Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai

Article

Federated Continual Learning for Privacy-Preserving, Reliable and Interpretable Multi-Center Corneal Diseases Diagnosis

Hongming Piao and Dapeng Oliver Wu*

Department of Computer Science, City University of Hong Kong, Hong Kong
* Correspondence: dpwu@ieee.org

How To Cite: Piao, H.; Wu, D. Federated Continual Learning for Privacy-Preserving, Reliable and Interpretable Multi-Center Corneal Diseases Diagnosis. *Transactions on Artificial Intelligence* **2026**, 2(1), 119–130. https://doi.org/10.53941/tai.2026.100008

Received: 2 December 2025
Revised: 30 December 2025
Accepted: 9 March 2026
Published: 1 April 2026

Abstract: Federated continual learning (FCL) is a distributed training framework that allows for learning from sequences of tasks on different centers under privacy-preserving. Although FCL has been extensively studied in fields such as image recognition and image segmentation, it remains unexplored in multi-center corneal diseases diagnosis, where data is inherently distributed and asynchronous while data privacy, reliability, and interpretability are urgently required. Therefore, this paper proposes Powderless for multi-center corneal diseases diagnosis, which can effectively transfer corneal diseases knowledge through prompt aggregation and prompt selection across various sequentially learned tasks from different centers under privacy. To further enhance diagnosis performance, ensure detection reliability, and improve interpretability, we design three key components: a multi-modal ensemble mechanism, an energy-based uncertainty estimation module, and a decision explanation module grounded in causal intervention. Comprehensive experimental results on the keratitis dataset demonstrate that our method achieves significant improvements compared to the base model, single-modality version, and local training in terms of both accuracy and the alignment between accuracy and uncertainty.

Keywords: federated continual learning; multi-center corneal diseases diagnosis; energy-based uncertainty estimation; interpretability

1. Introduction

Corneal diseases, particularly keratitis, are leading causes of preventable blindness globally, yet their diagnosis remains challenged by heterogeneous clinical presentations and fragmented multi-center data. While deep learning has shown promise [1], existing models rely on centralized training, which conflicts with privacy regulations (e.g., GDPR [2]) and fails to adapt to dynamic data distributions across time and clinical centers. Specifically, as shown in Figure 1, multi-center corneal diagnosis exacerbates three key challenges: (1) data is inherently non-iid across both time dimension and clinical center dimension, demanding effective knowledge transfer in both dimensions while avoiding catastrophic forgetting; (2) multi-modal corneal data need to complement each other; (3) clinical trust requires reliability and interpretability.

Federated Continual Learning (FCL) addresses the first challenge by merging federated learning for privacy-preserving decentralized training and continual learning for training under dynamic data distributions. Studies on federated continual learning is generally divided into two main paradigms: rehearsal-based and rehearsal-free methods. Rehearsal-based approaches utilize a memory buffer to store raw data samples or task-specific prototypes from prior tasks, which are then replayed alongside model decomposition (e.g., FedWEIT [3]), knowledge distillation (e.g., GLFC [4] and CFeD [5]), or regularization (e.g., Fedspace [6]) techniques. Nevertheless, memory buffers impose extra storage burdens and raise privacy concerns on centers. To mitigate this issue, rehearsal-free methods are proposed. TARGET [7] and FedCIL [8] generate pseudo samples by training additional generative models but research on model inversion attacks [9] shows that additional generative models pose privacy risks. With the development of vision foundation models, prompt-tuning is used to fully utilize

the generalization ability of vision foundation models in order to achieve rehearsal-free without pseudo samples. Fed-CPrompt [10] leverages a global prompt pool consisting of task-specific prompts following CODAPrompt [11], as well as a contrastive continual loss to address non-iid distribution among tasks. Powder [12] proposes a task-relevance-based prompt aggregation and prompt selection strategy to achieve knowledge transfer across time and center dimensions while balancing privacy and communication overhead.

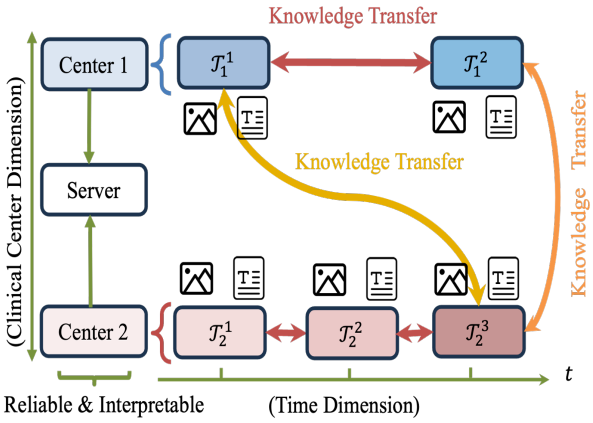


Figure 1. Challenges in multi-center corneal diagnosis.

However, due to the second and third challenges, existing FCL methods are far from usable in the field of multi-center corneal diseases diagnosis. Therefore, we propose Powderless, which adds three major modules to Powder. (1) It incorporates a multi-modal ensemble mechanism to fully utilize multi-modal corneal data; (2) It designs an energy-based uncertainty estimation module to enable physician intervention when uncertainty is high, thereby improving reliability; (3) It designs a decision explanation module grounded in causal intervention to provide reasons for decisions, thereby improving interpretability. Please refer to Table 1 for comparisons of Powderless and existing FCL methods.

Table 1. Comparisons of Powderless and existing FCL methods in terms of four desirable properties.

Method	Privacy-Preserving	Knowledge Transfer	Reliability	Interpretability
FedWEIT [3]	✗	✗	✗	✗
GLFC [4]	✗	✗	✗	✗
CFeD [5]	✗	✗	✗	✗
Fedspace [6]	✗	✗	✗	✗
TARGET [7]	✗	✗	✗	✗
Fed-CPrompt [10]	✓	✗	✗	✗
Powder [12]	✓	✓	✗	✗
Powderless	✓	✓	✓	✓

In conclusion, our contributions are threefold:

- The first work to apply FCL to multi-center corneal diseases diagnosis, explicitly explaining the importance of this scenario and addressing its three core challenges: privacy-preserving knowledge transfer, multi-modal complementarity, diagnostic reliability and interpretability.
- A targeted solution to each challenge via Powderless' modular design: (1) task-relevance-based prompt aggregation and prompt selection for privacy-preserving knowledge transfer; (2) ensemble mechanism to utilize multi-modal corneal data; (3) energy-based uncertainty estimation to enhance diagnostic reliability; (4) causal intervention to explain decision grounds for interpretability.
- Comprehensive evaluation on the Keratitis dataset shows that Powderless outperforms base models, single-modality variants and local training across key metrics including accuracy and uncertainty-accuracy alignment, providing solid evidence of its effectiveness for multi-center corneal diseases diagnosis.

