Towards exploring continual learning for toxicologic pathology in pharmaceutical drug discovery

Anonymous CVPR submission

Paper ID *****

Abstract

Machine learning has shown recent promise in advancing toxicologic pathology image analysis during non-clinical safety assessment stages of drug development. However, there exist notable challenges towards implementing these models in real-world scenarios, where datasets are sequentially acquired over extended periods across weeks or months in an Investigational New Drug study. A particular issue that arises in this process is the need for continuous adaptation to new data classes relevant to new tissue types being scanned, or new animal models being incorporated, and so on. Data retention is often hindered by privacy concerns, legal constraints, and storage limitations. Furthermore, existing deep networks are prone to catastrophic forgetting when trained on new tasks, resulting in a substantial loss of previously acquired knowledge. Therefore, there is an urgent need for algorithms that are resilient to forgetting and capable of generalizing to new data without the necessity of retaining large volumes of past examples or datasets. To address these challenges, we introduce a novel replay methodology that leverages generative models, augmented by a knowledge regularization approach utilizing attention embeddings from prior tasks. Our method integrates attention-based regularization, which prioritizes the relative spatial importance of features, with generative latent replay. This synergistic approach enables the model to retain and reinforce critical information from previous tasks while adapting to new data. We empirically demonstrate the superior continual learning performance of our method in non-stationary data environments, as evidenced by its application to a representative toxicologic pathology image analysis task.

1. Introduction

A key component of non-clinical safety evaluations in pharmaceutical drug discovery pertains to analysis of tissue morphologies in animal models subsequent to administration of the candidate drug in an Investigational New Drug (IND) validation process. Toxicologic pathology is a specialized field that integrates the study of toxicology (the science of poisons) and pathology (the study of disease) to understand and evaluate the effects of various agents on living organisms. This discipline is crucial in the regulatory safety assessment process for predicting human and animal responses to drugs, chemicals, and therapeutic devices, including their potential to cause adverse physiological and functional effects. Toxicologic pathologists analyze tissue-level alterations to distinguish between incidental and test article-related findings, crucial for accurate safety assessments in drug discovery. They identify alterations at the cellular, subcellular, and molecular levels, considering factors such as aging, genetics, and nutrition. This process integrates toxicology and pathology to evaluate drug safety and efficacy, predict human toxicities, and identify safety liabilities early in development. Approximately 70% of the toxicity-related attrition occurs in the preclinical phase [5], making it essential to establish a preclinical safety margin for viable drug candidates from the perspective of regulatory guidance and the downstream feasibility of the candidate.

The advent of deep learning has catalyzed significant advancements in toxicologic pathology research, enabling unprecedented analytical capabilities. However, the integration of these technological advances into clinical practice is fraught with substantial hurdles, including the continuous accumulation of large datasets over time and formidable constraints related to privacy, storage, and data quality, which impede the preservation of historical data. Contemporary transfer learning techniques, despite their potential, are often plagued by the phenomenon of catastrophic forgetting [4]. This occurs when a model's performance on initially learned tasks dramatically declines as it is exposed to new tasks, thereby compromising its reliability in deployment environments and hindering its ability to adapt continually to evolving data streams. Moreover, the storage of even meticulously curated subsets of past samples is often untenable due to stringent privacy regulations, logistical

impediments, and ethical considerations [14]. These challenges underscore the need for innovative approaches that enable machine learning systems to learn continuously from streaming data while preserving and leveraging knowledge acquired from previous tasks.

In this paper, we introduce a novel, storage-free approach to continual learning in the analysis of animal histology images sourced from prior drug safety studies, designed to overcome the aforementioned challenges. Our methodology synergistically integrates two core components. We employ a generative model to facilitate incremental-time replay from past data distributions. This model learns a latent space representation of previously encountered data, enabling the sampling of synthetic data points that are interspersed with new data during the learning of subsequent tasks. This strategy obviates the need for explicit storage of past data, thereby circumventing privacy and storage constraints. Concurrently, we preserve a snapshot of the most salient regions from prior classes by constructing class-specific attention embeddings. These embeddings encapsulate the most informative spatial features pertinent to the model's decision-making process for a given task at time point t. During the learning of new tasks at time $t+1$, these attention embeddings are employed for regularization, ensuring that the model retains and leverages the most critical aspects of previously acquired knowledge.

This dual framework introduces a novel pathway for integrating the most relevant regions into the model's decision-making process when learning representations for tasks at a given time point. The representations thus formed encapsulate a snapshot of previously encountered classes, which is presented to the model alongside sampled data points from the latent space learned via the generative model. While knowledge distillation methods have been exploited in prior continual learning systems, the augmentation of learned representations of prior tasks using the most contextually important subsets of inputs, or a latent-space replay for past data distributions, remains largely unexplored. Drawing inspiration from the notion that interpretability techniques encode the subsets of the image space most influential in model inference, we posit that prioritizing and amplifying the preservation of such informative spatial features during incremental learning sessions on novel classes or data can significantly enhance the model's ability to retain the most vital subsets of previously seen data. Our approach is underpinned by a robust theoretical framework and is meticulously evaluated on a range of digital pathology datasets, encompassing both histology images and toxicologic pathology data. Through extensive experimentation, we demonstrate that our method not only mitigates catastrophic forgetting but also enhances the model's adaptability, robustness, and generalization capabilities in dynamic real-world clinical settings. Furthermore, we show that our approach outperforms state-of-the-art continual learning techniques, thereby establishing its potential as a vital tool for toxicologic pathology and beyond.

