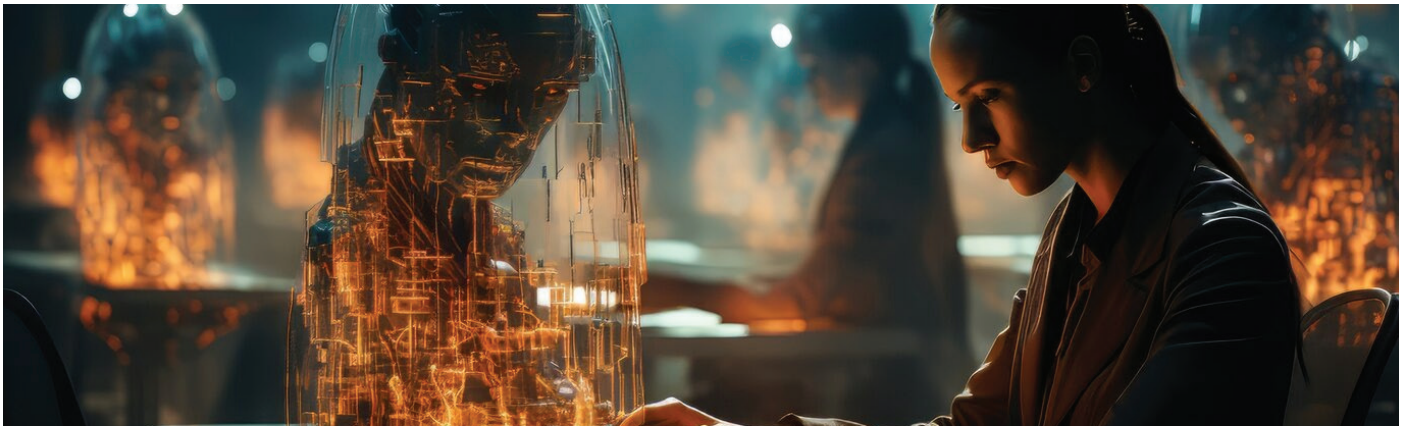

Combined Semi-Supervised Learning and Active Learning (SSL&AL) Framework for the Limited Labeled Data

Ms. T. Thanushika *(thanushika19991116@gmail.com) and* ***Dr. (Ms.) R. Nirthika*** *(nirthika@univ.jfn.ac.lk), Department of Computer Science, University of Jaffna.*

Ms. Thanushika Thedsanamoorthysarma is a final-year undergraduate student in the Department of Computer Science at the University of Jaffna, Sri Lanka. Her research interests include medical image analysis, deep learning, and machine learning.

Dr. Nirthika Rajendran is a Lecturer in the Department of Computer Science at the University of Jaffna, Sri Lanka. Her academic and research focus lies in the areas of medical image analysis, computer vision, deep learning, and machine learning.

Machine learning models in areas like Medical Diagnostics, Autonomous Driving, and NLP typically rely on large labeled datasets for high accuracy. However, gathering extensive labeled data is often costly and time-consuming, especially in specialized fields requiring expert annotation. This challenge has prompted the exploration of methods like semi-supervised learning (SSL) and active learning (AL) to improve model performance with minimal labeled data.

SSL shows promise by leveraging unlabeled data to boost performance where labeled data is scarce [1]. AL complements SSL by selectively querying the most informative samples for labeling, thus enhancing learning efficiency [2]. Combining SSL and AL into an SSL&AL framework enables high accuracy with minimal labeling. This article explores SSL&AL's design and applicability in fields requiring efficient data use, showcasing its ability to maintain robust performance with reduced labeled data.

Semi-Supervised Learning (SSL)

Semi-supervised learning (SSL) uses both labeled and unlabeled data to enhance model performance without extensive labeling. In medical imaging, pseudo-labeling is a common SSL technique, assigning labels to high-confidence unlabeled samples, thus reducing manual labeling demands [3].

