Interpreting Kolmogorov-Arnold Networks in Neuroimaging: A Path-Based Attribution Framework

Anonymous authors
Paper under double-blind review

Abstract

Explainability aspects of most classification models are learnt through instance-specific analysis. However, in understanding diseases, it is important to consider population-wide analysis in order to identify affected regions that are consistently seen across cohorts of diseased population. In this study, we report utility of Kolmogorov-Arnold Networks (KANs) in understanding population-wide characteristics seen in subjects affected by Alzheimer's disease (AD). KANs offer enhanced interpretability through learnable activation functions on network edges. Thus, the learned functions reflect the characteristics of the entire span of training data. In a KAN network trained for classification, attributions through the network can be traced to understand how specific inputs influence the output label. In this study, we propose a path-based attribution framework that generates global importance maps by tracing exhaustive information flow through all potential paths. Our method scores edges using L2 norms of the learned spline and base functions. Subsequently, these scores are propagated through the network to compute path-attributions. This approach scales linearly with network depth, and is only dependent on model training and does not need further analysis on data post-hoc. Evaluation on three public AD neuroimaging datasets (OASIS, ADNI, Mendeley, totally comprising 7428 acquisitions), were carried out on 3D brain volumes as well as 2D brain slices. The corresponding KAN test accuracies are 93.24%, 81.85%, and 91.25% on OASIS, ADNI, and Mendeley datasets, respectively. Alongside, improved performance via metrics such as Insertion AUC, Deletion AUC and Sufficiency, is also demonstrated. The generated attribution maps identify clinically meaningful regions including the body and genu of corpus callossum, corona radiata, bilateral caudate nuclei, medial prefrontal cortex and temporal lobe structures, aligned with established AD pathology literature. By providing voxel-level global attributions as network-intrinsic properties, our framework addresses a critical gap in medical AI interpretability and supports clinical validation of AI-assisted AD diagnosis systems.

1 Introduction

The intersection of medical imaging and computational neuroscience has been driven forward by the advent of artificial intelligence and deep learning. Imaging modalities such as magnetic resonance imaging (MRI), such as structural MRIs (sMRI) and functional MRIs (fMRI), among several others, provide spatial and temporal data at a very high resolution. This allows researchers to investigate the anatomy of the brain at fine granularity, assisting in early diagnosis and understanding of disorders Lundervold & Lundervold (2019). Deep learning models achieve high accuracies on brain imaging tasks, yet their opacity limits clinical adoption and regulatory approval in neuroscience Kelly et al. (2019), with model decisions often being inscrutable.

Interpretability of a network is crucial in medical analysis. Unlike consumer domains, medical decisions require transparent reasoning, and clinicians must be able to trace model decisions to anatomically or functionally meaningful biomarkers which can also aid in distinguishing between medically backed model decisions and erroneous correlations which can occur due to noisy or unclean data Wang et al. (2024) Salahuddin et al. (2022). Several attribution methods have been put forth to understand model decisions and generate input attributions. Localized attribution methods such as gradient saliency maps, Layer-wise

Relevance Propagation, Integrated Gradients, and occlusion-based methods, among others, have contributed insights to brain imaging by highlighting which voxels or regions contribute to model predictions. However, these approaches often produce localized, sample-specific explanations, making it difficult to identify which features or regions are consistently important across training data. While such local attributions can indicate potential importance for individual cases, such as tumor detection, they may provide limited clarity into patterns seen across brain populations for neurodegenerative diseases such as Alzheimer's disease (AD) and dementia, among others.

Local attributions reveal how features influence specific cases, but they may not capture population-level patterns that neural networks learn during training. Arriving at potentially thousands of local attributions can be difficult to interpret. Global feature attributions, which show how features influence the decision of a model across the entire training distribution, can differ significantly from averaged local attributions Ibrahim et al. (2019). This distinction is particularly important in medical imaging, for classification problems such as Alzheimer's detection, where clinicians and researchers need to understand not just why a model made a specific prediction, but which anatomical regions the model has learned to consistently rely upon across patient populations Munroe et al. (2024).

Global attributions can provide several advantages over localized attribution in medical imaging contexts. It enables identification of biomarkers that are present across the patient distribution in training data, it also aids model debugging by revealing biases or spurious correlations. By accurately highlighting clinically meaningful features learnt by the model and demonstrating that models rely on these features alone for their decision, regulatory approval and clinical trust is gained Hill (2024). It also bridges the communication gap between the AI and medical world by providing population-level insights that align with medical knowledge. By attributing global features not as an aggregated property of the inputs, but as an intrinsic property of the trained network itself, we generate accurate attribution maps that highlight medically relevant regions of the brain.

