Explainable Dual-Feature Knowledge Distillation for Efficient Autism Spectrum Disorder Classification Using fMRI Data

Fatima Ez-Zahraa Bazay[1] FATIMAEZZAHRAA.BAZAY@UM5R.AC.MA
Houda El Mohamadi[1,2] HOUDA.ELMOHAMADI@UM5R.AC.MA
Mohammed El Hassouni[3] MOHAMED.ELHASSOUNI@GMAIL.COM
Rachid Jennane[2] RACHID.JENNANE@UNIV-ORLEANS.FR
Ahmed Drissi El Maliani[1] A.ELMALIANI@UM5R.AC.MA
[1] *LRIT, Faculty of Sciences in Rabat, Mohammed V University in Rabat, Rabat, Morocco*
[2] *University of Orleans, Institut Denis Poisson- UMR CNRS 7013, Orleans, 45067, France*
[3] *FLSH, Mohammed V University in Rabat, Morocco*

Editors: Under Review for MIDL 2026

Abstract

Deep learning models for Autism Spectrum Disorder (ASD) classification from functional Magnetic Resonance Imaging (fMRI) data face two critical barriers to clinical deployment: high computational costs and lack of interpretability. We propose Dual-Feature Knowledge Distillation (DFKD), a framework that transfers both predictive accuracy and explainability from large teacher models to compact student networks. DFKD leverages Dual-Feature Saliency Extraction (DFS-Ex) to capture complementary texture and shape features from fMRI-derived glass brain visualizations, guiding student training through spatial attention alignment. Evaluation on ABIDE across eight teacher-student pairs demonstrates that DFKD achieves up to 97.95% accuracy with 8.6× compression, notably improving ResNet101-GhostNet from 93.46% baseline to 97.95%, consistently outperforming conventional distillation methods (Kullback-Leibler divergence Knowledge Distillation (KL-Div), Intermediate Knowledge Distillation (I-KD)). Grad-CAM visualizations confirm DFKD-trained students inherit interpretable attention patterns, focusing on diagnostically relevant brain regions. Our approach enables deployment of efficient, transparent models in resource-constrained clinical environments.

Keywords: Knowledge Distillation (KD), Explainability, Saliency maps, Texture Shape features, fMRI data, ASD

1. Introduction

Autism Spectrum Disorder (ASD) affects approximately 1 in 59 children worldwide (Baio et al., 2018). Early diagnosis is crucial, yet current assessment relies on subjective behavioral evaluations. Functional MRI reveals altered connectivity in ASD, and deep learning shows promise for automated classification (Heinsfeld et al., 2018). However, models face two barriers: (1) high computational demands impractical for clinical environments, and (2) lack of interpretability undermining trust.

To address these challenges, knowledge distillation (KD) (Hinton et al., 2015) has emerged as an effective model compression strategy, transferring knowledge from large teachers to compact students. Recent work includes intermediate distillation (Song et al., 2024) and contrastive distillation (Tian et al., 2020). However, conventional KD addresses

only computational efficiency while neglecting explainability. Prior research documents explainability methods (EXMs), with saliency maps (Selvaraju et al., 2017) visualizing decision-influential regions. While recent works (Sun et al., 2025; El Mohamadi et al., 2025) integrate EXMs into KD, existing approaches use unimodal maps inadequate for capturing complementary textural and structural Functional magnetic resonance imaging (fMRI) features.