2. Related Work

2.1. Federated Learning

Federated Learning (FL) constitutes a distributed model training paradigm anchored in privacy preservation. Within this framework, individual clients conduct local parameter updates using their private datasets, with only the trained model parameters transmitted to a central server for global model aggregation. This approach enables the development of high-performance global models while preventing the privacy risks associated with raw data sharing. FedAvg [13], a foundational algorithm in the FL domain, achieves global model aggregation by computing a weighted average of client-local models, where the weights are proportional to the size of each client's local dataset. Despite its widespread adoption, FL faces three critical challenges: mitigating catastrophic forgetting, facilitating effective knowledge transfer across clients, and reducing communication overhead under non-iid scenarios. To address the challenge of forgetting, existing approaches primarily incorporate regularization mechanisms either on model weights [14,15] or model outputs [16] to constrain the direction of parameter updates and preserve previously learned knowledge. For knowledge transfer, beyond the simplistic global aggregation strategy employed in FedAvg, personalized federated learning methods [17,18] are proposed to enhance knowledge sharing by designing client-specific aggregation strategies that prioritize relevant information exchange between clients. Regarding communication efficiency, research efforts focus on optimizing parameter transmission [19] and accelerating model convergence [20]. With the advancement of vision foundation models—such as the Vision Transformer (ViT) [21], recent studies demonstrate that parameter-efficient fine-tuning techniques [22,23] can effectively tackle the aforementioned FL challenges [24,25].

2.2. Continual Learning

Continual Learning is dedicated to equipping a model with the capability to acquire knowledge of new tasks sequentially while mitigating the forgetting of previously learned ones. Two core challenges in this field stand out: catastrophic forgetting and cross-temporal knowledge transfer. Regarding the first challenge, catastrophic forgetting, replay-based approaches [26–29] focus on optimizing the utilization of a memory buffer that preserves samples from prior tasks. They also develop sampling strategies to identify and store the most representative samples within this buffer. For the more demanding rehearsal-free setting, regularization-based methods [30,31] restrict or impose penalties on the update of parameters that are critical to previous tasks. In contrast, optimization-based methods directly regulate the optimization process of current tasks through techniques such as gradient projection [32,33], meta-learning [34,35], or the learning of robust representations [36,37]. For the second challenge, knowledge transfer, PR [38] leverages Bayesian principles to learn task-specific posteriors based on a shared meta-model. CUBER [39] classifies regularization into various transfer categories by analyzing the angles between task gradients and the sample space. D-TS [40] implements selective knowledge distillation for new tasks using a dynamically expanding teacher module. With the advancement of visual foundation models featuring robust representational capacities and the rise of Visual Prompt Tuning [22], several prompt-based approaches have demonstrated notable efficacy. L2P [41] proposes a key-query similarity mechanism to select prompts from a pre-constructed prompt pool, which are then used to adapt the visual foundation model to diverse tasks in the continual learning paradigm. Building on L2P [41], DualPrompt [42] splits the prompt pool into task-specific and task-invariant components. CODAPrompt [11] converts the prompt selection process into a differentiable procedure via an attention mechanism and appends task-specific prompt segments to the pool, thereby achieving state-of-the-art performance.

2.3. Multi-Center Medical Diagnosis

Multi-center medical diagnosis faces fundamental challenges in data acquisition, where labeling medical data requires expert knowledge. Collaboration between institutions could address this challenge, but sharing medical data to a centralized location faces various legal, privacy, technical, and data-ownership challenges, especially among international institutions. To address these concerns, federated learning emerges as a promising solution, enabling collaborative, decentralized training of models across multiple clients without transferring data between them. The federated learning framework provides a foundation for training large-scale multi-variate time series models [43] on critical care data [44] and leveraging fundus data from multiple institutions to improve diagnostic generalization at under-resourced hospital [45]. Advanced federated learning approaches are developed for specific medical applications, such as the federated stain normalization method for computational pathology [46], which proposes BottleGAN, a generative model that can computationally align the staining styles of many laboratories and can be trained in a privacy-preserving manner to foster federated learning in computational pathology. In this paper, we focus on corneal diseases.

2.4. Corneal Diseases Diagnosis

AI applications for specific corneal diseases achieve remarkable success. For keratoconus detection, [47] proposes a hybrid deep learning method using seven different corneal maps including anterior and posterior elevation, anterior and posterior curvature, anterior and posterior sagittal, and corneal thickness. [48] develops a Transformer-based CNN architecture that analyzes raw Scheimpflug corneal deformation sequences, achieving high sensitivity without the need for reconstructed topographic maps. For infectious keratitis diagnosis, a convolutional neural network using anterior segment photos is developed to differentiate between bacterial keratitis, fungal keratitis, non-infectious corneal lesions, and normal corneas [49]. Furthermore, [50] introduces a knowledge-enhanced multi-modal classifier that integrates slit-lamp images with clinical metadata to improve the differential diagnosis of bacterial and fungal keratitis. For diabetic neuropathy diagnosis, [51] employs a convolutional neural network with data augmentation for automated quantification of the corneal sub-basal nerve plexus. Additionally, [52] trains a ViT-based model for the binary classification of CCM images, demonstrating superior performance in detecting diabetic peripheral neuropathy compared to traditional CNN architectures.

3. Method

The proposed Powderless includes a prompt pool with aggregation and selection mechanism, Multi-modal Ensemble Mechanism (MEM), Energy-based Uncertainty Estimation (EUE) and Decision Explanation (DE). Please refer to Figure 2 for the overall framework.