In the subsequent sections, we delve into the intricacies of our proposed methodology, elucidate the experimental setup, present a comprehensive analysis of the results, and discuss the broader implications of our work. We conclude by highlighting avenues for future research, emphasizing the potential of our approach to revolutionize continual learning in data-constrained environments. Our approach is found to improve continual learning for healthcare, ensuring that models can adapt to new data without compromising the knowledge gained from previous tasks, thus maintaining high performance and reliability.

2. Prior Work

While digital pathology in the clinical realm has seen a surge of research seeking to integrate various machine learning approaches towards enhanced understanding of pathology imaging data, such approaches in preclinical safety evaluations have been relative nascent. There has been an emerging interest in evaluating deep learning for regulatory toxicologic pathology applications [22], with some recent studies documenting the automated identification of histopathologic lesions in whole slide images (WSIs) from regulatory toxicity studies, utilizing pixel-based manual annotations. Kuklyte et al. [11] employed various convolutional neural network (CNN) architectures to detect and quantify histopathologic lesions in toxicity studies. Their methodology involved training algorithms with pixel-based annotations of a substantial number of histopathologic lesions across five different organs. A comparable approach was adopted to identify proliferative lesions in the positive control group of Tg-RasH2 mouse carcinogenicity studies [16]. Shimazaki et al. [20] developed a customized U-Net architecture trained with pixel-based annotations of seven distinct liver lesions. In contrast, Zehnder et al. utilized an unsupervised approach [23], leveraging generative adversarial networks (GANs) and autoencoder architectures trained on normal liver samples to minimize the anomaly score for normal data. Consequently, exposing the model to abnormal data results in reconstruction and discrimination failures, leading to elevated anomaly scores. With the emergence of foundation models for digital pathology, extensions towards self-supervised representation learning pertinent to regulatory toxicity evaluations have been proposed for the discovery of morphomolecular signatures [7], and also for more general applications towards building downstream tasks for preclinical safety assessments [6].

Efforts to address steep declines in model validation performance with non-stationary learning schedules has seen the emergence of methods: a) based on preserving weights

like EWC [10], or freezing parameter subsets and progressively expanding for new data[17], which causes computational and memory overheads; b) knowledge-based regularization using distillation [13] for providing a glimpse of learnt representations from past tasks to enable the model finetuning during incremental adaptation to account for new and old tasks' knowledge; c) rehearsal methods that use generative models to recreate batches of previously seen data points [8]. Consolidation methods have struggled to scale to complex real-world datasets in clinical imaging [8], and replay approaches often require a storage of original inputs from past classes or tasks in a buffer for batch-mixing during new task learning. Recently, buffer-free methods have been explored in medical imaging [15, 21], primarily using stored features or separate replay modules, sharing of which causes regulatory complexities in hospital systems [21]. At the time of writing, we came across a recent work proposing a latent replay regime similar to ours[12], but uses non-public datasets and focuses mainly on domain incremental learning without considering the role of attention in regularization of future learning by prioritising salient features of past classes.

Most explainability methods image analysis seek to identify subregions of images most influential to specific model decisions. Usage of these ideas in continual learning regimes was explored in [2], but was limited to using image-level combinations for creating past subsets. We propose to directly use information from explainability workflows to enhance learnt task representations towards enforcing stronger distillation-driven regularization. To accomplish a buffer-free replay of samples pertinent to prior tasks, we retain a latent representation of the classes from past tasks to allow incremental time sampling using a generative model as an alternative to replay using retained original samples from past learning stages. To our knowledge, the question of continual learning for toxicologic pathology evaluations in drug discovery has not been explored despite the clear existence of data acquisition regimes that are temporally spaced in the non-clinical stages of pharmaceutical development. This is crucial from an image analysis standpoint as the deployment of machine learning for the preclinical discovery workflows requires adaptation of algorithms over the weekly or monthly arrivals of data, as well as from different tissue types and animal models in the course of an Investigational New Drug evaluation.

Our work proposes a continual learning methodology for the development of robust and adaptable models for toxicologic pathology image analysis, with implications for improving the accuracy of safety assessments and enhancing the understanding of tissue responses to pharmacological, chemical, or environmental agents. By mitigating catastrophic forgetting and enhancing the model's ability to generalize to new data, our approach has the potential to facilitate more widespread adoption of deep learning based image analysis for regulatory toxicologica pathology workflows, which would help accelerate non-clinical safety evaluation stages in preclinical development and enable expedited timelines for drug development in pharmaceutical settings.