Our work utilizes the inherent functional transparency of the recently proposed Kolmogorov-Arnold Networks (KANs). KANs, as originally proposed in Liu et al. (2025), are designed to replace every scalar weight in a multi-layer perceptron (MLP) with a learnable univariate spline function on each edge, as opposed to a node. They leverage the Kolmogorov-Arnold Representation Theorem, which states that "any continuous high-dimensional function can be decomposed into a finite sum of univariate continuous functions", which has made it possible to express the transformation from an input to the output decision as an explicit graph of simple, human interpretable univariate functions on edges that are simply added at each node. Due to the functions on each edge being explicit and univariate, it is possible to analyze the nature of each function, and thereby quantify the influence of the function locally. Specifically, a learned function on an edge consists of a base component and a spline component, expressed as $\phi(x) = w_b b(x) + w_s spline(x)$, where $b(x)$ is typically the SiLU activation function and $spline(x)$ is a linear combination of basis splines. This structure makes it possible to identify the individual contributions of these functions, their interactions with the rest of the network, and their association with each individual input.

In neuroimaging, specifically in understanding the contribution of each individual voxel corresponding to a certain part of the brain, input based attribution can prove to be particularly useful. By tracing the decision of the model back to a particular input/set of inputs, it can be made possible to identify and rank the features influencing the model. This makes it possible to construct faithful, global attribution maps at input-level granularity by analyzing the flow of information through the network, which tell us about the parts of the brain the model has learnt to differentiate between.

In this work, we propose a path-based attribution approach for KANs to generate population-level importance maps as a property of the trained network, by tracing information flow along explicit computational paths, unique to KAN's architecture, thereby scoring input importances based on the information flow. We also evaluate its utility and reliability for neuroimaging and clinical neuroscience applications. We demonstrate that this approach allows for faithful, global attribution maps that provide clinically and anatomically useful information. We list our primary contributions as follows:

i) A path-based attribution framework for KANs that aggregates importance across computational paths, with edge scoring done with L2 norms of the function.

ii) A scalable implementation of path-attribution for high-dimensional neuroimaging data.

iii) Voxel-level global-attribution maps applied through KANs for Alzheimer's detection.

iv) A comprehensive clinical evaluation showing anatomically meaningful regions associated with Alzheimer's disease, supported by ROI analyses.

2 Related Literature

Several attribution methods have been proposed to understand the functioning of neural networks as "black boxes". Gradient-based approaches, Simonyan et al. (2013) calculate output gradients with respect to inputs, and Sundararajan et al. (2017) integrate gradients along a baseline path to address gradient saturation. Propagation-based approaches such as Layerwise Relevance Propagation (LRP) Bach et al. (2015) have been applied to Alzheimer's detection, Böhle et al. (2019) using LRP on ADNI sMRI, producing maps concentrated in temporal lobe and hippocampus. However, the authors note a high inter-patient variability in patterns, which is a limitation in generalizing patterns across the data population. Class Activation Mapping (CAM), particularly Eigen-CAM and Grad-CAM have also been adopted for model decision explanations, but fail to provide faithful explanations, often cover unnecessarily large or imprecise regions, exhibit low granularity in highlighting salient features, and suffer from instability and biases due to architectural factors such as Global Average Pooling and gradient explosion Bae et al. (2020). Game-theoretic approaches such as LIME, DeepLIFT Shrikumar et al. (2017) are also popular, with SHAP being implemented by Qiu et al. (2022) on the NACC dataset and OASIS cohort and found distinct patterns unavailable through averaging local explanations, which were negative hippocampal values for cognitively normal subjects, positive for AD patients. The problem with trying to arrive at a global explanation from these local attribution methods is it conflates intrinsic feature importance and feature distribution in data Covert et al. (2020). A feature may accumulate high average SHAP scores simply because it varies widely in training data, not because the model relies on it for decisions. This theoretical insight explains the empirical observation: across all local attribution methods (LRP, IG, Grad-CAM), importance scores fluctuate dramatically between patients with identical diagnoses, reflecting both model variance and data noise.