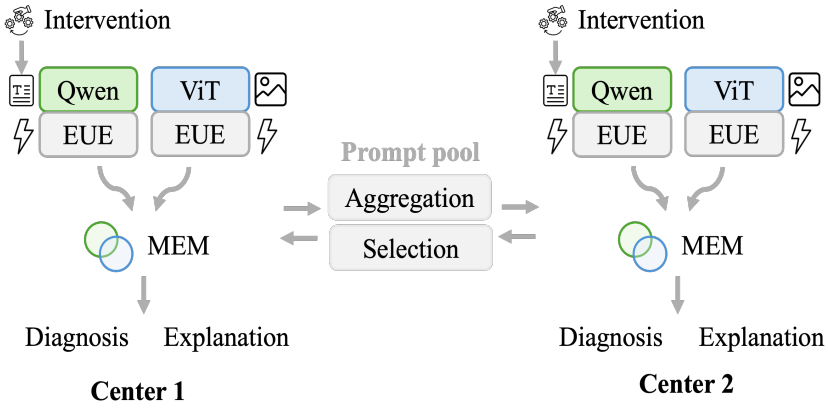


Figure 2. The overall framework of Powderless.

3.1. Preliminaries

Problem Statement. Suppose there are N corneal diseases diagnosis tasks $\mathcal{T} = \{\mathcal{T}^1, ..., \mathcal{T}^N\}$. In the scenario with C clinical centers sharing a single server, each center $c \in \{1, 2, ..., C\}$ can learn tasks sequentially at their own pace. We denote $\mathcal{T}_c^{t_c}$ as the current t_c-th task being learned by the c-th center. Each center sequentially learns N_c tasks. The model parameters of the c-th center during training the t_c-th task are denoted as $\boldsymbol{\theta}_c^{t_c}$. Besides, we use $|\cdot|$ to represent the size of sets, $[\cdot]_{i,j}$ to represent the ij-th entry of a matrix.

Prompt-tuning. Prompt-tuning operates on the multi-head self-attention layers (MHA) [53]:

$$\begin{aligned} \text{MHA}(\mathbf{h}_Q, \mathbf{h}_K, \mathbf{h}_V) &= \text{Concat}(\mathbf{h}_1, \ldots, \mathbf{h}_m)\mathbf{W}^O \\ \text{where} \quad \mathbf{h}_i &= \text{Attention}(\mathbf{h}_Q\mathbf{W}_i^Q, \mathbf{h}_K\mathbf{W}_i^K, \mathbf{h}_V\mathbf{W}_i^V), \end{aligned} \tag{1}$$

where m is the number of heads and $\mathbf{W}^O$, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are projection matrices. $\mathbf{h}_Q$, $\mathbf{h}_K$, $\mathbf{h}_V$ are the same in ViT [21]. Prompt-tuning concatenates trainable prompts to input tokens or input of MHA layers, which is equivalent to concatenate the same prompt parameter $\mathbf{p}$ to $\mathbf{h}_Q$, $\mathbf{h}_K$ and $\mathbf{h}_V$, namely $\text{MHA}([\mathbf{p}; \mathbf{h}_Q], [\mathbf{p}; \mathbf{h}_K], [\mathbf{p}; \mathbf{h}_V])$. During the training, $\mathbf{p}$ is optimized by gradient descent while other parameters are frozen.

CODAPrompt. For the generation of prompts, existing prompt-based FCL methods [10,54] mainly follow CODAPrompt [11], a state-of-the-art prompt-based CL method. During the training of $\mathcal{T}_c^{t_c}$, there is a set of task-specific prompts $\mathbf{P}_c^{t_c} \in \mathbb{R}^{M \times L \times D}$ for $\mathcal{T}_c^{t_c}$, where M is the length of the set, L is the length of a prompt, D is the output dimension of a ViT encoder. The prompts from existing tasks collectively form the global prompt pool $\mathbf{P}_g \in \mathbb{R}^{M_g \times L \times D}$, where $M_g = M \times N$. N is the number of existing tasks. For each sample x, its prompt $\mathbf{p} \in \mathbb{R}^{L \times D}$ is generated by a weighted sum of the prompts,

$$\mathbf{p} = \sum_{m}^{M_g} \alpha_m [\mathbf{P}_g]_m \,, \tag{2}$$

where the weights $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_{M_g}\}$ are achieved through query-key similarity:

$$\begin{aligned} \alpha = \{&\gamma(q(x) \odot [\mathbf{A}_g]_1, [\mathbf{K}_g]_1), \gamma(q(x) \odot [\mathbf{A}_g]_2, [\mathbf{K}_g]_2), \\ &\ldots, \gamma(q(x) \odot [\mathbf{A}_g]_{M_g}, [\mathbf{K}_g]_{M_g})\}, \end{aligned} \tag{3}$$

where $\mathbf{K}_g \in \mathbb{R}^{M_g \times D}$ and $\mathbf{A}_g \in \mathbb{R}^{M_g \times D}$ are trainable keys and trainable attention weights corresponding to each prompt in $\mathbf{P}_g$ respectively. $\odot$ is the Hadamard product. $\gamma(\cdot, \cdot)$ is the cosine similarity. Due to the one-to-one correspondence between $\mathbf{A}_g$, $\mathbf{K}_g$, $\mathbf{P}_g$, we use $\mathbf{P}_g \in \mathbb{R}^{M_g \times (2+L) \times D}$ to represent them together as the global prompt pool for simplicity.

Energy-Based Model (EBM). The core of EBM is an energy function $E_{\boldsymbol{\theta}_c^{t_c}} : \mathbb{R}^D \to \mathbb{R}$, which is parameterized by $\boldsymbol{\theta}_c^{t_c}$. The learning objective is letting $E_{\boldsymbol{\theta}_c^{t_c}}$ assign low energies to observed variables while giving high energies to unobserved ones. With $E_{\boldsymbol{\theta}_c^{t_c}}$, any probability density $p(x)$ for $x \in \mathbb{R}^D$ in an EBM can be written as

$$p_{\boldsymbol{\theta}_c^{t_c}}(x) = \frac{\exp(-E_{\boldsymbol{\theta}_c^{t_c}}(x))}{Z(\boldsymbol{\theta}_c^{t_c})}, \tag{4}$$

where $Z(\boldsymbol{\theta}_c^{t_c}) = \int_x \exp(-E_{\boldsymbol{\theta}_c^{t_c}}(x))$ denotes the normalizing constant, also known as the partition function. An EBM can be represented by using any function as long as the function can generate a single scalar given some input x. Following [55], we assume $E_{\boldsymbol{\theta}_c^{t_c}}$ is represented by a deep neural network.