3. Methodology

3.1. Datasets

We utilized the toxicologic pathology imaging dataset curated by Serna et al. [19] to validate our continual learning protocol, as it effectively models the practical requirement of multi-task tissue recognition over extended periods using the multi-magnification whole slide image (WSI) dataset contained therein. The dataset comprises nine preclinical rat studies with tissue processing conducted at three distinct contract research organization (CRO) laboratories. All studies were performed on Wistar Han rats and Sprague-Dawley rats. The original acquisition protocols, adhering to Swiss regulations and approved by the Cantonal Ethical Committee for Animal Research, ensured ethical compliance and rigorous experimental standards. The WSIs in this study encompass organs both with and without histopathological findings, and include scans from Hamamatsu (ndpi) and Aperio (svs) scanners. The studies were intentionally selected to incorporate significant staining variations, thereby facilitating the development and validation of algorithms robust to real-world settings where diverse scanning, staining, and storage protocols may be encountered during investigational new drug studies. The associated metadata for the organs included in each WSI were derived from internal protocols specifying the grouping of different organs per WSI.

The resulting dataset comprises N=320 whole slide images, featuring the following organs: liver, kidney, thyroid gland, parathyroid gland, urinary bladder, salivary gland, mandibular lymph node, and others (negative class). The liver and kidney were prioritized due to their critical relevance in preclinical safety assessments. Additionally, organs frequently embedded with them, such as submandibular lymph nodes and salivary glands (embedded with liver), and urinary bladder (embedded with kidneys), were included. The thyroid and parathyroid glands were also incorporated to evaluate the detection of small, closely situated organs. To enhance the positive selection of desired organs, various confounding organs from nine organ sets were included as a negative class for model training (class "other"): adrenal glands, aorta, and ureters; lung and heart; stomach and intestine; skeletal muscle, sciatic nerve, mammary gland, and skin; prostate and seminal vesicles; testis and epididymis; eye and harderian glands; bone with bone marrow; and spinal cord. This comprehensive approach ensures a robust and versatile dataset for advancing regulatory tox-

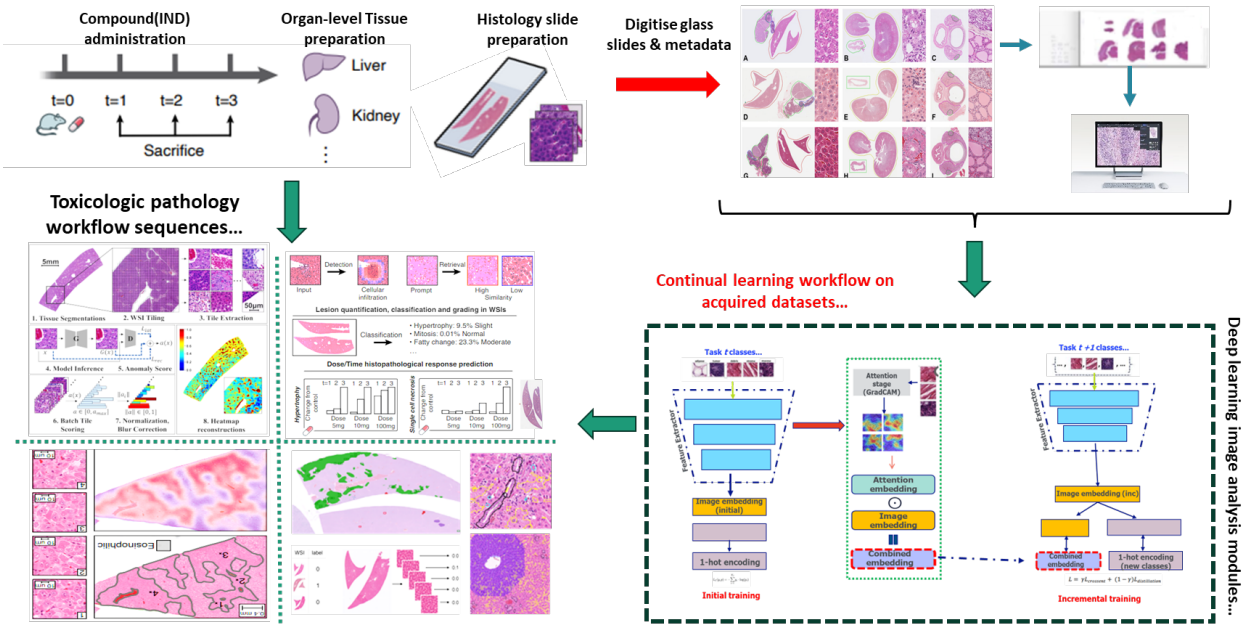

Figure 1. An abridged, representative toxicologic pathology and image analysis workflow in a pharmaceutical drug development context, with digitisation and deep learning approaches built in for preclinical safety evaluations.

icologic pathology image analysis. It is noted that the confounding tissue presence with the main tissues under study are not treated as separate independent classes in the continual learning settings proposed, since the simulation of such a setup requires us to have multiple tasks that are temporally spaced, but the tasks themselves need to include well-defined classes being studied per task. This is a shortcoming of incremental task learning protocols for continual adaptation to new data streams in toxicologic pathology studies, and can be alleviated potentially through compute-efficient adaptations of online learning or similar approaches, which would be a non-trivial exercise given the significant memory and computational overheads involved in processing of large whole slide images typically involved in such pathology studies and the requirement for several sequential analysis steps to be conducted over a large number of animal models per study.