To solve this problem, global attribution methods that highlight locations of important features seen throughout training data have been proposed. SAGE (Shapley Additive Global importancE) Covert et al. (2020) directly measures the contribution of features to the model's prediction by introducing features into arbitrary subsets and calculates their Shapley values, but scales exponentially with dimensionality. Chan & Veas (2024) use a gradient-based approach to rank features based on gradient descent updating a feature's weight. Wu et al. (2020)'s framework learns global explanations by identifying which feature detectors are most critical for each class, then testing how well those detectors respond to semantic concepts, with Graziani et al. (2020) extending this to medicine, proposing a framework to explain attribute pixel-level values to user-defined concepts. Global Attribution Mapping Ibrahim et al. (2019) captures heterogenous patterns across subpopulations of local explanations clustered on distance metrics to arrive at a singular global explanation. Although faithful, this relies on local attributions computed on input data, on which the algorithm is applied, which is time-consuming and computationally expensive. As the algorithm also computes the distance matrix for all input samples, the time complexity scales quadratically with up number of input samples, thereby making it unsuitable for large datasets. We aim to introduce an approach that offers voxel-level global attribution maps and considers feature importances as properties learned by the network itself. We implement this framework to neuroscience, or specifically, Alzheimer's detection to generate voxel-level global attribution maps.

2.1 KANs for attribution

With KANs being proposed as interpretable networks, there have been applications in medicine. Dong et al. (2025) used a KAN-based Graph Convolutional Networks (GCNs) for Alzheimer's disease diagnosis on the ADNI dataset, with their GCN-KAN achieved 62.6% over standard GCNs at 57.4%. They score each ROI's importance by summing the absolute spline coefficients from the KAN layer edges connected to that ROI. Although their GCN-KAN identified the hippocampus, inferior parietal gyrus, and amygdala as relevant

regions for AD, the approach does not work on voxel-level granularity. Knottenbelt et al. (2025) proposed CoxKAN, a KAN-based framework for interpretable survival analysis, using KANs with Cox proportional hazards regression. They discovered symbolic formulae for hazard functions, and pruned the network using the L1-norm of the function on each edge of the KAN.

With KANs being proposed as networks that be analyzed easier than typical DNNs, they have a structure where the flow of information throughout the network can be interpreted and input contributions can be quantified. By tracing the decision of the model looking at the flow of information back to a particular input/set of inputs, it can be made possible to identify and rank the features influencing the model. In Liu et al. (2025), the authors utilize L1 norms to score each layer. The individual L1 norm for each activation $\phi_{i,j}$ is calculated, which are then summed to give the score for each layer Φ, which is defined as follows:

$$|\phi_{i,j}|_1 = \frac{1}{N_p} \sum_{s=1}^{N_p} \left| \phi_{i,j} \left(x_j^{(s)} \right) \right|, \quad |\Phi|_1 = \sum_{i=1}^{n_{\text{out}}} \sum_{j=1}^{n_{\text{in}}} |\phi_{i,j}|_1$$

Here, $\phi_{i,j}$ is the learnable univariate function on the edge from input j to output i, $x_j^{(s)}$ denotes the j-th feature value from the s-th data sample, and N_p is the total number of samples.

KAN 2.0 Liu et al. (2024) states that this would not effectively capture the global relationships of the edge with the rest of the network, due to L1 norms capturing just the local information. To solve this, the authors propose a recursive scoring method utilizing standard deviations of activations on the edges and nodes.

In KAN 2.0, they consider an L-layer KAN with width $[n_0, n_1, \cdots, n_L]$, $E_{l,i,j}$ as the standard deviation of activations on the (l, i, j) edge, and $N_{l,i}$ as the standard deviation on the (l, i) node. They score nodes as $A_{l,i}$ and edges as $B_{l,i,j}$ recursively as follows: all output node scores are set to 1; i.e., $A_{L,i} = 1$ for all i. For $l = L, L-1, ..., 1$:

$$B_{l-1,i,j} = A_{l,j} \frac{E_{l,j,i}}{N_{l+1,j}}, \qquad A_{l-1,i} = \sum_{j=0}^{n_l} B_{l-1,i,j}$$

This iterative computation propagates attribution scores from the output layer to the input layer, capturing the activity of each node. We are looking to score each input with respect to the output, i.e., how much each input contributes to the prediction of the model. We are also looking to find which combinations of edges carry an input's influence to the output, so a path-based attribution is required. Hence, an attribution method that propagates edge scores throughout the network to compute the contribution of each input is necessary.

3 Methodology

3.1 Edge-Scoring

KANs have a structure in which each edge comprises both a linear base term and a nonlinearly modulated spline component, with addition or multiplication at the nodes. A meaningful edge-scoring metric must therefore reflect the combined influence of these two components under data activations. For each edge from input node i to output node o, the contribution depends on both the learned base weight and the projection of the input through a spline basis expansion whose coefficients are learned during training.