3.2. Prompt Aggregation and Prompt Selection

It can be seen from the preliminaries that how to generate prompts is the core of prompt-tuning in FCL. In order to transfer more useful knowledge with fewer parameters and privacy protection, as shown in Figure 3, we utilize a two-step prompt aggregation and selection following [12].

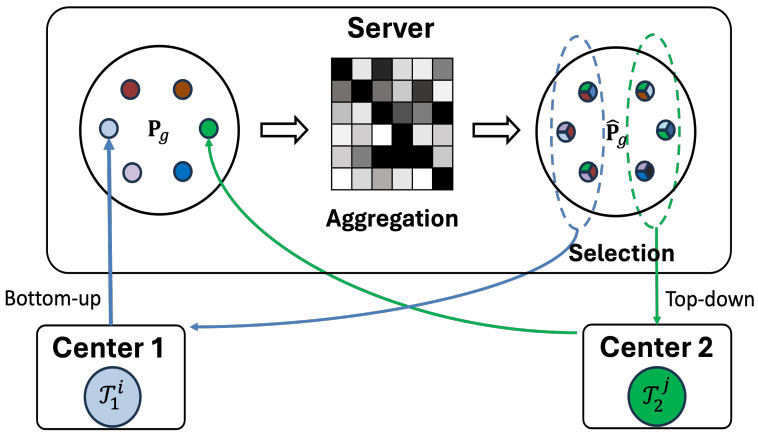


Figure 3. Prompt aggregation and prompt selection.

Prompt Aggregation. Initially, following [12], we estimate a dual task correlation matrix $\mathbf{G}_g^{\text{task}} \in \mathbb{R}^{M_g \times M_g}$ to represent the dual dimension correlations between existing tasks, where $[\mathbf{G}_g^{\text{task}}]_{i,j}$ represents the similarity between i-th task and j-th task. We update this matrix when new tasks emerge in the system. With $\mathbf{G}_g^{\text{task}}$, the prompts corresponding to each task in the global prompt pool $\mathbf{P}_g \in \mathbb{R}^{M_g \times (2+L) \times D}$ are aggregated with other prompts based on task correlation to transfer the most relevant knowledge, that is:

$$\begin{aligned} \hat{\mathbf{P}}_g &= \mathbf{G}_g^{\text{task}} \cdot \mathbf{P}_g, \\ \text{where } [\hat{\mathbf{P}}_g]_n &= \sum_{m}^{M_g} [\mathbf{G}_g^{\text{task}}]_{n,m} [\mathbf{P}_g]_m, \end{aligned} \tag{5}$$

Prompt Selection. Since transferring the entire prompt pool in each round would incur increasing communication and computational overhead, we select the top-k relevant tasks (including itself) for each task $\mathcal{T}_c^{t_c}$ based on the dual task correlation matrix $\mathbf{G}_g^{\text{task}}$, that is:

$$\mathbf{P}_{t_c}^{local} = \{\ [\hat{\mathbf{P}}_g]_s \mid s \in \text{top-}k([\mathbf{G}_g^{\text{task}}]_{t_c})\ \}, \tag{6}$$

where $[\mathbf{G}_g^{\text{task}}]_{t_c}$ is the correlation of $\mathcal{T}_c^{t_c}$ with other tasks. Then, the server transmits the aggregated prompts to the client where task $\mathcal{T}_c^{t_c}$ is located. The prompts for sample x are calculated by aggregation with query-key similarity introduced in Section 3.1, but with the local prompt pool $\mathbf{P}_c^{local} = \{\mathbf{P}_i^{local} | 1 \le i \le t_c\}$.

Although only a subset of the global prompt pool is transmitted between the center and server to mitigate the communication overhead, the existence of the first-step aggregation allows the center to transfer knowledge from the entire global prompt pool, which has also been filtered through task correlation. Additionally, as new tasks emerge, previous tasks continuously gain knowledge from the most relevant new tasks, thus achieving backward knowledge transfer. Moreover, due to the first-step aggregation, we avoid transmitting prompts containing task-specific knowledge unaltered to the center, and the transmitted prompts have a high relevance to the tasks on the center, which alleviates privacy concerns to some extent.

3.3. Multi-Modal Ensemble Mechanism (MEM)

For both the image and text modality classification models, the final two steps of the classification head are identical. First, $p = \text{softmax}(\mathbf{h})$, where $\mathbf{h} \in \mathbb{R}^F$ represents the model's logits, F is the number of classes, and $p \in \mathbb{R}^F$ is the probability distribution over the classes. Then, the class with the highest probability is selected as the final prediction. To fully leverage the multi-modal data and enable information complementarity between the modalities, we design the following ensembling mechanism: $p = \text{softmax}(\mathbf{h}_{\text{image}} + \mathbf{h}_{\text{text}})$, where $\mathbf{h}_{\text{image}} \in \mathbb{R}^F$ and $\mathbf{h}_{\text{text}} \in \mathbb{R}^F$ are the logits from the image model and the text model, respectively.

3.4. Energy-Based Uncertainty Estimation (EUE)

Given a classification task $\mathcal{T}_c^{t_c}$ with F categories, we use an open-world softmax [55] classifier with $F+1$ categories. The $F+1$-th score represents open-world uncertainty, where the classifier should be able to produce high uncertainty scores to abnormal input, which in turn can lower the confidence on the original F categories' prediction. Combined with MEM in Section 3.3 with $p \in \mathbb{R}^{F+1}$, $\mathbf{h}_{\text{image}} \in \mathbb{R}^{F+1}$, $\mathbf{h}_{\text{text}} \in \mathbb{R}^{F+1}$ and $p = \text{softmax}(\mathbf{h}_{\text{image}} + \mathbf{h}_{\text{text}})$, we obtain the final prediction as follows:

$$f_{\boldsymbol{\theta}_c^{t_c}}(x) = \begin{cases} \text{argmax}_{1 \le i \le F} p[i] & \text{if } p[F+1] < \beta \\ \text{physician intervention} & \text{else} \end{cases}, \tag{7}$$

where β is the uncertainty threshold that requires physician intervention. In order to achieve significant classification performance while allowing $p[F+1]$ to encode uncertainty, the open-world softmax classifier is trained with the following energy-based objective function:

$$\min_{\boldsymbol{\theta}_c^{t_c}} \mathbb{E}_{\bar{p}(x)} \left[-\log p[y]\right] + \lambda_1 \mathcal{L}_{\text{dual}} + \lambda_2 \mathbb{E}_{\bar{p}_{\boldsymbol{\theta}_c^{t_c}}(x)} \left[-\log p[F+1]\right], \tag{8}$$

where λ_1 and λ_2 are hyperparameters. The first term is the maximum log-likelihood objective for the classification task with F categories using the ground-truth label y. The second term is the dual distillation loss following [12] to reserve the transferred knowledge. The third term can also be seen as maximum log-likelihood objective—for input sampled from $\bar{p}_{\boldsymbol{\theta}_c^{t_c}}(x)$. Note that $\bar{p}_{\boldsymbol{\theta}_c^{t_c}}(x)$ denotes the abnormal input distribution conditioned on $\boldsymbol{\theta}_c^{t_c}$, which is approximated by Stochastic Gradient Langevin Dynamics (SGLD) [56]. Specifically, the SGLD sampling process follows

