Unmasking Dementia Detection by Masking Input Gradients: A JSM Approach to Model Interpretability and Precision

Yasmine Mustafa and Tie Luo
Computer Science Department, Missouri University of Science and Technology
{yam64, tluo}@mst.edu

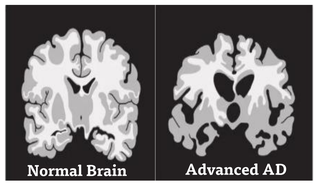
> Motivation

➢ Can we trust a machine for diagnosis?

Artificial intelligence (AI) has made significant strides in recent years but applying it in medicine comes with unique challenges. Despite high accuracy, AI systems in healthcare are often met with skepticism due to concerns about explainability and reliability.

These issues are critical because medical decisions have profound impacts on patient lives. Our study addresses this by focusing on trustworthy medical AI, emphasizing two key aspects:

- **Explainability:** AI diagnoses must be clear and understandable. If patients and doctors can't see why an AI made a decision, they are less likely to trust it.
- **Reliability:** AI systems must consistently base predictions on valid patterns rather than irrelevant factors or biases.

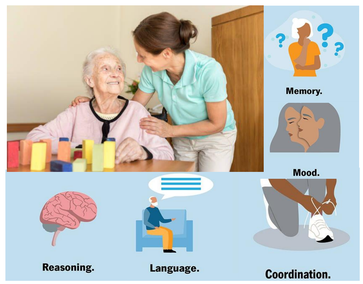


We propose a new approach that enhances trust in AI for diagnosing Alzheimer's disease (AD). AD is an irreversible neurodegenerative disease that impact the lives of those who have it by affecting not only memory, but also mood, coordination, reasoning and language. It's challenging for doctors to detect early patterns of AD until dementia has already occurred.

We **support early intervention** by using neural networks to detect early patterns of AD. Challenges of AD diagnosis involves:

- Fusion of different types of data like MRI, CT, etc.
- Guiding neural networks to areas with dementia

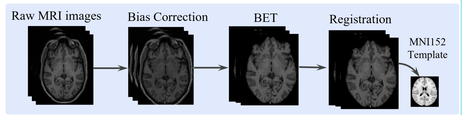
Objective: Our aim is to use Jacobian-Augmented Loss (JAL) to guide neural networks during training, ensuring they learn from the right cues while avoiding those that could lead to errors.

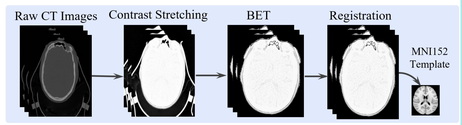


> Methodology

1 Preprocessing

- To minimize spatially varying intensity bias we apply bias field correction to MRI images and contrast stretching for CT images.
- We eliminate non-brain areas using the brain extraction tool (BET).
- Finally, we register MRI and CT scans to the MNI152 template, a standardized brain template used in neuroscience for spatial normalization and comparison across subjects.



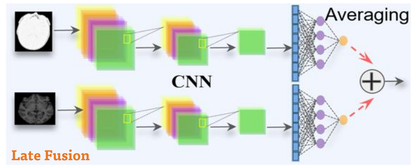


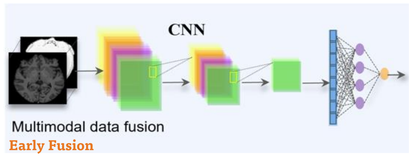
2 Data and Fusion Approaches

In view of the multimodal nature of our study, we incorporated two data fusion techniques:

1)**Late Fusion:** late fusion adopts a dual-branch structure, treating each modality independently and subsequently aggregating their predictions through an averaging mechanism.

2)**Early Fusion:** early fusion involves concatenating the input images as well as their corresponding JSM maps, which allows the model to glean correlations between the two modalities and concurrently debug predictions for the input holistically.





3 Classifier Convolution Neural Networks (CNN)

Contains two convolutional layers coupled with batch normalization, ReLU activation, dropout, and max pooling operations, which capture spatial hierarchies and correlation patterns in the input data. The convolutional layers use a kernel size of 3x3x3, stride and padding of 1, with a dropout of rate 0.2 for regularization. The max-pooling operation reduces spatial dimensions to 2x2x2.

4 Model Debugging using Jacobian Maps

Jacobian maps, or Jacobian Saliency Maps (JSM) are unique in that they convey information about the volumetric changes found in the brain in comparison with a healthy brain template (in our Case MNI152).