In the SSL&AL framework, SSL applies pseudo-labeling after training a CNN on labeled data. As shown in Figure 1, data augmentation is used with both “weak” and “strong” transformations to increase model robu-

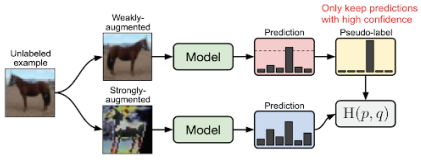

Figure 1: *Illustration of the SSL framework. (Image source: [4])*

stness. Only high-confidence pseudo-labeled samples are added to the training set, effectively expanding the labeled dataset and boosting classification accuracy.

Active Learning (AL)

Active learning (AL) enhances labeling efficiency by selecting the most informative samples for annotation, crucial in scenarios with limited labeled data [2]. Figure 2 demonstrates that AL prioritizes low-confidence samples, often near decision boundaries, to improve model accuracy.

In the SSL&AL framework, AL utilizes a cluster-based sampling approach. Using K-means++ clustering (Figure 2), samples are grouped, and low-confidence samples within each cluster are identified by probability scores. By prioritizing these boundary samples for labeling, the SSAL framework minimizes labeling needs while optimizing model performance.

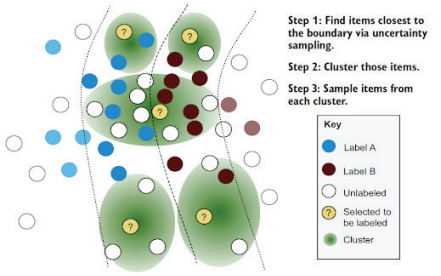

Figure 2: *Illustration of the AL framework. (Image source: [5])*

Semi-Supervised Learning and Active Learning (SSL&AL)

SSL and AL are combined in the SSL&AL framework to provide an effective, iterative categorization model. While AL chooses the most informative samples for labeling, maximizing resources, and increasing accuracy with little labeled data, SSL uses unlabelled data to improve learning [1].

The SSL&AL framework alternates between SSL and AL phases: starting with a CNN model trained on labeled data, the SSL phase adds high-confidence pseudo-labeled samples to the dataset, while the AL phase targets low-confidence samples for annotation via clustering and confidence sampling. This cycle incrementally improves classification accuracy by efficiently utilizing both labeled and unlabeled data.

Experiment on the BCCD Dataset

The SSL&AL framework was tested on the BCCD dataset [3], a standard in blood cell classification, chosen for its diverse cell images that support model generalization. Starting with limited labeled data, the framework expanded the dataset iteratively using SSL and AL, resulting in significant improvements in classification accuracy, even with minimal labeled samples. Each iteration refined decision boundaries, enhancing the model's ability to classify new data and underscoring SSL&AL's efficiency in achieving high accuracy with limited labeling.

Discussion and Conclusion

The SSL&AL framework's success on the BCCD dataset underscores its potential in medical diagnostics, where large labeled datasets are costly. Combining SSL and AL, the framework efficiently uses unlabeled data and selectively annotates informative samples, optimizing resources [1, 2]. Its iterative approach continuously improves model accuracy, making SSL&AL a scalable solution for complex classification tasks. Future enhancements to SSL and AL could extend SSL&AL's adaptability to diverse and complex medical imaging applications.

References

[1]. I. Nakamura, H. Ida, M. Yabuta, W. Kashiwa, M. Tsukamoto, S. Sato, S. Ota, N. Kobayashi, H. Masauzi and K. Okada, "Evaluation of two semi-supervised learning methods and their combination for automatic classification of bone marrow cells," Scientific reports, vol. 12, no. 1, p. 16736., 2022.

[2]. X. Li, X. Wang, X. Chen, Y. Lu, H. Fu and Y. C. Wu, "Unlabeled data selection for active learning in image classification.," Scientific Reports, vol. 14, no. 1, p. 424, 2024.

[3]. M. J. Mahmood, P. Raj, D. Agarwal, S. Kumari and P. Singh, "SPLAL: Similarity-based pseudo-labeling with alignment loss for semi-supervised medical image classification.," Biomedical Signal Processing and Control, vol. 89, p. 105665, 2024.

[4]. K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. Cubuk, A. Kurakin and C. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence.," Advances in neural information processing systems, vol. 33, pp. 596-608, 2020.

[5]. G. Min-soo, https://wikidocs.net/book/11816, 2024.