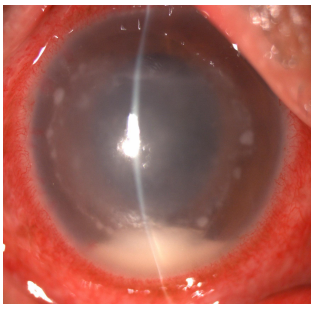
$$x_{t+1} = x_t - \frac{\alpha}{2} \frac{\partial E_{\boldsymbol{\theta}_c^{t_c}}(x_t)}{\partial x_t} + \sqrt{\alpha}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{9}$$

where $t \ge 1$ denotes the SGLD iteration and $x_1 = x$. α is the step size and ϵ is a random noise sampled from a normal distribution. In practice, we use hidden states rather than original inputs as x_t because they are smoother for the estimation of gradients [57].

3.5. Decision Explanation (DE)

To meet the interpretability requirements for multi-center corneal diseases diagnosis, we propose using causal intervention [58, 59] as an influence function to assess the impact of features in the text modality input on the

prediction results. Specifically, we sequentially remove features from the text modality input, as illustrated in Figure 4b, with replacement, and calculate the difference between the probability distribution of the prediction results after removing each feature $p' \in \mathbb{R}^F$ and the distribution before removal $p \in \mathbb{R}^F$. This difference $\Delta p = p - p'$ is regarded as the influence of that feature on the prediction. The influence of all features is regarded as the reasons for decision.



(a) Image


```
{
 "edema": {"Transparency": "Moderate"},
 "infil": {"Graininess": "Coarse", "Distribution Uniformity":
           "Uneven", "Density": "Medium Density", "Abnormal
           Color": "No Abnormality", "Edge Morphology":
           "Uniform Edges", "Surface Dryness/Wetness": "Wet",
           "Satellite Lesions": "No Satellite Lesions"},
 "epithelial defect": {"Terminal Bulging": null}
}
```


(b) Text

Figure 4. Image and text samples from the Keratitis dataset.

4. Experiment

4.1. Dataset

Our Keratitis dataset comprises 2834 samples from 621 patients, categorized into four classes: Amoeba, Bacterial, Fungal, and Hsk. Detailed statistical results are presented in Table 2. As shown in Figure 4, each sample includes a corneal image, textual descriptions, and its corresponding unique category. The textual descriptions consist of: the degree of edema reflected by transparency; the Infil attribute characterized by seven features including granularity, distribution uniformity, density, abnormal color, edge morphology, surface moisture, and satellite lesions; the presence or absence of an epithelial defect manifested as terminal enlargement.

Table 2. Keratitis dataset statistics.

Class Name	Number of Patients	Number of Samples
Amoeba	48	371
Bacterial	109	508
Fungal	153	824
Hsk	311	1131

4.2. Experimental Setup

At the task level, we divide the 4 categories into a total of 6 distinct binary classification tasks, and assign three continual learning tasks to each of the two clinical centers. To ensure no overlap of data between the training and test sets, as well as across different tasks, we first randomly sample 20% of the data from each category of the complete dataset as a public test set while the remaining 80% is used as the training set. For each task, 30% of the data is sampled from the corresponding category in the training set in a non-overlapping manner between tasks. At the backbone level, for the image modality, we employ ViT-B/16 as the backbone and attach a two-layer MLP with ReLU activation as the classification head. For the text modality, we use Qwen3-Embedding-0.6B as the backbone, and compute the cosine similarity between the input text embedding and each category embedding as the classification head. At the methodology level, for the image modality, we insert prompts at layers 3–5, with $M = 10$, $L = 8$, and $D = 768$. For the text modality, prompts are inserted between layers 1–18, with $M = 10$, $L = 8$, and $D = 1024$. At the training level, each task is trained for 10 epochs using the Adam optimizer with a learning rate of 0.005. Prompt aggregation and prompt selection are performed only at the beginning of each task. λ_1 is set to 1, λ_2 to 0.5, α to 1, and use x_t in the first layer. We compare our methods with vanilla fine-tuning and state-of-the-art FCL methods GLFC, Fed-CPrompt and Powder.

4.3. Evaluation Metrics

Final Accuracy (FA). This metric measures the final accuracy after the whole FCL process, computed as $\text{FA} = \frac{1}{|\mathcal{T}|}\sum_{\mathcal{T}_c^t \in \mathcal{T}} a_c^t$, where $\mathcal{T}$ denotes all tasks during the FCL process, a_c^t denotes the accuracy of task $\mathcal{T}_c^t$ after the whole FCL process. a_c^t is calculated by $\frac{1}{|\mathcal{T}_c^t|}\sum_{(x,y)\in\mathcal{T}_c^t} 1(f_{\theta_c^{N_c}}(x) = y)$.

Final Expected Calibration Error (FECE). This metric measures the final reliability after the whole FCL process, computed as $\text{FECE} = \frac{1}{|\mathcal{T}|}\sum_{\mathcal{T}_c^t \in \mathcal{T}} e_c^t$. Expected Calibration Error (ECE) e_c^t approximates the expectation of the difference between accuracy and confidence, which can reflect how well the prediction confidence aligns with the true accuracy. Specifically, the confidence estimates made on all test samples are partitioned into L equally spaced bins, and the difference between the average confidence and accuracy within each bin I_l is calculated by $\sum_{l=1}^{L} \frac{1}{|\mathcal{T}_c^t|} |\sum_{(x,y)\in I_l} p(f_{\theta_c^{N_c}}(x)|x) - \sum_{(x,y)\in I_l} 1(f_{\theta_c^{N_c}}(x) = y)|$.

