Journal of Advanced Health Informatics Research (JAHIR)
Vol. 2, No. 2, 2024, pp. 95-101
ISSN: 2985-6124, DOI: 10.59247/jahir.v2i2.294

A Comprehensive Review of Knowledge Distillation for Lightweight Medical Image Segmentation

Asmat Burhan[1], Purwono Purwono[2]
[1]*College of Nursing, International Pogram Ph.D Nursing, Taipei Medical University, Taiwan*
[2]*Department of Informatics, Universitas Harapan Bangsa, Purwokerto, Indonesia*

ARTICLE INFO

Article history:

Received June 03, 2024
Revised August 29, 2024
Published September 19, 2024

Keywords:

Knowledge Distillation;
Medical Image Segmentation;
Model Compression;
Lightweight Deep Learning;
Comprehensive Review;

ABSTRACT

Medical image segmentation plays a crucial role in computer-aided diagnosis by enabling precise identification of anatomical and pathological structures. While deep learning models have significantly improved segmentation accuracy, their high computational complexity limits deployment in resource-constrained environments, such as mobile healthcare and edge computing. Knowledge Distillation (KD) has emerged as an effective model compression technique, allowing a lightweight student model to inherit knowledge from a complex teacher model while maintaining high segmentation performance. This review systematically examines key KD techniques, including Response-Based, Feature-Based, and Relation-Based Distillation, and analyzes their advantages and limitations. Major challenges in KD, such as boundary preservation, domain generalization, and computational trade-offs, are explored in the context of lightweight model development. Additionally, emerging trends, including the integration of KD with Transformers, Federated Learning, and Self-Supervised Learning, are discussed to highlight future directions in efficient medical image segmentation. By providing a comprehensive analysis of KD for lightweight segmentation models, this review aims to guide the development of deep learning solutions that balance accuracy, efficiency, and real-world applicability in medical imaging.

Corresponding Author:

Asmat Burhan, Taipei Medical University, Taiwan
Email: d432113004@tmu.edu.tw

1. INTRODUCTION

Medical image segmentation has been recognized as a key element in various medical disciplines, particularly in improving the accuracy of diagnosis and the effectiveness of therapeutic interventions. With the utilization of artificial intelligence (AI), the process of extracting and identifying anatomical and pathological structures can be automated, resulting in more precise segmentation than conventional methods. The application of AI-based segmentation enables faster and more objective image analysis, so that subjective interpretations from medical personnel can be minimized [1]. Computer-aided diagnosis (CAD) systems have been developed to support AI-based clinical decision-making, which has the potential to improve healthcare efficiency and disease detection accuracy [2].

Although significant progress has been made in the development of deep learning models for medical image segmentation, limitations in computational efficiency are still a major challenge. Conventional deep learning models tend to have complex architectures with a very large number of parameters, requiring high

computational power in both the training and inference stages [3]. The need for expensive computing infrastructure limits the implementation of these models, especially in clinical environments with limited resources. Delays in data processing due to the complexity of the model often hamper applications in real-time scenarios that require fast and accurate responses.

Another challenge is found in the application of deep learning-based segmentation models in mobile healthcare. Models that have large size and high power consumption cannot be efficiently implemented on low-power devices such as smart phones or edge devices [4]. This limits the adoption of AI in distributed computing-based health systems, which could potentially expand access to AI-based diagnosis to regions with limited infrastructure. Optimization strategies that can reduce model complexity without compromising segmentation accuracy are needed.

One solution that has been developed to overcome this limitation is Knowledge Distillation (KD), a technique that enables knowledge transfer from complex teacher models to lighter student models [5]. With this approach, smaller models can be trained to replicate the performance of larger models through a knowledge distillation mechanism that includes soft labels and feature representations. This technique has proven to be effective in reducing the number of model parameters, thereby reducing computational power requirements without significantly degrading segmentation accuracy. Several variants of KD, such as logit-based distillation, feature-based distillation, and attention-based distillation, have been developed to further improve model compression efficiency in medical image segmentation tasks.

This paper presents a comprehensive analysis of KD in medical image segmentation, focusing on its effectiveness in lightweight model development. Various distillation approaches are systematically reviewed, including their advantages, limitations, and current research trends. Challenges in the trade-off between accuracy and computational efficiency are also discussed. This review is expected to provide insights into the development direction of deep learning-based lightweight models as well as future research opportunities to improve the efficiency and reliability of AI in healthcare.

