INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE MINAS GERAIS (IFMG)
CAMPUS BAMBUÍ
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Cauê Silva Dias

AVALIAÇÃO DE ARQUITETURAS DE DEEP LEARNING PARA ESCLEROSE MÚLTIPLA: DE BASELINES CONVOLUCIONAIS A REDES HÍBRIDAS VIT E PROTOTÍPICAS

Bambuí - MG
2026

CAUÊ SILVA DIAS

AVALIAÇÃO DE ARQUITETURAS DE DEEP LEARNING PARA ESCLEROSE MÚLTIPLA: DE BASELINES CONVOLUCIONAIS A REDES HÍBRIDAS VIT E PROTOTÍPICAS

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Orientadora: Prof. [ME] Natália Camillo do Carmo

Bambuí - MG
2026

Catalogação na Fonte Biblioteca IFMG - Campus Bambuí

D541a Dias, Cauê Silva.
Avaliação de arquiteturas de deep learning para esclerose múltipla: de baselines onvolucionais a redes híbridas Vit e prototípicas. / Cauê Silva Dias. – 2026.
15 f.; il.: color.

Orientadora: Prof. [ME] Natália Camillo do Carmo.
Trabalho de Conclusão de Curso (graduação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG, Curso Bacharelado em Engenharia de Computação, 2026.

1. Multiple sclerosis. 2. Hybrid architecture. 3. Vision transformer. I. Carmo, Natália Camillo do. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG. III. Título.

CDD 004.22

Elaborada por Douglas Bernardes de Castro- CRB-6/2802

Cauê Silva Dias

EVALUATING DEEP LEARNING ARCHITECTURES FOR MULTIPLE SCLEROSIS: FROM CONVOLUTIONAL BASELINES TO HYBRID VIT AND PROTOTYPICAL NETWORKS

Undergraduate thesis submitted to the Bachelor's Degree Program in Computer Engineering at the Federal Institute of Education, Science and Technology of Minas Gerais (IFMG) – Bambuí Campus, in partial fulfillment of the requirements for obtaining the degree of Bachelor of Computer Engineering.

Approved on May 06, 2026, by the examining committee:

Prof. [ME] Natália Camillo do Carmo – IFMG – *Campus* Bambuí – (Advisor)
Prof. [DR] Felipe Lopes de Melo Faria – Campus Bambuí
Prof. [DR] Álvaro Antônio Fonseca de Souza – Campus Bambuí

Evaluating Deep Learning Architectures for Multiple Sclerosis: From Convolutional Baselines to Hybrid ViT and Prototypical Networks

Cauê Silva Dias
cauedias52@gmail.com
Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG)
Campus Bambuí
Bacharelado em Engenharia de Computação
Orientadora: Prof. [ME] Natália Camillo do Carmo

06 de Maio de 2026

ABSTRACT

The diagnosis of multiple sclerosis (MS) by magnetic resonance imaging (MRI) remains a challenge due to the complex spatial distribution of the lesions. Although Convolutional Neural Networks (CNNs) are effective at identifying local textures, they often fail to capture the broader spatial relationships between distant areas of the brain. In this study, we propose and compare three deep learning architectures for automated MS classification: a standalone ResNet18 baseline, a hybrid CNN-ViT model trained with cross-entropy loss, and a CNN-ViT variant trained with prototypical loss for metric learning. To ensure a reliable performance assessment, all models were validated using a 5-fold cross-validation and an independent test set of 60 subjects. The hybrid CNN-ViT architecture achieved 86.67% accuracy and 90.00% specificity, consistently outperforming the CNN-only baseline. The prototypical variant achieved 83.33% accuracy and 80.00% specificity, offering a complementary metric learning perspective. Additionally, we analyzed the model's attention maps to interpret its decision-making process, which highlighted both its ability to focus on relevant neurological features and its current limitations with non-brain tissues. These results indicate that combining self-attention mechanisms with convolutional layers improves classification performance and interpretability, providing a robust framework for automated MS detection even with limited datasets.

Keywords: Multiple Sclerosis. Hybrid Architecture. Vision Transformer. MRI Classification. Metric Learning. Prototypical Networks. Deep Learning.

1 INTRODUCTION

Multiple sclerosis can be defined as a chronic autoimmune disease of the central nervous system characterized by inflammatory and neurodegenerative processes that lead to structural damage in the tissues of the brain and spinal cord (Haki *et al.*, 2024; Ferguson *et al.*,

1997; Oh; Bar-Or; Marrie, 2023). Early diagnosis of the disease is crucial to enable therapeutic interventions capable of slowing its progression, preserving the functional abilities of the patient, and promoting a better quality of life (Multiple Sclerosis International Federation, 2020).

