Knowledge Distillation of Vision Transformers for Multiple sclerosis Lesion Classification in Brain MRI imaging*

No Author Given

No Institute Given

Abstract. Multiple sclerosis lesion detection in brain MRI remains a challenging task due to lesion heterogeneity, class imbalance, and variability in imaging protocols to detect the progression of this lesion. In this work, a new study using knowledge distillation is presented, employing Vision Transformers for multiple sclerosis lesion detection in brain MRI, focusing on transferring knowledge from a powerful Pyramid Vision Transformer teacher model to a lightweight MobileViT student model. Experimental results show that Pyramid Vision Transformer, as the most performant teacher, significantly, increasing the accuracy of student models from 91.25% to 94.64% and F1-score from 89.4% to 93.19%, achieving a 4% gain on a limited dataset. This study shows that using a powerful model like PVT as a teacher in a knowledge distillation framework can effectively improve the performance of smaller models such as MobileViT, even when training data is limited or imbalanced. By transferring rich feature representations, the approach enables lightweight models to achieve high accuracy and generalization, making them suitable for deployment in resource-constrained healthcare settings.

Keywords: Multiple sclerosis lesion, Brain MRI images, Vision Transformer, Knowledge distillation, deep learning, medical diagnosis.

1 Introduction

Multiple Sclerosis (MS) is a chronic, inflammatory, and demyelinating disease of the central nervous system, characterized by the formation of lesions in the brain and spinal cord. Early and accurate diagnosis of MS is crucial for effective clinical management and timely treatment, which can significantly slow disease progression and improve patient outcomes. Magnetic Resonance Imaging (MRI) plays a pivotal role in the diagnosis and monitoring of MS, as it provides high-resolution, non-invasive visualization of brain tissue and lesion patterns.

Despite the growing interest in leveraging deep learning for automated multiple sclerosis (MS) lesion detection in MRI, several key challenges continue to affect advancement in this field. A high challenge lies in the heterogeneity of MS lesions, which vary widely in size, shape, intensity, and anatomical location

* Funded by the Algerian DGRSDT as part of the PRFU project

across patients and disease stages. These lesions typically affect small, scattered areas of the brain, often dispersed across different regions, making data collection more challenging and resulting in severe class imbalance within the datasets. This imbalance complicates the training process and negatively impacts model generalization. Furthermore, significant variability in MRI acquisition protocols from different scanner types and imaging sequences introduces inconsistencies that limit the generalizability and stability of models across different datasets.

A key issue in this area is the longitudinal nature of MS, where tracking lesion evolution over time is clinically essential. However, the availability of robust longitudinal datasets continues to be limited. Finally, the high computational demands of training and deploying advanced deep learning models pose a barrier for many institutions, particularly in under-resourced environments, restricting the scalability and adoption of such methods in routine clinical workflows.

In this context, knowledge distillation emerges as a powerful technique for transferring the representational learning capabilities of large, high-performing models (teachers) to smaller, more efficient models (students). This approach is particularly valuable in medical imaging tasks such as MRI lesion detection, where deploying resource-intensive models is often impractical. By leveraging the distilled knowledge from teacher models, student models can achieve competitive performance with significantly reduced computational costs and memory requirements. Despite its potential, there remains a scarcity of research exploring knowledge distillation for lesion detection in MRI, especially in the context of complex neurological conditions like multiple sclerosis.

2 Related work

MS lesions differ significantly in size, shape, location, and intensity, often appearing hyperintense on FLAIR and hypointense on T1-weighted images. Their similarity to non-pathological features and variability across stages complicate accurate detection. Inconsistent visibility and anatomical diversity further challenge multimodal analysis and feature fusion, requiring advanced methods to handle structural complexity and lesion heterogeneity effectively [1]. Additionally, the inherent class imbalance, particularly between common lesions and rare yet clinically significant subtypes like paramagnetic rim lesions (Rim+) linked to chronic active inflammation in MS, which is important for diagnosis, further hinders reliable model training [2]. These factors collectively make MS lesion detection a complex task requiring sophisticated models capable of capturing fine-grained, multimodal features and accounting for lesion heterogeneity.