2. FUNDAMENTALS OF KNOWLEDGE DISTILLATION IN MEDICAL IMAGE SEGMENTATION

Knowledge Distillation is a model optimization technique that aims to transfer knowledge from large and complex teacher models to lighter student models, while maintaining optimal accuracy [5]. In this approach, the student model is trained using the output of the teacher model, which typically includes soft labels and probability distributions that are more informative than the hard labels from the annotated data [6]. KD has been widely used as a model compression method in various deep learning tasks, including medical image segmentation. When compared to other techniques such as pruning and quantization, KD has the advantage of retaining richer information from the original model, so that the student model is still able to produce accurate predictions even though its complexity has been reduced [7]. While pruning removes connections and neurons that are considered unimportant in the network, and quantization reduces the precision of the numerical representation of the model, KD works in a more adaptive way by utilizing the knowledge representation of the teacher model, which allows the performance of the student model to remain high even if its size has been significantly compressed [8].

Figure 1 [9], shows the KD architecture for brain tumor segmentation, where the teacher model (teacher encoder) transfers knowledge to the lighter student model (student encoder). The teacher model generates feature maps and predictions that are used to train the student model through a loss mechanism.

In the context of medical image segmentation, KD contributes to improving model efficiency by reducing the number of Floating Point Operations (FLOPs), which directly lowers computational power requirements. This reduction in FLOPs allows the model to execute faster without significantly compromising the segmentation accuracy [10]. The size of the model can be reduced substantially, making it easier to implement on resource-constrained devices, such as edge computing and mobile healthcare systems [4]. The main advantage of KD in medical image segmentation lies in its ability to generate more efficient models without losing crucial information needed in the inference process. The application of KD can accelerate the process of automated diagnosis as well as enable the integration of deep learning-based models in portable medical systems, ultimately contributing to the increased accessibility of AI-based healthcare services.

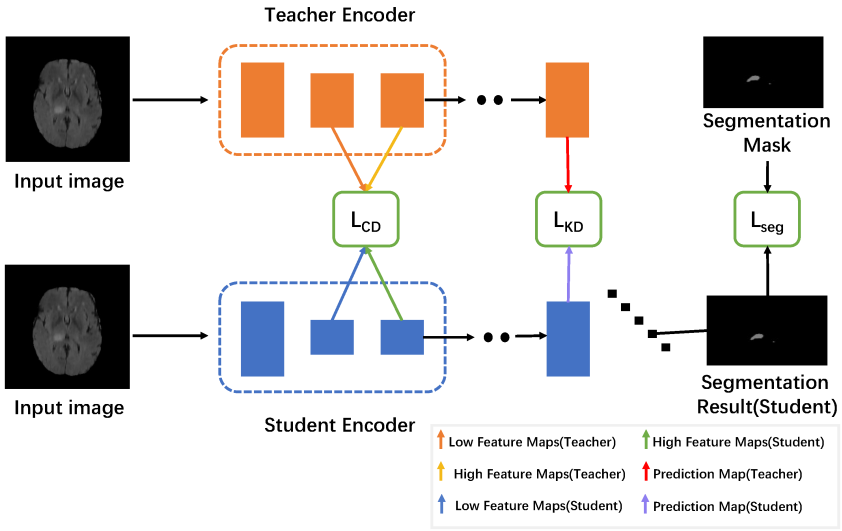

Fig. 1. Brain Tumor Segmentation Using Knowledge Distillation

3. KNOWLEDGE DISTILLATION TECHNIQUES FOR LIGHTWEIGHT MEDICAL IMAGE SEGMENTATION

3.1. Response-Based Distillation

In Knowledge Distillation, the main process is to transfer logits or probabilistic outputs from the teacher model to the student model. These logits contain finer information compared to hard labels, allowing the student model to learn complex patterns more efficiently despite having a lighter architecture. This approach allows the student model to not only mimic the final output of the teacher model, but also understand richer inter-class probability distributions, so that the generalizability of the model is maintained despite the reduced complexity [11].