3.2. Problem Definition

We explore a setting where a model is to be trained in an M-stage protocol, where every stage is a classification task with classes as $X_t = \{X_{t,i}\}_{i=1}^{Q_t}, t \in [1, M]$, with each X being a class representing its samples (i.e., $x_t \in X_t$), and Q_t being the number of classes at stage t. At stage t, the classifier from previous stage $t-1$ is incrementally optimized over the classes at current stage. The objective of the learning is that, given a small threshold ϵ, after the t stage optimization, the reductions in the inference accuracies over validation sets from all previous stages $D = \{D_1, .., D_{t-1}\}$ meet the following criteria:

$$\forall d \in D, d < \epsilon$$

We design this as an incremental learning experiment with four classes in the initial training stage and four subsequent classes in the incremental stage ($M = 2, Q_1 = Q_2 = 4$). This study is modelled as a sequential learning task over sets of classes as above, with a proportion of classes being learnt as **base classes** during an initial training stage. Next, the remaining classes are learnt as **incremental classes** in a subsequent learning stage, leading to a multistage learning system over a temporal interval. The base classes are the *liver (LI), kidney (KD), Thyroid gland (TG)* and *urinary gland (UG)*. The incrementally learnt classes include *parathyroid gland (PG), salivary gland (SL), mandibular lymph nodes (MLN)* and *Other classes (BG)*. The former are used to optimize for the initial task (task 1) and the latter are used to adapt the model trained over base classes for the incremental task (task 2), thus simulating a continual learning scenario. In the first stage, where task 1 is performed, a Resnet-50 derived model is used to conduct a classification task for the initial set of classes. The result of the first stage using a cross-entropy loss is $p = \text{softmax}(z) \in \mathbb{R}^{Q_1}$, where z is the set of logits. The classification loss in the first stage is defined as:

$$L_C(y, p) = -\sum_{i=1}^{Q_1} y_i \cdot \log(p_i) \quad (1)$$

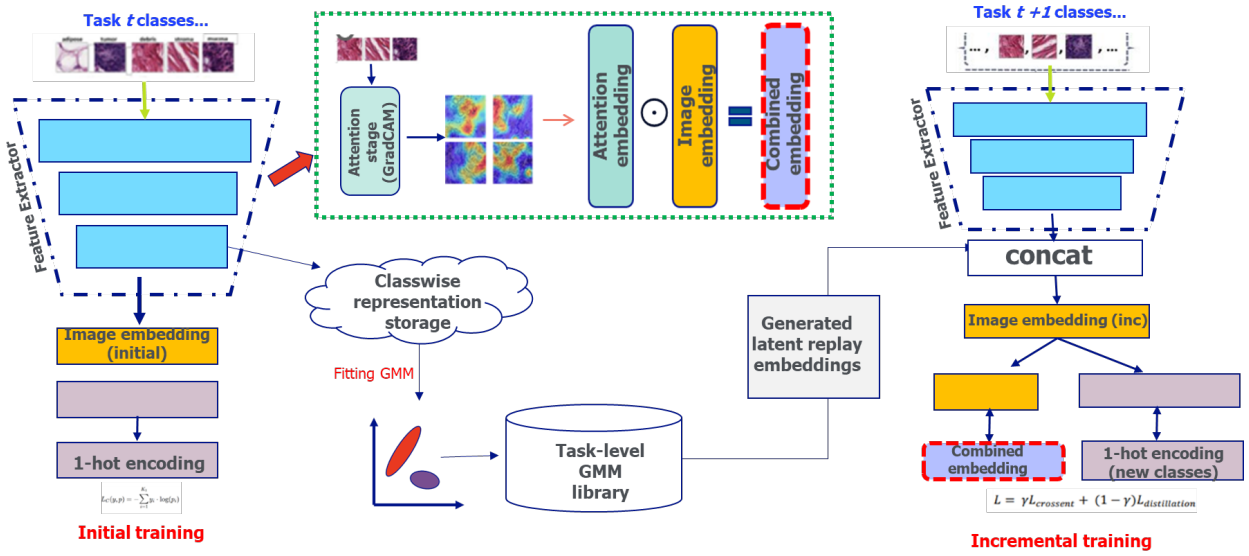

Figure 2. Our overall pipeline for computing attention embeddings and creating the GMM library for the latent replay operations during incremental learning over subsequent dataset arrivals corresponding to the new tissue classes.

where p_i is the predicted probabilities of classes in the initial task, y_i is the corresponding ground truth represented as a one-hot encoding.