Recent paper [3] demonstrate that artificial intelligence, using ML and DL, enhances MS diagnosis and prognosis by enabling accurate lesion detection, subtype classification, and differentiation from other conditions. Models predict CIS-to-MS conversion with 67.6–92.9% accuracy and forecast disability progression by integrating clinical and imaging data, where lesion burden predicts short-term and gray matter damage long-term outcomes. Another paper [4] proposes a two-step method for Multiple Sclerosis lesion segmentation in MRI, combining a

modified expectation-maximization algorithm for brain tissue classification with a FLAIR-based thresholding and refinement process to detect lesions. Where [5] proposes a multiscale, segmentation-based approach for MS lesion detection in 3D multichannel MRI. Using hierarchical segmentation and a decision forest classifier trained on expert labels, it captures regional features effectively. The method achieved 0.74 sensitivity, 0.96 specificity, and 0.96 accuracy, showing strong alignment with expert annotations. A recent study [6] introduces a highly accurate and efficient framework combining SWIN Transformer and MobileNetV3-small for feature extraction with CatBoost, XGBoost, and Random Forest for classification. Achieving 99.8% average accuracy and 0.07 loss on the Kaggle MS dataset, the method shows strong potential for early, interpretable, and clinically effective MS detection.

Several recent methods have improved MS lesion segmentation using 3D CNNs. QSMRim-Net [2] combines deep residual 3D CNNs with radiomic features and DeepSMOTE to handle class imbalance, achieving 0.976 accuracy and 0.70 F1 score. A 3D U-Net in [7], trained with manual annotations and augmentation, reached 85% accuracy, while an ensemble of 3D U-Nets in [8] achieved a 0.70 ± 0.12 F1 score, enhancing robustness and generalization.

While many studies enhance MS lesion segmentation through architecture, augmentation, or loss design, they often depend on large, resource-heavy models. In contrast, knowledge distillation has emerged as a promising technique in the field of medical imaging, allowing smaller models to leverage the capabilities of larger models while maintaining efficiency.

Knowledge distillation (KD) has emerged as a powerful technique to enhance brain MRI diagnosis by transferring knowledge from complex teacher models to lightweight student models, maintaining high accuracy while reducing computational cost. A study [9] using 357 MRI scans achieved 98.10% accuracy with a multi-teacher KD strategy. FM-LiteLearn [10] improved tumor feature representation using image integration, with T-ResNet18 showing a 9.4% classification accuracy boost. CASD [11] applied self-distillation to refine feature extraction in multi-modal glioma grading. Studies addressing limited 3D brain imaging data [12–14] also report success. For example, [12] used CNN-LSTM to reach 85.96% accuracy and a 3.83% improvement in Alzheimer's detection. KD-FMV [14] improved transparency, with brain tumor classification accuracy of 98.77% (student: 97.48%) and Alzheimer's classification reaching 99.46%. CReg-KD [13] boosted performance across models like ResNet and DenseNet ($\sim$93–94% accuracy). In privacy-sensitive contexts, Fed-Brain-Distill [15] and Fed-SPD [16] combine KD with federated learning, achieving up to 94.38% accuracy with reduced training time and model size. Additional work [17, 18] uses multi-teacher strategies and attention mechanisms, achieving 95.85% accuracy on ACDC. KD with Vision Transformers [19–21] further improves efficiency and performance, including QViT_28 achieving AUC 0.812 and accuracy 0.693, closely matching ViT_28 while preserving quantum advantages.

In Alzheimer's disease detection, the Res-Transformer and ResU-Net combination [22] enhanced skip connections and stability, achieving 96.9% accuracy

via KD. RTAB [23] distilled temporal features from dual-stream ViTs, guiding models with subtle longitudinal cues, achieving 0.899 accuracy and 0.917 recall on MIRIAD. Additionally, [24] demonstrated ViTs' feasibility under low-data constraints, reaching 79.7% accuracy on ADNI1 and 82.0% on ADNI2, outperforming models trained without distillation.

Despite its growing success in tasks such as tumor segmentation and Alzheimer's disease classification, to the best of our knowledge, no prior work has investigated the use of knowledge distillation for MS lesion detection in brain MRI. This highlights a critical gap in the literature and motivates our study, which is the first to explore this approach for enhancing performance and efficiency in MS diagnosis.

3 Proposed approach

This section describes the proposed approach employed in our study as depicted in Fig 1, comprising three main steps: (1) transfer learning using large pre-trained models, (2) transfer learning using small models, and (3) knowledge distillation to transfer knowledge from the large model to the smaller one. Our objective was to evaluate various large models to identify the most effective one for serving as the teacher model based on its performance metrics rather than its size or architectural complexity. This ensures that the distilled knowledge comes from the model with the strongest generalization capability, regardless of how resource-intensive it may be.