In spinal image segmentation, the student model based on Mobile U-Net can gain insights from the teacher model equipped with a channel attention mechanism. Through the KD process, the student model not only receives manual annotation-based learning, but also gains additional information from the feature representation that has been learned by the teacher model. The student model is able to achieve efficient segmentation with lower complexity, enabling implementation on resource-constrained systems, such as edge computing devices and mobile healthcare applications [11].

3.2 Feature-Based Distillation

In addition to transferring logits, KD can also be done by transferring feature representations from the teacher model to the student model. In this approach, information obtained from the intermediate layers of the teacher model is utilized to improve the student model's understanding of the spatial structure and contextual relationships in the data [12]. By transferring features at different levels of abstraction, the student model can absorb more complex information despite having a simpler architecture, resulting in more accurate segmentation.

The SPARK method improves the quality of the intermediate feature representation by considering the spatial position, allowing the student model to obtain more precise segmentation insights. This approach allows the student model to better understand the segmentation pattern through the spatial distribution obtained from the teacher model [13]. The Inter-Subspace Knowledge Distillation (ISKD) framework, which optimizes feature transfer by utilizing feature subspaces and inter-class similarity. With this technique, structural and contextual information contained in the teacher model can be more effectively absorbed by the student model, thus improving segmentation accuracy without significantly increasing model complexity [14].

3.3 Relation-Based Distillation

In KD, in addition to transferring logits and feature representations, relational knowledge between different parts of the input data can also be learned by the student model from the teacher model. This knowledge transfer focuses on understanding the relationships between different features in the medical image, so that the student model not only mimics the output of the teacher model, but also understands deeper structural and contextual

interrelationships. This approach is particularly important in multi-class segmentation, where the relationships between objects in the image need to be considered to produce more accurate predictions [15].

The Dual-Stage Progressive Knowledge Distillation (DSP-KD) framework applies a two-stage stepwise distillation approach to maximize mutual information between teacher and student model representations [11]. With this strategy, the student model is able to absorb more information about the relationships between classes in the image, thus improving performance in multi-class skin disease segmentation tasks. This approach demonstrates that by considering the relationships between features in the data, KD can not only reduce model complexity but also maintain better contextual understanding, enabling more accurate segmentation in complex medical classification tasks.

Table 1. Comparison of Knowledge Distillation Techniques

KD Technique	Transfer Method	Excellence	Weaknesses
Response-Based KD	Output logit transfer	Mudah diterapkan	Does not capture intermediate features
Feature-Based KD	Intermediate feature transfer	Richer representation	Requires teacher-student mapping
Relation-Based KD	Transfer relationship between features	Understand spatial relations	More complex computation

Table 1 presents a comparison of the three main techniques in KD applied in medical image segmentation. Response-based techniques transfer probabilistic output from the teacher model to the student model, feature-based techniques allow the student model to capture a richer representation of the teacher network, while relation-based techniques focus on understanding the relationships between features in the image.

4. CHALLENGES IN KNOWLEDGE DISTILLATION FOR LIGHTWEIGHT SEGMENTATION

4.1 Boundary Awareness

One of the main challenges in KD is the limitation in effectively transferring boundary information. Lighter student models often have difficulty in detecting object boundaries with high precision, which can lead to less accurate segmentation, especially in complex and thin structures. This happens because most KD methods focus on transferring global feature representations or class probability distributions, without considering finer boundary features [16].

As a result, the resulting models often suffer from degradation in boundary discrimination, which affects the segmentation performance, especially in cases where the differences between classes depend on very small spatial details. KD techniques that retain boundary information more effectively are needed, such as by adding an attention-based distillation mechanism or utilizing explicit strategies to sharpen boundary features in student models. With a more boundary-aware approach, lighter segmentation models can achieve higher accuracy without losing critical structural details in medical images [15].

4.2 Domain Generalization

One of the main challenges in developing lightweight models for medical image segmentation is the limitation in cross-domain generalization, which is the ability of the model to maintain good performance when applied to medical datasets that differ from the training dataset. Conventional deep learning models tend to have a greater capacity to capture variations between datasets, but in models that have been compressed using KD, this ability is often degraded. This is due to the loss of domain-specific information during the distillation process, making the student model more sensitive to different data distributions [17].