We quantify edge importance using the L2 magnitude of the combined spline and base activations. This choice reflects the fact that KAN edges are smooth and differentiable functions, and measuring their integrated L2 "energy" captures how strongly and consistently an edge contributes across the input domain. The L2 metric is also smoother and more stable under small perturbations than L1 Rosca et al. (2020), whose sparsity-biased behavior can cause it to ignore sustained contributions that are meaningful in functional models. Overall, the L2 norm provides a more reliable and functionally aligned estimate of edge importance in KANs.

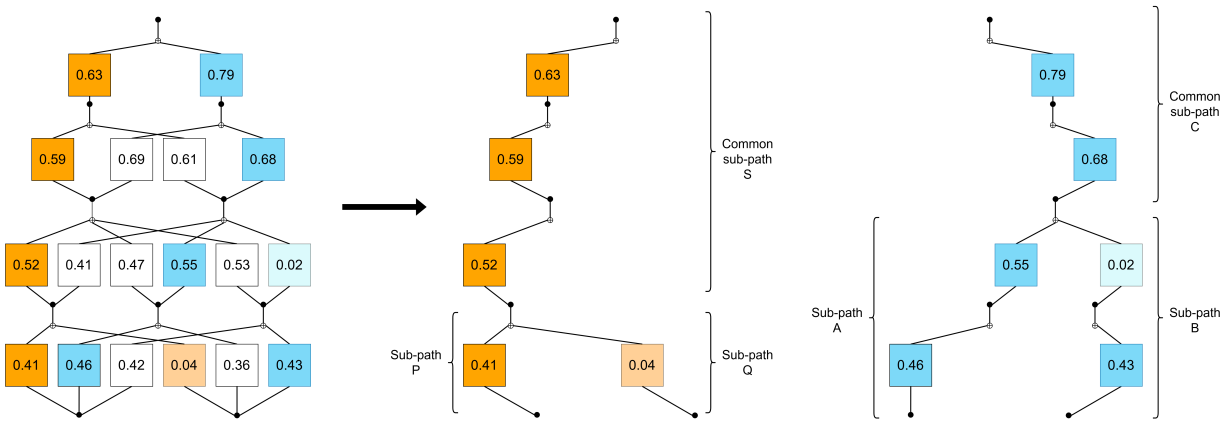


Figure 1: The proposed path factorization avoids double counting in KAN attribution. From left to right: Flow of information is bottom-to-top; A multi-layer KAN showing two different sets of subpaths (orange and blue) to the output node, with example function scores on edges. Decomposition shows how shared subpath S factors identically into both paths, with distinguishing contributions arising only from unique sub-paths A, and B; P and Q. The multiplicative factorization ensures that overlapping edges contribute uniformly without additive redundancy, preserving the relative importance of each input.

Formally, the L2 edge score $E_{i,o}$ is:

$$E_{i,o} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\sum_{c=1}^{C} B_{b,i,c}\, W_{o,i,c}^{\text{spline}} \;+\; W_{o,i}^{\text{base}}\,\sigma(x_{b,i})\right)^2}\,, \tag{1}$$

where $B_{b,i,c}$ is the activation of the c-th spline basis for input i in sample b, $W_{o,i,c}^{\text{spline}}$ is the associated spline weight, $W_{o,i}^{\text{base}}$ is the base weight from i to o, and $\sigma(\cdot)$ is the base activation function (e.g. SiLU). The inner sum accumulates both spline and base contributions for each example, and the outer mean-square followed by the square root computes their L2 magnitude.

This measures each edge's activation energy across data, ensuring that edges with consistently strong influence are ranked highly, even if their contributions arise jointly from spline and base terms. Compared to the L1 scoring used in Liu et al. (2025), this L2 measure avoids underestimating important edges that display a lower variance, penalizes sporadic large activations, and provides a more faithful signal for path-level attribution.

3.2 Path-based Attribution

The formula on Equation 1 captures just the local influence of the function on the edge, and hence, does not account for the contributions of edges downstream in the network. Even if the edge has a strong local influence, it can be canceled out by an edge later on in the network Liu et al. (2025). To account for downstream contribution of edges towards scoring the contribution of the input, we propose a path-based attribution method. To prevent an explosion of the edge-score, we normalize the score with the tanh function. We score the paths as the absolute of the tanh of the product of edge-scores on the path. The input is scored as a sum of all paths from the respective input node to the output node. If there are several output nodes, all the scores from the specific input node to all the output nodes are averaged. Formally, the importance score of an input feature i in an L-layer Kolmogorov-Arnold Network (KAN) with n_O output nodes can be expressed as the average over all output neurons of the sum over all possible paths $\mathcal{P}_{i\to o}$ from input i to output neuron o:

$$\text{imp}_i = \frac{1}{n_O}\sum_{o=1}^{n_O}\sum_{p\in\mathcal{P}_{i\to o}}\prod_{e\in p}\alpha_e,\ \text{where } \alpha_e = \tanh\!\big(E_e\big). \tag{2}$$