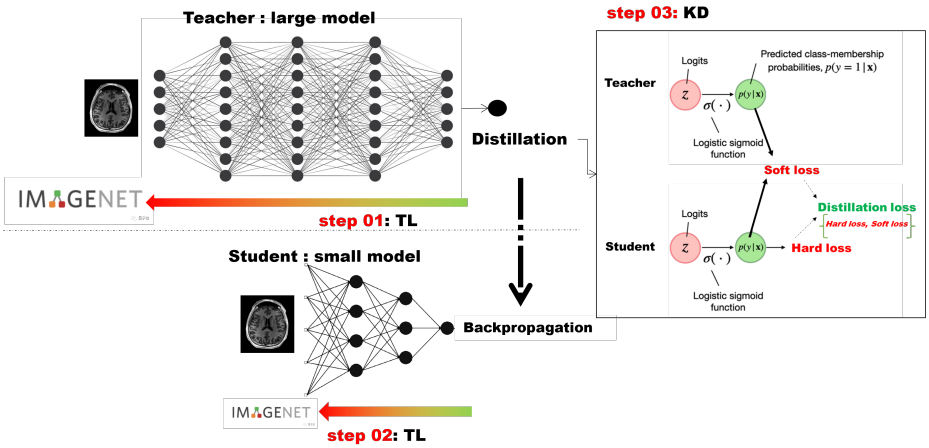

Fig. 1. The proposed distillation approach: Student Learning from Hard Labels and Teacher's Soft Predictions (Eq. 3, and 4)

3.1 Step1 : TL of Teacher Model (Large model)

In this study, transfer learning is used to fine-tune the Pyramid Vision Transformer (PVTV2B1) model for the diagnosis of brain MRI lesion. PVT is a hierarchical transformer architecture tailored for vision tasks, which efficiently balances local and global feature extraction by progressively reducing spatial resolution through its layers. This design allows it to capture rich contextual information while maintaining a moderate number of parameters (approximately 14 million in the V2B1 variant). PVTv2 is an improved extension of the original PVTv1, addressing some of its limitations by introducing overlapping patch embeddings (instead of non-overlapping ones) and convolutional feed-forward networks in place of the standard MLP blocks. These enhancements lead to better feature continuity, improved spatial understanding, and overall stronger performance in downstream tasks.

To adapt PVT to our target brain MRI dataset,the deeper layers of each model are fine-tuned, representing approximately $\frac{1}{4}$ of the total layers. The shallow layers learn local visual features and the deeper layers capture semantic information suitable for classification. This approach enabled us to leverage learned representations from large-scale natural image data while focusing training on high-level features relevant to lesion detection.

3.2 Step 2 : TL of Student Model (Small model)

The second stage involves applying transfer learning to a smaller, lightweight vision transformer model optimized for resource-constrained environments. In this study, we selected MobileViT v2-50 due to its compact architecture and proven efficiency in visual recognition tasks. Specifically, we used a reduced variant of MobileViT v2-50 with approximately 10.2% fewer parameters, making it even more suitable for efficient deployment. we choosed in this step an improved MobileViT v2, which is an enhanced version of the original MobileViT v1, which initially combined convolutional layers with lightweight transformer blocks. The v2 version introduces several improvements, including separable convolutions, inverted residual blocks, and more efficient attention mechanisms, resulting in better accuracy–efficiency trade-offs and improved performance on mobile and edge devices.

While its lightweight architecture offers a high advantage of inference speed and memory usage, its limited capacity can underperform when trained directly on the target dataset.

3.3 Step 03: Knowledge distillation from large to small model

To bridge performance gaps between large and small models, knowledge distillation trains a smaller student model using both ground truth labels and softened teacher logits. This combined loss approach enables the student to imitate the teacher's patterns efficiently. Its success depends on logit softening and effective student training.

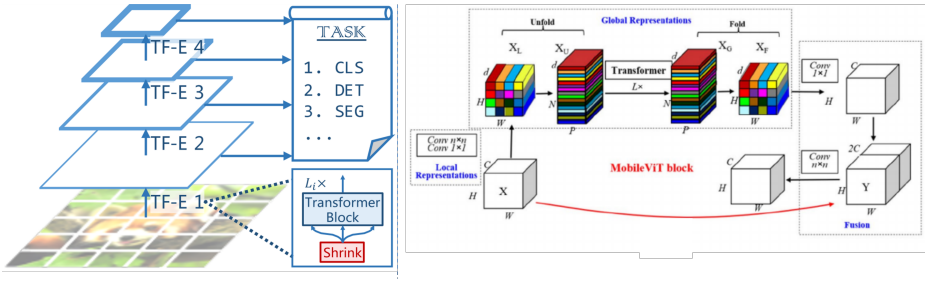

Fig. 2. Architectural vision transformers: Teacher: PVT(left) and student :MobileViT(right)

