ConceptIL: Concept Incremental Learning for Interpretable Medical AI

Hangqi Zhou[1], Bomin Wang[1], and Xiahai Zhuang[1⋆]

School of Data Science, Fudan University, Shanghai, 200433, China

Abstract. Despite providing concept-based explanations for decision-making in medical AI, concept bottleneck models (CBMs) suffer from prediction distortions under insufficient concept annotations. Inspired by real-world scenarios where experts continually learn diagnostic signs, we introduce the framework of concept incremental learning (ConceptIL). Such learning framework acquires knowledge of new concepts from sequential data with limited annotation of concepts, and further promotes interpretability of AI decision-making. However, achieving this ConceptIL can be challenging due to the limited access of historical data and imbalanced distribution of concepts. To address these, we propose an analytical concept learning scheme, which optimizes parameters through a recursive formulation. This ConceptIL method achieves concept prediction accuracy theoretically proven to be equivalent to that of joint training. Additionally, we propose a dynamic weighting strategy to mitigate the concept imbalance problem. Results on a skin disease dataset show that our method matches the diagnostic performance of models trained on the full concept set, demonstrating the effectiveness of our incremental concept learning. Moreover, our work obtains the best performance compared to competing incremental learning algorithms, establishing a new state-of-the-art for ConceptIL.

Keywords: Explainable-AI · Concept Learning · Incremental Learning · Analytical Learning.

1 Introduction

In recent years, interpretable artificial intelligence has attracted growing interest in medical applications, driven by the demand for transparent and trustworthy clinical decision-making in healthcare [12,15,14]. A promising approach of making interpretable decisions is to use concept bottleneck models (CBMs) and other variants [7,3,16]. Ideally, given sufficient expert-labeled concepts, CBMs are capable of achieving strong performance in both concept prediction and downstream tasks. However, in real-world clinical practice, it is extremely labor-intensive and time-consuming for experts to annotate a large set of concepts in medical images. For instance, dermatologists need to select concept terms from 48 clinical concepts for each image in the SkinCon dataset [2]. Typically,

⋆ Corresponding author. E-mail: zxh@fudan.edu.cn

a dataset only has a limited number of labeled concepts. This scarcity has been both theoretically and empirically proved to cause inherent prediction distortions in CBMs [16,18]. Post-hoc CBMs [16,13] and label-free CBMs [8] leverage vision-language models to bypass this limitation. However, they mainly study how to build interpretable models under a fixed concept set. Specially, Shang *et al.* proposed Res-CBM [13] to discover residual concepts from the candidate concept bank within a single task, but the number of residual concepts is still pre-defined. Moreover, interpretability of these methods is constrained by the concept misalignment from pre-trained models.

To address these limitations, we introduce the framework of **concept incremental learning (ConceptIL)**. In this learning framework, models learn new concepts from sequentially arriving data with limited concept annotation, and further leverages all observed concepts to promote interpretable decision-making. This reflects real-world scenarios where experts gradually recognize evolving diagnostic signs to re-constitute their knowledge, based on the stepwise acquisition of patient data, updates of clinical guidelines, and newly introduced devices [5]. ConceptIL poses two challenges. First, similar to conventional incremental learning (IL) [6,10,11,1], the demand for preserving data privacy limits the access of previous data. Second, upcoming concepts may have limited positive samples. Such imbalanced distribution of concepts make it difficult to learn emerging concepts.

To address these challenges, inspired by recent advances in analytical learning [21,22,20], we propose an analytical concept learning scheme, introducing an analytical concept predictor (ACP) that optimize its parameters via a recursive formulation. It is theoretically proved that ACP provides results equivalent to those obtained from joint training over all data, which is beneficial for incremental concept learning. Moreover, instead of saving historical samples from old tasks, ACP stores the inverse correlation matrix for each concept, thereby preserving data privacy. Furthermore, we propose a dynamic weighting strategy to address the challenge of concept imbalance. Evaluations on a skin disease dataset demonstrate that the proposed ConceptIL method matches the diagnostic performance of models trained with full concept annotation. In addition, this method outperforms competing incremental learning methods across two incremental settings.