The diagnosis is made with the result of clinical signs first and then validated with MRI images. Differences in resolution, acquisition protocols, and sequences used directly impact image quality and, consequently, diagnostic accuracy. Furthermore, the detection of small or atypical lesions can go unnoticed in visual assessments, requiring tools capable of identifying subtle and recurring patterns in large datasets (Rocca *et al.*, 2024).

Radiology specialists still outperform automated methods in MS detection, mainly in terms of precision and clinical interpretation of images (Commowick *et al.*, 2018). Nevertheless, the diagnostic imaging process is often time-consuming, subject to variability among specialists and, in many contexts, limited by the availability of qualified professionals (Commowick *et al.*, 2023).

In MS research, deep learning approaches, especially 2D and 3D CNNs applied to MRI, have shown promising results for lesion segmentation and disease analysis (Belwal; Singh, 2025; Daqqaq; Alhasan; Ghunaim, 2024). However, existing studies consistently report limitations related to heterogeneous experimental designs, variability in imaging protocols, limited availability of standardized public datasets, and insufficient external clinical validation, which restrict model generalization and clinical applicability. Recent studies have explored hybrid neural network architectures that combine CNNs with attention mechanisms or transformers, demonstrating improved performance and interpretability in complex medical imaging and clinical decision-support tasks (Djoumessi; Mensah; Berens, 2025; Eguia *et al.*, 2024).

In this paper, three deep learning frameworks for automated multiple sclerosis classification are proposed and compared from MRI images. The first is a standalone ResNet18 serving as a convolutional baseline. The second integrates a ResNet18 backbone with a Vision Transformer (ViT), using the convolutional layers to extract local feature representations and a Transformer encoder to model global contextual dependencies through self-attention, trained with weighted cross-entropy loss. The third architecture reuses the same CNN-ViT backbone but replaces the classification head with a prototypical learning strategy, training the network to produce L2-normalized embeddings in a metric space where classification is performed by nearest-prototype assignment. The effectiveness of all three architectures is evaluated using metrics such as accuracy, precision, recall, specificity, F1 score, and confusion matrix analysis. Model performance is assessed through 5-fold stratified cross-validation in the training set and validated in a fixed independent test set comprising 30% of the total dataset. Additionally, attention map visualization is performed to provide interpretability of the learned feature representations and demonstrate the model's focus regions within the brain MRI.

2 METHODS

The methodology consisted of developing and training three deep learning models to classify brain magnetic resonance images according to the presence or absence of MS, following the workflow illustrated in Fig. 1. MRI data were collected from two main datasets, one publicly available dataset with images of MS patients and healthy controls[1], and the other, the data were provided by the Functional and Molecular Neuroimaging Program, IRCCS Istituto delle Scienze Neurologiche di Bologna (Fiscone *et al.*, 2024)[2]. Data were analyzed to verify integrity, quality, and consistency. Both datasets were combined into a single cohort of 199 subjects, comprising 50 healthy controls (subjects 001–099) and 149 MS patients (subjects 100–199). The preprocessing steps included loading NIfTI (.nii.gz) volumes, selecting the central axial slice of each 3D image, applying min-max intensity normalization to the range [0, 1], and resizing all images to a resolution of 224×224 pixels. The dataset was then divided into training, validation, and test subsets: 30% of the subjects (60 subjects) were held out as a fixed independent test set using stratified random splitting (random_state=42), and the remaining 70% (139 subjects) were used for training and validation through 5-fold stratified cross-validation.

To address class imbalance, class weights were computed inversely proportional to class frequencies and applied to the loss function. Training data augmentation was performed on the fly using random horizontal flipping (probability 0.5), random vertical flipping (probability 0.3), random rotation within $\pm 10^{\circ}$ (probability 0.3), and Gaussian blur with a 3×3 kernel (probability 0.3). Augmentation was applied with higher probability to healthy subjects (0.75) compared to MS patients (0.50) to further compensate for class imbalance. Validation and test images were not augmented. Images were normalized using ResNet standard statistics (mean=0.485, std=0.229) and resized to 224×224 pixels prior to model input.

The choice of ResNet18 as the convolutional backbone for local feature extraction is motivated by both architectural and computational considerations. ResNet18 employs residual connections of the form $y = F(x, \{W_i\}) + x$, where the identity shortcut enables stable gradient propagation across 18 layers without vanishing gradient degradation, allowing the network to learn discriminative low- and mid-level features, such as edges, textures, and intensity gradients, relevant to MS lesion characterization in T1-weighted MRI. By truncating the forward pass at the output of the third residual block (layer3), the backbone produces spatially rich feature maps of dimension 14×14×256, where each of the 196 spatial positions encodes a 256-dimensional descriptor capturing local convolutional activations over a receptive field that covers a substantial portion of the 224×224 input. This spatial resolution deliberately preserves sufficient structural detail for the subsequent tokenization step: each 14×14 spatial cell is treated as an independent patch token, enabling the Transformer encoder to reason over 196 distinct local descriptors without the spatial collapse that would result from using the full ResNet18 output