Logits Extraction and Softening Instead of directly using the final output probabilities of the teacher, we extract logits from the last dense layer before activation. These logits are softened using a temperature parameter T to control the smoothness of the output distribution produced by the teacher. Let us denote the following:

- z_t: Logits (raw output vector) from the teacher model
- z_s: Logits from the student model
- T: Temperature parameter
- C: Number of classes
- p_t: Softened probability distribution from the teacher
- p_s: Softened probability distribution from the student

Logits Extraction: Logits z are the outputs of the neural network before any activation function (like softmax or sigmoid). For a classification model, the logits are usually taken from the final linear layer.

Softening with Temperature T: To extract soft labels, we apply the sigmoid function to the logits with temperature

$$\mathbf{p}_t = Softmax\left(\frac{\mathbf{z}_t}{T}\right) = \frac{\exp\left(\mathbf{z}_t/T\right)}{\sum_{j=1}^{C}\exp\left(z_t^{(j)}/T\right)} \tag{1}$$

$$\mathbf{p}_s = Softmax\left(\frac{\mathbf{z}_s}{T}\right) = \frac{\exp\left(\mathbf{z}_s/T\right)}{\sum_{j=1}^{C}\exp\left(z_s^{(j)}/T\right)} \tag{2}$$

where $\mathbf{p}_t$ and $\mathbf{p}_s$ are the softened probability distributions of the teacher and the student, respectively.

To analyze the effect of distillation between student and teacher, we experimented with different configurations of the temperature with increasing parameter T: T=3, T= 10, and T= 100 which evaluate the similarities between classes that are masked by the hard labels. Higher temperatures produce softer probability distributions, helping the student model to learn more representations from the teacher predictions that are hidden by hard labels.

Training Strategy: The student model is optimized by minimizing a weighted combination of hard loss and distillation loss, controlled by a factor α. The distillation loss is computed using the Kullback–Leibler (KL) divergence between the softened outputs of the teacher and student:

$$\mathcal{L}_{distill} = T^2 \cdot KL(\mathbf{p}_t \,\|\, \mathbf{p}_s) = T^2 \sum_i p_t^{(i)} \log \left(\frac{p_t^{(i)}}{p_s^{(i)}} \right) \tag{3}$$

We experimented with different values of the temperature T and the weighting coefficient α, which balances the hard loss and distillation loss. Adjusting α allowed us to control how much the student model relies on the teacher's knowledge versus the original ground truth in the distillation loss (from the teacher's soft outputs) and the standard hard loss (from ground truth labels). By tuning α with different configurations, we aimed to explore how much the student should rely on the teacher's knowledge versus the original labels, and how closely the student model's learned distribution could align with that of the teacher. The student is trained by minimizing a weighted sum of both loss components:

$$\mathcal{L}_{Total} = \alpha \cdot \mathcal{L}_{Distill} + (1 - \alpha) \cdot \mathcal{L}_{Hard} \tag{4}$$

Gradients from this combined loss are used to update the student's parameters,improving it to match the teacher's performance while maintaining computational efficiency.

3.4 Dataset used

The dataset used in this study was originally hosted on Kaggle and is now available upon request or with authorized access [25]. It consists of grayscale brain MRI images curated for the task of multiple sclerosis (MS) detection and classification (see Fig. 3). All images are preprocessed and resized to a standard

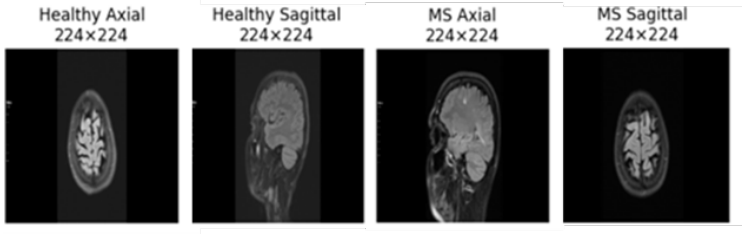

Fig. 3. Samples images from each class with labels and image size.

input shape of 224×224×1, representing single-channel grayscale slices, suitable for convolutional and transformer-based deep learning models.

The dataset (accessed on 19 feb 2024).contains 3427 MRI images across four classes: Healthy Axial, Healthy Sagittal, MS Sagittal, and MS Axial. It was split

once into training, validation and test, and this same split is used consistently across all experiments to ensure consistent comparison of the obtained results. The dataset exhibits a class imbalance with more healthy samples (see figure 3).