When implemented in diverse medical environments, such as different types of imaging (CT, MRI, or dermoscopy) or variations in image acquisition techniques, lightweight models often suffer from significant accuracy degradation. To overcome this, several approaches have been developed, such as KD-based domain adaptation, where the student model learns not only from the teacher model but also from a broader domain representation. In addition, augmentation strategies based on adversarial learning or multi-source distillation methods can be used to improve the generalization ability of models on heterogeneous medical datasets. By improving the capacity of lightweight models in the face of inter-domain variations, the adoption of KD-based segmentation models in clinical environments can be expanded more effectively.

4.3 Computational Trade-offs

One of the main challenges in applying KD for medical image segmentation is balancing between the reduction of computational load and the maintenance of high segmentation accuracy. Although KD has proven

to be effective in generating lighter student models with fewer parameters, the trade-off between computational efficiency and segmentation quality remains a complex issue. Model size reduction often results in the loss of important information in the feature representation, which can lead to a decrease in accuracy, especially in segmentation tasks with complex spatial details.

The resulting lightweight models must still be able to run optimally on resource-constrained devices, such as edge devices and mobile healthcare systems, without compromising inference speed. To address this, various approaches have been developed, including the use of efficient architectures such as depthwise separable convolutions, attention-based distillation strategies, and quantization optimization to speed up inference. With the combination of these techniques, KD models can be designed to be more balanced in terms of computational efficiency and segmentation accuracy, enabling wider applicability in clinical environments where fast, real-time processing is required [18] [19].

5. CASE STUDIES AND APPLICATIONS

5.1 Diabetic Foot Ulcer Segmentation

Although specific studies on KD for segmentation of Diabetic Foot Ulcers (DFU) are limited, KD principles can be applied to improve the accuracy of lightweight models. DFU segmentation is challenging due to variations in wound size, shape and texture, as well as indistinct boundaries, which often degrade the performance of compressed models.

To overcome this, KD-based feature distillation and boundary-aware distillation approaches can be applied to preserve the details of the wound texture and boundaries [15]. The attention-based distillation method can be used to strengthen the representation of features that are more relevant in distinguishing the wound area from healthy tissue [20]. With this strategy, an efficient DFU segmentation model can be developed, enabling implementation on mobile healthcare and edge computing devices to support real-time wound diagnosis and monitoring.

5.2 Tumor Segmentation

KD has been shown to be effective in tumor segmentation, especially in the case of glioma, where lightweight models such as compressed U-Net can achieve accuracy comparable to larger models. By applying KD, the student model can learn from the feature representation generated by the teacher model, enabling a reduction in model complexity without significant performance loss [21].

In addition to reducing storage size, KD also substantially speeds up inference time, making it an ideal solution for implementation in edge computing-based systems or clinical devices with computational limitations. Thus, KD enables the development of more efficient tumor segmentation models, supporting the application of AI in imaging-based diagnosis with high speed and accuracy.

5.3 Segmentasi Vertebrata

A two-stage segmentation approach has been applied to segment individual vertebrae in CT images, integrating the KD method to improve the performance of lightweight models. This strategy allows the student model to gradually absorb complex structural information from the teacher model, resulting in more precise segmentation despite using a more efficient architecture [15].

Through this approach, the lightweight model is able to achieve a significant improvement in segmentation accuracy, while reducing the computational load required during inference. These results show that KD can not only reduce model complexity but also still maintain high segmentation quality, making it an effective solution for the application of vertebra segmentation in resource-constrained clinical environments.

6. EMERGING TRENDS IN KNOWLEDGE DISTILLATION FOR MEDICAL IMAGE SEGMENTATION

6.1 Knowledge Distillation and Transformers

Vision Transformers (ViTs) have started to be integrated with Knowledge Distillation (KD) to leverage their powerful modeling capacity while overcoming high computational challenges. Although ViTs excel in capturing spatial and global relationships in images, the large computational complexity is a major obstacle in their application, especially for resource-constrained medical segmentation [22].

Hybrid approaches that combine Convolutional Neural Networks (CNNs) and Transformers have shown potential in improving segmentation performance. In this scenario, KD is used to transfer knowledge from a Transformer-based teacher model to a lighter student model, such as a CNN-based model or a hybrid lightweight Transformers architecture. With this method, the spatial representation advantage of Transformers

can be retained in a more efficient model, enabling high-quality segmentation with lower computational requirements.

