Machine Unlearning for Pathological Image Classification Models

Min Jun Kim, Jae Hyun Cho, Hyun Jun Yook, Young Seon Kim, Su Yeon Kim, and Youn Kyu Lee*
Department of Computer Science and Engineering
Chung-Ang University
Seoul, Republic of Korea
{alswns8123, wogus2031, hyunjun6, thsu1084, suyeonkim, younkyul}@cau.ac.kr

***Abstract*—Deep learning models for pathological classification need to delete or reclassify existing classes according to changed clinical criteria. To address these issues, we propose a novel framework that guides the class to be deleted toward a non-pathological class while robustly maintaining the performance of the classification models.**

***Index Terms*—Machine Unlearning, Medical AI, Pathological Classification**

I. Introduction

Recent deep learning models have achieved high accuracy in anatomical segmentation and pathological classification by learning complex patterns in medical imaging data such as MRI and CT, showing applicability in clinical practice [1]. However, the 2017 revision of the diagnostic criteria for multiple sclerosis excluded spinal MRI from the essential requirements [2]. There is a need to modify classes that are no longer clinically valid [3]. In this case, classification models need to reflect new diagnostic criteria by deleting or reclassifying existing classes [4]. To address this issue, the classification models can be retrained from scratch by excluding invalid classes. However, securing large-scale datasets is challenging in the medical domain, which can lead to data scarcity issues when retraining [5]. Hence, training with small datasets may cause overfitting and degrade generalization performance [6].

To address these limitations, class-wise machine unlearning has been proposed to selectively remove invalid class from classification model [7], [8]. However, this method has a limitation, by optimizing the invalid class with a random class, it tends to misclassify invalid class as visually similar existing classes [9]. In particular, medical imaging data exhibit variability within the same class due to differences in scanning techniques or anatomical structures, while the images from different classes appear highly similar in visual features [3]. Therefore, removing a invalid class from trained classification models can impair the prediction performance on the existing classes, leading to overall performance degradation [9].

In this paper, we propose a novel unlearning framework that can remove the influence of invalid classes in pathological classification models. Our proposed framework defines an invalid class as a 'ghost class' and reduces misclassification by guiding them toward an existing non-pathological class. Specifically, our proposed framework identifies the parameter set that is affected by the ghost class and optimizes only the corresponding parameters by defining the non-pathological class as the 'target class.' This process effectively removes the influence of the ghost class while preserving the classification performance of the existing classes.

In this paper, our contributions are as follows:

1) Proposal of a novel unlearning framework that guides invalid class in pathological classification models to be classified as a pre-specified class.
2) Proposal of a robust framework that maintains existing pathological classification model performance after unlearning invalid class.

II. Related work

Machine unlearning refers to a technique that selectively removes data or a class from a trained model [7]. While retraining from scratch is a reliable approach to completely remove the influence, it requires a significant computational burden [9]. Several unlearning methods have been proposed to address this issue in classification models [8]. However, they are limited in completely removing the influence of a class on the model, as they rely on indirect approaches such as modifying a single layer or the model's final outputs [8].

III. Method

In this paper, we propose a novel unlearning framework to remove clinically invalid classes from pathological classification model. We define an invalid class as a 'ghost class' and remove it from the classification model, guiding its subsequent inputs toward a pre-specified 'target class.'

As shown in Fig. 1, the proposed framework is composed of the Identifying Salient stage and the Controlled Weight Update stage. The Identifying Salient stage identifies the parameter set that significantly affects the ghost class prediction. The Controlled Weight Update stage updates the classification model using the loss function L_{guide}, which guides predictions toward the target class. This update selectively targets only the identified parameter set. Our proposed framework enables the classification model to maintain high performance on the existing classes while guiding inputs that are visually similar to the ghost class toward the target class. The detailed description of the framework is as follows.

*Corresponding Author

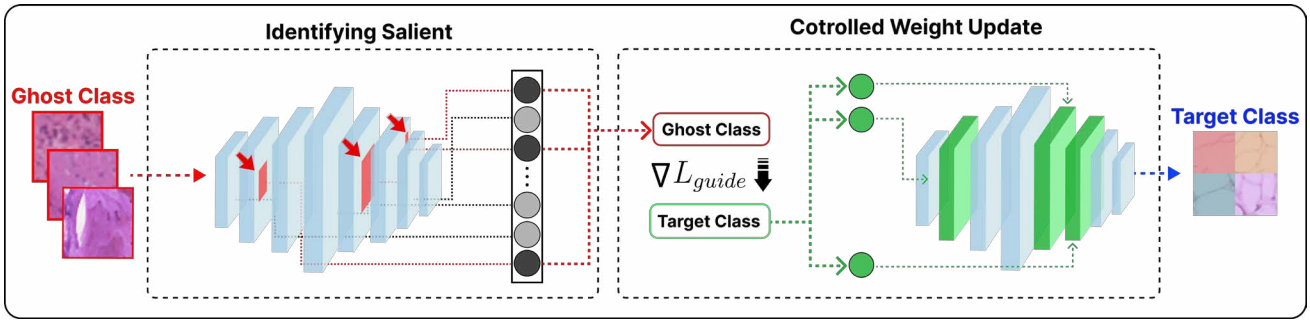

Fig. 1. Overview of unlearning framework for removing clinically invalid classes.