Despite its success in model compression, knowledge distillation faces fundamental limitations that hinder its adoption in clinical medical imaging. First, capacity gap challenges (Mirzadeh et al., 2020): when the teacher-student capacity difference is large (e.g., $>10\times$ compression), naive distillation often fails as compact students lack sufficient representational capacity to absorb complex teacher knowledge (Gou et al., 2021). This is particularly problematic in medical imaging where large models (ResNets, transformers) achieve state-of-the-art performance but deployment requires ultra-lightweight architectures (MobileNets, ShuffleNets) for resource-constrained clinical workstations. Second, cross-architecture transfer difficulties: transferring knowledge between heterogeneous architectures (e.g., transformers to CNNs) encounters fundamental incompatibilities in feature representations. Transformers process global context via self-attention over patch embeddings while CNNs extract hierarchical local features through convolutions (Touvron et al., 2021). Third, and most critically for clinical deployment, the black-box nature of distilled models: conventional KD methods transfer predictive capabilities through soft targets or intermediate features but provide no mechanism to ensure that students inherit interpretable decision strategies from teachers (Koh et al., 2020). This results in compact models that may achieve high accuracy yet focus on spurious correlations or irrelevant image regions, undermining clinical trustworthiness. Recent attempts to integrate explainability into KD (Zagoruyko and Komodakis, 2017; Park et al., 2019; Heo et al., 2019) rely on attention transfer or relational knowledge but operate at abstract feature levels without grounding in human-interpretable spatial saliency. Moreover, existing explainability-aware distillation methods (Sun et al., 2025) employ unimodal saliency representations that fail to capture the hierarchical, multi-scale nature of medical image features. These include texture patterns encoding fine-grained anatomical details and shape structures representing global connectivity patterns. These limitations create a critical gap: clinical deployment demands models that are simultaneously efficient (lightweight for real-time inference), accurate (maintaining diagnostic performance), and interpretable (providing evidence-based explanations), yet no existing KD framework addresses all three requirements.

Building on these foundations, we propose Dual-Feature Knowledge Distillation (DFKD), a novel framework that synergizes dual-feature saliency maps with knowledge transfer. Our method incorporates the Dual-Feature Saliency Extraction (DFS-Ex) mechanism, which fuses early-layer texture features and late-layer shape features from glass brain visualizations to provide richer feature representations. Unlike traditional KD that transfers only logits (Hinton et al., 2015) or intermediate features (Romero et al., 2015), we explicitly incorporate explainability as an additional knowledge stream through: (i) dual saliency map extraction from teacher and student networks at complementary depths, capturing both fine-grained texture patterns from early layers and coarse-grained shape structures from late layers; and (ii) explainability-aligned loss functions enforcing spatial consistency between saliency maps, ensuring that students attend to diagnostically relevant brain regions such

as the posterior cingulate cortex, medial prefrontal cortex, and temporo-parietal junction. These are key nodes of networks implicated in ASD pathophysiology (Di Martino et al., 2014). This dual-level approach addresses the critical gap in existing explainability-driven KD methods that rely on unimodal saliency representations (Sun et al., 2025), failing to capture the hierarchical and complementary nature of features learned by deep networks (Zeiler and Fergus, 2014). We validate DFKD on the ABIDE dataset (ABIDE Consortium, 2014) across eight carefully selected teacher-student architecture pairs spanning convolutional and transformer-based models. Our contributions include:

- A dual-feature saliency extraction mechanism (DFS-Ex) tailored for fMRI-derived glass brain images, capturing complementary texture and shape features.
- An explainability-driven distillation loss that enforces spatial consistency between teacher and student saliency maps, ensuring interpretability preservation.
- Comprehensive evaluation on eight teacher-student pairs, achieving up to 97.9% accuracy with 8.6× model compression while maintaining diagnostic explainability.

The remainder of this paper is organized as follows: Section 2 details the proposed DFKD methodology. Section 3 describes experimental settings including dataset and training protocols. Section 4 presents comparative results and discusses explainability analysis. Section 5 concludes with implications and future directions.

2. Proposed Approach

Figure 1 illustrates the DFKD framework comprising three stages: (1) independent teacher-student training, (2) dual-feature saliency extraction via DFS-Ex, and (3) explainability-driven distillation enforcing attention alignment.

2.1. Baseline Model Training

We begin by training teacher and student architectures independently on fMRI-derived glass brain visualizations generated from resting-state functional connectivity data. We select eight teacher-student pairs combining large-capacity teachers with compact students, spanning both convolutional and transformer architectures. All models are initialized from ImageNet weights and adapted to the neuroimaging domain through supervised learning.

For binary ASD vs. Typical Controls (TC) classification, we optimize models using standard binary cross-entropy. Given an input image $\mathbf{x}_i$ and its corresponding label $\ell_i \in \{0, 1\}$, the optimization objective is:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\ell_i \log \hat{p}_i + (1 - \ell_i) \log(1 - \hat{p}_i)] \tag{1}$$

where $\hat{p}_i = \sigma(\mathbf{z}_i; \Theta)$ represents the predicted probability under parameters Θ, $\sigma(\cdot)$ is the sigmoid activation, and $\mathcal{B}$ denotes the training batch.