6.2 Knowledge Distillation in Federated Learning

Federated Learning (FL) has been developed as a solution to address data privacy concerns in training artificial intelligence models by leveraging data from multiple decentralized sources without the need to share raw data. In the context of medical image segmentation, FL allows models to be trained collaboratively across different medical institutions without violating privacy regulations, such as HIPAA or GDPR [23] [24].

The integration of KD with FL has the potential to improve the performance of lightweight models while maintaining the security and efficiency of information transfer between nodes. The FedLPPA framework has introduced strategies based on personalized prompts and dual-decoders, which enable model adaptation to heterogeneous data from different institutions [25]. With this approach, KD can be used to transfer richer knowledge from global models to lighter local models, increasing the generalizability of models without the need to share data between medical institutions.

CONCLUSION

KD offers a promising approach in the development of lightweight models for medical image segmentation, by addressing computational and efficiency challenges without compromising accuracy. By utilizing KD, lighter models can be trained to maintain comparable performance to more complex models, thus enabling implementation in clinical environments with limited resources.

The integration of KD with current technologies, such as Vision Transformers and Federated Learning, further expands its scope of application in the medical world. This combination not only improves computational efficiency, but also strengthens the privacy aspect and adaptability of the model in various clinical scenarios. With the continuous development of KD methods and innovations in model architecture, this approach is expected to become a key solution in optimizing medical image segmentation for real-world applications.

REFERENCES

[1] A. M. Breesam, S. R. Adnan, and S. M. Ali, "Segmentation and Classification of Medical Images Using Artificial Intelligence: A Review," *Al-Furat Journal of Innovations in Electronics and Computer Engineering*, vol. 3, no. 2, pp. 299–320, Jul. 2024, doi: 10.46649/fjiece.v3.2.20a.29.5.2024.

[2] S. Mishra, H. K. Tripathy, and B. Acharya, "A Precise Analysis of Deep Learning for Medical Image Processing," 2021, pp. 25–41. doi: 10.1007/978-981-15-5495-7_2.

[3] Md. B. Hossain, N. Gong, and M. Shaban, "Computational Complexity Reduction Techniques for Deep Neural Networks: A Survey," in *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, IEEE, Sep. 2023, pp. 1–6. doi: 10.1109/AIBThings58340.2023.10292477.

[4] J. Wong and Q. Zhang, "Deep Knowledge Distillation Learning for Efficient Wearable Data Mining on the Edge," in *2023 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, Jan. 2023, pp. 1–3. doi: 10.1109/ICCE56470.2023.10043546.

[5] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *Int J Comput Vis*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.

[6] S. Zhang, C. Chen, Q. Xie, H. Sun, F. Dong, and S. Peng, "Distribution Unified and Probability Space Aligned Teacher-Student Learning for Imbalanced Visual Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2414–2425, Apr. 2024, doi: 10.1109/TCSVT.2023.3311142.

[7] J. Kim, S. Chang, and N. Kwak, "PQK: Model Compression via Pruning, Quantization, and Knowledge Distillation," in *Interspeech 2021*, ISCA: ISCA, Aug. 2021, pp. 4568–4572. doi: 10.21437/Interspeech.2021-248.

[8] J. Kim, "Quantization Robust Pruning With Knowledge Distillation," *IEEE Access*, vol. 11, pp. 26419–26426, 2023, doi: 10.1109/ACCESS.2023.3257864.

[9] Y. Qi, W. Zhang, X. Wang, X. You, S. Hu, and J. Chen, "Efficient Knowledge Distillation for Brain Tumor Segmentation," *Applied Sciences*, vol. 12, no. 23, p. 11980, Nov. 2022, doi: 10.3390/app122311980.

[10] Z. Zheng and G. Kang, "Model Compression with NAS and Knowledge Distillation for Medical Image Segmentation," in *2021 4th International Conference on Data Science and Information Technology*, New York, NY, USA: ACM, Jul. 2021, pp. 173–176. doi: 10.1145/3478905.3478940.