[1] Publicly dataset from Imperial College London
[2] MRI dataset for susceptibility-based radiomic feature extraction

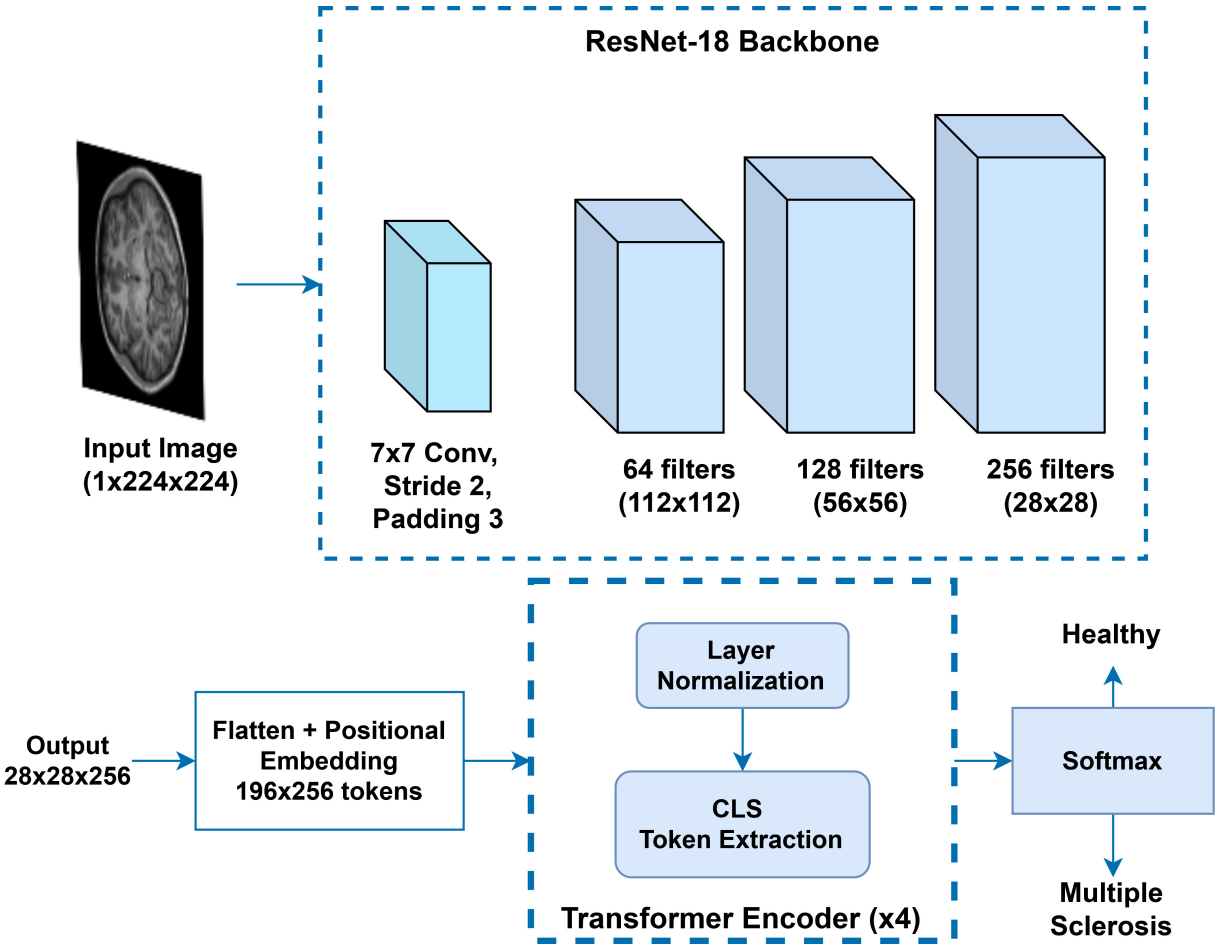

Figura 1 – Proposed Hybrid CNN-ViT architecture for MS classification, with a ResNet-18 backbone and a Transformer encoder with CLS token extraction for final classification.

after global average pooling. Furthermore, ResNet18 offers a parameter-efficient inductive bias for local feature extraction; its convolutional filters inherently capture translation-equivariant patterns, which complements the global, permutation-invariant self-attention of the Transformer encoder, achieving a structured division of labor between local and global representation learning within a single unified architecture.

The third architecture reuses the same ResNet18 backbone and Transformer encoder described above, but replaces the conventional classification head with a metric learning strategy based on prototypical networks (Djoumessi; Mensah; Berens, 2025; Azad *et al.*, 2024). The model is trained to produce 256-dimensional L2-normalized embeddings, where classification at inference time is performed by assigning each sample to the nearest class prototype computed from the full training set.