Our contributions can be summarized as follows:

- We introduce the framework of concept incremental learning (ConceptIL), where models learn new concepts from sequential data with limited annotation of concepts, and further enhances interpretability of AI decision-making. To the best of our knowledge, this is the first attempt to investigate such a problem for medical AI.
- We propose an analytical concept learning scheme to address the challenges in ConceptIL. This scheme adopts an analytical concept predictor (ACP) to solve parameter optimization via a recursive formulation. Additionally, we propose a dynamic weighting strategy to address concept imbalance.

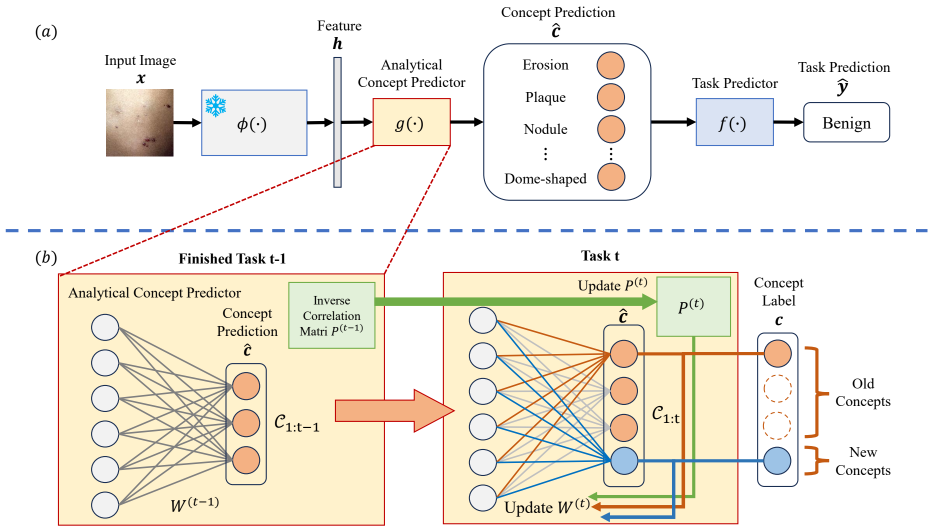

Fig. 1. (a) The overall framework of the proposed analytical concept learning scheme. It comprises a frozen feature extractor $\phi(\cdot)$, an analytical concept predictor (ACP) $g(\cdot)$, and a task predictor $f(\cdot)$. Both predictors are linear layers. (b) Illustration of how parameters W of ACP are updated from task $t-1$ to t. ACP stores the inverse correlation matrix P, which represents the memory of previous data without violating data privacy. $W^{(t)}$ and $P^{(t)}$ are updated via a recursive formulation.

- Experimental results across two ConceptIL settings show that our ConceptIL method achieves the diagnostic performance comparable to that of models trained on the full concept set. Furthermore, this method obtains the best results among competing incremental learning algorithms.

2 Method

2.1 Overall Framework of Concept Incremental Learning

In the T-task incremental learning, a network is trained on T sequential tasks without access to data from previous tasks. In ConceptIL, each task includes annotations for only a small subset of concepts. Let $\mathcal{C}_t$ be the set of annotated concept indices in task t, and $K_t = |\mathcal{C}_t|$ be the number of annotated concepts. The set of all concepts observed from task 1 to task t can be written as $\mathcal{C}_{1:t} = \bigcup_{i=1}^{t} \mathcal{C}_i$, with the total number of concepts $K_{1:t} = |\mathcal{C}_{1:t}|$. In task t, each image $\mathbf{x} \in \mathbb{R}^{d_x}$ has a one-hot task label $\mathbf{y} \in \mathbb{R}^{d_y}$ and a concept vector $\mathbf{c} \in \mathbb{R}^{K_{1:t}}$, where only the entries indexed by $\mathcal{C}_t$ have ground-truth labels (0 or 1).