4.4. Experimental Results

Final Accuracy. As shown in Table 3, Powderless achieves a 9.08% improvement in FA compared to vanilla fine-tuning, and a more than 1.89% improvement compared to existing FCL methods. This illustrates the importance of FCL in multi-center corneal diseases diagnosis while showing that EUE even brings a slight performance improvement compared with the vanilla powder.

Table 3. The FA of single-modal and multi-modal version of different methods.

Method	Text-Only			Image-Only			Ensemble		
	Client 1↑	Client 2↑	Average↑	Client 1↑	Client 2↑	Average↑	Client 1↑	Client 2↑	Average↑
Vanilla Fine-tuning	31.31	45.99	38.65	50.96	60.52	55.74	51.11	60.74	55.93
Vanilla Fine-tuning + EUE	31.31	45.99	38.65	53.62	62.04	57.83	53.47	62.26	57.87
GLFC	31.31	45.99	38.65	50.28	62.29	56.47	53.85	63.23	58.91
Fed-CPrompt	54.97	43.15	49.16	52.16	63.98	57.97	55.53	65.10	60.23
Powder	59.53	48.48	54.01	52.73	62.91	57.82	55.54	64.21	59.88
Powderless (Local)	55.24	42.95	49.10	52.58	63.77	58.18	54.51	64.64	59.58
Powderless	55.69	43.82	49.76	52.88	64.64	58.76	56.28	65.73	61.01

Knowledge Transfer. As shown in Table 3, Powderless improves FA by 2.40% compared to local training, with significant transfer effects across different clinical centers. This demonstrates the effectiveness of task-relevance-based prompt aggregation and prompt selection.

Multi-modal Complementarity. As shown in Table 3, Powderless achieves 9.45% improvement in FA compared to single-modal vanilla fine-tuning, and 3.83% improvement compared to single-modal Powderless training. This demonstrates the effectiveness of the MEM.

Reliability. As shown in Table 4, Powderless reduces the FECE by 5.92% compared to vanilla fine-tuning, and by 11.15% compared to using only Powder. This demonstrates the effectiveness of EUE.

Table 4. The FECE of single-modal and multi-modal version of different methods.

Method	Text-Only			Image-Only			Ensemble		
	Client 1↑	Client 2↑	Average↑	Client 1↑	Client 2↑	Average↑	Client 1↑	Client 2↑	Average↑
Vanilla Fine-tuning	38.74	43.59	41.17	28.29	20.28	24.29	32.24	26.89	29.57
Vanilla Fine-tuning + EUE	38.74	43.59	41.17	27.23	19.68	23.46	33.72	29.22	31.47
GLFC	38.74	43.59	41.17	20.32	20.23	20.30	33.03	23.37	28.31
Fed-CPrompt	45.40	37.19	41.28	20.48	20.20	20.26	33.18	23.39	28.24
Powder	47.13	36.01	41.57	20.25	18.96	19.61	36.72	25.89	31.31
Powderless (Local)	45.58	35.67	40.63	19.81	19.15	19.48	30.96	22.10	26.53
Powderless	44.73	36.53	40.63	20.10	19.86	19.98	32.66	22.98	27.82

Interpretability. As shown in Figure 5, Powderless illustrates the degree of influence of each text feature on the final decision. This will help to demonstrate whether the decision is reasonable and to identify the model weaknesses.

Scalability. We construct a setting with two centers, each containing six tasks, by equally splitting the training sets of the six distinct binary classification tasks mentioned in Section 4.2 into two parts. As shown in Table 5, as the number of tasks increases, the FA of Powderless and the baselines decreases, but Powderless consistently maintains its lead. Additionally, we evaluated the average single-task training time and inference time of Powderless and the baselines in Table 6. Powderless achieves better performance while keeping inference time at the same level

as the baselines. Although SGLD introduces additional training overhead, this is acceptable in medical scenarios where the number of centers and task frequency are typically low.

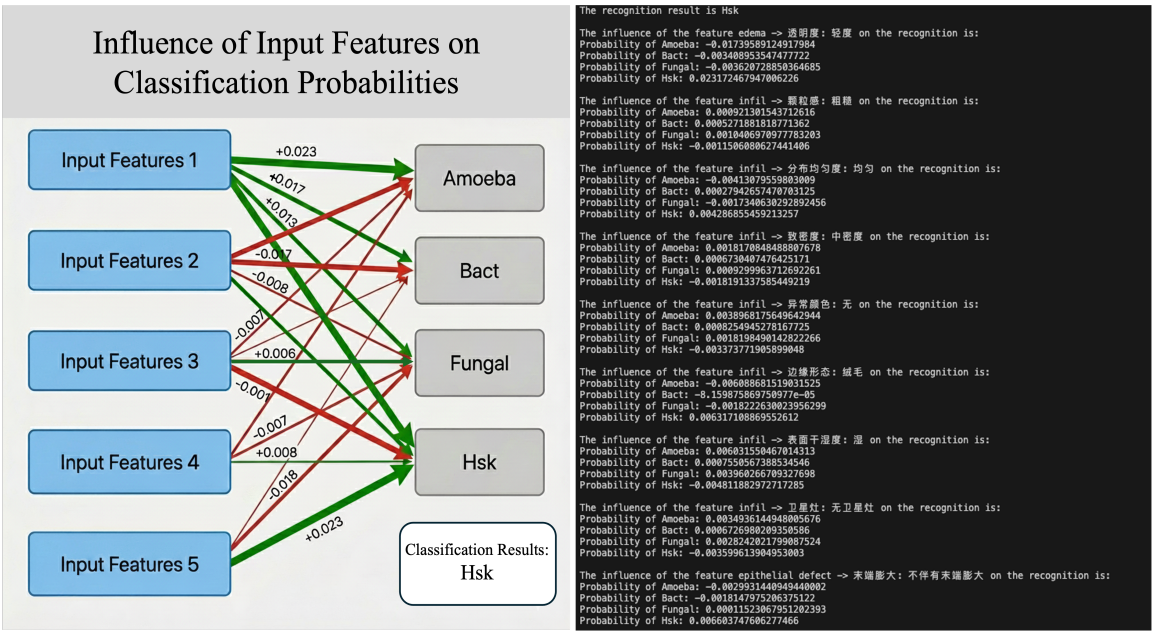


Figure 5. The influence of input features on the decision.

