job creation in the tech sector. Additionally, it includes provisions to strengthen the healthcare system and improve access to essential health services. The move comes as the country strives to address both economic recovery and public health concerns

Pre-defined Labels

[Politics, Economy, Technology]

Sensor Data

Temperature	Humidity	Moisture	Rainy	Cloudy
25	60	0.7	1	0
22	75	0.95	0	0
18	90	0.5	1	1
30	40	0.45	0	0

Pre-defined Labels

[Rainy, Cloudy]

FIGURE 2. Example of multi-label data across different modalities.

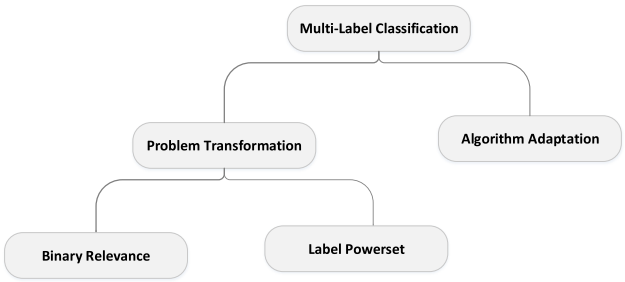


FIGURE 3. The taxonomy of multi-label classification techniques.

unique combination of labels forms a distinct class [6]. In essence, LP strategies treat every unique label combination as a different class and train a multi-class classifier to predict the correct label combination for every instance. This method explicitly models label relations while taking into account the joint distribution of labels. The LP approach, however, has scaling problems because the number of possible label combinations can increase exponentially as more labels are added.

Tsoumakas et al. worked on BR's adherence to the One-Versus-All (OVA) approach, in which a separate binary dataset is created for each label [25]. The Classifier Chains (CC) model [22] utilizes N binary classifiers that are interlinked, with each classifier incorporating the labels predicted by the previous classifiers as additional features. This approach considers label relations randomly and exhibits linear complexity with respect to the number of labels. The Probabilistic Classifier Chains (PCC) model is a Bayes optimal approach to forming classifier chains that outperforms the original CC model at the expense of computational time [26]. Antenreiter et al. implement a BR strategy in two stages [27]. In the first stage, the model learns from the data, and in the second stage it carries out meta-learning.

The Vanilla LP method [8] constructs a single-label dataset, considering each possible combination of labels as a separate class. A multi-class algorithm is then employed for further processing. Pruned Problem Transformation (PPT) or Pruned Sets (PS) algorithms aim to reduce the complexity of LP by focusing on the most crucial label combinations [28]. They achieve this by pruning examples with less frequent label sets. Tsoumakas et al. introduce an ensemble-based algorithm which utilizes random projections of the label space that construct multiple LP classifiers, each trained with a random subset of k labels [23].

C. ALGORITHM ADAPTATION

Unlike problem transformation, which converts the multi-label problem into multiple single-label or multi-class classification tasks, algorithm adaptation focuses on modifying existing classification algorithms or developing new ones explicitly designed for multi-label data [24]. This approach recognizes the intrinsic complexity of multi-label problems and seeks to take advantage of label relations during model training and prediction.

An adaptation of the C4.5 algorithm [29] to the Multi-Label Learning (MLL) setting has been introduced [9]. The original C4.5 algorithm was developed to generate decision trees using the concept of entropy, and is capable of handling both continuous and discrete attributes. To adapt it for MLL, the algorithm was modified to enable multiple labels in the leaves, and the definition of entropy was modified to consider both membership and non-membership of each class. The Multi-layer Multi-Perceptron (MMP) algorithm [30] associates each label with a separate perceptron, and the performance of the entire ensemble is taken into account when updating each perceptron, in contrast to BR (Binary Relevance). The study demonstrates that MMP exhibited superior performance compared to BR in text classification tasks. Predictive Clustering Tree (PCT) [31] is a flexible framework used to perform prediction tasks by defining a distance metric and prototype. It has been successfully applied to various tasks, including predicting tuples of variables and hierarchical multi-label classification, where each label represents a component of the target tuple. PCTs are generated top-down, with data partitioned into clusters to minimize intra-cluster variation at each node. The Multi-Label Paired Comparisons (ML-PC) method [32] utilizes two probabilistic binary classifiers to distinguish between each pair of overlapping classes. Wan and Xu define a set of linear classifiers optimized to minimize a measure that evaluates the average fraction of label pairs that are reversely ordered for each instance [33]. Multi-label k Nearest Neighbours (kNN) [34] uses lazy learning to determine the k nearest neighbors before computing a membership counting vector which indicates the number of neighbors belonging to each possible class. Using the statistical information thus derived from the label sets of the neighbors, the set of labels for the unseen instance is determined based on the maximum a posteriori (MAP) principle.

It must be mentioned, however, that these methods are not always used in isolation. Many studies use an ensemble of learners. Vateekul and Kubat, for example, employ an ensemble of decision trees to automate the categorization of multi-label text documents described by thousands of features [35]. Read et al. use an ensemble of pruned sets to identify the most relevant relations between labels [21]. This pruned sets approach operates by treating sets of labels as a single label in order to reduce the complexity of the output space.

Given that this review is about exploring the interplay between multi-label classification and lifelong learning, we do not provide an exhaustive review of all multi-label algorithms. Interested readers are referred to [6], [36], [37], and [38] for more comprehensive reviews of multi-label algorithms.

III. FUNDAMENTALS OF LIFELONG LEARNING

A. WHAT IS LIFELONG MACHINE LEARNING (LML)?

The concept of lifelong machine learning emerged from the realization that ML algorithms often struggle to adapt to

dynamic and evolving environments [2]. Lifelong machine learning approaches aim to develop algorithms that can continuously learn and adapt to new data, tasks and environments, without forgetting previously acquired knowledge. The term lifelong learning is often used synonymously with continual learning [39]. Although both these approaches focus on the idea of learning over time, they are distinct in scope and objectives. Lifelong learning encompasses a broader vision of learning across diverse tasks and domains over an extended period. In addition to knowledge retention, lifelong learning systems are capable of discovering new tasks and using accumulated past knowledge to help future learning [2]. Continual learning is a specific area within this broader context that only deals with the challenges of learning continuously from new data while preserving knowledge of past tasks.

According to Thrun's definition of lifelong machine learning [40], a learning system undergoes a sequential learning process where it accumulates knowledge from N previously encountered tasks. Subsequently, when presented with the $(N+1)th$ task, the system leverages the knowledge it has acquired from the preceding N tasks to facilitate the learning process for the $(N+1)th$ task. This approach allows the system to benefit from prior experiences and effectively transfer learned knowledge to new tasks, thus promoting continuous learning and adaptation over time. This definition emphasizes the utilization of previously acquired knowledge as a means to enhance the system's performance on new tasks and enable effective knowledge transfer. However, this definition, while insightful, introduces some ambiguity regarding the precise interpretation of the terms "task" and "knowledge". That is, one must reflect on whether a task is defined by a specific problem instance, a distinct learning objective, or a combination of both. Additionally, the notion of "knowledge" remains somewhat vague. It is unclear whether the term knowledge refers to generalizable insights, or to the underlying model parameters acquired during the learning process. Due to the lack of explicit clarification of these fundamental concepts in Thrun's definition, there have been attempts in the literature to provide a more concise definition.

Chen and Liu define lifelong machine learning as a continuous learning process: "At any time point, the learner performed a sequence of N learning tasks, $T_1, T_2, \ldots, T_N$. These tasks can be of the same type or different types and from the same domain or different domains. When faced with the $(N+1)th$ task T_{N+1} (which is called the new or current task) with its data D_{N+1}, the learner can leverage past knowledge in the knowledge base to help learn T_{N+1}. The objective of LML is usually to optimize the performance on the new task T_{N+1}, but LML can optimize any task by treating the rest of the tasks as previous tasks. A Knowledge base is constructed to maintain the knowledge learned and accumulated from the previous task. When learning T_{N+1} is complete, the knowledge base is updated with the knowledge gained from learning T_{N+1}.

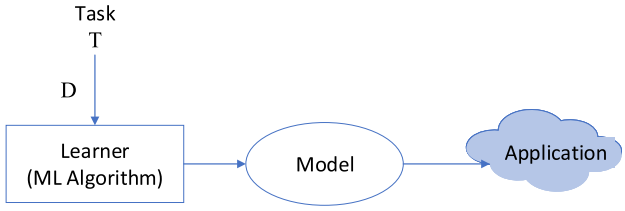


FIGURE 4. Classical machine learning architecture as depicted by [2].

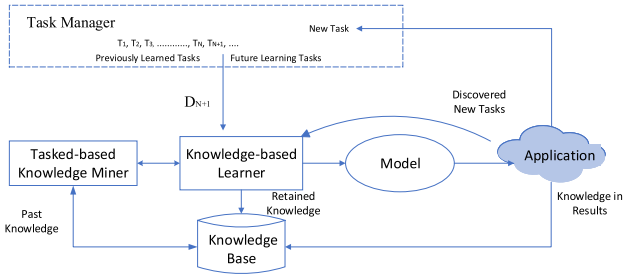


FIGURE 5. Lifelong learning paradigm [2].