Fig. 1(a) illustrates the overall framework of the analytical concept learning scheme. Given $\mathbf{x}$, this ConceptIL method first uses a frozen feature extractor $\phi(\cdot)$ to obtain the image embedding $\mathbf{h} \in \mathbb{R}^{d_h}$ from $\mathbf{x}$: $\mathbf{h} = \phi(\mathbf{x})$. After that, the analytical concept predictor (ACP) $g(\cdot)$ maps the image embedding to the concept prediction $\hat{\mathbf{c}} = g(\mathbf{h}) \in \mathbb{R}^{K_{1:t}}$. Finally, the task predictor $f(\cdot)$ generates the final task prediction: $\hat{\mathbf{y}} = f(\hat{\mathbf{c}})$. Here, we use one linear layer for both predictors. Note that the parameters of $g(\cdot)$ and $f(\cdot)$ continually grow as the number of encountered concepts increases. After completing T tasks of incremental concept learning, the network is expected to predict all $K_{1:T}$ concepts, and leverage these concepts to make the final decision with enhanced interpretability.

2.2 Analytical Concept Predictor

The analytical concept predictor (ACP) models concepts as Gaussian variables $\mathbf{c} = \mathcal{N}(g(\mathbf{h}), \sigma^2)$ with mean $g(\mathbf{h})$ and variance σ^2, and formulates the optimization of $g(\cdot)$ as a least-squares (LS) problem. Therefore, it can achieve fast and exemplar-free ConceptIL through a recursive formulation. Specifically, in task t, the stacked matrices of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{h}$, and $\mathbf{c}$ are denoted as $X^{(t)} \in \mathbb{R}^{N_t \times d_x}$, $Y^{(t)} \in \mathbb{R}^{N_t \times d_y}$, $H^{(t)} \in \mathbb{R}^{N_t \times d_h}$, and $C^{(t)} \in \mathbb{R}^{N_t \times K_{1:t}}$, respectively, where N_t denotes the number of training samples in task t. Let $W = [W_1, W_2, \cdots, W_{K_{1:t}}] \in \mathbb{R}^{d_h \times K_{1:t}}$ be the parameters of the ACP: $\hat{\mathbf{c}} = g(\mathbf{h}) = \mathbf{h}^T W$, and $W_j \in \mathbb{R}^{d_h \times 1}$ is responsible for predicting the j-th concept. Additionally, ACP stores an inverse correlation matrix $P_j \in \mathbb{R}^{d_h \times d_h}$ for each W_j, which represents the memory of previous knowledge from up to task $t-1$.

Fig. 1(b) summarizes the update of $W_j^{(t)}$ and $P_j^{(t)}$ in ACP. Given $H^{(t)}$ and $C_j^{(t)}$, where $C_j^{(t)}$ represents the labels for the j-th concept, $W_j^{(t)}$ and $P_j^{(t)}$ can be iteratively updated from $W_j^{(t-1)}$ and $P_j^{(t-1)}$ as follows:

$$K_j^{(t)} = P_j^{(t-1)} (H^{(t)})^T (H^{(t)} P_j^{(t-1)} (H^{(t)})^T + I)^{-1}, \tag{1}$$

$$W_j^{(t)} = \begin{cases} W_j^{(t-1)} + K_j^{(t)} (C_j^{(t)} - H^{(t)} W_j^{(t-1)}), & j \in \mathcal{C}_t, \\ W_j^{(t-1)}, & j \notin \mathcal{C}_t. \end{cases} \tag{2}$$

$$P_j^{(t)} = \begin{cases} P_j^{(t-1)} - K_j^{(t)} H^{(t)} P_j^{(t-1)}, & j \in \mathcal{C}_t, \\ P_j^{(t-1)}, & j \notin \mathcal{C}_t. \end{cases} \tag{3}$$

We set $W_j^{(0)} = \mathbf{0}$, $P_j^{(0)} = \lambda I$, where λ is the regularization parameter.