This phase validates teacher accuracy and quantifies the teacher-student gap, enabling measurement of distillation effectiveness.

2.2. Dual-Feature Saliency Extraction

Explainability integration requires capturing interpretable features at multiple abstraction levels. We propose DFS-Ex, a dual-stream saliency extraction mechanism that operates on early and late network layers to capture complementary visual cues. inspired by texture-shape explainability methods (El Mohamadi and El Hassouni, 2023). Early layers encode localized texture patterns (fine-grained activations corresponding to specific brain regions) while late layers encode global shape structures that represent large-scale connectivity patterns. This hierarchical decomposition aligns with the known architecture of deep networks (Zhou et al., 2016), where representational complexity increases with depth.

For convolutional architectures, we define two extraction points: the output of the first residual block (texture layer ℓ_{early}) and the penultimate feature map before the classification head (shape layer ℓ_{late}). Consider a feature tensor $\mathbf{H}^{(\ell)} \in \mathbb{R}^{N\times D\times W\times H}$ at layer ℓ, where N is batch size, D is feature dimensionality, and $W \times H$ are spatial dimensions. We compute the saliency mask via:

$$\mathbf{R}^{(\ell)}_{n,w,h} = \begin{cases} 1, & \text{if } \max_d \mathbf{H}^{(\ell)}_{n,d,w,h} > \mu^{(\ell)} + \kappa \cdot \sigma^{(\ell)} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mu^{(\ell)}$ and $\sigma^{(\ell)}$ are the mean and standard deviation of activations across the spatial-channel dimensions, and κ controls sensitivity (adjusted based on architecture characteristics).

For transformer architectures, spatial importance is derived from attention mechanisms rather than convolutional activations. We extract attention maps from multiple transformer blocks and aggregate them:

$$\mathbf{W}_{\text{agg}} = \frac{1}{|\mathcal{L}_{\text{attn}}|} \sum_{\ell\in\mathcal{L}_{\text{attn}}} \mathbf{W}^{(\ell)} \tag{3}$$

where $\mathbf{W}^{(\ell)}$ represents the attention weights at block ℓ, and $\mathcal{L}_{\text{attn}}$ is the set of selected attention layers.

The final saliency representation combines texture and shape information through element-wise weighted summation:

$$\mathbf{G} = \eta \cdot \mathbf{R}^{(\ell_{\text{early}})} + (1-\eta) \cdot \mathbf{R}^{(\ell_{\text{late}})} \tag{4}$$

where η balances the contribution of texture and shape features. This fused map $\mathbf{G}$ encodes spatially localized importance, serving as the explainability knowledge to be transferred during distillation.

2.3. Explainability-Driven Knowledge Distillation

The distillation phase leverages the extracted saliency maps $\mathbf{G}_{\text{T}}$ and $\mathbf{G}_{\text{S}}$ (where subscripts T and S denote teacher and student, respectively) to guide knowledge transfer. By aligning these attention distributions, we ensure that compact student models focus on diagnostically relevant brain regions, such as the default mode network and salience network, mirroring the teacher's interpretable decision patterns.