The updating can involve inconsistency checking, reasoning, and meta-mining of additional higher-level knowledge" [2]. While this expanded definition has a broader scope and effectively enhances the conceptualization of a task, it falls short of providing a clear definition of knowledge. Fig. 4 and Fig. 5 illustrate the difference between classical machine learning and the lifelong machine learning paradigm.

Parisi et al. highlight the importance of avoiding catastrophic forgetting and the need for efficient mechanisms to consolidate and transfer knowledge across different tasks and experiences [41]. Their comprehensive review offers valuable insights into the field of lifelong machine learning and provides a broader perspective on the challenges it involves, along with potential solutions to them.

B. LIFELONG LEARNING TECHNIQUES AND ALGORITHMS

Learning in a continual manner presents some challenges, with catastrophic forgetting, as previously mentioned, being a particularly significant problem [3], [40]. As such, research in this field for the past three decades has been geared towards mitigating catastrophic forgetting. Existing work on lifelong learning can be grouped into three categories: regularization, rehearsal, and parameter isolation approaches (shown in Fig. 6. Our categorization is inspired by the work of De Lange et al. [42]. As with traditional ML techniques, a strict allocation of works into these three categories is not always practical, and hence some researchers provide a fourth category to represent works that combine multiple techniques [39] In this section, the effectiveness of these approaches are examined to provide insights into their practical implementation and limitations.

1) REGULARIZATION

Regularization is a broad term used to refer to a set of techniques that can prevent overfitting in neural networks by imposing constraints on updating model parameters. In the

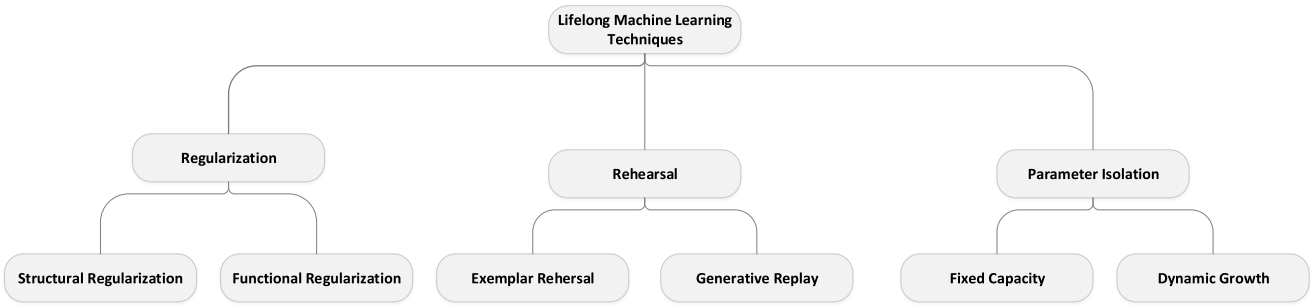


FIGURE 6. Lifelong learning techniques.

lifelong learning context, regularization techniques aim to minimize forgetting by preventing the model from overfitting a new task while preserving representations learned from previous tasks.

Elastic Weight Consolidation (EWC) [43] alleviates catastrophic forgetting by imposing a quadratic penalty on the discrepancy between the parameters learned from old and new tasks, which aids in slowing down learning for task-relevant weights in order to preserve previously acquired knowledge. Synaptic Intelligence (SI) [44] enables estimating the significance of individual synapses (parameters) for solving a learned task. This algorithm penalizes alterations to the most relevant synapses, facilitating the learning of new tasks with less forgetting. Chaudhry et al. generalize EWC and SI by creating an objective function that utilizes both Fisher information-based importance and an additional optimization-path-based importance score [45]. The latter perspective involves calculating distances within the induced Riemann manifold and optimizing the importance score based on the optimization trajectory. Aljundi et al. compute the importance of neural network parameters in an unsupervised and online manner, preventing important knowledge related to previous tasks from being overwritten when learning new tasks [46]. Learning without Forgetting (LwF) [47] involves the use of Convolutional Neural Networks (CNNs). The network with predictions from previously learned tasks is forced to be similar to the network handling the current task through knowledge distillation (the transfer of knowledge from a large, highly regularized model to a smaller one). The LwF algorithm optimizes a set of shared parameters across all tasks (θ_s) while optimizing the parameters of the new task (θ_n). It also imposes an additional constraint to ensure that predictions on the samples of the new task using θ_s and the parameters of old tasks θ_o do not undergo significant shifts, in order to retain θ_o's memory.

Regularization methods offer a means of mitigating catastrophic forgetting under specific circumstances. Nevertheless, they introduce additional loss terms to safeguard consolidated knowledge. As highlighted in Parisi et al., this may result in a trade-off between performance on old and new tasks when neural network resources are limited [41].

2) REHEARSAL

Rehearsal, also known as replay, has emerged as a promising approach to addressing catastrophic forgetting. It involves the selective reintroduction of past data samples during the training of a model on new tasks [48]. These techniques aim to preserve learned knowledge by re-exposing the model to previous experiences. Replay techniques can be broadly classified into two categories, namely Exemplar Rehearsal and Generative Replay. Exemplar Rehearsal involves storing a buffer of selected past data samples that are representative of the previously observed distribution. This minimizes the amount of memory required to store all previous samples. Generative replay techniques leverage generative models, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), to generate synthetic data samples which resemble the distribution of previously encountered tasks. These synthetic samples are then combined with the current task's data during training. This provides a means of supplementing the training process with data that resemble past experiences. While these are the two primary sub-categories of rehearsal, they can be combined to form a hybrid method. As the name implies, hybrid approaches combine real and synthetic data samples for rehearsal. By utilizing a mixture of real past data samples and generated samples, such approaches aim to provide a diverse and representative training experience for the model and to allow the model to leverage the benefits of both approaches.

Gradient Episodic Memory (GEM) [49] is a good example of exemplar replay. This approach facilitates the positive transfer of knowledge to previous tasks. GEM employs episodic memory to store a subset of observed examples from each task, which helps mitigate catastrophic forgetting. While minimizing the loss on the current task, it ensures that the losses on the episodic memories of previous tasks are treated as inequality constraints, allowing for their decrease without an increase. Building on the foundational work on GEM, significant improvements have been made in terms of the computational and memory costs involved in optimization under the constraints of gradient updating [50]. Deep Generative Replay (DGR) [51] is a typical example

of the generative replay approach. It involves a dual-model architecture comprising a deep generative model and a task solver. This design allows training data from previously learned tasks to be sampled; these data can then be interleaved with data from new tasks. Consequently, there is no need to explicitly revise old training samples for experience replay, leading to reduced working memory requirements.

3) PARAMETER ISOLATION

Parameter isolation approaches to lifelong learning focus on modifying the underlying architecture to accommodate lifelong learning scenarios (this is also referred to as the architectural approach in the literature [39], [41]). This category can be sub-divided into two sub-categories: fixed capacity and dynamic growth. Fixed capacity approaches attempt to maintain a fixed parameter size throughout the learning process. Such methods have faced criticism due to their reliance on an overparametrized network [39]. On the other hand, dynamic growth architectures address this limitation by allowing the network to dynamically expand its capacity to accommodate new tasks. In a neural network setting, these architectures can grow by adding new neurons, modules, or layers when learning new tasks. This expansion capability enables them to maintain their performance on previous tasks while efficiently incorporating new information.

Bayesian Neural Networks [52] are a classic example of methods that maintain a fixed capacity. They work by guiding task-specific information through the architecture. The key idea is to adjust the system so that a given maximum number of units in the neural network are highly active at any time. This reduces the overlap between activities performed for different tasks. Such methods also employ a measure of uncertainty to refine the model, creating specific binary masks for each task that help pinpoint the relevant parts of the model's complex weight distributions.

Progressive Neural Networks (PNN) [53] aim to preserve a network trained on previous knowledge while expanding the architecture with new sub-networks to accommodate new information. This is a good example of a method that uses dynamic growth architecture. A pool of pre-trained models is retained, one for each learned task. When a new task (T_{N+1}) is presented, a new neural network is created, and lateral connections with the existing tasks are learned to facilitate knowledge transfer. Mallya et al. use a mechanism that is similar to PNN, except it introduces a gating mechanism to automate the selection of a suitable model from an ensemble of learners [54].