Theorem 1. *The $W_j^{(t)}$ and $P_j^{(t)}$ obtained from Eqs.* (1) *to* (3) *satisfy that*

$$W_j^{(t)} = \arg\min_{W} \|\bar{C}_j^{1:t} - \bar{H}^{1:t} W_j\|_F^2 + \lambda \|W_j\|_F^2, \tag{4}$$

$$P_j^{(t)} = (\lambda I + \sum_{1 \le i \le t, j \in \mathcal{C}_i} {H^{(i)}}^T H^{(i)})^{-1}, \tag{5}$$

where $\bar{H}^{1:t}$ denotes the accumulated features of images that have available annotation for the j-th concept, and $\bar{C}^{1:t}$ represents the corresponding accumulated concept labels.

Proof. We only have to consider the case where $j \in \mathcal{C}_t$. For $P_j^{(t)}$, note that $P_j^{(t)} = \big((P_j^{(t-1)})^{-1} + (H^{(t)})^T H^{(t)}\big)^{-1}$. Then Eq. (5) can be proved using the Woodbury matrix identity:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}. \tag{6}$$

Let $A = (P_j^{(t-1)})^{-1}$, $B = (H^{(t)})^T$, $C = I$, $D = H^{(t)}$, we can directly obtain Eq. (3). For $W_j^{(t)}$, note that $\hat{W}_j^{(t)} = P_j^{(t)}(H^{1:t})^T\bar{C}_j^{1:t}$ satisfies Eq. (4), we only have to show that $W_j^{(t)} = \hat{W}_j^{(t)}, \forall 1 \le t \le T$, which can be proved using the similar approach in [22].

Theorem 1 indicates that the parameters of ACP after task t are equivalent to those obtained from joint training over on all concept annotations. Moreover, ACP stores the inverse correlation matrix instead of original samples from previous tasks, thereby preserving data privacy. Given $d_h = 1024$, ACP needs around 2MB memory for each concept.

2.3 Dynamic Weighting Strategy

In real-world clinical scenarios, some concepts have only limited number of positive samples, and such distribution imbalance poses a significant challenge in ConceptIL. To address this challenge, we further propose a dynamic weighting strategy by modifying Eq. (1) to the following equation:

$$K_j^{(t)} = P_j^{(t-1)}(H^{(t)})^T(H^{(t)}P_j^{(t-1)}(H^{(t)})^T + \Gamma^{-1})^{-1}, \tag{7}$$

where $\Gamma = \mathrm{diag}([\omega_1, \cdots, \omega_{N_t}]) \in \mathbb{R}^{N_t \times N_t}$ is the sample-wise weight matrix. This is equivalent to modeling $\mathbf{c}$ as Gaussian variables with different variances for each $\mathbf{c}_i$. In this case, the modified $W_j^{(t)}$ and $P_j^{(t)}$ correspond to the solution of the following weighted least-squares problem:

$$W_j^{(t)} = \arg\min_{W} \mathrm{tr}[(\bar{C}_j^{1:t} - \bar{H}^{1:t}W_j)^T\Gamma(\bar{C}_j^{1:t} - \bar{H}^{1:t}W_j)] + \lambda\|W_j\|_F^2, \tag{8}$$

$$P_j^{(t)} = (\lambda I + \sum_{1 \le i \le t, j \in \mathcal{C}_i} {H^{(i)}}^T \Gamma H^{(i)})^{-1}. \tag{9}$$

We set ω_i as the weighted average of inverse concept frequencies, thereby assigning higher weight to images containing concepts with less frequency.