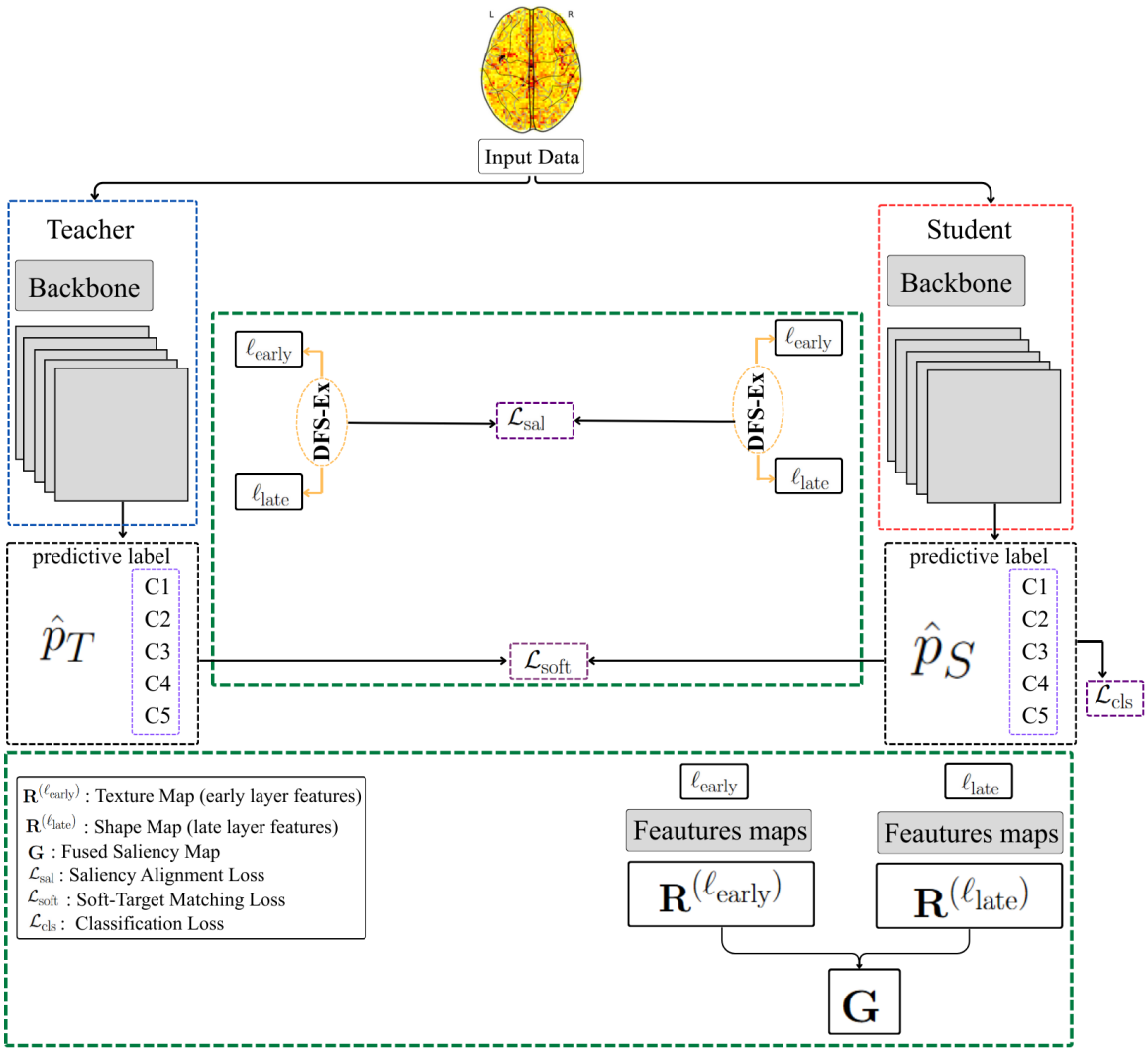

Figure 1: Overview of the DFKD framework. Teacher and student models process glass brain images, extracting dual-feature saliency maps (DFS-Ex) from early (texture) and late (shape) layers. The fused saliency maps guide knowledge distillation through spatial alignment ($\mathcal{L}_{\text{sal}}$), soft-target matching ($\mathcal{L}_{\text{soft}}$), and classification ($\mathcal{L}_{\text{cls}}$).

Our training objective integrates three loss components with complementary roles.

Soft Label Matching. Following Hinton et al. (Hinton et al., 2015), we transfer dark knowledge by matching temperature-scaled distributions:

$$\mathcal{L}_{\text{soft}} = T^2 \cdot \mathbb{D}_{\text{KL}} \left[\text{softmax}(\mathbf{z}_{\text{S}}/T) \parallel \text{softmax}(\mathbf{z}_{\text{T}}/T)\right] \tag{5}$$

where $\mathbf{z}_{\mathrm{T}}$ and $\mathbf{z}_{\mathrm{S}}$ denote teacher and student logits, T is the temperature hyperparameter, and $\mathbb{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence. Temperature scaling softens the probability distribution, revealing subtle inter-class relationships.

Saliency Alignment. The core contribution of DFKD is enforcing spatial alignment between teacher and student saliency maps. We minimize the ℓ_2 distance between their attention distributions:

$$\mathcal{L}_{\mathrm{sal}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \|\mathbf{G}_{\mathrm{T}}(\mathbf{x}_i) - \mathbf{G}_{\mathrm{S}}(\mathbf{x}_i)\|_2^2 \tag{6}$$

This loss guides the student network to attend to the same brain regions as the teacher, preserving interpretability while transferring predictive knowledge.