Table 5. The FA of multi-modal version of different methods with 3 and 6 tasks on each center.

Method	3-Task	6-Task
GLFC	58.91	53.47
Fed-CPrompt	60.23	53.85
Powder	59.88	55.53
Powderless	61.01	55.53

Table 6. The training and inference overhead of multi-modal version of different methods.

Method	Training Time (s)	Inference Time (s)
GLFC	100.21	19.41
Fed-CPrompt	105.02	20.74
Powder	104.83	20.68
Powderless	2560.26	20.92

Hyperparameter Sensitivity Analysis. As shown in Figure 6, we conduct a sensitivity analysis on the four hyperparameters λ_1, λ_2, α, and β. The results indicate that the impact of λ_1, λ_2 and α on FA does not exceed 10%. FA gradually increases as β increases but this also requires more physician intervention.

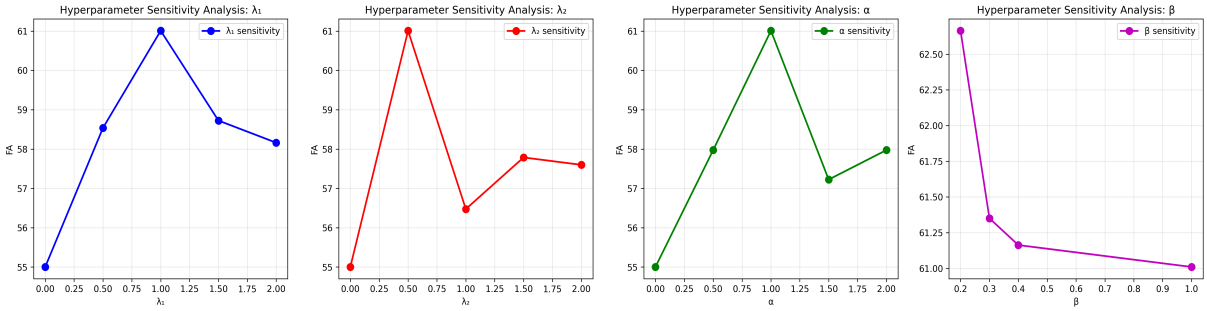


Figure 6. Hyperparameter sensitivity analysis.

5. Conclusions

This paper proposes Powderless, a federated continual learning method for multi-center corneal diseases diagnosis. It achieves knowledge transfer across time and clinical center dimensions under privacy-preserving conditions through task-relevance-based prompt aggregation and prompt selection. By employing an ensemble mechanism, it facilitates mutual enhancement among multi-modal data. Additionally, it enhances diagnostic reliability and interpretability via energy-based uncertainty estimation and causal intervention. For future work, we will continue to explore interpretability on the visual side.

Author Contributions

H.P.: Conceptualization, investigation, writing, and revision. D.W.: Conceptualization, investigation, writing, and revision. All authors have read and agreed to the published version of the manuscript.

Funding

This paper is partially supported by Hong Kong Innovation and Technology Commission (ITC) grant #MHP/034/22.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest. Given the role as Editor-in-Chief, Dapeng Oliver Wu had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author(s) used nano-banana to provide figure references. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

1. Li, Z.; Xie, H.; Wang, Z.; et al. Deep learning for multi-type infectious keratitis diagnosis: A nationwide, cross-sectional, multicenter study. *NPJ Digit. Med.* **2024**, *7*, 181.
2. Voigt, P.; Von dem Bussche, A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017.
3. Yoon, J.; Jeong, W.; Lee, G.; et al. Federated continual learning with weighted inter-client transfer. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 12073–12086.
4. Dong, J.; Wang, L.; Fang, Z.; et al. Federated class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10164–10173.
5. Ma, Y.; Xie, Z.; Wang, J.; et al. Continual federated learning based on knowledge distillation. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; Volume 3.
6. Shenaj, D.; Toldo, M.; Rigon, A.; et al. Asynchronous Federated Continual Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5054–5062.
7. Zhang, J.; Chen, C.; Zhuang, W.; et al. Addressing Catastrophic Forgetting in Federated Class-Continual Learning. *arXiv* **2023**, arXiv:2303.06937.
8. Qi, D.; Zhao, H.; Li, S. Better generative replay for continual federated learning. *arXiv* **2023**, arXiv:2302.13001.
9. Carlini, N.; Hayes, J.; Nasr, M.; et al. Extracting training data from diffusion models. In Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23), Anaheim, CA, USA, 9–11 August 2023; pp. 5253–5270.
10. Bagwe, G.; Yuan, X.; Pan, M.; et al. Fed-CPrompt: Contrastive Prompt for Rehearsal-Free Federated Continual Learning. *arXiv* **2023**, arXiv:2307.04869.