3 Experiments

3.1 Experimental Setup

Datasets and incremental settings: We use the Fitzpatrick17k (`F17k`) [4] dataset for evaluation, which incorporates 48 expert-labeled concepts from the

Table 1. Performance comparison on the `F17k` dataset across two incremental settings, $T = 5$ and $T = 15$. Results from five different random seeds are reported as mean $\pm$ std. "Ours w/o weight" represents results of our methods without the dynamic weighting strategy. On average, models have observed 17.6 concepts in total when $T = 5$, and 29.4 concepts when $T = 15$. $\bar{A}$ denotes the average incremental performance, and BFC represents the performance gain from the framework of ConceptIL. ACC and F1 for concept prediction and diagnosis are denoted as C-ACC, C-F1, T-ACC, and T-F1, respectively. **Bold** text indicates the best results.

Setting	Method	$\bar{A}$ C-ACC(%)	$\bar{A}$ C-F1(%)	$\bar{A}$ T-ACC(%)	$\bar{A}$ T-F1(%)	BFC T-ACC(%)	BFC T-F1(%)
$T = 5$	Finetune	**92.62**$_{\pm 1.13}$	62.16$_{\pm 2.52}$	54.11$_{\pm 1.94}$	45.06$_{\pm 4.07}$	3.14$_{\pm 3.03}$	4.55$_{\pm 5.62}$
	SI [17]	92.61$_{\pm 1.14}$	62.12$_{\pm 2.55}$	53.52$_{\pm 1.86}$	44.52$_{\pm 3.86}$	0.68$_{\pm 2.37}$	1.33$_{\pm 5.76}$
	EWC [6]	92.57$_{\pm 1.17}$	62.03$_{\pm 2.58}$	53.30$_{\pm 0.85}$	44.17$_{\pm 1.98}$	1.14$_{\pm 2.07}$	4.06$_{\pm 5.87}$
	ER [11]	91.64$_{\pm 0.32}$	63.84$_{\pm 3.70}$	53.56$_{\pm 1.49}$	44.13$_{\pm 4.29}$	0.62$_{\pm 1.34}$	0.82$_{\pm 2.95}$
	DER [1]	90.96$_{\pm 1.47}$	61.11$_{\pm 3.37}$	51.52$_{\pm 1.24}$	38.12$_{\pm 3.38}$	-0.48$_{\pm 2.28}$	-2.46$_{\pm 5.52}$
	SimpleCIL [19]	90.17$_{\pm 1.33}$	58.16$_{\pm 1.69}$	58.22$_{\pm 3.00}$	53.23$_{\pm 4.97}$	2.68$_{\pm 3.29}$	4.49$_{\pm 5.79}$
	Ours w/o weight	92.61$_{\pm 1.06}$	64.04$_{\pm 2.06}$	59.89$_{\pm 2.45}$	55.77$_{\pm 3.41}$	8.29$_{\pm 3.47}$	12.99$_{\pm 4.81}$
	Ours	91.70$_{\pm 1.10}$	**66.24**$_{\pm 1.72}$	**60.53**$_{\pm 2.63}$	**56.82**$_{\pm 3.26}$	**9.41**$_{\pm 4.26}$	**13.03**$_{\pm 5.43}$
$T = 15$	Finetune	89.36$_{\pm 1.04}$	59.84$_{\pm 2.19}$	58.97$_{\pm 1.94}$	53.69$_{\pm 2.52}$	1.15$_{\pm 2.60}$	0.68$_{\pm 2.87}$
	SI [17]	**89.40**$_{\pm 1.05}$	59.24$_{\pm 2.28}$	57.93$_{\pm 1.12}$	52.29$_{\pm 1.63}$	0.76$_{\pm 1.50}$	0.15$_{\pm 2.21}$
	EWC [6]	89.32$_{\pm 1.10}$	58.84$_{\pm 2.34}$	56.32$_{\pm 1.12}$	49.99$_{\pm 1.99}$	0.74$_{\pm 1.59}$	1.42$_{\pm 2.04}$
	ER [11]	88.08$_{\pm 1.37}$	59.67$_{\pm 2.17}$	58.73$_{\pm 1.58}$	53.87$_{\pm 2.77}$	1.02$_{\pm 1.86}$	1.18$_{\pm 2.82}$
	DER [1]	87.93$_{\pm 1.05}$	56.30$_{\pm 2.25}$	51.80$_{\pm 0.88}$	39.90$_{\pm 2.21}$	-4.15$_{\pm 1.49}$	-8.41$_{\pm 3.89}$
	SimpleCIL [19]	87.14$_{\pm 1.45}$	54.54$_{\pm 2.47}$	61.69$_{\pm 1.19}$	58.94$_{\pm 1.44}$	1.52$_{\pm 1.42}$	3.29$_{\pm 3.24}$
	Ours w/o weight	89.38$_{\pm 1.12}$	61.83$_{\pm 2.41}$	64.55$_{\pm 1.52}$	62.03$_{\pm 2.12}$	5.74$_{\pm 1.72}$	**7.60**$_{\pm 1.58}$
	Ours	88.16$_{\pm 1.16}$	**62.95**$_{\pm 2.47}$	**64.59**$_{\pm 1.65}$	**62.34**$_{\pm 2.27}$	**5.83**$_{\pm 2.09}$	7.54$_{\pm 1.87}$