In this section, we have provided an overview of lifelong learning techniques and highlighted notable algorithms that play a pivotal role in this domain. Understanding the foundational principles and algorithms of lifelong learning sets the stage for exploring their practical applications, which will be the primary focus of the upcoming section. For a more in-depth study of lifelong learning techniques, refer to [2], [39], and [41]

IV. REVIEW FINDINGS

So far, we have provided an overview of multi-label classification and lifelong learning in isolation. Here, we explore how these areas come together, and share insights about multi-label lifelong learning. Drawing inspiration from the definition of lifelong learning established by Chen and Liu [2], we propose a tailored definition for this nascent subfield. Formally, we define multi-label lifelong learning as an ongoing and continuous learning paradigm in which the learner engages in a sequence of N learning tasks, denoted as $T_1, T_2, \ldots, T_N$. In each task T_i, the learner is presented with a dataset D_i, whereby each instance of the dataset is associated with a label set y. The label set y for each instance is defined such that each label can take on values in the set $\{0, 1\}$, indicating the absence (0) or presence (1) of the corresponding label.

Recall that no comprehensive review of the literature on multi-label lifelong learning has yet been published. While Chen and Liu [2], Parisi et al. [41] and several others have published reviews which focus solely on lifelong learning, and Gibaja et al. [36] have focused on multi-label classification, there exists a growing body of work that explores the integration of lifelong learning in a multi-label setting. Our aim is to identify the extent and scope of this work and shed light on the challenges and opportunities presented by this emerging research area.

A. RESEARCH METHODOLOGY

The scoping review methodology employed in this study is grounded in the five-stage framework proposed by Arksey and O'Malley [55]. By adhering to this systematic and rigorous process, this study ensures complete transparency, facilitating replication of the search results and the reliability of our findings. Arksey and O'Malley's framework encompasses five essential stages, each executed to explore and synthesize the existing literature on multi-label lifelong machine learning. These stages are described in greater detail below.

1) STAGE 1: RESEARCH QUESTIONS

At this stage, the initial research questions establishing the foundation for the entire review are developed. Techniques that acknowledge the distinctive attributes of multi-label data and offer effective mechanisms to address issues of catastrophic forgetting in lifelong learning are analyzed, and measures of performance are evaluated. Thus, the research questions are:

1) Which lifelong learning algorithms have been proposed to deal with multi-label classification?
2) Which metrics and datasets have been used to evaluate the performance of these algorithms?

2) STAGE 2: IDENTIFYING RELEVANT STUDIES

Here, a comprehensive search is undertaken to identify all relevant studies related to multi-label lifelong machine learning. We employed rigorous and exhaustive search

TABLE 1. Development of search strategy.

	Concept 1	AND Concept 2	AND Concept 3
	Multi-label	lifelong	machine learning
OR	Multi-output	life long	learning
OR		life-long	
OR		Continuous	
OR		Continual	

techniques to gather a wide range of primary literature from diverse sources, consulting a bibliometric expert to help us craft key search terms. These terms were used in Boolean search queries to establish a search strategy (see Table 1). Table 1 is organized into three columns, representing different concepts that are logically connected by the operator 'AND'. This dictates that each search query must contain terms from Concept 1, Concept 2, and Concept 3 simultaneously to meet the criteria. Concept 1 includes the terms ''multi-label'' and ''multi-output''. Concept 2 encompasses variations of the term lifelong learning, such as ''lifelong'', ''life long'', and ''life-long'', ''continuous'', and ''continual''. Concept 3 is consistent with the term ''machine learning''. Additionally, the rows are interconnected by the operator 'OR', indicating that any of the variations within a single concept can be used interchangeably. For example, a valid search query might combine ''multi-label'' (from Concept 1), ''lifelong'' (from Concept 2), and ''machine learning'' (from Concept 3). This formulation provided comprehensive and precise search queries, ensuring that the search process was both thorough and relevant search process.

We utilized Google Scholar as the principal search engine for the retrieval of pertinent academic works. In addition, we retrieved articles from Scopus and Web of Science, but these were identified as duplicates of articles retrieved from Google Scholar, and were therefore excluded from the analysis.

3) STAGE 3: STUDY SELECTION

In this stage, we applied inclusion and exclusion criteria to filter the retrieved studies. Only those that met the predetermined criteria were retained for further examination. This screening process guaranteed that the selected studies were directly pertinent to the research questions, eliminating any potential bias or ambiguity. Our primary focus was on papers addressing lifelong learning and multi-label classification in conjunction, in order to align with our main research questions. Table 2 shows the inclusion and exclusion criteria for sources in our review, which define the scope and boundaries of our work. The column titled 'Inclusion' highlights the criteria applied when selecting articles, which were mainly English academic sources focusing on multi-label lifelong learning. In contrast, the exclusion criteria removed articles that solely centered on lifelong learning in education, those discussing lifelong learning or multi-label classification independently, non-academic sources or those without a focus on machine learning, and sources which were not available in the English language.

TABLE 2. Inclusion and exclusion criteria.

Criteria	Inclusion	Exclusion
Population	Papers focusing on multi-label lifelong machine learning	Papers focusing solely on lifelong learning in education
Context	Fusion of both domains	Papers focusing only on lifelong learning or multi-label classification independently
Types of Sources	Academic papers, conference proceedings, and journal articles	Non-academic sources or sources not focused on machine learning
Timeframe	No specific timeframe	Not Applicable
Language	English-language sources	Non-English-language sources

4) STAGE 4: DATA CHARTING AND COLLATION

In this stage, summaries of the retrieved articles were created and analyzed. Table 3 provides a brief summary of the studies included in this process. It lists the different approaches and their underlying algorithms. Each row refers to a specific approach, categorized by the lifelong learning technique employed (i.e Replay, Parameter Isolation, Regularization), the multi-label technique applied (mostly Algorithm Adaptation), the base algorithm used, and a brief description of how the method works.

B. STAGE 5: SUMMARY OF FINDINGS

In this section, we focus on articles that directly address our research questions. Of the 109 papers assessed for eligibility, 13 presented lifelong learning algorithms in a multi-label setting. Some of these benchmarked their models on openly available datasets. The PRISMA diagram for the review is shown in Fig. 7

1) TECHNIQUES AND ALGORITHMS

Here, as a direct answer to Research Question 1, we describe the techniques and algorithms used by papers' authors. Roseberry et al. introduce a self-adapting algorithm which learns from multi-label drifting data streams [56]. Their study focuses not only on learning continuously from multi-label data, but also on adapting to different concept drifts. As the authors establish, the problem of learning from multi-label data is further exacerbated when the statistical properties of the data change over time as a result of drift. Traditional learning methods tend to struggle to maintain accuracy and adapt to evolving data distributions. The proposed approach leverages the k Nearest Neighbors (kNN) algorithm, which is known for its simplicity and effectiveness in performing classification tasks. The kNN algorithm uses the notion of proximity when making predictions by comparing the test instance to the k closest examples in the training set. However, unlike traditional kNN, the proposed method introduces a self-adjusting mechanism that dynamically adapts the value of k based on the properties of the incoming data. This

TABLE 3. Summary of identified multi-label lifelong machine learning algorithms.

Algorithm	LL Technique / Method	Multi-label Technique / Method	Base Algorithm	Datasets	Purpose
[56] MLSAkNN	**	Algorithm Adaptation	KNN	Yeast, Others	Introduces a self-adapting approach for learning from multi-label drifting data streams
[57] PRS	Replay	Algorithm Adaptation	Neural Network	MSCOCO, NUS-WIDE	Maintains a reservoir that represents a true random sample of the data stream
[58] OCDM	Replay	Algorithm Adaptation	Neural Network	MSCOCO, NUS-WIDE	Treats the updating of its replay buffer as an optimization problem
[59] MLCA	Parameter Isolation	Algorithm Adaptation	Gaussian Kernel Function	Yeast, Others	Synergizes Adaptive Resonance Theory (ART) and Bayesian techniques for MLC
[60] LTM*	Replay	Problem Transformation (BR)	Multiple	MSCOCO, NUS-WIDE	lifelong topic modeling approach designed to uncover hidden topics in a text corpus
[61] BAT-OCDM	Replay	Algorithm Adaptation	Neural Network	ALPI	Modifies OCDM by introducing separate memory for each task
[62] AGCN	Parameter Isolation	Algorithm Adaptation	Neural Network	MSCOCO, NUS-WIDE	Utilizes an Augmented Correlation Matrix (ACM) to dynamically capture label dependencies
[63] DFSL	Parameter Isolation	Algorith Adaptation	Neural Network	ESC-50, AudioSet	Few-shot continual learning framework for audio classification
[64] CPG	Parameter Isolation	Algorithm Adaptation	Neural Network	DAGM	Deep lifelong learning for defect detection in manufacturing pipelines
[65] KRT	Regularization	Algorithm Adaptation	Neural Network	MSCOCO, PASCAL-VOC	Knowledge Restore and Transfer (KRT) tailored for multi-label class-incremental learning
[66] CIFDM	Parameter Isolation	Algorithm Adaptation	Neural Network	Yeast, MIR-FLICKR25K	A framework for multi-label stream learning
[67] CDSH	Regularization	Algorithm Adaptation	Neural Network	PASCAL-VOC, MIR-FLICKR25K, MSCOCO, NUS-WIDE	Continual Deep Semantic Hashing (CDSH) for learning binary codes of multi-label images
[68] DLFL	Replay	Algorithm Adaptation	Neural Network	PASCAL-VOC, MSCOCO, NUS-WIDE	a disentangled label feature learning (DLFL) framework for multi-label learning