During training, the prototypical loss is computed by measuring the Euclidean distance between each embedding and the per-batch class prototypes, converting distances to temperature-scaled logits, and applying cross-entropy:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(-\tau \cdot d(\hat{\mathbf{e}}_i,\, \mathbf{p}_{y_i}))}{\sum_c \exp(-\tau \cdot d(\hat{\mathbf{e}}_i,\, \mathbf{p}_c))} \tag{1}$$

where $\hat{\mathbf{e}}_i$ is the L2-normalized embedding of sample i, $\mathbf{p}_c$ is the L2-normalized mean embedding of class c within the batch, $\tau = 10.0$ is a fixed temperature, and $d(\cdot, \cdot)$ is the Euclidean distance. This formulation encourages same-class embeddings to cluster while separating inter-class representations, which is particularly relevant for small and imbalanced medical imaging datasets (Haruna *et al.*, 2025). The training procedure employed AdamW (lr $= 10^{-4}$, weight decay 0.02) with gradient clipping (norm 1.0), maintaining the same backbone fine-tuning strategy as the cross-entropy variant.

3 RESULTS

The standalone CNN model achieved an accuracy of 0.83 in the fixed test set, with a precision of 0.81, a recall of 0.86, and a specificity of 0.80, giving an F1 score of 0.83. During the training of k-fold cross-validation, the validation accuracies exhibited variability across the folds, ranging from 0.75 to 0.96 in the final epochs. The hybrid CNN-ViT model demonstrated mean validation accuracy of 0.82 $\pm$ 0.05, precision of 0.81 $\pm$ 0.07, recall of 0.85 $\pm$ 0.11, and specificity of 0.81 $\pm$ 0.02, resulting in F1 score of 0.82 $\pm$ 0.06. Although the mean validation accuracy was slightly lower than the ResNet18 test performance, the hybrid architecture achieved superior results on the fixed test set, with accuracy of 0.8667, precision of 0.89, recall of 0.83, specificity of 0.90, and F1 score of 0.86. A summary of the accuracy and precision metrics for all three proposed architectures is presented in Table 1.

The CNN-ViT with prototypical loss demonstrated mean validation accuracy of

Tabela 1 – Quantitative Comparison of Proposed Models

Architecture	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
CNN (ResNet18)	83.33	81.25	86.67	80.00
Hybrid CNN-ViT	86.67	89.29	83.33	90.00
CNN-ViT + Metric Learning	83.33	81.25	86.67	80.00

0.7918 ± 0.0908, precision of 0.8160 ± 0.0616, recall of 0.7494 ± 0.1303, and specificity of 0.8153 ± 0.0905, resulting in an F1 score of 0.7779 ± 0.0880. On the fixed independent test set, this model achieved an accuracy of 0.8333, precision of 0.8125, recall of 0.8667, specificity of 0.8000, and F1 score of 0.8387. These results indicate that the prototypical formulation maintains competitive test performance relative to the cross-entropy baseline (ResNet18), while exhibiting higher cross-validation variance, which reflects the additional sensitivity of metric learning to the batch composition and the limited dataset size. The higher recall (0.8667) compared to the standard CNN-ViT (0.8333) suggests that the prototypical model is more sensitive toward the positive class (MS patients), a clinically relevant property in screening contexts.

A qualitative inspection of representative test examples further illustrates the model's behavior. As shown in Fig. 2, the network correctly classified a patient with multiple sclerosis, indicating its ability to identify pathological patterns associated with lesion presence. However, in the misclassified case, a healthy subject was predicted as diseased. Visual analysis suggests that high-intensity structures near the cortical borders and skull interface, as well as contrast variations in peripheral regions, may have influenced the attention mechanism. These edge-related patterns can resemble lesion-like hyperintensities, potentially leading the model to focus on non-pathological anatomical variations. This behavior indicates that peripheral artifacts, intensity inhomogeneities, or dataset-specific biases may still affect the decision process, highlighting the need for improved spatial regularization or lesion-focused attention mechanisms.

The attention maps generated by the Vision Transformer component across different subjects, depicted in Fig. 3, reveal additional insight into the model's spatial reasoning. These maps are produced by extracting the self-attention weights of the CLS token with respect to all 196 patch tokens from the last Transformer encoder layer, averaging across all 8 attention heads, reshaping the resulting vector into a 14×14 spatial grid, and upsampling it to 224×224 pixels via bilinear interpolation. In Fig. 3, the attention weights are concentrated in anatomically plausible regions, particularly in the superior frontal and parietal white matter, with a secondary focus in the inferior periventricular region, both areas commonly associated with MS lesion distribution, with peak activation values approaching 0.8–1.0. This pattern suggests that the self-attention mechanism successfully directed the model toward diagnostically relevant features within the cerebral parenchyma.