SkinCon dataset [2]. We focus on the binary classification task of distinguishing between *Benign* and *Malignant* images. For T tasks, we first randomly split the `F17k` dataset into training, validation, and test sets with a split ratio of 60%, 20%, and 20%, respectively. After that, we randomly divide the training and validation sets into T subsets, one for each task. Additionally, we randomly select 5 concepts and ensure that the chosen concepts are consistent across all methods. Note that only concepts with at least one positive sample in the training dataset can be selected, and concepts from different tasks may overlap. This setting is intended to better reflect real-world clinical scenarios, where diagnostic signs or concept annotations used across different datasets may overlap rather than be completely disjoint. In each task $t = 1, 2, \ldots, T$, we train and validate models on the current subset using the concepts annotated for that task, and evaluate the models on the entire test set using **all observed concepts**. We set $T = 5$ and $T = 15$ as two incremental settings for evaluation.

Implementation details: We used PyTorch on an NVIDIA GeForce RTX 3090 GPU for all experiments. We chose the pre-trained CLIP-RN50 [9] as the feature extractor $\phi(\cdot)$. To avoid concept leakage, in each task we first trained the concept predictor $g(\cdot)$ on the ground-truth concept annotations, and then trained the task predictor $f(\cdot)$ using the predicted concepts as input. In general, training of both $g(\cdot)$ and $f(\cdot)$ lasts 20 epochs in each task, with batch size set to 128. We random cropping, resizing, and horizontal flipping for data augmen-

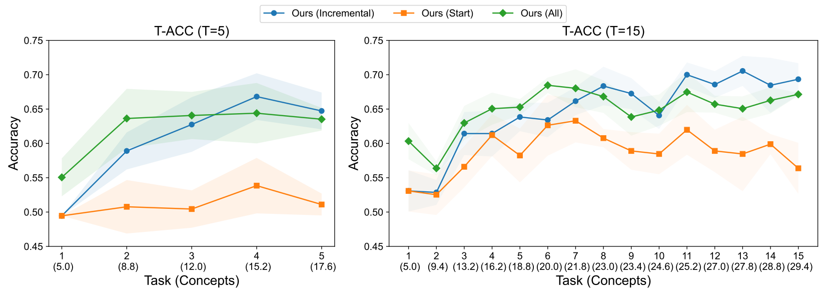

Fig. 2. The task-wise performance across five different random seeds. Here, "Ours (Incremental)" denotes models trained within the framework of ConceptIL. The number in parentheses below the x-axis indicates the total number of concepts observed by the "Ours (Incremental)" models after task t (referred to as $K_{1:t}$), averaged over five runs. For comparison, "Ours (Start)" represents the same models that only have access to only K_1 concepts for all T tasks. Similarly, "Ours (ALL)" denotes the same models trained with annotations of all $K_{1:T}$ concepts throughout training.