** The MLSAkNN model operates by storing previously encountered examples, utilizing these instances during the inference phase. The base algorithm (kNN) is non-parametric, and 'learning' in the traditional sense does not occur. Consequently, categorizing MLSAkNN under any of the three commonly-recognized learning techniques would be a misrepresentation, given its distinct operational mechanism that diverges from standard parametric learning models.

adaptive nature allows the model to respond effectively to concept drift and changing label distributions, enabling continuous learning in dynamic environments. Moreover, the conventional k Nearest Neighbors (kNN) algorithm typically employs a criterion such as majority voting to assign a class to a test instance. However, this method encounters limitations in the context of multi-label scenarios, as individual neighbors may possess distinct label sets. To determine which labels to assign to a given test instance, the algorithm tallies the frequency of each label's occurrence within the closest neighbors identified from the training dataset. Employing this count, it proceeds to compute the likelihood and posterior probability associated with each label, utilizing Bayesian principles. Label presence or absence is determined by these probabilities.

Masuyama et al. combine the principles of Adaptive Resonance Theory (ART) [69] and the Bayesian technique for label probability computation to effectively group instances with similar label patterns into clusters, thereby identifying relations within the data and reducing the size of the output space to be learnt [59]. ART is a theory of cognitive information processing that underpins the development of some neural networks. ART-based algorithms attempt to solve the stability-plasticity dilemma by being competitive and self-organizing. By leveraging ART, the approach devised by Masuyama et al. adaptively and continually generates prototype nodes corresponding to the given data, and the generated nodes are used as classifiers. Meanwhile, a Bayesian approach is used to independently track label occurrences for each label and compute corresponding probabilities, in a similar manner to [56]. As a result, each classifier outputs a set of probabilities that is used to determine label occurrence and hence to simultaneously assign multiple labels to a given instance. The authors use a probability threshold of 0.5, with values lower than this threshold considered to indicate the absence of a label. This approach enables the effective handling of a growing number of labels, ensuring the algorithm's adaptability and scalability in multi-label scenarios. The proposed algorithm, called MLCA, can learn continuously and exhibits competitive classification performance on synthetic and real-world multi-label datasets.

The Augmented Graph Convolutional Network (AGCN) [62], [70] approach has been proposed to solve the problem of

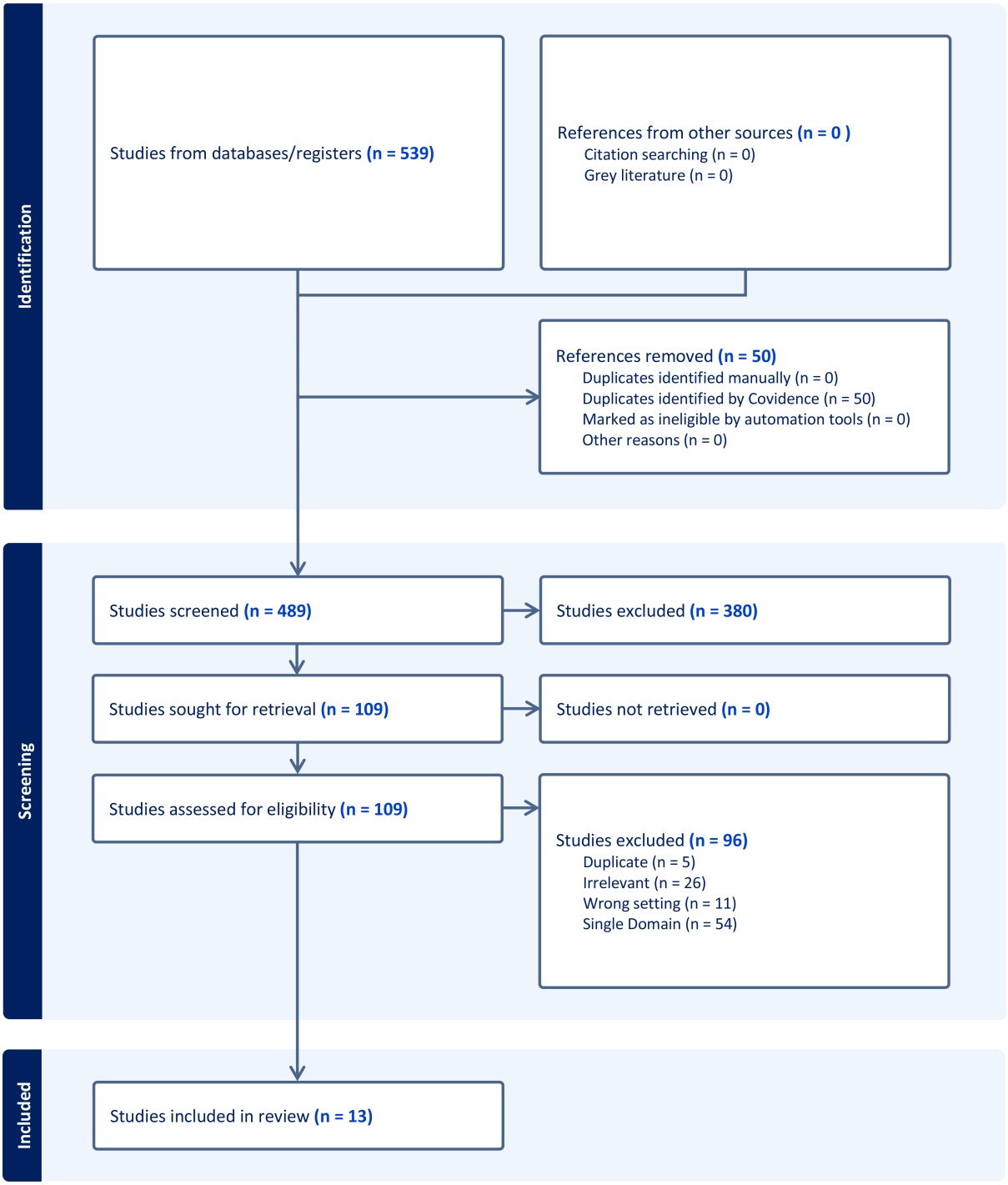


FIGURE 7. PRISMA diagram for scoping review.

catastrophic forgetting of old classes when training a model on data with different partial labels in image recognition problems. The proposed method builds an Augmented Correlation Matrix (ACM) across sequential partial-label tasks and captures label dependencies with a dynamic augmented structure to yield effective label representations. This facilitates knowledge transfer and adaptation across different image recognition tasks in a lifelong learning setting. By incorporating graph convolutional networks with a relationship-preserving loss function, AGCN effectively captures the relationships and dependencies among labels and images, enabling accurate and efficient multi-label recognition.

Wang et al. introduce a few-shot continual learning framework for audio classification [63]. Few-shot learning allows for the recognition of novel classes based on only a few

labeled data at the time of inference. By efficiently utilizing small amounts of labeled data during inference, this approach enables fast and efficient model updates. The approach is an adaptation of the Dynamic Few-Shot Learning (DFSL) framework, which uses a CNN classifier to extract features and an attention mechanism to exploit past knowledge. To train a model capable of predicting multiple concurrent classes, the categorical cross entropy loss in DFSL is replaced with binary cross entropy. This shift to binary cross entropy loss is crucial, as it allows the model to evaluate each class label as an independent binary classification problem. In essence, rather than predicting a single label from a set of mutually exclusive labels, the model assesses the presence or absence of each class label independently.

Kim et al. focus on alleviating catastrophic forgetting in neural networks when learning from imbalanced datasets by proposing a technique called Partition Reservoir Sampling (PRS) [57]. This approach is a modified version of the Reservoir Sampling technique, which is widely used to efficiently store a stream of data and sample it. Reservoir Sampling maintains a reservoir of elements that represent a true random sample of the data seen so far in the stream. As the stream flows, each element in the reservoir has an equal probability of being replaced. In PRS, the authors extend this method to create balanced training partitions that ensure each mini-batch contains a representative sample of both majority and minority classes before it is rehearsed. It achieves this by caching a running statistic of all observed examples and uses label frequencies to set the target proportion of classes in memory. This technique allows the model to learn from imbalanced data effectively and mitigates the impact of class imbalance during lifelong learning.