Table 1. Dataset Split: Training, Validation, and Test Sets for Healthy and MS Classes

Class	Train	Validation	Test	Total
Healthy	1289	323	404	2016
MS	903	226	282	1411
Total	2192	549	686	3427

4 Experimental results

In our experiments, we designed a binary classification task to distinguish between healthy individuals and patients with multiple sclerosis (MS) using brain MRI images. The original dataset labels were recoded such that images from healthy controls (both axial and sagittal). To ensure balanced class distribution, stratified split is applied on the training data into training and validation sets, using 80% for training and 20% for validation as illustrated in Table 1.

A knowledge distillation framework is implemented using hybrid vision transformer (ViT) models pretrained on ImageNet. Only the final transformer stack was fine-tuned, while earlier layers were frozen to retain pretrained features. Additional dense layers, dropout, and batch normalization were added for improved performance. Models were trained using the Adam optimizer (lr = $1e{-}3$), binary cross-entropy loss, and evaluated using accuracy, precision, recall, and F1-score. Mixed-precision training with XLA compilation accelerated training.

4.1 Results on step 01: Transfer learning on large models

In this experiment, the PVT-V2B1 (14M) model is evaluated as large teacher model to assess its effectiveness on the MS lesion classification task. The results obtained from this evaluation are summarized in Table 2.

Table 2. Transfer learning performance of PVT_V2B1 on the MS dataset

Model	Params	Test/Val	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PVTV2B1	14M	Val	96.54	96.83	94.69	95.75
		Test	**96.50**	**96.40**	**96.04**	**95.71**

Based on the obtained results in the table bellow (see Table 2) , PVT V2 B1 emerged as the best-performing larger model, achieving a test accuracy of

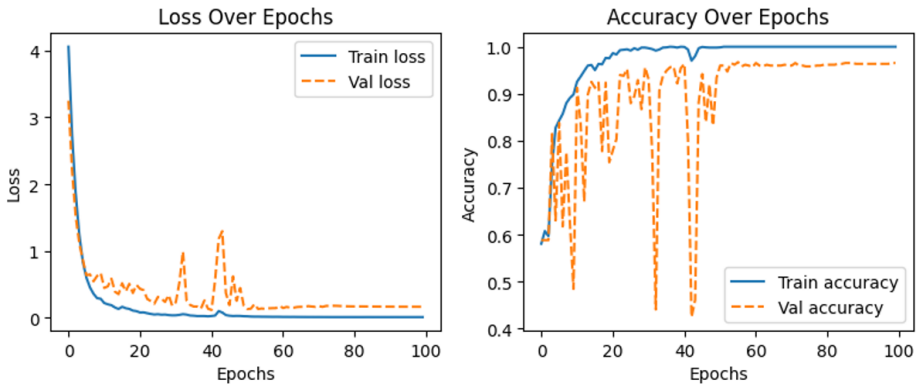

Fig. 4. the performance of the selected model to serve as teacher: PVT

96.5%, precision of 96.4%, recall of 96.0%, and an F1-score of 95.7%. However, PVT V2 B1 showed early instability (see Fig. 4) due to class imbalance, causing decreasing validation accuracy and convergence issues at epoch 40.

4.2 Results of step 02: Transfer learning on smaller models

In a second experiment, a lightweight MobileViT-V2-050 model is evaluated , under the same parameter tuning to assess their performance on MS lesion classification. The results are illustrated in the table bellow (see Table 3) The

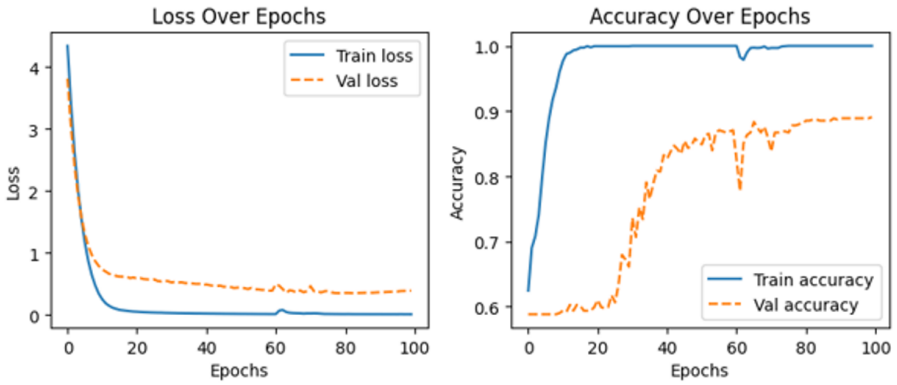

Fig. 5. TL on smaller lightweight vision transformer: MobileViTv2

divergence between training and validation curves in Fig. 5 reveals a generalization gap, indicating overfitting. While training loss consistently declines and accuracy nears 100%, validation loss and accuracy stop improving early whith

significantly lower performance that training. The persistent gap in validation metrics suggest limited generalization to unseen data.