tation during training. We employed AdamW optimizer with a weight decay of 0.01. The learning rate was initialized to 0.02 and decayed following a cosine scheduler. The regularization hyperparameter was set λ to 0.1. Note that for our proposed analytical concept learning scheme, the training of $g(\cdot)$ only need one epoch. In inference, it's runtime remains as fast as the backbone. For evaluation, we used accuracy (ACC) and F1 score for both concept prediction and classification task, referred to as C-ACC, C-F1, T-ACC, and T-F1, respectively. Following the established protocol [10], we evaluated the average incremental performance $\bar{A}$: $\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A_t$, where A_t represents the test results after task t. Moreover, to investigate the effectiveness from the framework of concept incremental learning, we evaluated the benefit from concept incremental learning (BFC): $BFC = \frac{1}{T}\sum_{t=1}^{T}(A_t - A_t^0)$, where A_t^0 represents the performance of the same model trained only with the initial concept set for all tasks. We conducted five experiments with different random seeds.

3.2 Results

Comparison with IL methods: We chose 5 IL methods for comparison, including EWC [6], SI [17], ER [11], DER [1], and SimpleCIL [19]. EWC and SI represent the regularization-based IL methods, ER and DER are replay-based approaches that store original samples from previous tasks for rehearsal. SimpleCIL is the state-of-the-art IL method that employs the pretrained-model to extract image embeddings, and uses the averaged image embeddings as weights of the predictor. Besides, we reported performance of the backbone CBM [7] without any IL strategies, denoted as Finetune. As the results in Table 1 show, without dynamic weighting strategy, our proposed method achieved comparable C-ACC to Finetune, and obtained the best average incremental performance

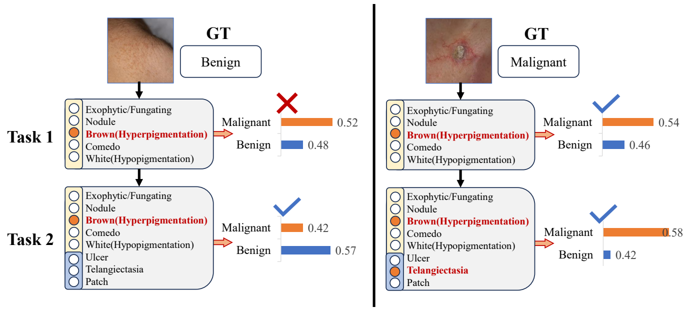

Fig. 3. Illustration of how ConceptIL improves diagnosis accuracy. GT denotes the ground-truth task label of images. Bold red highlighting indicates the presence of the corresponding concept in the image. The circle with a blue background represents a newly observed concept in task 2.

among other IL methods in terms of C-F1, T-ACC, and T-F1. Moreover, when the dynamic weighting strategy is applied, C-F1 further increased from 64.04 to 64.24 in $T = 5$, and 61.83 to 62.95 in $T = 15$, which demonstrates the effectiveness of this strategy.

Effectiveness of ConceptIL framework: The results of BFC in Table 1 demonstrate the effectiveness of ConceptIL. In the setting of $T = 5$, even models that directly finetune on the upcoming concept labels have seen increased performance in T-ACC and T-F1. In addition, one can see that the proposed method achieves the highest BFC in terms of T-ACC and T-F1, verifying its effectiveness. Furthermore, as Fig. 2 shows, in the early stage of ConceptIL (before the third task for $T = 5$ and the seventh task for $T = 15$, respectively), when the number of seen concepts grew rapidly, performance of ConceptIL increased fast. At the end of ConceptIL, our method achieved the performance comparable to that of the same models trained with all observed concepts.