[11] X. Zeng *et al.*, "DSP-KD: Dual-Stage Progressive Knowledge Distillation for Skin Disease Classification," *Bioengineering*, vol. 11, no. 1, p. 70, Jan. 2024, doi: 10.3390/bioengineering11010070.

[12] V. Gorade, S. Mittal, D. Jha, and U. Bagci, "Rethinking Intermediate Layers Design in Knowledge Distillation for Kidney and Liver Tumor Segmentation," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, IEEE, May 2024, pp. 1–6. doi: 10.1109/ISBI56570.2024.10635141.

[13] X. Shi, Y. Li, J. Cheng, J. Bai, G. Zhao, and Y.-W. Chen, "Knowledge Distillation Using Segment Anything to U-Net Model for Lightweight High accuracy Medical Image Segmentation," in *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, IEEE, Oct. 2024, pp. 1073–1076. doi: 10.1109/GCCE62371.2024.10760506.

[14] qiaoyi chen and X. Yu, "Feature denoising distillation for medical image segmentation," in *Fourth International Conference on Image Processing and Intelligent Control (IPIC 2024)*, K. Du and A. bin Mohd Zain, Eds., SPIE, Aug. 2024, p. 42. doi: 10.1117/12.3038513.

[15] Y. Wen, L. Chen, S. Xi, Y. Deng, X. Tang, and C. Zhou, "Towards Efficient Medical Image Segmentation Via Boundary-Guided Knowledge Distillation," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 2021, pp. 1–6. doi: 10.1109/ICME51207.2021.9428395.

[16] L. Lin *et al.*, "FedLPPA: Learning Personalized Prompt and Aggregation for Federated Weakly-supervised Medical Image Segmentation," *IEEE Trans Med Imaging*, pp. 1–1, 2024, doi: 10.1109/TMI.2024.3483221.

[17] O. S. EL-Assiouti, G. Hamed, D. Khattab, and H. M. Ebied, "HDKD: Hybrid data-efficient knowledge distillation network for medical image classification," *Eng Appl Artif Intell*, vol. 138, p. 109430, Dec. 2024, doi: 10.1016/j.engappai.2024.109430.

[18] L. Xu, Z. Wang, W. Song, Y. Ji, and C. Liu, "SPARK: Cross-Guided Knowledge Distillation with Spatial Position Augmentation for Medical Image Segmentation," 2025, pp. 431–445. doi: 10.1007/978-981-97-8496-7_30.

[19] X. Qi *et al.*, "Exploring Generalizable Distillation for Efficient Medical Image Segmentation," *IEEE J Biomed Health Inform*, vol. 28, no. 7, pp. 4170–4183, Jul. 2024, doi: 10.1109/JBHI.2024.3385098.

[20] Y. Zhang, S. Li, and X. Yang, "Knowledge Distillation with Active Exploration and Self-Attention Based Inter-Class Variation Transfer for Image Segmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10097262.

[21] P. Liang, J. Chen, Q. Chang, and L. Yao, "RSKD: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in Vision Transformer models," *Knowl Based Syst*, vol. 293, p. 111664, Jun. 2024, doi: 10.1016/j.knosys.2024.111664.

[22] X. Qi, G. Yang, Y. He, W. Liu, A. Islam, and S. Li, "Contrastive Re-localization and History Distillation in Federated CMR Segmentation," 2022, pp. 256–265. doi: 10.1007/978-3-031-16443-9_25.

[23] D. Qing and L. Qi, "Research on Lightweight Spine X-ray Image Segmentation Algorithm Based on Knowledge Distillation," in *Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing*, New York, NY, USA: ACM, Jan. 2024, pp. 142–146. doi: 10.1145/3665689.3665713.

[24] L. Serrador, F. P. Villani, S. Moccia, and C. P. Santos, "Knowledge distillation on individual vertebrae segmentation exploiting 3D U-Net," *Computerized Medical Imaging and Graphics*, vol. 113, p. 102350, Apr. 2024, doi: 10.1016/j.compmedimag.2024.102350.

[25] L. Lin *et al.*, "FedLPPA: Learning Personalized Prompt and Aggregation for Federated Weakly-supervised Medical Image Segmentation," *IEEE Trans Med Imaging*, pp. 1–1, 2024, doi: 10.1109/TMI.2024.3483221.