Optimizing Class Distribution in Memory (OCDM) [58] is a memory-based technique that dynamically maintains a representative distribution of classes in memory, in a similar manner to [57]. As this is a rehearsal based technique, a small amount of previously seen data is stored in a replay buffer. Unlike [57], the proposed OCDM method formulates the memory update mechanism as an optimization problem. All the observed mini-batches of data are first added directly into memory until it is full. When a new observation is introduced, the memory is updated such that the new distribution of samples is closest to a given target distribution. This is achieved by minimizing the distance between the two distributions (the target distribution and the distribution of the selected dataset), measured using a metric such as KL divergence.

Pham et al. propose a lifelong topic modeling approach to facilitate the discovery of hidden topics in a text corpus by exploiting prior domain knowledge [60]. This approach uses a probability-based close domain metric to select valuable knowledge that a model has learnt from the past. This is used to produce more topics associated with the current domain. The proposed metric measures the closeness of two domain datasets, which is then used to select data that enhance the current task's learning. Knowledge of hidden topics is derived from the closeness of the domains. The approach enables the use of this knowledge to enrich features for a multi-label text classifier. The multi-label learning algorithm proposed by Zhang et al. [71] is used. This comprises two main steps. First, for each label in the dataset, the algorithm conducts a clustering analysis on both the positive and negative instances associated with that label. Through this analysis, it constructs features that are specific to each label, ensuring that the characteristics unique to each label are captured. In the second step, it uses these label-specific features to develop a set of binary classifiers, each tailored to the distinct attributes of a different label.

BAT-OCDM [61] investigates Domain Incremental Learning and presents a scenario for lifelong learning whereby a model adapts to handle a stream of machines with distribution shifts. Domain Incremental Learning refers to a lifelong learning scenario whereby the set of labels in the output do not change across tasks. Instead, changes occur in the distribution of the input data from one task to the other [61]. The proposed approach modifies the OCDM algorithm [58] by using a separate memory for each task. This is done to ensure balance on both labels and tasks, as OCDM does not guarantee the retention of all previously seen tasks. Tests on real data sourced from the packaging industry are conducted to demonstrate the feasibility of addressing this significant problem. The goal is to predict a list of distinct and mutually inclusive alarms that are likely to occur in the future. This is modeled as a multi-label classification problem.

Chen et al. propose the use of deep lifelong learning (learning with densely connected neural networks) for defect detection in manufacturing pipelines [64]. The approach allows the model to learn to detect new defect types while maintaining its ability to detect old defect types without retraining on previous data. Each task is formulated as a binary classification problem for each defect type. When a new defect type is discovered in the manufacturing process, a new binary classification task is formulated by collecting the dataset for the defect. The proposed Compact, Picking and Growing (CPG) algorithm then learns the new task. The CPG algorithm's learning process involves identifying crucial weights within the pre-existing deep neural network model learned from past tasks, compacting the model to free up weights for upcoming tasks, and enlarging the network's size if the performance target has not been met.

Knowledge Restore and Transfer (KRT) [65] is a framework tailored for multi-label class-incremental learning (MLCIL). This framework incorporates two key modules: the Dynamic Pseudo-Label (DPL) module, which restores knowledge from old labels, and the Incremental Cross-Attention (ICA) module, which is designed to preserve task-specific knowledge and transfer old knowledge to the new model. Through the application of this proposed method, the authors report significant improvements in recognition performance and effectively mitigate the issue of forgetting in multi-label class-incremental learning tasks.

Continual and interactive feature distillation for multi-label stream learning (CIFDM) [66] is a proposed framework in which new labels emerge continuously in changing environments and are assigned to previous data. The framework utilizes knowledge from previous tasks to learn new knowledge and avoid catastrophic forgetting, and consists of three components: an interactive knowledge compression function, a knowledge bank, and a pioneer module. The compression function compresses and transfers new knowledge to the bank. The knowledge bank stores the compressed knowledge along with its associated label set. When a new task with novel labels is encountered, the knowledge base comes into play, initializing a pioneer module which is intended to learn the new information from incoming examples.

Song et al. propose a method called Continual Deep Semantic Hashing (CDSH) for learning the binary codes of multi-label images with increasing labels [67]. This method consists of two hashing networks, one for hashing the increasing semantics of data into semantic codes and the other for mapping images to the corresponding semantic code. CDSH incorporates empirically verified loss, and a special regularization design to ensure that old labels remain unchanged during encoding. The authors also theoretically demonstrate that their method improves the probability of the old data's code remaining unchanged after the model is updated.

Jia et al. propose a Disentangled Label Feature Learning (DLFL) framework to learn a disentangled representation for each label [68]. The framework introduces the One-Specific-Feature-for-One-Label (OFOL) mechanism to address the limitations of the One-Shared-Feature-for-Multiple-Labels (OFML) mechanism commonly used in multi-label classification. The framework includes a feature disentanglement module, which contains learnable semantic queries and a Semantic Spatial Cross-Attention (SSCA) sub-module. The SSCA sub-module localizes the label-related spatial regions and aggregates located region features into the corresponding label feature to achieve feature disentanglement.

While all these approaches have been proposed to handle multi-label data, some of the mechanisms they use to identify interrelationships among labels and assign multiple labels per instance are not presented transparently. For example, Kim et al. [57] and Liang et al. [58] emphasize sample selection and dealing with imbalance, but fall short of providing practical insights into how the multi-label problem is solved. It is crucial to note that both methods are rooted in neural network architectures, as the choice of activation and loss functions within these networks is key. These functions are tailored to optimize distinct objectives, and their selection can substantially influence outcomes — potentially leading to results that could be misinterpreted if not contextualized correctly. In addition, it is obvious from Table 3 that the majority of the multi-label lifelong learning algorithms (ten out of thirteen) use neural networks as their base algorithm. For these algorithms, it is reasonable to expect that the output size matches the number of labels, with each neuron generating a prediction for each label. In terms of the activation function, a common approach is to produce separate probabilities for each neuron; one example of this is the coupling of the sigmoid function with binary cross entropy loss, as employed by Wang et al. [63]. Finally, although some of the methods discussed have been used for specific applications such as equipment monitoring [61], audio classification [63] and fault detection [64], it is important to note that the core algorithm in each case is not limited to any specific domain. Each of these core algorithms can be applied to any dataset with the same modality as those described in their respective studies.

2) EVALUATION METRICS AND DATASETS

To answer our second research question, we identified the metrics and datasets used to evaluate the algorithms discussed earlier. None of the works we have reviewed propose new metrics specifically for the evaluation of the methods under consideration. Existing assessment metrics encompass either lifelong learning metrics, multi-label metrics, or a fusion of both. Consequently, we now proceed to examine these metrics.

a: MULTI-LABEL METRICS

To accurately assess a multi-label classification model's performance, it is essential to understand its capabilities and limitations. To achieve this, a diverse range of evaluation metrics are used, each serving a specific purpose in evaluating the model's overall effectiveness in handling multiple labels.

Tsoumakas et al. propose two distinct categories of metrics for evaluating multi-label learning methods: example-based metrics and label-based metrics [25]. Example-based metrics are calculated for each test example and then averaged across the entire test set. In the label-based approach, individual metrics are computed for each label based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and then averaged to obtain an overall value. This averaging can be performed using either the macro method (shown in (2)) whereby metrics are calculated for each label and then averaged across all categories, or the micro method (shown in (3)), which considers predictions for all instances together and calculates the measure across all labels by aggregating TP, TN, FP, and FN values.

$$\text{Macro } B = \frac{1}{N}\sum_{i=1}^{N} B(TP_i, TN_i, FP_i, FN_i) \tag{2}$$

$$\text{Micro } B = B\left(\sum_{i=1}^{N} TP_i, \sum_{i=1}^{N} TN_i, \sum_{i=1}^{N} FP_i, \sum_{i=1}^{N} FN_i\right) \tag{3}$$

where,

N = Number of labels

B = Any binary evaluation measure

From the above formulas, a micro and macro average can be calculated for metrics such as Recall, Precision,

TABLE 4. Multi-label evaluation metrics.

Metric	Brief Description	Equation
Exact Match Ratio	Proportion of instances in which all predicted labels match the true labels exactly	$\frac{1}{n}\sum_{i=1}^{n} Z_i = Y_i$
Hamming Loss	Fraction of misclassified labels per instance	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\lvert Z_i \Delta Y_i\rvert$
Ranking Loss	Ranking quality of the predicted labels compared to the true labels	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lvert Y_i\rVert\overline{Y_i}\rvert}\lvert E\rvert$
Jaccard Index	Measures the similarity between predicted and true label sets	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i \bigcap Z_i\rvert}{\lvert Y_i \bigcup Z_i\rvert}$
Subset Accuracy	Ratio of instances where all predicted labels are a subset of the true labels to total number of instances	$\frac{1}{n}\sum_{i=1}^{n} Z_i \subset Y_i$

Notation for Multi-label Evaluation Metrics
n: Number of instances in the dataset.
Z_i: Set of labels predicted for the i-th instance.
Y_i: Set of true labels for the i-th instance.
q: Total number of possible labels.
$|\cdot|$: Cardinality of a set.
Δ: Symmetric difference between two sets.
$|Y_i \parallel \overline{Y_i}|$: Total number of label pairs where one label is in Y_i and the other is not, for the i-th instance.
E: Set of incorrectly ordered label pairs.
$\bigcap$: Intersection of two sets.
$\bigcup$: Union of two sets.
$\subset$: Subset relation.