Unified Objective. The complete DFKD loss combines saliency alignment, soft label matching, and hard label classification:

$$\mathcal{L} = \alpha \mathcal{L}_{\mathrm{sal}} + \beta \mathcal{L}_{\mathrm{soft}} + (1 - \alpha - \beta)\mathcal{L}_{\mathrm{cls}} \tag{7}$$

where α and β are weighting coefficients that balance the three objectives. By simultaneously optimizing all three components, DFKD produces compact student models that maintain both diagnostic accuracy and interpretable attention patterns inherited from the teacher.

3. Experimental Settings

3.1. Dataset and Preprocessing

We evaluate DFKD on the Autism Brain Imaging Data Exchange (ABIDE) (ABIDE Consortium, 2014), the largest publicly available multi-site repository of resting-state fMRI data for ASD research, comprising 1,112 subjects across 17 international sites. To ensure data quality and consistency, we focus on the New York University (NYU) Langone Medical Center site, which provides 174 subjects (75 ASD, 99 TC) with standardized acquisition protocols. This site-specific evaluation controls for scanner variability and acquisition parameter differences that can introduce confounds in multi-site analyses (Di Martino et al., 2014).

All scans undergo standardized preprocessing using the Configurable Pipeline for the Analysis of Connectomes (CPAC) (Craddock et al., 2013), which implements the following pipeline: (1) skull stripping to remove non-brain tissue; (2) slice timing correction to account for interleaved acquisition; (3) motion correction via rigid-body realignment; (4) spatial normalization to MNI152 standard space using nonlinear registration; (5) nuisance signal regression to remove physiological confounds (white matter, cerebrospinal fluid, motion parameters); and (6) temporal band-pass filtering (0.01–0.1 Hz) to isolate low-frequency fluctuations characteristic of resting-state networks.

Following preprocessing, we apply a single-volume image generator (Ahmed et al., 2020) to extract representative axial slices from the 4D fMRI volumes. These slices are then converted to glass brain visualizations via Nilearn (Abraham et al., 2014) and resized to 224×224 pixels for model input. The dataset is partitioned into training, validation, and testing sets in a 70:15:15 ratio with stratification.

3.2. Implementation Details

We investigate 8 teacher-student pairs: DeiT-Base, Wide ResNet-50-2, ConvNeXt-Tiny, ResNet101, ResNet152 as teachers with ShuffleNetV2 (0.5x/1.0x), MobileNetV3-Small, RegNetY-400MF, GhostNet as students (Table 1). DFKD uses $\alpha = 0.4$, $\beta = 0.3$, $T = 6$. We compare against KL-divergence KD (Hinton et al., 2015), I-KD (Song et al., 2024), and student-only training.

Evaluation uses accuracy, loss, and F1-score on the test set. Experiments run on Linux with NVIDIA V100 GPU (32GB VRAM) and 84GB RAM using PyTorch 2.5.

4. Results and Discussion

Table 1 presents a comprehensive comparison of DFKD against baseline approaches across 8 teacher-student pairs, evaluated on accuracy, cross-entropy loss, and F1-score.

Table 1: Performance comparison of DFKD against baseline and traditional KD methods.