Table 3. Evaluation of transfer learning performance using smaller models on MS dataset (in %)

Model	params	Test/val	Accuracy (%)	Precision (%)	Recall (%)	Fscore (%)
MobileViTV2	1.37M	Val	89.07	90.29	82.30	86.11
		Test	91.25	90.51	87.94	89.21

4.3 Results of step 03: knowledge distillation:

The third experiment focused on applying knowledge distillation to compress and optimize model performance for MS lesion classification. A high-capacity PVT_V2-Large model served as the teacher, while lightweight architectures MobileViT-V2-050 is selected as student models. Various distillation settings were explored, including different temperature values ($T \in \{3, 10, 100\}$) to control output softness and alpha values ($\alpha \in [0.3, 0.9]$) to balance student and distillation losses. Two loss function combinations, including binary cross-entropy and Kullback–Leibler divergence, were tested.

Table 4. Performance Comparison of Student Models With and Without Knowledge Distillation (in %)

Student Model	T	α	Accuracy	Precision	Recall	F1-score
MobileViT_v2_050 (With Distillation)						
	3	0.5	**94.46**	**94.20**	**92.20**	**93.19**
	3	0.7	92.71	95.67	86.17	90.67
	10	0.7	90.52	92.22	84.04	87.94
	10	0.5	92.57	92.62	89.01	90.78
	3	0.4	92.42	93.89	87.23	90.44
	100	0.7	88.78	89.88	81.91	85.71
MobileViT_v2_050 (Without Distillation)			91.25	90.51	87.94	89.21

According to the results obtained (see Table 4), Knowledge distillation significantly enhanced lightweight model performance in MS lesion classification, with MobileViT-V2-050 showing a 4% F1-score improvement (93.19% vs. 89.21%) using $T = 3$ and $\alpha = 0.5$. Moderate temperatures ($T = 3, 10$) improved generalization, while higher values degraded performance. Lower alpha values (0.4–0.5) improve the distillation loss, resulting in better loss-accuracy balance and more effective training outcomes.

5 Discussion

MS lesion detection is challenging due to their variability in size, shape, and location, often requiring expert-guided segmentation. This study leverages Vision Transformers and transfer learning, selecting the most performing model as a teacher for knowledge distillation to train smaller, efficient student models. The approach avoids complex 3D segmentation, data augmentation, or deep architectures like U-Net while maintaining high accuracy. Notably, as seen in Table 2, PVTv2 demonstrate the efficiency and practicality of the proposed approach leveraging knowledge distillation technique to enhance the performance of transfer learning on smaller models like MobileViTv2 with just one millions of parameters. An improvement of 4% was achieved on a limited 2D imbalanced dataset using MobileViT as the student model (see Table 4), which has 10.2 times fewer parameters than the teacher model (PVT), suggesting that a larger size gap between teacher and student leads to better distillation performance and reduces the student's generalization gap between training and validation.

Table 5. Performance of our proposed MS lesion detection approaches on Kaggle multiple sclerosis dataset

Ref	Model	Approach	Performance (%)	Parameters (M)	Validation
[25]	Hybrid CNN-ML	Combination of DenseNet201, ResNet50, and classical classifiers (SVM and KNN)	ACC: 97	∼45	10-fold CV
[26]	ExMPLPQ	Patch-based handcrafted feature extraction using multi-parameter LPQ + INCA feature selection + Fine kNN classifier	ACC: 98.22	–	10-fold CV
Ours	PVT_v2	Teacher model via KD	ACC: 96.5, F1: 95.71	14	20% split
	MobileViT_v2	Student model via KD	ACC: 94.46, F1: 93.19	1.48	20% split

Unlike the hybrid method used by [25], which combines DenseNet201, ResNet50, and traditional classifiers like SVM and KNN, relying on multiple deep networks and added classifier overhead, and the handcrafted feature-based method in [26], which uses patch-based local phase quantization (LPQ) and Fine kNN with 10-fold cross-validation, our proposed approach achieves comparable performance

with significantly reduced complexity using a lightweight MobileViTV2 student model. It requires 30× fewer parameters than the combined DenseNet201 and ResNet50 model. Moreover, our method operates end-to-end without any segmentation step or handcrafted feature extraction, making it more suitable for practical deployment and large-scale screening in clinical or resource-constrained environments.