Receiver Operating Characteristic (ROC), Area Under Curve (AUC), Accuracy, and F1 score. In addition to these metrics, in Table 4 we provide brief definitions of other evaluation metrics discussed in the literature [6], [36], and [37]. These metrics are included here in order to provide readers with an understanding of how the performance of such algorithms is quantified and compared.

1) **Exact Match Ratio** computes the proportion of instances in which all predicted labels match the true labels exactly. It provides a more stringent evaluation metric than subset accuracy as it requires all labels to be predicted correctly for an instance to be considered correctly classified.
2) **Hamming Loss** measures the fraction of misclassified labels per instance by comparing predicted labels with true labels.
3) **Ranking Loss**Evaluates the ranking quality of the predicted labels compared to the true labels.
4) **Jaccard Index** measures the similarity between predicted and true label sets using the ratio of their intersection to their union. This assesses the model's ability to handle label sets with varying degrees of overlap.
5) **Subset Accuracy** evaluates the ratio of the total number of instances where all predicted labels are a subset of the true labels to the number of instances in the dataset. It is a less strict metric than the Exact Match Ratio.

The Exact Match Ratio has an all-or-nothing approach, which is useful in scenarios where label accuracy is paramount. However, its strict nature might not be suitable for applications in which partial matches are still informative. It fails to acknowledge the partial correctness of the predictions, which can be useful in many real-world scenarios. While subset accuracy and Hamming loss are less strict, critical misclassifications may be overlooked. In cases where missing certain labels is a more critical problem than misclassifying others (i.e., where there are important labels that should not be missed), these metrics are not particularly helpful.

b: LIFELONG LEARNING METRICS

The more frequently used metrics [41] for assessing the quality of lifelong learning methods are as follows:

1) **Forgetting** measures the extent to which an algorithm loses its performance on previously learned tasks while learning new ones. It evaluates the preservation of knowledge acquired during earlier tasks and indicates the potential interference between new and old knowledge during model updates.
2) **Forward Transfer** measures the impact of previously learned tasks on the performance of new tasks. It measures whether knowledge learned from earlier tasks has positive effects on learning new tasks.
3) **Backward Transfer** assesses the influence of new tasks on the performance of previously learned tasks. It measures whether learning a new task improves or degrades performance on earlier tasks.
4) **Interference** measures the extent to which the acquisition of new knowledge hinders or interferes with the retention and adaptation of existing knowledge, leading to a degradation of performance on earlier tasks.

Some methods, including [57], [58], [62], [65], and [68], are evaluated using both micro and macro averaging techniques in addition to mean average precision (mAP). Specifically, the overall precision (OP), overall recall (OR) and overall F1 (OF1) micro averages are determined. For the macro averages, per-class precision (PC), per-class recall (CR), and per-class F1 (CF1) are used. No attempts are made to assess the model's ability to learn new tasks without forgetting, or to transfer past knowledge to current tasks. Some researchers [56], [63], [64], [66] further limit the scope of evaluation by using either a macro or micro averaging technique, rather than both. Dalle et al. combine both multi-label metrics and lifelong metrics [61]. While such a combination offers valuable insights into certain aspects of trained models, it still falls short of comprehensively capturing the intricate dynamics associated with learning multiple labels across sequential tasks.

c: DATASETS

Various datasets have been used to assess the effectiveness of the algorithms reviewed in this paper. Table 5 lists the datasets most commonly used in the retrieved articles.

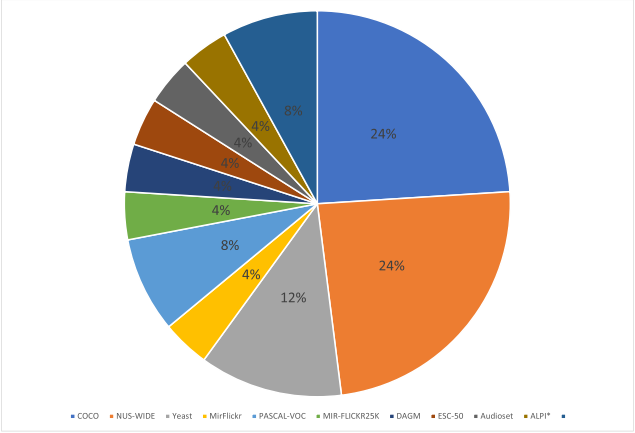


FIGURE 8. Proportion of datasets and their usage in the reviewed literature.

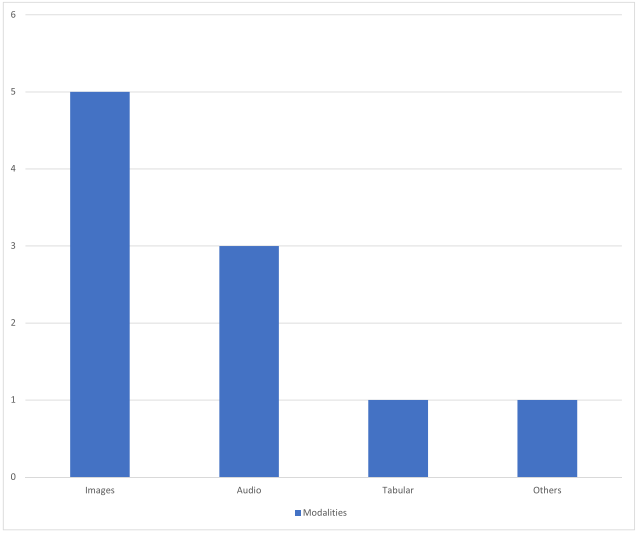


FIGURE 9. Modalities of the multi-label datasets. The vertical axis represents the number of datasets with the given modality.

These include MSCOCO, NUS-WIDE, Yeast, PASCAL-VOC, MIR-FLICKR25K, DAGM, ESC-50, AudioSet, and ALPI, each with a distinct label count and modality (image, tabular, or audio). MSCOCO and PASCAL-VOC, for example, are image datasets used in object detection and scene segmentation, while Yeast is a tabular dataset derived from gene expression data in yeast cells. The table also includes audio datasets like ESC-50 and AudioSet, which focus on environmental sounds and a wide range of real-world audio clips, respectively. The usage proportion of the identified datasets, as well as their modalities, are also shown in Fig. 8 and Fig. 9 respectively. The bar chart shows the count of different data modalities used across the reviewed literature. We observe that imagery is the most frequent data modality with a count of 5, indicating that images are the most commonly used type of data in this context. Audio follows with a count of 3, suggesting that it is also significant but less frequently used than images. Tabular data has a count of 1, showing that it is present in the literature but not as prominently utilized as image or audio data. Lastly, the 'others' category also has a count of 1; this category includes any data modalities that do not fall into the first three categories.

3) APPLICATIONS AND USE-CASES

Multi-label lifelong learning has a wide range of practical applications that capitalize on its ability to continuously adapt to evolving data and label distributions. Chen et al. present an interesting approach for detecting defects in manufacturing pipelines [64]. In traditional defect detection methods, models are trained to identify specific types of defects. However, as the manufacturing process develops, new types of defects may emerge that are not mutually exclusive with older defects. Models trained on older defect types struggle to identify these new types, leading to inefficiencies. Multi-label lifelong learning allows classifiers to identify old types of defects while incrementally learning to detect new ones, even when they occur simultaneously. In large-scale image retrieval, images are converted into compact binary codes known as 'hashes' to facilitate efficient storage and searching. Song et al. present a useful approach for the lifelong hashing and retrieval of multi-label images [67]. The applications presented by Dalle et al. [61] and Wang et al. [63] can be broadly categorized under audio classification. The former focuses on monitoring equipment in a manufacturing setting, specifically on keeping track of alarms emitted by various machines. Multi-label LML appears to be a viable method for addressing challenges in the following established domains:

1) **Medical diagnostics and healthcare** In the medical field, patient health data can be dynamic and multi-dimensional, and often requires multiple labels for accurate diagnosis and treatment planning. Multi-label lifelong learning can be leveraged to build robust medical diagnosis systems that adapt to changing patient profiles, integrate new medical knowledge, and make accurate predictions across various medical conditions.
2) **Environmental monitoring** Environmental monitoring involves the analysis of diverse parameters such as air quality, water pollution, and climatic conditions. Multi-label lifelong learning can be applied to build predictive models that continuously update to capture environmental changes, making it possible to provide real-time warnings which facilitate proactive measures to address ecological challenges.
3) **Finance and trading** Multi-label lifelong learning enables trading algorithms to incorporate new patterns and market trends without compromising knowledge of historic market dynamics.
4) **Robotics and autonomous systems** Robotic systems can continuously learn from real world experiences, allowing them to operate in dynamic and changing environments.