ConceptIL for interpretable decision-making: To illustrate how the inclusion of new concepts enhances interpretability of decision-making, we report case studies in Fig. 3. In task 1 with limited concepts, one can see that the two images share the same concepts annotations despite having different task labels. Therefore, the model failed to make a correct prediction for the left image. In contrast, with the help of the newly observed concept `Telangiectasia` in task 2, the model is able to distinguish them. This observation aligns with the theoretical proof in [18].

4 Conclusion

In this paper, we introduce concept incremental learning (ConceptIL), a framework that learns new concepts from sequentially arriving data with partial concept annotations, and further utilizes them to enhance interpretable decision-making for medical AI. ConceptIL faces two challenges: limited access to the historical samples and concept imbalance. To address these, we propose an analytical concept learning scheme, which optimizes parameters through a recursive formulation, generating results equivalent to joint training without storing previous data. Experimental results demonstrate that our method matches the diagnostic performance of models trained on the full concept set, highlighting the effectiveness of ConceptIL. Furthermore, our approach outperforms competing incremental learning methods across two incremental settings.

Acknowledgments. This work was funded by the Science and Technology Commission of Shanghai Municipality (25TS1412100) and the National Natural Science Foundation of China (62372115).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark Experience for General Continual Learning: A Strong, Simple Baseline. In: *Advances in Neural Information Processing Systems* **33**, 15920–15930 (2020)
2. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R.A., Zou, J.: SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In: Neural Information Processing Systems (2022)
3. Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lió, P., Jamnik, M.: Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. In: Advances in Neural Information Processing Systems. vol. 35, pp. 21400–21413 (2022)
4. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021)
5. Jay, R., Davenport, C., Patel, R.: Clinical reasoning—the essentials for teaching medical students, trainees and non-medical healthcare professionals. British Journal of Hospital Medicine **85**(7), 1–8 (2024)
6. Kirkpatrick, J., et al.: Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences* **114**(13), 3521–3526 (2017)
7. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept Bottleneck Models. In: Proceedings of the 37th International Conference on Machine Learning. pp. 5338–5348 (2020)
8. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-Free Concept Bottleneck Models. In: International Conference on Learning Representations (2023)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)

10. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
11. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience Replay for Continual Learning. In: *Advances in Neural Information Processing Systems* **32** (2019)
12. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
13. Shang, C., Zhou, S., Zhang, H., Ni, X., Yang, Y., Wang, Y.: Incremental residual concept bottleneck models (2024)
14. U.S. Food and Drug Administration: Transparency for machine learning-enabled medical devices: Guiding principles (2024)
15. van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., Viergever, M.A.: Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis **79**, 102470 (2022)
16. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=nA5AZ8CEyow
17. Zenke, F., Poole, B., Ganguli, S.: Continual Learning Through Synaptic Intelligence. In: *International Conference on Machine Learning*. pp. 3987–3995 (2017)
18. Zhang, R., Du, X., Yan, J., Zhang, S.: The decoupling concept bottleneck model. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(2), 1250–1265 (2025)
19. Zhou, D.W., Cai, Z.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. International Journal of Computer Vision **133**, 1012–1032 (2025)
20. Zhuang, H., Chen, Y., Fang, D., He, R., Tong, K., Wei, H., Zeng, Z., Chen, C.: Gacl: Exemplar-free generalized analytic continual learning. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 83024–83047. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2641
21. Zhuang, H., Lin, Z., Toh, K.A.: Blockwise recursive moore–penrose inverse for network learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems **52**(5), 3237–3250 (2021)
22. Zhuang, H., Weng, Z., Wei, H., Xie, R., Toh, K.A., Lin, Z.: Acil: Analytic class-incremental learning with absolute memorization and privacy protection. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 11602–11614. Curran Associates, Inc. (2022)