3.3. Latent Replay

Many continual learning (CL) approaches employ a replay buffer to selectively retain samples from previous tasks, which are then mixed with data from current classes in a process known as rehearsal. This technique is designed to mitigate catastrophic forgetting, where the model loses the ability to perform well on previously learned tasks as it adapts to new ones. However, storing actual samples from past tasks can be problematic due to memory constraints and potential privacy concerns, particularly in sensitive domains such as medical imaging. To address these challenges, we propose a novel latent replay method that performs rehearsal without the need to store any actual samples from previous tasks. Instead, our approach involves sampling from a learned latent space that represents past classes. This is achieved by constructing Gaussian Mixture Models (GMMs) for each class during each training session. The GMMs capture the statistical distribution of the data, allowing us to generate synthetic samples that closely resemble the original data without retaining any of the actual samples. In the context of deep learning models, the initial layers are typically responsible for extracting low-level features from the input data, such as edges, textures, and simple patterns. These features are generally applicable across a wide range of tasks and datasets. Through pre-training on initial datasets, the weights of these early layers stabilize and can be effectively repurposed for various applications, including complex tasks like medical image processing. For instance, Srivastava et al. [21] demonstrated the efficacy of transfer learning in medical imaging, where pre-trained models were fine-tuned for specific medical tasks, achieving state-of-the-art performance. Conversely, the layers closer to the classification head of the model are tasked with extracting high-level, discriminative features that are specific to the classes and tasks at hand. These layers require fine-tuning to optimize the model's accuracy for the target task. To leverage this hierarchical feature extraction process, we propose to extract fine-grained features from an intermediate layer of the model, which we designate as the replay layer.

Our method involves training generators using the activations from this replay layer. Instead of storing raw histopathology samples in a buffer memory, we only store the generators. This approach significantly reduces memory requirements and mitigates potential privacy violations, as the generators do not contain any actual patient data. During the initial training phase for each class, we retain the layer preceding the last batch normalization layer of a ResNet50 backbone as the averaged representation of every class per task stage. This averaged representation serves as a compact and informative summary of the class-specific features, which can be used to generate synthetic samples for rehearsal during subsequent training sessions. By doing so, we ensure that the model maintains its performance on previously learned tasks while adapting to new ones, all without the need to store any actual samples from past tasks. A GMM is constructed using these representations per class for a task session, similar to the protocol proposed by [12].

After each task session, we learn session-specific GMMs per class using aggregated features obtained during training from the ResNet50 backbone. For the $\mathbf{t}^{\text{th}}$ session and $\mathbf{i}^{\text{th}}$ class, the generator includes K multivariate Gaussians $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_K$ in the mixture with probability density function (PDF) as:

$$p(\mathbf{f}_t) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{f}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (2)$$

where $\boldsymbol{\mu}_k = \frac{1}{N_t} \sum_{n=1}^{N_t} \mathbf{f}_t^n$ and $\boldsymbol{\Sigma}_k = \frac{1}{N_t} \sum_{n=1}^{N_t} (\mathbf{f}_t^n - \boldsymbol{\mu}_k)(\mathbf{f}_t^n - \boldsymbol{\mu}_k)^T$ Here, N_t is the number of samples in session t for a class, $\mathcal{N}(\mathbf{f}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the PDF of the k^{th} component with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The value of $K \in [1, K_{\max}]$ is estimated using Bayesian Information Criterion approaches in cluster analysis [3]. For each K, parameters of the mixture are computed by an EM algorithm [1]. Finally, the fitted model B_t^i enables generation of random samples for session i and class t. Over sequential task sessions, a library of session-specific GMMs are accumulated to allow the creation of representations from any past classes and sessions or task stages for replay.

3.4. Attention Guidance Regularization

The optimization of neural architectures has predominantly relied on the empirical risk minimization framework facilitated by gradient-based weight adaptation, conceptualizing the learning dynamics as analogous to the hill-climbing problem. In this paradigm, the reduction of loss, defined as the discrepancy between actual and predicted labels, is akin to descending a gradient slope [10]. Theoretically, this process can be interpreted as the movement of gradient values within a conservative field, suggesting that the transition from a loss value L_1 to a lower minimum L_2 is independent of the specific parameter changes during the process and is solely contingent on the initial and final parameter sets [4].

In practical terms, this implies that during transfer learning across new distributions, there is no inherent regularization governing how parameter configurations may adapt to optimize for the new task. Conversely, it is computationally infeasible to control shifts in the majority of the network's parameters. A potential compromise involves prioritizing input features that exert a disproportionate influence on model outcomes, thereby indirectly enforcing regularization on the most critical model parameters. This insight forms the basis of our approach, which incorporates attention embeddings into the continual learning process.

While the computation of such embeddings and their integration into sequential learning processes has been sparsely explored in the literature, we investigated a relatively straightforward paradigm. In this approach, we assess the suitability of attention embeddings as auxiliary branches appended to the standard pre-softmax logits retained from previous learning instances.