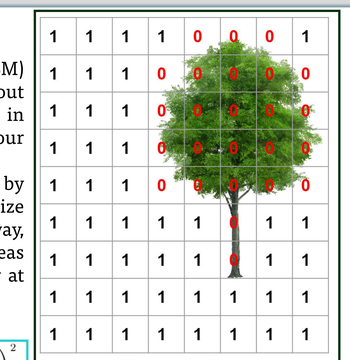
We use JSM to debug the CNN training by incorporating it in the loss function. We penalize the CNN at areas with low importance. This way, after training, the gradients will be high at areas with volumetric changes in the brain and low at areas with no volumetric change.

Jacobian-Augmented Loss Function (JAL)

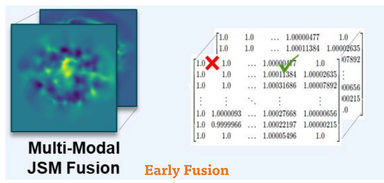
$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = -\underbrace{\sum_{k=1}^{K} y_k \log(\hat{y}_k)}_{\text{Normal loss term}} + \lambda \underbrace{\sum_{d=1}^{D}\sum_{p=1}^{P}\left(w_{dp}\frac{\partial}{\partial x_{dp}}\log\left(\sum_{k=1}^{K} e^{\hat{y}_k}\right)\right)^2}_{\text{Debugging Term}}$$

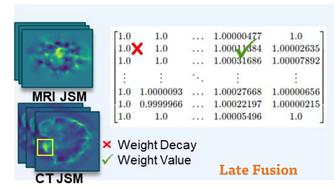
$$w_{dp} = \begin{cases} \text{feature\_weight}, & \text{if } JSM_{dp} \neq 1 \\ \text{debug\_weight}, & \text{otherwise} \end{cases}$$

Gradients with respect to the input are multiplied by a weight matrix that gives a feature weight of 0 to important areas and a debug weight of 1 for less important areas



Simpler Example: If we want the model to detect trees, we set 1 for irrelevant areas to denote higher penalty and set 0 for relevant areas to denote low penalty.





> Results

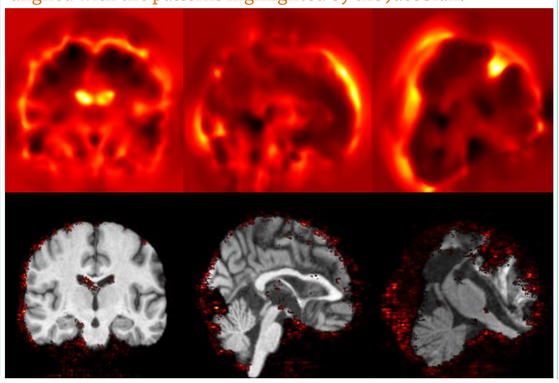
Fusion	Loss Function	Sensitivity				Specificity				Accuracy			
		CN	MCI	MLD	SEV	CN	MCI	MLD	SEV	CN	MCI	MLD	SEV
Early	w/o JAL	80	79	74	81	84	90	84	89	80	84	80	89
	w/ JAL	**90**	**95**	**94**	**92**	**99**	**92**	**99**	**90**	**99**	**87**	**87**	**93**
Late	w/o JAL	88	88	86	88	88	87	87	85	89	86	88	87
	w/ JAL	**95**	**93**	**93**	**92**	**100**	**92**	**92**	**92**	**100**	**92**	**93**	**97**

1 Classification Performance (Ablation Study)

We classify AD into four of its stages: cognitively normal (CN), mild cognitive impairment (MCI), mild dementia (MLD), and severe dementia (SEV). **The table above** compares the results with and without using JAL. It demonstrates how using JAL improves the performance of the classifier for all classes in terms of accuracy, sensitivity, and specificity.

2 Explainability and Reliability

We examined the similarity between the volumetric changes characterized by Jacobian and the decision-making process of the neural network. By plotting gradients overlaid on their corresponding input images, we observed that they closely aligned with the patterns highlighted by the Jacobian.



3 Comparison with the Literature

Paper	No. of Classes	Sensitivity	Specificity	Accuracy
Salami et al 2022	2	86	85	88
Massalimova et al 2021	3	96	96	96
Lazli et al 2019	2	92	92	91
Basheer et al 2021	2	82	Not reported	92
Castellano et al 2021	2	Not reported	Not reported	80
Our Work	4	93	95	91
Our Work	4	94	94	95

> Conclusion

Jacobian-Augmented Loss Function (JAL)

- Rectifies erroneous predictions and identifies important regions
- Provides substantial accuracy improvement (by up to 10%) and greater model interpretability in identifying significant brain areas that lead to diagnostic predictions.
- Works seamlessly with our multimodal data fusion methods.

> References

- **Y. Mustafa** and T. Luo. Unmasking dementia detection by masking input gradients: A JSM approach to model interpretability and precision. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2024.
- Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017, August). Right for the right reasons: training differentiable models by constraining their explanations. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 2662-2670).

Computer Science