Models		Params (M)		T Performance			S Baseline			S KL-Div			S I-KD			S DFKD (Ours)		
Teacher	**Student**	**T**	**S**	Acc	Loss	F1	Acc	Loss	F1	Acc	Loss	F1	Acc	Loss	F1	Acc	Loss	F1
DeiT-Base	ShuffleNetV2-0.5x	86M	1.4M	0.9693	0.1148	0.9693	0.9295	0.1975	0.9292	**0.9415**	**0.1758**	**0.9415**	0.9115	0.3651	0.9117	0.9217	0.2126	0.9217
Wide ResNet-50-2	MobileNetV3-Small	68.9M	2.5M	0.9762	0.0990	0.9762	0.9404	0.1734	0.9406	0.8474	0.3478	0.8481	0.8705	0.3005	0.8698	**0.9534**	**0.1571**	**0.9534**
DeiT-Base	ShuffleNetV2-1.0x	86M	2.3M	0.9693	0.1148	0.9693	0.9564	0.2376	0.9564	0.9495	**0.1530**	0.9494	0.9549	**0.2248**	0.9550	**0.9607**	**0.1313**	**0.9607**
ConvNeXt-Tiny	MobileNetV3-Small	28.1M	2.5M	0.9709	0.1628	0.9709	0.9404	0.1734	0.9406	**0.9411**	0.1916	**0.9409**	0.8818	0.2777	0.8813	**0.9736**	**0.1296**	**0.9736**
ResNet101	GhostNet	44.5M	5.2M	0.9790	0.0964	0.9790	0.9346	0.2331	0.9345	**0.9746**	**0.0983**	**0.9746**	**0.9604**	**0.1287**	**0.9604**	**0.9795**	**0.0924**	**0.9795**
ResNet152	ShuffleNetV2-0.5x	60.2M	1.4M	0.9744	0.1046	0.9745	0.9295	0.1975	0.9292	0.9265	**0.1969**	0.9266	0.9090	0.3717	0.9089	0.9181	0.2216	0.9180
ResNet101	ShuffleNetV2-1.0x	44.5M	2.3M	0.9790	0.0964	0.9790	0.9564	0.2376	0.9564	0.9526	**0.1474**	0.9524	0.9440	0.2590	0.9440	**0.9587**	**0.1363**	**0.9587**
ResNet152	RegNetY-400MF	60.2M	4.3M	0.9744	0.1046	0.9745	0.9612	0.1776	0.9612	0.9305	0.1911	0.9300	**0.9899**	**0.0351**	**0.9899**	**0.9780**	**0.1075**	**0.9780**

Teacher models generally achieve higher accuracy (96.93%–97.90%) compared to student baselines (92.95%–96.12%), reflecting the expected performance gap between large-capacity and compact architectures. This initial evaluation establishes a reference point to quantify the effectiveness of knowledge transfer.

We compare DFKD against two established distillation methods: KL-Divergence KD (Hinton et al., 2015), which transfers knowledge through temperature-scaled soft target probabilities, and Intermediate KD (I-KD) (Song et al., 2024), which additionally distills intermediate feature representations. As shown in Table 1, DFKD consistently outperforms both baselines across the majority of configurations.

Baseline Knowledge Distillation Results. KL-Div shows mixed results: improvements in some configurations (e.g., DeiT→ShuffleNetV2-0.5x: 94.15% vs. 92.95% baseline) but severe degradation in others (Wide ResNet→MobileNetV3: 84.74% vs. 94.04% baseline). This instability suggests that soft-target matching alone is insufficient when large capacity gaps exist between teacher and student, particularly in cross-family transfers where architectural mismatches hinder direct probability alignment. I-KD demonstrates competitive performance on larger students (ResNet152→RegNetY: 98.99%) but degrades significantly on highly compressed configurations (ResNet152→ShuffleNetV2-0.5x: 90.90%), indicating that naive intermediate feature alignment can overwhelm small-capacity networks with excessive representational constraints.

DFKD Performance Gains. DFKD achieves the most notable improvement on ResNet101→GhostNet (97.95% vs. 93.46% baseline, +4.49% absolute gain), demonstrating effective knowledge transfer even with 8.6× compression. Similarly strong gains appear in

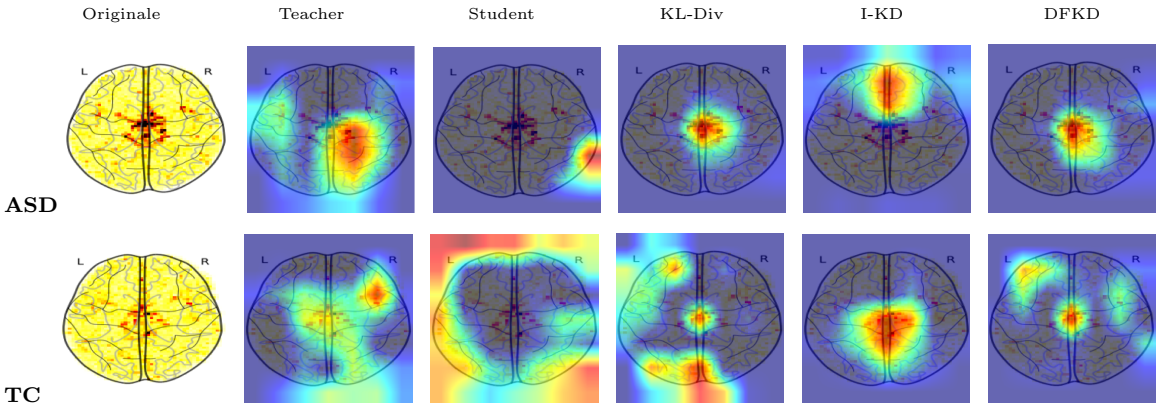

Figure 2: Grad-CAM visualizations for ResNet101→GhostNet. DFKD aligns student attention with teacher patterns.