The architectural configuration of the Transformer encoder directly influences this capacity for structured spatial reasoning. Each of the 196 patch tokens encodes a 16×16-pixel

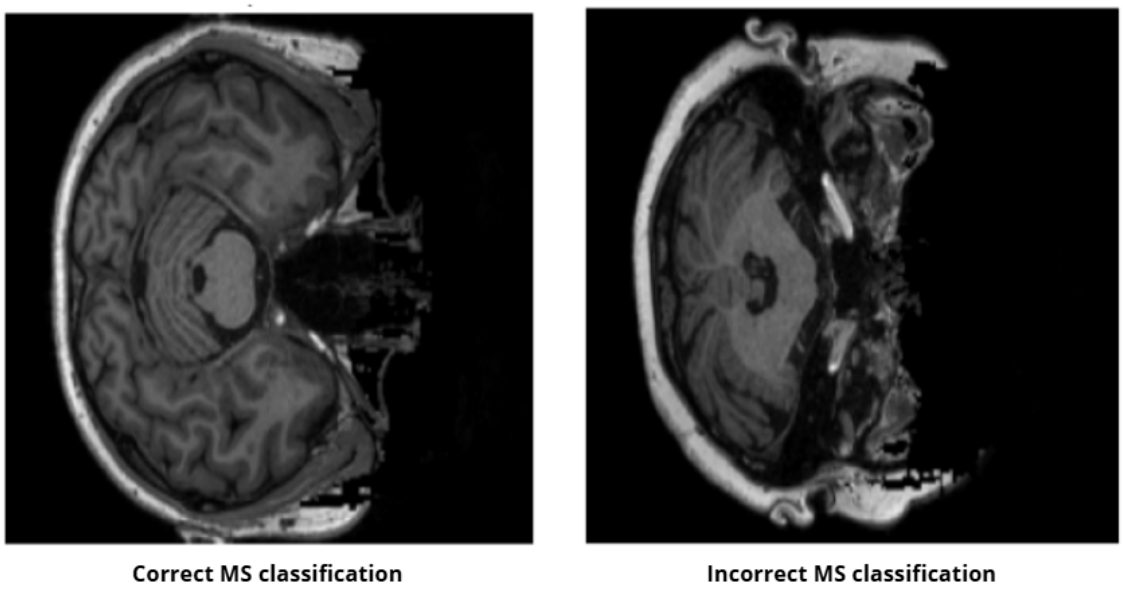

Figura 2 – Representative test set predictions from the CNN-ViT model. Left: MS patient correctly classified as diseased. Right: healthy subject misclassified as diseased (false positive).

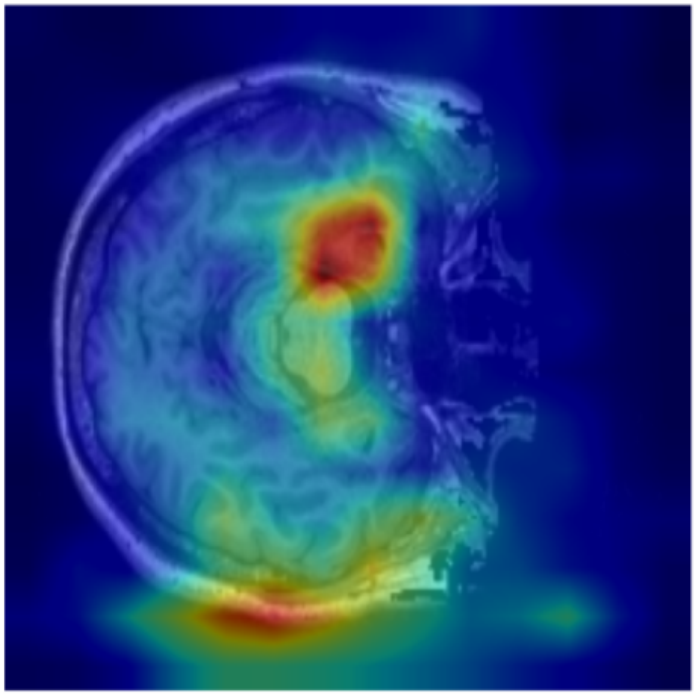

Figura 3 – Attention map with attention focused on anatomically plausible regions for MS classification.