11. Smith, J.S.; Karlinsky, L.; Gutta, V.; et al. CODA-Prompt: COntinual Decomposed Attention-based Prompting for Rehearsal-Free Continual Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 11909–11919.
12. Piao, H.; Wu, Y.; Wu, D.; et al. Federated continual learning via prompt-based dual knowledge transfer. In Forty-first International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
13. McMahan, H.B.; Moore, E.; Ramage, D.; et al. Federated learning of deep networks using model averaging. *arXiv* **2016**, arXiv:1602.05629.
14. Li, T.; Sahu, A.K.; Zaheer, M.; et al. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
15. Shoham, N.; Avidor, T.; Keren, A.; et al. Overcoming forgetting in federated learning on non-iid data. *arXiv* **2019**, arXiv:1910.07796.
16. Lee, G.; Jeong, M.; Shin, Y.; et al. Preservation of the Global Knowledge by Not-True Distillation in Federated Learning. *arXiv* **2021**, arXiv:2106.03097.
17. Ma, X.; Zhang, J.; Guo, S.; et al. Layer-wised model aggregation for personalized federated learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10092–10101.
18. Chen, H.Y.; Zhong, J.; Zhang, M.; et al. Federated Learning of Shareable Bases for Personalization-Friendly Image Classification. *arXiv* **2023**, arXiv:2304.07882.
19. Chen, Y.; Sun, X.; Jin, Y. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 4229–4238.
20. Karimireddy, S.P.; Kale, S.; Mohri, M.; et al. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5132–5143.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Jia, M.; Tang, L.; Chen, B.C.; et al. Visual prompt tuning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 709–727.
23. Wu, Y.; Piao, H.; Huang, L.K.; et al. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv* **2025**, arXiv:2501.13198.
24. Feng, C.M.; Li, B.; Xu, X.; et al. Learning Federated Visual Prompt in Null Space for MRI Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8064–8073.
25. Yang, F.E.; Wang, C.Y.; Wang, Y.C.F. Efficient model personalization in federated learning via client-specific prompt generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19159–19168.
26. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; et al. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2001–2010.
27. Aljundi, R.; Lin, M.; Goujaud, B.; et al. Gradient based sample selection for online continual learning. In Proceedings of the 2019 Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
28. Wang, Q.; Wang, R.; Wu, Y.; et al. Cba: Improving online continual learning via continual bias adaptor. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 19082–19092.
29. Wu, Y.; Huang, L.K.; Wang, R.; et al. Meta Continual Learning Revisited: Implicitly Enhancing Online Hessian Approximation via Variance Reduction. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
30. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526.
31. Zenke, F.; Poole, B.; Ganguli, S. Continual learning through synaptic intelligence. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3987–3995.
32. Lopez-Paz, D.; Ranzato, M. Gradient episodic memory for continual learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
33. Chaudhry, A.; Khan, N.; Dokania, P.; et al. Continual learning in low-rank orthogonal subspaces. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 9900–9911.
34. Javed, K.; White, M. Meta-learning representations for continual learning. In Proceedings of the 2019 Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
35. Gupta, G.; Yadav, K.; Paull, L. Look-ahead meta learning for continual learning. In Proceedings of the NeurIPS 2020, Virtual, 6–12 December 2020; pp. 11588–11598.

36. Mirzadeh, S.I.; Farajtabar, M.; Pascanu, R.; et al. Understanding the role of training regimes in continual learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 7308–7320.
37. Mirzadeh, S.I.; Farajtabar, M.; Gorur, D.; et al. Linear mode connectivity in multitask and continual learning. *arXiv* **2020**, arXiv:2010.04495.
38. Henning, C.; Cervera, M.; D'Angelo, F.; et al. Posterior meta-replay for continual learning. In Proceedings of the NeurIPS 2021, Virtual, 6–14 December 2021; pp. 14135–14149.
39. Lin, S.; Yang, L.; Fan, D.; et al. Beyond not-forgetting: Continual learning with backward knowledge transfer. In Proceedings of the NeurIPS 2022, New Orleans, LA, USA, 28 November–9 December 2022; pp. 16165–16177.
40. Ye, F.; Bors, A.G. Dynamic self-supervised teacher-student network learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5731–5748.
41. Wang, Z.; Zhang, Z.; Lee, C.Y.; et al. Learning to prompt for continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 139–149.
42. Wang, Z.; Zhang, Z.; Ebrahimi, S.; et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 631–648.
43. Luo, Y.; Chen, N.; Huang, Z.; et al. Pyramid Transformer for Multivariate Time Series Anomaly Detection in IoUT. In Proceedings of the 2024 IEEE International Conference on Smart Internet of Things (SmartIoT), Shenzhen, China, 14–16 November 2024; pp. 533–539.
44. Burger, M.; Sergeev, F.; Londschien, M.; et al. Towards foundation models for critical care time series. *arXiv* **2024**, arXiv:2411.16346.
45. Raj, G.M.; Morley, M.G.; Eslami, M. Federated Learning for Diabetic Retinopathy Diagnosis: Enhancing Accuracy and Generalizability in Under-Resourced Regions. In Proceedings of the 2024 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 11–13 October 2024.
46. Wagner, N.; Fuchs, M.; Tolkach, Y.; et al. Federated stain normalization for computational pathology. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022; pp. 14–23.
47. Al-Timemy, A.H.; Mosa, Z.M.; Alyasseri, Z.; et al. A hybrid deep learning construct for detecting keratoconus from corneal maps. *Transl. Vis. Sci. Technol.* **2021**, *10*, 16–16.
48. Fassbind, B.; Langenbucher, A.; Streich, A. Automated cornea diagnosis using deep convolutional neural networks based on cornea topography maps. *Sci. Rep.* **2023**, *13*, 6566.
49. Satitpitakul, V.; Puangsricharern, A.; Yuktiratna, S.; et al. A Convolutional Neural Network Using Anterior Segment Photos for Infectious Keratitis Identification. *Clin. Ophthalmol.* **2025**, *19*, 73–81.
50. Maehara, H.; Ueno, Y.; Yamaguchi, T.; et al. Artificial intelligence support improves diagnosis accuracy in anterior segment eye diseases. *Sci. Rep.* **2025**, *15*, 5117.
51. Williams, B.M.; Borroni, D.; Liu, R.; et al. An artificial intelligence-based deep learning algorithm for the diagnosis of diabetic neuropathy using corneal confocal microscopy: A development and validation study. *Diabetologia* **2020**, *63*, 419–430.
52. Ben, R.C.; Petropoulos, I.; Malik, R.; et al. Vision transformers for automated detection of diabetic peripheral neuropathy in corneal confocal microscopy images. *Front. Imaging* **2025**, *4*, 1542128.
53. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In Proceedings of the 2017 Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
54. Halbe, S.; Smith, J.S.; Tian, J.; et al. HePCo: Data-Free Heterogeneous Prompt Consolidation for Continual Federated Learning. *arXiv* **2023**, arXiv:2306.09970.
55. Wang, Y.; Li, B.; Che, T.; et al. Energy-based open-world uncertainty modeling for confidence calibration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9302–9311.
56. Welling, M.; Teh, Y.W. Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Washington, DC, USA, 28 June–2 July 2011; pp. 681–688.
57. Bengio, Y.; Mesnil, G.; Dauphin, Y.; et al. Better mixing via deep representations. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 552–560.
58. Meng, K.; Bau, D.; Andonian, A.; et al. Locating and editing factual associations in gpt. In Proceedings of the NeurIPS 2022, New Orleans, LA, USA, 28 November–9 December 2022; pp. 17359–17372.
59. Wang, S.; Zhou, Q.; Wu, K.; et al. Interventional Root Cause Analysis of Failures in Multi-Sensor Fusion Perception Systems. In Proceedings of the Network and Distributed System Security (NDSS) Symposium 2025, San Diego, CA, USA, 24–28 February 2025.