In this context, our results are projected onto other methods that used different datasets for subjective evaluation. The comparative analysis in Table 6 includes comprehensive dataset specifications to contextualize performance differences arising from data characteristics,and it is noteworthy that many advanced MS lesion detection methods use complex 3D segmentation, manual annotations, and ensemble learning, making them computationally intensive.

Table 6. Comparison of State-of-the-Art MS Lesion Detection Methods Utilizing MRI Data

Ref	Year	Model	Approach	Dataset	Size	Access	Performance
[5]	2009	3D Segmentation	Multiscale graph partitioning w/decision forest	Multichannel MRI	91	Private	ACC: 98%
[2]	2022	QSMRim-Net	3D ResNet + radiomic fusion + DeepSMOTE	Cornell MS-QSM	688	Private	ACC: 97.6%, F1: 70%
[7]	2022	3D U-Net	Supervised segmentation + data augmentation	CLAIMS	440	Private	ACC: 85%
[6]	2024	Hybrid CNN-ViT Ensemble	Swin Transformer + MobileNet RF	Kaggle MS 2023	765	Public	ACC: 99%
[8]	2024	3D U-Net Ensemble	Aggregated ensemble predictions	In-house dataset	491	Private	F1:70%

Techniques like 3D U-Net [5, 7, 8], QSMRim-Net [2], and 2D hybrid CNN-Transformer models [6] achieve high accuracy but rely on heavy data processing, manual input, and complex pipeline architectures, limiting their practical deployment.

By using transfer learning from ImageNet with knowledge distillation approach and without any segmentation method, this approach reduces reliance on manual annotations and complex preprocessing, while still preserving critical diagnostic in-sight, making it highly practical for real-world deployment.

6 Conclusion

This study presents a novel approach for multiple sclerosis (MS) lesion detection in brain MRI using knowledge distillation with vision transformers (ViTs), without relying on explicit segmentation models. By representations from large pretrained PVT teacher models to smaller student models, our method significantly improves performance, achieving a 4% gain improvement. Our approach has practical value in healthcare, particularly in resource-constrained environment. The ability to distill powerful attention representations into lightweight

models like MobileViT enables faster, more accessible MS detection tools and supports early intervention, all critical in managing disease progression. For future work, we aim to expand this framework by investigating knowledge distillation across various model architectures, focusing on the relationship between the complexity of teacher and student models and the optimal size difference needed to achieve higher performance. Additionally, we plan to evaluate the approach on larger, more diverse datasets and explore its integration into clinical decision support systems for longitudinal MS monitoring.

References

1. Bagnato, F., Sati, P., Hemond, C.C., Elliott, C., Gauthier, S.A., Harrison, D.M., et al.: Imaging chronic active lesions in multiple sclerosis: a consensus statement. Brain **147**(9), 2913–2933 (2024)
2. Zhang, H., Nguyen, T.D., Zhang, J., Marcille, M., Spincemaille, P., Wang, Y., et al.: QSMRim-Net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps. NeuroImage (2022)
3. Rocca, M.A., Preziosa, P., Barkhof, F., Brownlee, W., Calabrese, M., De Stefano, N., et al.: Current and future role of MRI in the diagnosis and prognosis of multiple sclerosis. Lancet Reg. Health Eur. **44** (2024)
4. Cabezas, M., Oliver, A., Roura, E., Freixenet, J., Vilanova, J.C., Ramió-Torrentà, L., et al.: Automatic multiple sclerosis lesion detection in brain MRI by FLAIR thresholding. Comput. Methods Programs Biomed. **115**(3), 147–161 (2014)
5. Akselrod-Ballin, A., Galun, M., Gomori, J.M., Filippi, M., Valsasina, P., Basri, R., Brandt, A.: Automatic segmentation and classification of multiple sclerosis in multichannel MRI. IEEE Trans. Biomed. Eng. **56**(10), 2461–2469 (2009)
6. Ismail, K.A.A., Dutta, A.K., Sait, A.R.W.: Ensemble learning-based multiple sclerosis detection technique using magnetic resonance imaging. J. Disabil. Res. **3**(6), 20240078 (2024)
7. La Rosa, F., Beck, E.S., Maranzano, J., Todea, R.A., van Gelderen, P., de Zwart, J.A., et al.: Multiple sclerosis cortical lesion detection with deep learning at ultra-high-field MRI. NMR Biomed. **35**(8), e4730 (2022)
8. Wiltgen, T., McGinnis, J., Schlaeger, S., Kofler, F., Voon, C., Berthele, A., et al.: LST-AI: A deep learning ensemble for accurate MS lesion segmentation. NeuroImage: Clin. **42**, 103611 (2024)
9. Anantathanavit, R., Raswa, F.H., Thaipisutikul, T., Wang, J.C.: Lightweight brain tumor diagnosis via knowledge distillation. In: Int. Conf. on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1–6 (2024)
10. Tan, S., Cai, Y., Zhao, Y., Hu, J., Chen, Y., He, C.: FM-LiteLearn: A lightweight brain tumor classification framework integrating image fusion and multi-teacher distillation strategies. In: Int. Conf. on AI in Healthcare (2024)
11. Li, J., Zhang, L., Zhong, K., Qian, G.: A discrepancy-aware self-distillation method for multi-modal glioma grading. Knowl.-Based Syst. **295**, (2024)
12. Li, Y., Luo, J., Zhang, J.: Classification of Alzheimer's disease in MRI images using knowledge distillation framework: an investigation. Int. J. Comput. Assist. Radiol. Surg. **17**(7), 1235–1243 (2022)
13. Yang, Y., Guo, X., Ye, C., Xiang, Y., Ma, T.: Creg-kd: Model refinement via confidence regularized knowledge distillation for brain imaging. Med. Image Anal. **89**, 102916 (2023)