A. Identifying Salient

The Identifying Salient stage defines a Saliency Score, which quantifies the influence of each parameter on the prediction of the ghost class. These scores are then used to identify the significant parameters. To calculate the Saliency Score, the gradients of the classification model parameters are calculated for all data samples in the ghost class. The computed gradients serve as Saliency Scores, quantifying the influence of each parameter of the classification model on the ghost class prediction. A Binary Mask (BM) is constructed based on the median of the Saliency Score, indicating the relative importance of parameters. The BM has the same dimensions as each parameter matrix of the classification model and consist of binary values, where positions with Saliency Scores exceeding the threshold are set to 1 and all others are set to 0. This process enables the identification of parameters that are important for predicting the ghost class for updating.

B. Controlled Weight Update

The Controlled Weight Update stage updates the classification model while maintaining the performance of the classification models on the existing classes and guiding the predictions of ghost class data toward the target class. The guidance loss L_{guide} is calculated as follows.

L_{guide} is a loss function that trains the classification model to classify ghost class data as the target class. To compute this loss, we take all data (x_i) from the ghost class and re-assign their ground truth class (y_{ghost}) to the target class (y_{target}). These re-assigned class are then fed into the classfication model to compute the loss. The gradients of this loss are used as L_{guide}, which trains the classification model to predict ghost class data as the target class. As shown in Figure 1, the gradient of the calculated L_{guide} selectively updates only the masked parameters θ_1 instead of updating all parameters θ while the unmasked parameters θ_0 are not updated. The specific process is as follows.

$$\theta' = \theta - \eta \left(BM \odot \nabla_\theta L_{guide}(\theta_1 \text{ or } \theta_0) \right) \quad (1)$$

Eq. 1 represents the process where the entire classification model parameters θ are updated to new parameters θ'. Through the element-wise product $(\odot)$ operation between BM and the gradient of L_{guide}, only masked θ_1 are selectively updated, and the update strength is controlled through η (learning rate). Through this stage, the classification model can effectively remove the influence of the ghost class while preserving classification performance on the existing classes.

IV. Conclusion

In this paper, we propose a novel unlearning framework for removing ghost class from pathological classification models. The proposed framework removes the influence of ghost class while minimizing misclassification possibilities, and robustly maintaining the performance of the classification model. In our future work, we plan to evaluate the applicability across various medical fields.

Acknowledgment

This work was supported partly by Korea Foundation for Women In Science, Engineering and Technology (WISET) grant, funded by the Ministry of Science and ICT(MSIT) under the Team Research Program for female engineering students; and partly by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-16066331).

References

[1] V. K. Prasad, A. Verma, P. Bhattacharya, S. Shah, S. Chowdhury, M. Bhavsar, S. Aslam, and N. Ashraf, "Revolutionizing healthcare: a comparative insight into deep learning's role in medical imaging," *Scientific Reports*, vol. 14, no. 30273, 2024.

[2] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, and J. A. Cohen, "Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria," *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, 2018.

[3] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proceedings of the ACM conference on health, inference, and learning.*, pp. 151–159, ACM, 2020.

[4] D. Roy, P. Panda, and K. Roy, "Tree-CNN: A hierarchical deep convolutional neural network for incremental learning," *Neural networks*, vol. 121, pp. 148–160, 2020.

[5] S. Piffer, L. Ubaldi, S. Tangaro, A. Retico, and C. Talamonti, "Tackling the small data problem in medical image classification with artificial intelligence: a systematic review," *Progress in Biomedical Engineering*, vol. 6, no. 3, p. 032001, 2024.

[6] L. Brigato and L. Iocchi, "A close look at deep learning with small data," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2490–2497, IEEE, 2021.

[7] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation," *arXiv preprint arXiv:2310.12508.*, 2023.

[8] T. Baumhauer, P. Schöttle, and M. Zeppelzauer, "Machine unlearning: linear filtration for logit-based classifiers," *Machine Learning*, vol. 111, pp. 3203–3226, 2022.

[9] Y. Wang, A. Ebrahimpour-Boroojeny, and H. Sundaram, "On the Necessity of Output Distribution Reweighting for Effective Class Unlearning," *arXiv preprint arXiv:2506.20893*, 2025.