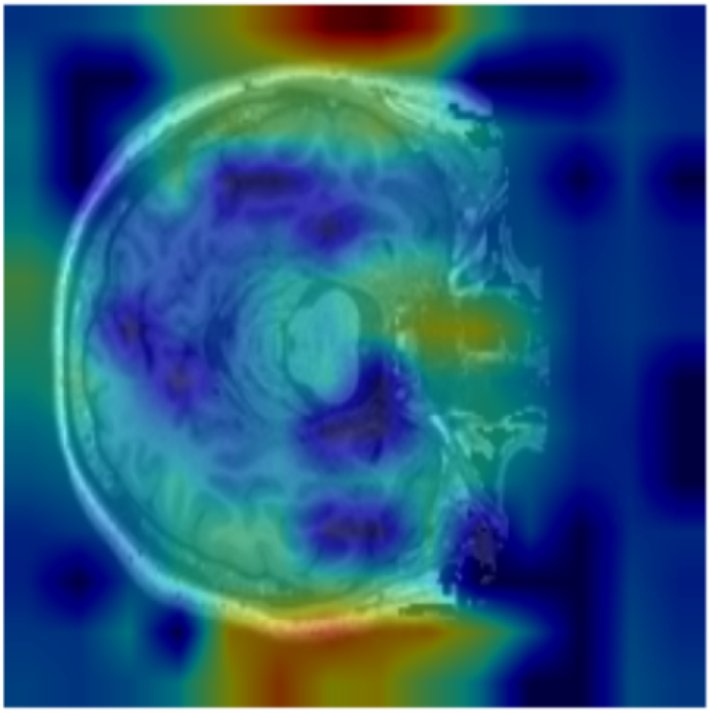

Figura 4 – Attention map with attention focusing on regions outside the brain.

region of the 14×14 feature map produced by the ResNet18 layer3, carrying a 256-dimensional feature representation. The separable 2D positional embeddings, composed of independent learnable height and width embeddings of dimension 14×256, inject spatial awareness into the token sequence, enabling the model to distinguish patches at different anatomical positions. Across the 4 stacked Transformer encoder layers with 8 attention heads each and a feedforward dimension of 2048, the model progressively refines its inter-token attention patterns, allowing it to assign high attention weights to spatially coherent lesion-like regions rather than responding to isolated intensity artifacts.

However, Fig. 4 illustrates a contrasting failure mode, in which high-intensity activations are concentrated outside the brain boundary, at the skull interface and in regions external to the cranial vault. This dispersed and anatomically implausible pattern suggests that the self-attention mechanism failed to suppress non-informative peripheral signals, likely because the 196 patch tokens include regions corresponding to the skull, background, and image borders, which were not excluded by any brain extraction preprocessing step. This limitation could be addressed in future work through skull-stripping preprocessing or through attention regularization techniques that penalize high weights assigned to anatomically non-informative regions.

4 DISCUSSION

The performance gains observed in the hybrid architecture suggest that the integration of Transformer-based attention mechanisms addresses inherent limitations of purely convolutional approaches for neuroimaging analysis (Umirzakova *et al.*, 2025; Aslam *et al.*, 2022).

While convolutional layers excel at extracting local textural patterns and edge features characteristic of MS lesions, the self-attention mechanism enables the model to establish relationships between spatially distant brain regions, which is particularly relevant given the multifocal nature of MS pathology. Recent systematic reviews have demonstrated that hybrid CNN-ViT architectures effectively mitigate the limitations of each component by combining global context understanding with precise local feature extraction (Azad *et al.*, 2024; Haruna *et al.*, 2025; Kuang *et al.*, 2025). The substantial improvement in precision (89.29% vs. 81.25%) indicates that the model became more conservative in positive predictions, likely due to the Transformer's ability to contextualize local features within the global anatomical structure before classification. This reduction in false positive rate from 20% to 10% is clinically significant, as it directly impacts the positive predictive value in real-world screening scenarios where MS prevalence is relatively low. Meta-analyses of deep learning approaches for MS detection have shown that 2D-3D CNN architectures achieve high diagnostic performance with accuracies ranging from 91% to 98% (Umirzakova *et al.*, 2025), positioning our hybrid model's 86.67% accuracy within a competitive range while demonstrating superior generalization compared to the baseline ResNet18.

The CNN-ViT architecture exhibited accelerated convergence compared to the conventional CNN model. Training loss for the hybrid model decreased from initial values of 0.58–0.77 at epoch 1 to 0.18–0.22 by epoch 40, while validation loss stabilized below 0.45 within the first 10 epochs across most folds. The conventional CNN required approximately 20–25 epochs to achieve comparable validation loss levels, with training loss converging to values between 0.23 and 0.28 after 40 epochs. These findings indicate that the integration of Vision Transformer components with positional embeddings enhances both convergence speed and generalization capability on unseen test data.