While the generative latent replay stage allows sampling from a latent space relying only on the mean and variance of the learnt representations, the incorporation of a notion of feature importance towards relaying the most salient parts of a data distribution is crucial towards enabling past tasks' regularization during incremental training on new classes. This is achieved by 'attentive distillation', where we reinforce past task embeddings for distillations with task-relevant attention vectors. We use class activation mapping [18] for interpreting the model outputs in terms of a heatmap of most relevant regions of the images used for prediction. Such a step is carried out for the top-25% of the instances in classes in the initial task, in the interest of a trade-off between the expressiveness of the model representation and real-time memory constraints. For individual images per class in the training data subset, we use the attention module to compute corresponding activation maps.

The class activation heatmaps so obtained are then vectorized corresponding to the image representations of such instances as projected in the penultimate fully-connected layers of the model. The attention embeddings are then cast into the dimensionality of the image representation logits obtained in pre-softmax layers using average-pooling operations, and a dot product operation is performed between the attention vectors and relevant image embeddings. This enables boosting of the most salient subsets of the image space in the final representation. Finally, the class-specific embeddings that reflect both the image representation and corresponding activations are obtained for usage in the distillation-based regularization steps. Considering the initial task logits prior to the superimposition of the attention embeddings as w_{old}, the inclusion of the attention embedding a_{old} results in the overall old class logits as z_{old}, which are weighted and used in the distillation objective for incremental new tasks. In subsequent sessions, a distillation term is used in the objective to allow inclusion of past knowledge in the optimization (y' are final class probabilities for new task classes prior to softmax operation):

$$L_{distillation}(z_{old}, y') = -\sum_{i=1}^{N} \text{softmax}(\frac{z_{old}}{T}) \cdot \log(\text{softmax}(\frac{y_i'}{T})) \quad (3)$$

Logits and predictions are scaled with a temperature hyperparameter T in a softening process. It helps reduce the disparity between the class label with the highest confidence score in the probability vector with respect to other class labels and helps better reflect inter-class relationships at the representation learning stage. Considering the overall logit vector for old classes, after weighting as z_{old}, class-specific logits are weighted to obtain a sum of class-weighted logits as:

$$z_{old} = \sum_{i=1}^{K_1} u_i \cdot z_i \quad (4)$$

Table 1. Accuracy (%) for task 1/stage 1 classes, after task 1 is trained for, and after task 2 is incrementally added in stage 2. The difference in accuracies on validation sets of task 1 classes represents forgetting on task 1 classes due to task 2 addition. AM - Attention Modulated, wKD - weighted Knowledge Distillation, FT- simple Finetuning, '+' indicates combined usage

Stage	Stage 1 (for task 1, T1)					Stage 2 (for task 2, T2)					ΔAcc
	LI	KD	TG	UG	Avg(T1)	LI	KD	TG	UG	Avg(T2)	T2-T1
Our(AM+wKD)	82.45	80.77	78.36	79.44	80.26	78.65	76.84	74.22	76.91	76.66	3.61
Our(AM+KD)	82.45	80.77	78.36	79.44	80.26	73.62	72.46	70.81	70.21	71.77	8.49
Our (KD)	82.45	80.77	78.36	79.44	80.26	66.37	61.78	59.34	62.51	62.50	17.76
Our (AM+FT)	82.45	80.77	78.36	79.44	80.26	65.03	58.14	56.77	60.31	60.06	20.20
Ours (FT)	82.45	80.77	78.36	79.44	80.26	45.11	37.95	36.53	40.91	40.12	40.14
LwF.ewc [9]	82.45	80.77	78.36	79.44	80.26	61.75	58.37	53.82	57.26	57.80	22.46
Priv [21]	82.45	80.77	78.36	79.44	80.26	65.47	63.04	56.71	61.07	61.57	18.69
PCL [12]	82.45	80.77	78.36	79.44	80.26	65.29	63.78	56.32	60.02	61.35	18.91
MT [15]	82.45	80.77	78.36	79.44	80.26	65.98	63.17	58.43	62.33	62.48	17.78
DGR [21]	82.45	80.77	78.36	79.44	80.26	63.18	59.82	56.74	60.12	59.97	20.29

The logits from individual classes $z_i, i \in [1, Q_1]$ are calculated by averaging pre-softmax probability values (after sigmoid activation) for examples from each of Q_1 classes. The weights $(u_1, u_2, \ldots, u_{Q_1})$ are computed as inverse of class-specific accuracy on validation sets of the initial classes. The idea is to boost logits from classes which are inherently difficult to learn for the model (lower the class-specific accuracy, higher the class weight). This reduces the disparity among classes in their contribution towards the overall sessional representation vector to be saved as an imprint of stage 1 learning. Overall, the net incremental task training objective for learning beyond initial sessions then becomes ($\gamma = 0.5$):