14. Jiang, Y., Zhao, X., Wu, Y., Chaddad, A.: A knowledge distillation-based approach to enhance transparency of classifier models. arXiv preprint arXiv:2502.15959 (2025)
15. Gohari, R.J., Aliahmadipour, L., Valipour, E.: FedBrain-Distill: Communication-efficient federated brain tumor classification using ensemble knowledge distillation on non-IID data.(2024)
16. Wu, B., Shi, D., Aguilar, J.: Brain tumors classification in MRIs based on personalized federated distillation learning with similarity-preserving. Int. J. Imaging Syst. Technol. **35**(2), e70046 (2025)
17. Nabavi, S., Hamedani, K.A., Moghaddam, M.E., Abin, A.A., Frangi, A.F.: Multiple Teachers-Meticulous Student: A domain adaptive meta-knowledge distillation model for medical image classification. (2024)
18. Matcha, N., Ramanarayanan, S., Al Fahim, M., GS, R., Ram, K., Sivaprakasam, M.: SFT-KD-recon: Learning a student-friendly teacher for knowledge distillation in magnetic resonance image reconstruction. In: Medical Imaging with Deep Learning (2024)
19. Ferdous, G.J., Sathi, K.A., Hossain, M.A., Hoque, M.M., Dewan, M.A.A.: LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. IEEE Access **11**, 20337–20350 (2023)
20. EL-Assiouti, O.S., Hamed, G., Khattab, D., Ebied, H.M.: HDKD: Hybrid data-efficient knowledge distillation network for medical image classification. Eng. Appl. Artif. Intell. **138**, 109430 (2024)
21. Boucher, T., Mazomenos, E.B.: Distilling knowledge into quantum vision transformers for biomedical image classification. (2025)
22. Song, Y., Wang, J., Ge, Y., Li, L., Guo, J., Dong, Q., Liao, Z.: Medical image classification: Knowledge transfer via residual U-Net and vision transformer-based teacher-student model with knowledge distillation. J. Vis. Commun. Image Represent. (2024)
23. Chen, K., Wang, Y., Zhou, Y., Wang, H.: DS-ViT: Dual-stream Vision Transformer for cross-task distillation in Alzheimer's early diagnosis. (2024)
24. Kunanbayev, K., Shen, V., Kim, D.S.: Training ViT with limited data for Alzheimer's disease classification: An empirical study. In: Int. Conf. on Med. Image Comput. Comput.-Assist. Interv., pp. 334–343 (2024)
25. Tatli, S., Macin, G., Tasci, I., Tasci, B., Barua, P.D., Baygin, M., et al.: Transfer-transfer model with MSNet: An automated accurate multiple sclerosis and myelitis detection system. Expert Syst. Appl. **236**, 121314 (2024), https://www.kaggle.com/datasets/buraktaci/multiple-sclerosis (accessed Feb 19, 2025)
26. Macin, G., Tasci, B., Tasci, I., Faust, O., Barua, P.D., Dogan, S., Tuncer, T., Tan, R.-S., Acharya, U.R.: An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: ExMPLPQ. Appl. Sci. **12**(10), 4920 (2022), https://doi.org/10.3390/app12104920