The prototypical variant exhibited different training dynamics compared to the cross-entropy models. Training loss values were considerably lower from the first epoch (ranging from 0.02 to 0.18), reflecting the distinct scale of the prototypical loss, which operates on normalized embedding distances rather than raw logits. Validation accuracy across folds ranged from 0.75 to 0.89, with higher fold-to-fold variability (standard deviation of 0.09) than the cross-entropy CNN-ViT (standard deviation of 0.05). This increased variance is consistent with the sensitivity of metric learning to batch composition, since prototypes are estimated from each batch during training and may be unstable when class representation within a batch is uneven, a factor amplified by the limited dataset size. Despite this variability, the prototypical model achieved competitive test set performance (83.33% accuracy), demonstrating that the metric learning formulation can learn discriminative representations even without explicit class-boundary supervision (Haruna *et al.*, 2025). The advantage of this approach lies in its potential for generalization to new classes or unseen imaging protocols without retraining the classification head, a desirable property for future multi-center clinical deployment (Azad *et al.*, 2024).

Tabela 2 – Model comparison

Study	Year	Architecture	Dataset Size	Accuracy (%)	Precision (%)
Our Work (ResNet18 baseline)	2026	CNN (ResNet18)	199 subjects (60 test)	83.33	81.25
Our Work (ResNet18 + ViT)	2026	Hybrid CNN-ViT	199 subjects (60 test)	86.67	89.29
Our Work (ResNet18 + ViT + Proto)	2026	CNN-ViT + Metric Learning	199 subjects (60 test)	83.33	81.25
Daqqaq et al. (Daqqaq; Alhasan; Ghunaim, 2024)	2024	2D-3D CNN (various)	15 studies	91–98	99
Scientific Reports (Khattap *et al.*, 2024)	2024	KNN + Sea-horse Optimizer	76 subjects	97.97	90.76
PMC Study (Ekmekyapar; Taşcı, 2023)	2023	MobileNetV2	Not specified	99.05	98.43

To contextualize our findings within the current state-of-the-art, Table 2 presents a comparative analysis of recent deep learning approaches for MS classification. While our hybrid CNN-ViT architecture achieved 86.67% accuracy with 89.29% precision, other contemporary studies report significantly higher accuracies, with Daqqaq et al. (Daqqaq; Alhasan; Ghunaim, 2024) demonstrating 91–98% accuracy across 15 studies in their meta-analysis, and specialized architectures such as the KNN with Sea-horse Optimizer (Khattap *et al.*, 2024) and MobileNetV2 (Ekmekyapar; Taşcı, 2023) achieving 97.97% and 99.05% accuracy, respectively. However, these elevated performance metrics warrant careful interpretation. The majority of high-performing models were evaluated on homogeneous, single-center datasets without rigorous external validation, potentially leading to optimistic performance estimates that may not generalize to diverse clinical settings. In contrast, our study prioritized methodological rigor through 5-fold stratified cross-validation and evaluation on a completely independent test set, which likely provides a more realistic assessment of real-world performance. Furthermore, the balanced performance across precision (89.29%), recall (83.33%), and specificity (90.00%) is particularly noteworthy when compared to studies that report only accuracy metrics, as this balance is critical for clinical deployment where both false positives and false negatives carry significant consequences for patient management and healthcare resource allocation.

5 CONCLUSION

This study presented and compared three deep learning architectures for automated multiple sclerosis classification from T1-weighted MRI images: a standalone ResNet18 baseline, a hybrid CNN-ViT model trained with cross-entropy loss, and a CNN-ViT variant employing prototypical metric learning. The cross-entropy hybrid achieved 86.67% accuracy, 89.29% precision, 83.33% recall, and 90.00% specificity on an independent test set of 60 subjects, outperforming both the CNN baseline (83.33% accuracy) and the prototypical variant (83.33% accuracy). While the prototypical model did not surpass the cross-entropy hybrid in overall accuracy, its higher recall (86.67%) and its ability to learn compact metric representations without an explicit classification boundary suggest complementary strengths relevant to data-scarce clinical scenarios and potential transfer to unseen imaging protocols.

These results cannot definitively establish any of the models as clinically efficient for deployment in real-world diagnostic settings, but they demonstrate promising performance that validates the potential of hybrid CNN-ViT architectures in neuroimaging applications. The integration of self-attention mechanisms with convolutional feature extraction successfully captured both local lesion characteristics and global anatomical context, as evidenced by the superior generalization capability of the cross-entropy CNN-ViT compared to the standalone ResNet18 baseline. However, the limited dataset size of 199 subjects likely constrained the capacity of all models to learn robust, generalizable representations across the heterogeneous manifestations of MS pathology. The observed performance variability across cross-validation folds

further suggests that a more extensive and diverse training dataset would enable the hybrid architectures to achieve more consistent and potentially superior diagnostic performance. Future research should prioritize the acquisition of larger, multi-center datasets to validate model robustness across different imaging protocols and patient populations. Additionally, extending the current 2D approach to a 3D convolutional-transformer architecture that incorporates multiple MRI sequences, specifically T1-weighted, T2-weighted, and FLAIR images, would enable the model to leverage complementary information for more comprehensive lesion characterization.