$$L = \gamma L_{crossent} + (1 - \gamma) L_{distillation} \quad (5)$$

4. Experiments and Discussions

The experiment is split into two sequential tasks, labeled task 1 and task 2. The initial task proceeds with a cross-entropy loss and task 2, the incremental task utilizes a joint loss with a cross-entropy and a distillation term with a ResNet-50 feature extractor. The pre-softmax layer generates probability scores by a sigmoid operation. An 80:20 split is used for train:test split on datasets. Input images are resized to 224x224 and a batch size of 25 is used with a learning rate of 0.0001 and Adam optimization. task 1 models are trained for 250 epochs on a *(N,label)* set for all N frames. In task 2, models are trained for 250 epochs on *(N',label,logit)* tuples, where the logit is the attention-modulated version of logits obtained in pre-softmax layers after the initial task. We set T = 5.0 after grid search in $T \in [1, 10]$. Two 32 GB Nvidia V100 GPUs, 512MB RAM were used for training using the ResNet-50 base models (˜24.8 million parameters, implemented in Python 3.7.1 and Tensorflow 2.0), with an average training time of 102s per epoch observed in both tasks. In terms of the computational landscape, the primary usage of floating point operations stems from the model training procedures over the datasets from each of the tasks involved. The computation of the attention maps, and the generation of the class-level sampled embeddings from the GMMs in the library defined per task session, account for a minority of the computational budget.

4.1. Results

To design the incremental learning setup in the context of toxicologic pathology, Task 2 is initiated with four incremental classes following the completion of Task 1, where four initial classes were trained. The top 25% of samples per class are processed through the attention modules to generate attention maps. These maps are then vectorized, pooled, and transformed into embeddings with dimensions comparable to those obtained for corresponding images processed through the model. The dot product between image embeddings and their corresponding attention embeddings is aggregated over selected images per class to derive class-level embeddings. These embeddings are integrated into the net representation used for regularization in the distillation loss terms during the second stage of training for Task 2.

The inclusion of the most informative regions from the top quartile of correct predictions in prior tasks, along with sampling from the generators in the Gaussian Mixture Model (GMM) library constructed for the initial data, ensures robust task and class-level knowledge availability during incremental training. Notably, Task 1 classes exhibit consistent accuracies across compared baselines and proposed methods, as identical backbones are employed in a pure classification task for the initial stage.

In Table 1, evidence of catastrophic forgetting is observed, marked by a sharp decline in Task 1 validation accuracy following sequential optimization over Task 2 data (**ΔAcc**), when no intervention is applied to mitigate it. The implementation of our combined latent replay with attention-informed embeddings to preserve prior knowl-

edge effectively mitigates such steep performance declines. This is evident when comparing relative reductions in accuracy (Δ**Acc**) for methods that employ end-stage distillation alone. The dynamic interplay of attention-modulated weighted distillation with intermediate concatenations of generative representations significantly reduces accuracy declines, as measured by validation performance on Task 1 holdout sets before and after training on new task classes. We adapted recent buffer-free methods in clinical datasets to compare our joint approach of generative replay and attentive distillation in our class-incremental setting. Our method outperforms strategies such as feature retention and generative replay alone, indicating a non-trivial influence of enhancing the visibility of the most salient image-level features for past tasks' logit vectors used for regularization in incremental training phases for the new classes' data. We evaluated knowledge regularization settings involving class-weighted and non-weighted configurations, including versions that utilize attention embeddings, with the former demonstrating performance gains across attention and replay configurations.

Interestingly, measures to address past task forgetting also improve classification accuracies on new classes (Fig. 3), likely due to enhanced regularization causing forward transfer effects which are potentially a result of more optimal initializations of the parameter spaces during the initial optimization task. Overall, incorporating insights from spatial feature importance over past inputs, along with generative latent rehearsal, shows clear gains in the model's learning regime, enforcing generalization when optimized on new tasks' data in toxicologic pathology applications.

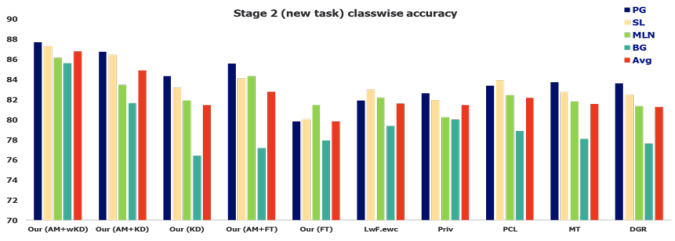

Figure 3. Model performance on new classes after incremental adaptation showing positive forward transfer effects.

5. Conclusion

We present novel methods for continual learning for regulatory safety applications in drug discovery, using a public toxicologic pathology dataset. Using attention embeddings with a generative replay of samples in concordance with previous data distributions is found to enhance the representative power of the most salient input subsets without requiring storage of actual prior instances over time, and shown to mitigate forgetting during cross-distillation based continual adaptations on incremental tasks. This work represents an early effort towards examining the question of tissue-level image analysis in drug development contexts through the lens of continual learning as a means towards performing efficient machine learning based drug safety assessments in the light of temporally-spaced arrivals of datasets during the toxicologic pathology process. In future, we will adapt our methods on systems reliant on transformer architectures and similar approaches, study computational complexities of scaling attention and generation modules on larger datasets and any resultant scaling properties that might become evident in the process, and integrate multimodal data relevant to histopathological evaluations in clinical and preclinical pathology workflows, including but not limited to, multi-omics data and associated metadata that may contain salient information about diagnostic states under study.