TABLE 5. Summary of datasets widely used for multi-label classification.

Dataset	Label Count	Modality	Brief Description
[72] MSCOCO	80	Image	MS COCO (Microsoft Common Objects in Context) is a widely used large-scale dataset for object detection, segmentation, and captioning tasks in computer vision. It contains a diverse collection of images, each annotated with multiple object instances and their corresponding labels
[73] NUS-WIDE	81	Image	A multi-label dataset that consists of a diverse collection of images, each associated with multiple relevant labels. The dataset covers a wide range of object categories and scenes, making it suitable for various real-world applications
[74] Yeast	14	Tabular	This dataset is derived from gene expression data and contains information about the presence or absence of some functional classes in yeast cells. Each data instance corresponds to a yeast gene, and the labels represent the functions associated with the gene
[75] PASCAL-VOC	20	Image	The PASCAL-VOC (Visual Object Classes) dataset is a widely used benchmark dataset, just like COCO. It consists of images from various real-world scenes
[76] MIR-FLICKR25K	24	Image	This multi-label dataset is mostly used in the field of Multimedia Information Retrieval (MIR). It presents a challenging and diverse set of images with a wide range of visual content and semantic labels
[77] DAGM	10	Image	Algorithm Adaptation
[78] ESC-50*	50	Audio	This dataset consists of environmental audio recordings comprising animal sounds, human sounds, and background noise
[79] AudioSet	527	Audio	A large collection of human-labeled sound clips that provides comprehensive coverage of real-world sounds
[80] ALPI	154	Audio	A sequence of alarms logged by packaging equipment in an industrial environment

Note: The actual datasets used in the multi-label lifelong experiments are sequentialized versions of the original.
*Originally single labeled, and then converted to multi-label for comparative analysis [63]

5) **Education** Educational platforms can progressively tailor contents based on the individual learner's needs.
6) **Smart homes** Lifelong learning techniques can enable IoT devices in smart homes to continuously learn and adapt to the resident's preferences.
7) **Cybersecurity and fraud detection** Lifelong learning can facilitate the learning of new attack patterns, identification of emerging threats, and the updating of defense mechanisms to enhance protection against evolving cyber threats.
8) **Transportation** Intelligent Transportation Systems and autonomous vehicles can continuously learn from traffic patterns, adapt to changing road conditions, and improve navigation.

It is important to note that this list of domains and applications, while not exhaustive, provides examples which highlight the versatility and potential impact of multi-label lifelong learning across different sectors.

C. DISCUSSION

Multi-label lifelong learning algorithms emerged from the multi-label classification and lifelong learning domains. Leveraging the strengths of both fields, these algorithms effectively navigate the complexities of handling multi-label data while accommodating lifelong learning scenarios.

The research landscape in the field of lifelong learning has predominantly emphasized continual learning, with a major focus on addressing catastrophic forgetting. However, lifelong learning encompasses a broader spectrum of challenges and scenarios, including cumulative learning across multiple tasks and domains. This highlights the need for a more balanced exploration of the various aspects of lifelong learning, considering such aspects as task identification and knowledge transfer over extended periods — features often associated with real-world lifelong learning applications.

Intuitively, providing solutions to real-world problems must involve both data and algorithms. As such, the methods presented here are not always used in isolation. An end-to-end solution would typically involve one or more of the two broad categories of problem transformation and algorithm adaptation. With regard to algorithm adaptation, traditional ML algorithms require substantive modification to work well with multi-label datasets. Neural networks, on the other hand, are easy to modify and adapt for multi-label problems. However, they fall short when it comes to explainability, which is a key requirement in many applications. Finally, we learnt that imbalance is an inherent property of most multi-label datasets, and as such most multi-label classification algorithms tend to include mechanisms for handling imbalance.

Lifelong machine learning represents an exciting research paradigm that is essential for the creation of AI systems that can adapt and learn continuously in dynamic environments. The lessons learned from this research shed light on the importance of mitigating catastrophic forgetting through the employment of memory-augmented models or the adaptation

of architectures to enhance lifelong learning performance. As the field continues to advance, addressing scalability challenges and exploring real-world applications will be critical for unlocking the full potential of lifelong machine learning in various domains.

V. CONCLUSION AND FUTURE WORK

This paper presents a scoping review of the algorithms, methods, and metrics used in lifelong learning within the context of multi-label classification. Through a scoping review framework, we identified areas of overlap and common ground between these two domains. By leveraging Arksey and O'Malley's scoping review methodology, we have systematically gathered a wide range of relevant literature, enabling us to gain new insights into the state of research in this field. Our review uncovers several knowledge gaps in the existing literature. First and foremost, with regard to application, we noticed an under-exploration of the potential of generative replay techniques for handling multi-label data. Generative replay methods that use generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have demonstrated their effectiveness in alleviating catastrophic forgetting in sequential learning tasks [51] by synthesizing and replaying past data samples during model updates. However, the current body of research predominantly focuses on exemplar replay (refer to Table 3), which necessitates memory buffers, leaving a substantial gap in terms of investigation of the applicability and potential challenges of generative replay in multi-label learning scenarios. The integration of generative replay into the multi-label context remains largely unexplored.

Moreover, existing multi-label lifelong learning algorithms often rely on modifying current standalone datasets to simulate various scenarios, such as class incremental or domain incremental learning. However, this approach may not fully capture the complexities and nuances of real-world environments. The effectiveness and generalizability of these algorithms in practical, dynamic settings remains uncertain. Furthermore, the absence of specific evaluation metrics tailored for multi-label lifelong learning has led to the utilization of standalone metrics in the existing literature. This remains a significant research challenge.

These identified gaps present exciting opportunities for future research to significantly advance the field and drive the development of more effective lifelong learning algorithms in multi-label settings. A promising future direction for multi-label lifelong learning research lies in the exploration and adaptation of generative replay techniques specifically for multi-label data. One potential limitation of addressing this gap is the computational complexity of generative models, which may hinder their scalability to large multi-label datasets. These models are known for their intensive computational requirements, and researchers may need to resort to specialized hardware such as high performance Graphics Processing units (GPUs) or even Tensor Processing Units (TPUs). Additionally, the quality of the generated samples is crucial for effective replay, and generating high-quality multi-label samples may pose a challenge. If the samples used for replay are not of good quality, repeatedly replaying these samples can reinforce the model's errors, leading to a compounding effect that degrades performance over time.With regard to datasets, it is essential to develop approaches that enable lifelong learning models to interact and adapt to real world scenarios directly, as opposed to using modified handcrafted static datasets. This includes devising strategies to incorporate real-world data streams, dynamic changes, and varying contexts into the learning process. This is a promising research direction with the potential to help us better assess the performance, robustness, and applicability of these algorithms in ever-evolving scenarios. Another promising research direction is the development of dedicated evaluation metrics that can effectively address the unique challenges and requirements posed by lifelong learning in a multi-label context.

REFERENCES

[1] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Develop.*, vol. 44, no. 1.2, pp. 206–226, Jan. 2000.

[2] Z. Chen and B. Liu, *Lifelong Machine Learning*, vol. 1. Cham, Switzerland: Springer, 2018.

[3] M. Mermillod, A. Bugaiska, and P. Bonin, "The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects," *Frontiers Psychol.*, vol. 4, p. 54654, Aug. 2013.

[4] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006.

[5] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2005, pp. 195–200.

[6] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[7] M.-L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 999–1008.

[8] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[9] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Freiburg im Breisgau, Germany: Springer, 2001, pp. 42–53.

[10] F. De Comite, R. Gilleron, and M. Tommasi, "Learning multi-label alternating decision trees from texts and data," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Leipzig, Germany: Springer, 2003, pp. 35–49.

[11] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2001, pp. 1–7.

[12] J. Fürnkranz, E. Hullermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, Nov. 2008.

[13] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[14] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. 15th ACM Int. Conf. Multimedia*, Sep. 2007, pp. 17–26.

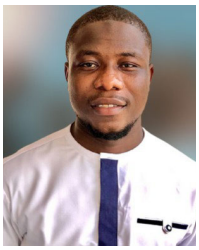
[15] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2002, pp. 1–8.

[16] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2005, pp. 274–281.

[17] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 381–389.