Wide ResNet→MobileNetV3 (+5.30%) and ConvNeXt→MobileNetV3 (+3.32%), validating DFKD's effectiveness across diverse teacher-student combinations. Even on configurations where baselines fail (Wide ResNet→MobileNetV3 KL-Div: 84.74%), DFKD recovers near-teacher performance (95.34%), highlighting its robustness to capacity mismatches and architectural heterogeneity.

DFKD achieves stable improvements by combining $\mathcal{L}_{\text{soft}}$, $\mathcal{L}_{\text{sal}}$, and $\mathcal{L}_{\text{cls}}$. Figure 2 confirms DFKD students align attention with teachers, focusing on diagnostically relevant regions. Large students (GhostNet, RegNetY) achieve near-teacher performance (97.95%), while compressed students show modest but stable gains. Cross-architecture transfer (DeiT→ShuffleNetV2) demonstrates DFKD bridges gaps between attention-based and convolutional models.

DFKD's generalizability enables lightweight models (1.4M–5.2M params) to achieve teacher-level performance with interpretable attention for clinical deployment.

5. Conclusion

We proposed DFKD, an explainability-driven knowledge distillation framework for fMRI-based autism diagnosis. By transferring dual-feature saliency maps alongside predictive knowledge, DFKD enables compact student models to achieve up to 97.95% accuracy with 8.6× compression and up to 4.8% absolute improvement over student baselines, while maintaining interpretable attention patterns. Evaluation on ABIDE demonstrates consistent improvements over conventional distillation methods, with Grad-CAM visualizations confirming alignment between student and teacher attention on diagnostically critical brain regions. Our approach enables deployment of efficient, transparent models in resource-constrained clinical environments.

References

ABIDE Consortium. Autism brain imaging data exchange. https://fcon_1000.projects.

nitrc.org/indi/abide/abide_I.html, 2014. Last accessed 2024/11/17.

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, et al. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14, 2014.

Md Rishad Ahmed, Yuan Zhang, Yi Liu, and Hongen Liao. Single volume image generator and deep learning-based asd classification. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3044–3054, 2020.

Jon Baio et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2014. *MMWR Surveillance Summaries*, 67(6):1–23, 2018.

Cameron Craddock, Yassine Benhajali, Carlton Carlton, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 7:27, 2013.

Adriano Di Martino, Chao-Gan Yan, Qingyang Li, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.

Houda El Mohamadi and Mohammed El Hassouni. Enhanced deep learning explainability for COVID-19 diagnosis from chest x-ray images by fusing texture and shape features. In *International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–6, 2023.

Houda El Mohamadi, Mohammed El Hassouni, and Rachid Jennane. Dual-domain explainability-driven data augmentation for enhanced covid-19 detection in chest x-rays. *Multimedia Tools and Applications*, pages 1–22, 2025.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Anibal S Heinsfeld, Alexandre R Franco, R Cameron Craddock, et al. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2018.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *AAAI*, volume 33, pages 3771–3778, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, et al. Concept bottleneck models. In *ICML*, pages 5338–5348, 2020.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, et al. Improved knowledge distillation via teacher assistant. *AAAI*, 34:5191–5198, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

Yucheng Song, Jincan Wang, Yifan Ge, Lifeng Li, Jia Guo, Quanxing Dong, and Zhifang Liao. Medical image classification: Knowledge transfer via residual u-net and vision transformer-based teacher-student model with knowledge distillation. *Journal of Visual Communication and Image Representation*, 102:104212, 2024.

Tianli Sun, Wei Lu, Xiongkuo He, et al. Explainability-based knowledge distillation. *Pattern Recognition*, 159:111095, 2025.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving cnns via attention transfer. In *ICLR*, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.