References

[1] Frank Dellaert. The expectation maximization algorithm. *College of Computing, Georgia Institute of Technology*, 2002. 6

[2] P. Dhar, R. V. Singh, K. C. Peng, Z. Wu, and R. Chellappa. Learning without memorizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019. 3

[3] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998. 6

[4] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 1, 6

[5] Cornelis ECA Hop. Compound attrition at the preclinical phase. *Attrition in the Pharmaceutical Industry: Reasons, Implications, and Pathways Forward*, pages 46–82, 2015. 1

[6] Guillaume Jaume, Simone de Brot, Andrew H Song, Drew FK Williamson, Lukas Oldenburg, Andrew Zhang, Richard J Chen, Javier Asin, Sohvi Blatter, Martina Dettwiler, et al. Deep learning-based modeling for preclinical drug safety assessment. *bioRxiv*, 2024. 2

[7] Guillaume Jaume, Thomas Peeters, Andrew H Song, Rowland Pettit, Drew FK Williamson, Lukas Oldenburg, Anurag Vaidya, Simone De Brot, Richard J Chen, Jean-Philippe Thiran, et al. Ai-driven discovery of morphomolecular signatures in toxicology. *bioRxiv*, 2024. 2

[8] R. Kemker and C. Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017. 3

[9] H.E. Kim, S. Kim, and J. Lee. Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018. 7

[10] 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. *Proceedings of the national academy of sciences*, pages 3521–3526, 2017. 3, 6

[11] Jogile Kuklyte, Jenny Fitzgerald, Sophie Nelissen, Haolin Wei, Aoife Whelan, Adam Power, Ajaz Ahmad, Martyna Miarka, Mark Gregson, Michael Maxwell, et al. Evaluation of the use of single-and multi-magnification convolutional neural networks for the determination and quantitation of lesions in nonclinical pathology studies. *Toxicologic Pathology*, 49(4):815–842, 2021. 2

[12] P. Kumari, D. Reisenbüchler, L. Luttner, N. S. Schaadt, F. Feuerhake, and D. Merhof. Continual domain incremental learning for privacy-aware digital pathology. *arXiv preprint arXiv:2409.06455*, 2024. 3, 5, 7

[13] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3

[14] Arijit Patra, Yifan Cai, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Multimodal continual learning with sonographer eye-tracking in fetal ultrasound. In *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2*, pages 14–24. Springer, 2021. 2

[15] H. Ravishankar, R. Venkataramani, S. Anamandra, P. Sudhakar, and P. Annangi. Feature transformers: privacy preserving lifelong learners for medical imaging. In *MICCAI 2019: 22nd International Conference*, Shenzhen, China, 2019. 3, 7

[16] Daniel Rudmann, Jay Albretsen, Colin Doolan, Mark Gregson, Beth Dray, Aaron Sargeant, Donal O'Shea D, Jogile Kuklyte, Adam Power, and Jenny Fitzgerald. Using deep learning artificial intelligence algorithms to verify n-nitroso-n-methylurea and urethane positive control proliferative changes in tg-rash2 mouse carcinogenicity studies. *Toxicologic Pathology*, 49(4):938–949, 2021. 2

[17] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[19] Citlalli Gámez Serna, Fernando Romero-Palomo, Filippo Arcadu, Jürgen Funk, Vanessa Schumacher, and Andrew Janowczyk. Mmo-net (multi-magnification organ network): A use case for organ identification using multiple magnifications in preclinical pathology studies. *Journal of Pathology Informatics*, 13:100126, 2022. 3

[20] Taishi Shimazaki, Ameya Deshpande, Anindya Hajra, Tijo Thomas, Kyotaka Muta, Naohito Yamada, Yuzo Yasui, and Toshiyuki Shoda. Deep learning-based image-analysis algorithm for classification and quantification of multiple histopathological lesions in rat liver. *Journal of Toxicologic Pathology*, 35(2):135–147, 2022. 2

[21] S. Srivastava, M. Yaqub, K. Nandakumar, Z. Ge, and D. Mahapatra. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 226–238, 2021. 3, 5, 7

[22] Oliver C Turner, Famke Aeffner, Dinesh S Bangari, Wanda High, Brian Knight, Tom Forest, Brieuc Cossic, Lauren E Himmel, Daniel G Rudmann, Bhupinder Bawa, et al. Society of toxicologic pathology digital pathology and image analysis special interest group article*: opinion on the application of artificial intelligence and machine learning to digital toxicologic pathology. *Toxicologic Pathology*, 48(2):277–294, 2020. 2

[23] Philip Zehnder, Jeffrey Feng, Reina N Fuji, Ruth Sullivan, and Fangyao Hu. Multiscale generative model using regularized skip-connections and perceptual loss for anomaly detection in toxicologic histopathology. *Journal of Pathology Informatics*, 13:100102, 2022. 2