[18] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Sydney, NSW, Australia: Springer, May 2004, pp. 22–30.
[19] W. Cheng and E. Hullermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Mach. Learn.*, vol. 76, nos. 2–3, pp. 211–225, Sep. 2009.
[20] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 834–843.
[21] J. Read, B. Pfahringer, and G. Holmes, "Multi-label classification using ensembles of pruned sets," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 995–1000.
[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.
[23] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Proc. Eur. Conf. Mach. Learn.* Warsaw, Poland: Springer, 2007, pp. 406–417.
[24] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.
[25] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Mining Multi-Label Data*. Berlin, Germany: Springer, 2010, pp. 667–685.
[26] W. Cheng, E. Hüllermeier, and K. J. Dembczynski, "Bayes optimal multilabel classification via probabilistic classifier chains," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 279–286.
[27] M. Antenreiter, R. Ortner, and P. Auer, "Combining classifiers for improved multilabel image classification," in *Proc. 1st Workshop Learn. Multilabel Data (MLD) Held Conjunction ECML/PKDD*, 2009, pp. 16–27.
[28] J. Read, "A pruned problem transformation method for multi-label classification," in *Proc. NZ Comput. Sci. Res. Student Conf.*, 2008, pp. 143–150.
[29] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[30] K. Cram and Y. Singer, "A family of additive online algorithms for category ranking," *J. Mach. Learn. Res.*, vol. 3, pp. 1025–1058, Feb. 2003.
[31] H. Blockeel, L. De Raedt, and J. Ramon, "Top-down induction of clustering trees," in *Proc. 15th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 1998, pp. 55–63.
[32] M. Petrovskiy, "Paired comparisons method for solving multi-label learning problem," in *Proc. 6th Int. Conf. Hybrid Intell. Syst. (HIS)*, Dec. 2006, p. 42.
[33] S.-P. Wan and J.-H. Xu, "A multi-label classification algorithm based on triple class support vector machine," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, 2007, pp. 1447–1452.
[34] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Proc. IEEE Int. Conf. Granular Comput.*, Jul. 2005, pp. 718–721.
[35] P. Vateekul and M. Kubat, "Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2009, pp. 320–325.
[36] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *WIREs Data Mining Knowl. Discovery*, vol. 4, no. 6, pp. 411–444, Nov. 2014.
[37] J. M. Moyano, E. L. Gibaja, K. J. Cios, and S. Ventura, "Review of ensembles of multi-label classifiers: Models, experimental study and prospects," *Inf. Fusion*, vol. 44, pp. 33–45, Nov. 2018.
[38] M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A survey of multi-label classification based on supervised and semi-supervised learning," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 3, pp. 697–724, Mar. 2023.
[39] M. Mundt, Y. Hong, I. Pliushch, and V. Ramesh, "A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning," *Neural Netw.*, vol. 160, pp. 306–336, Mar. 2023.
[40] S. Thrun, *Lifelong Learning: A Case Study*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 1995.
[41] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.
[42] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.
[43] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.
[44] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.
[45] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 532–547.
[46] R. Aljundi, M. Rohrbach, and T. Tuytelaars, "Selfless sequential learning," 2018, *arXiv:1806.05421*.
[47] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
[48] D. Isele and A. Cosgun, "Selective experience replay for lifelong learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1.
[49] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[50] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," 2018, *arXiv:1812.00420*.
[51] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[52] S. Ebrahimi, M. Elhoseiny, T. Darrell, and M. Rohrbach, "Uncertainty-guided continual learning with Bayesian neural networks," 2019, *arXiv:1906.02425*.
[53] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," 2016, *arXiv:1606.04671*.
[54] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7120–7129.
[55] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *Int. J. Social Res. Methodol.*, vol. 8, no. 1, pp. 19–32, Feb. 2005.
[56] M. Roseberry, B. Krawczyk, Y. Djenouri, and A. Cano, "Self-adjusting K nearest neighbors for continual learning from multi-label drifting data streams," *Neurocomputing*, vol. 442, pp. 10–25, Jun. 2021.
[57] C. D. Kim, J. Jeong, and G. Kim, "Imbalanced continual learning with partitioning reservoir sampling," in *Computer Vision—ECCV*. Glasgow, U.K.: Springer, 2020, pp. 411–428.
[58] Y.-S. Liang and W.-J. Li, "Optimizing class distribution in memory for multi-label online continual learning," 2022, *arXiv:2209.11469*.
[59] N. Masuyama, Y. Nojima, C. K. Loo, and H. Ishibuchi, "Multi-label classification via adaptive resonance theory-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8696–8712, Jul. 2023.
[60] T.-N. Pham, Q.-T. Ha, M.-C. Nguyen, and T.-T. Nguyen, "A probability-based close domain metric in lifelong learning for multi-label classification," in *Advanced Computational Methods for Knowledge Engineering*. Cham, Switzerland: Springer, 2020, pp. 143–149.
[61] D. Dalle Pezze, D. Deronjic, C. Masiero, D. Tosato, A. Beghi, and G. A. Susto, "A multi-label continual learning framework to scale deep learning approaches for packaging equipment monitoring," *Eng. Appl. Artif. Intell.*, vol. 124, Sep. 2023, Art. no. 106610.
[62] K. Du, F. Lyu, F. Hu, L. Li, W. Feng, F. Xu, and Q. Fu, "AGCN: Augmented graph convolutional network for lifelong multi-label image recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.
[63] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 321–325.
[64] C.-H. Chen, C.-H. Tu, J.-D. Li, and C.-S. Chen, "Defect detection using deep lifelong learning," in *Proc. IEEE 19th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2021, pp. 1–6.
[65] S. Dong, H. Luo, Y. He, X. Wei, and Y. Gong, "Knowledge restore and transfer for multi-label class-incremental learning," 2023, *arXiv:2302.13334*.
[66] Y. Wang, Z. Wang, Y. Lin, L. Khan, and D. Li, "CIFDM: Continual and interactive feature distillation for multi-label stream learning," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2121–2125.
[67] G. Song, K. Huang, H. Su, F. Song, and M. Yang, "Deep continual hashing for real-world multi-label image retrieval," *Comput. Vis. Image Understand.*, vol. 234, Sep. 2023, Art. no. 103742.

[68] J. Jia, F. He, N. Gao, X. Chen, and K. Huang, "Learning disentangled label representations for multi-label classification," 2022, *arXiv:2212.01461*.
[69] G. A. Carpenter and S. Grossberg, "Adaptive resonance theory," Dept. Cogn. Neural Syst., Center Adapt. Syst., Boston Univ., Boston, MA, USA, Tech. Rep. CAS/CNS TR-98-029, 2010.
[70] K. Du, L. Li, F. Lyu, F. Hu, Z. Xia, and F. Xu, "Class-incremental lifelong learning in multi-label classification," 2022, *arXiv:2207.07840*.
[71] M.-L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.
[72] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.
[73] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, Santorini, Greece, Jul. 2009, pp. 1–9.
[74] K. Nakai, "Yeast," UCI Mach. Learn. Repository, 1996, doi: 10.24432/C5KG68.
[75] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[76] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retr.*, New York, NY, USA, 2008, pp. 39–43.
[77] M. Wieler and T. Hahn, "Weakly supervised learning for industrial optical inspection," in *Proc. DAGM Symp.*, vol. 6, 2007, p. 11.
[78] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[79] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.
[80] D. Tosato, D. D. Pezze, C. Masiero, G. A. Susto, and A. Beghi, "Alarm logs in packaging industry (ALPI)," IEEE Dataport, 2020, doi: 10.21227/nfv6-k750.



MOHAMMED AWAL KASSIM received the B.Sc. degree in computer engineering from the University of Ghana, in 2021. He is currently pursuing the M.Sc. degree in computer science with a concentration in applied artificial intelligence with the University of Ottawa, Canada. Since 2022, he has been a Research Assistant with the School of Electrical Engineering and Computer Science, University of Ottawa. His research interests include machine learning, deep learning, and generative modeling. He was a recipient of the Vector Scholarship in Artificial Intelligence, an entrance award for top students pursuing master's degrees in AI in Ontario. He was also a recipient of the General Electric Scholarship, in 2017; and the Provost Honor Award at the University of Ghana, in 2021.

HERNA VIKTOR is a Full Professor with the School of Electrical Engineering and Computer Science (EECS), University of Ottawa, Canada. Since July 2022, she has been the Director of the School of EECS. She has over 20 years of experience in designing, implementing, and applying machine learning solutions in healthcare, cybersecurity, finance, and biomedicine. She is the author of more than 150 journal articles, conference papers, and book chapters, and serves on the editorial board of three journals. Her research interests include machine learning algorithms for sequential and temporal data, bias-free and discrimination-aware learning, and class imbalance. Her work has received widespread recognition and numerous awards.

WOJTEK MICHALOWSKI is a Professor Emeritus with the University of Ottawa, where he is a Full Professor of health informatics and decision support. He was the Vice-Dean (Research) and the Interim Dean with the Telfer School of Management, University of Ottawa. He is an Adjunct Research Professor with the Sprott School of Business, Carleton University, and an Affiliated Researcher with Montfort Hospital Research Institute. He has authored or coauthored over 160 referenced publications and has been invited to give numerous talks. His research interests include clinical decision support systems, computer-interpretable clinical practice guidelines for multi-morbid patients, and use of predictive modeling to support patient management.