REFERENCES

ASLAM, N. *et al.* Multiple Sclerosis Diagnosis Using Machine Learning and Deep Learning: Challenges and Opportunities. **Sensors**, MDPI, v. 22, n. 20, p. 7856, 2022. DOI: 10.3390/s22207856.

AZAD, R. *et al.* Advances in medical image analysis with vision Transformers: A comprehensive review. **Medical Image Analysis**, Elsevier, v. 91, p. 103000, 2024. DOI: 10.1016/j.media.2023.103000.

BELWAL, P.; SINGH, S. Deep Learning techniques to detect and analysis of multiple sclerosis through MRI: A systematic literature review. **Computers in Biology and Medicine**, v. 185, p. 109530, 2025. ISSN 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2024.109530.

COMMOWICK, O. *et al.* How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review. **NeuroImage: Clinical**, v. 39, p. 103491, 2023. DOI: 10.1016/j.nicl.2023.103491.

COMMOWICK, O. *et al.* Objective evaluation of multiple sclerosis lesion segmentation using a large data set. **Scientific Reports**, Nature Publishing Group UK London, v. 8, n. 1, p. 13650, 2018. DOI: 10.1038/s41598-018-31911-7.

DAQQAQ, T. S.; ALHASAN, A. S.; GHUNAIM, H. A. Diagnostic effectiveness of deep learning-based MRI in predicting multiple sclerosis: A meta-analysis. **Neurosciences Journal**, v. 29, n. 2, p. 77–89, 2024. DOI: 10.17712/nsj.2024.2.20230103. Disponível em: https://doi.org/10.17712/nsj.2024.2.20230103.

DJOUMESSI, K.; MENSAH, S. O.; BERENS, P. A hybrid fully convolutional CNN-Transformer model for inherently interpretable medical image classification, 2025. eprint: 2504.08481.

EGUIA, H. *et al.* Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review. **Journal of Medical Internet Research**, v. 26, p. e55315, 2024. DOI: 10.2196/55315.

EKMEKYAPAR, T.; TAŞCI, B. Exemplar MobileNetV2-Based Artificial Intelligence for Robust and Accurate Diagnosis of Multiple Sclerosis. **Diagnostics**, v. 13, n. 19, p. 3030, 2023. DOI: 10.3390/diagnostics13193030. Disponível em: https://doi.org/10.3390/diagnostics13193030.

FERGUSON, B. *et al.* Axonal Damage in Acute Multiple Sclerosis Lesions. **Brain**, Oxford University Press, v. 120, n. 3, p. 393–399, 1997. DOI: 10.1093/brain/120.3.393.

FISCONE, C. *et al.* Multiparametric MRI dataset for susceptibility-based radiomic feature extraction and analysis. **Scientific Data**, Nature Publishing Group UK London, v. 11, n. 1, p. 575, 2024.

HAKI, M. *et al.* Review of multiple sclerosis: Epidemiology, etiology, pathophysiology, and treatment. **Medicine**, LWW, v. 103, n. 8, p. e37297, 2024.

HARUNA, Y. *et al.* Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 144, p. 110057, 2025. DOI: 10.1016/j.engappai.2025.110057.

KHATTAP, M. G. *et al.* AI-based model for automatic identification of multiple sclerosis based on enhanced sea-horse optimizer and MRI scans. **Scientific Reports**, v. 14, p. 12104, 2024. DOI: 10.1038/s41598-024-61876-9. Disponível em: https://doi.org/10.1038/s41598-024-61876-9.

KUANG, H. *et al.* LW-CTrans: A lightweight hybrid network of CNN and Transformer for 3D medical image segmentation. **Medical Image Analysis**, Elsevier, v. 102, p. 103545, 2025. DOI: 10.1016/j.media.2025.103545.

MULTIPLE SCLEROSIS INTERNATIONAL FEDERATION. **Atlas of MS (3rd Edition)**. English. 3. ed. London, UK, 2020. Updated 30 September 2020.

OH, J.; BAR-OR, A.; MARRIE, R. A. Multiple Sclerosis: A Review of Current and Emerging Therapeutic Strategies. **Nature Reviews Neurology**, Nature Publishing Group, v. 19, n. 5, p. 305–320, 2023. DOI: 10.1038/s41582-023-00801-6.

ROCCA, M. A. *et al.* Current and future role of MRI in the diagnosis and prognosis of multiple sclerosis. **The Lancet Regional Health–Europe**, Elsevier, v. 44, 2024.

UMIRZAKOVA, S. *et al.* Deep learning for multiple sclerosis lesion classification and stratification using MRI. **Computers in Biology and Medicine**, Elsevier, v. 192, p. 110078, 2025. DOI: 10.1016/j.compbiomed.2025.110078.