Here, the product is taken over all edges e along the path p from $o_0 = i$ (input neuron) to $o_L = o$ (output neuron), where o_l denotes the neuron index at layer l and E_e is the edge-score at edge e. While the proposed path-based attribution framework is agnostic to the choice of local edge scoring function, in this work we adopt the edge scoring metric discussed in Sec. 3.1. This choice retains information that may be lost due to induced sparsity in the standard L1 norm and aligns with recent work in KAN-based edge scoring Liu et al. (2024) that utilizes component-wise activity. We deliberately use absolute magnitudes, rather than directional (signed) magnitudes, as the L2 norm captures the overall energy carried by an edge regardless of sign cancellations, such that strong but oscillatory activations are not underestimated. We are also looking to assign an importance score to each input feature x_i, where absolute contribution strength, rather than directionality, provides a more faithful and interpretable measure.

The non-linear functions on a KAN lies on its respective edge, with simple addition performed at the nodes. This means the influence of an input variable factorizes naturally along directed paths from itself to the output. While KAN computation graphs are not path disjoint, different inputs often traverse overlapping subnetworks due to shared spline edges. Under these architectural assumptions, where each edge multiplies its input by a learned function and each node sums incoming signals, the network output is expressed as a path expansion:

$$y = \sum_i x_i \sum_{P \in \mathcal{P}(i \to o)} \prod_{e \in P} \alpha_e. \tag{3}$$

Consider the illustrative toy-example in Fig. 1. For two paths P and Q that share a common subpath S, the factorization gives:

$$\prod_{e \in P} \alpha_e = \left(\prod_{e \in S} \alpha_e \right) \left(\prod_{e \in P \setminus S} \alpha_e \right), \qquad \prod_{e \in Q} \alpha_e = \left(\prod_{e \in S} \alpha_e \right) \left(\prod_{e \in Q \setminus S} \alpha_e \right) \tag{4}$$

where the shared product $\prod_{e \in S} \alpha_e$ is factored in identically in both paths as a common multiplicative scalar rather than accumulating additively. As a result, the distinguishing contribution of each path arises only from the edges unique to that path, i.e. from $\prod_{e \in P \setminus S} \alpha_e$ versus $\prod_{e \in Q \setminus S} \alpha_e$. Overlaps therefore are factored in uniformly, propagating the same multiplicative weight to every traversing path, avoiding bias due to double counting. The measure is non-negative and bounded, giving a faithful, overlap-neutral attribution that is consistent with the multilayer KAN architectures used in our experiments. The path scoring accounts for downstream effects of the edge on the rest of the network, where any cancellation upstream results in lower-relevance and lower activity and weights downstream. This is similar to relevance propagation in attribution methods such as DeepLIFT Shrikumar et al. (2017) and Layer-wise Relevance Propagation Bach et al. (2015), where upstream inactivity diminishes contributions to later layers. The input score is then converted to percentage-wise contribution to highlight individual contributions of each input node. This level of granularity is useful for domains such as medical imaging, where regional interpretability is desired. Although the path-based formulation in Eqn. (1) captures both local edge influence and downstream interactions, computing these scores by explicitly enumerating all input-output paths is not feasible for deep or moderately wide KANs, since the number of possible paths grows exponentially with network depth. Therefore, an efficient propagation mechanism is required to obtain the same path-aggregated importance values without enumerating paths directly. We introduce such an efficient matrix-based computation in the following section.

3.3 Efficient Path-based Attribution

For a computationally efficient attribution framework, we compute the same input-score using matrix multiplication. Let $\boldsymbol{\alpha}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ be the matrix of edge importances for layer l, where d_0 is the input dimensionality and n_O is the number of output neurons (here, one for binary classification). Then, the propagated importance matrix $\mathbf{M}$ is:

$$\mathbf{M} = \boldsymbol{\alpha}^{(L)} \cdot \boldsymbol{\alpha}^{(L-1)} \cdots \boldsymbol{\alpha}^{(1)} \quad \in \quad \mathbb{R}^{n_O \times d_0}, \tag{5}$$

The importance of the i-th input neuron simplifies to $\text{imp}_i = \frac{1}{n_O} \sum_{o=1}^{n_O} \mathbf{M}_{o,i}$.

This matrix-based calculation scales linearly with the number of edges, $\mathcal{O}\Big(\sum_{l=1}^{L} d_l d_{l-1}\Big)$, which is same as chained matrix multiplication compared to the exponential complexity of the naive path enumeration, $\mathcal{O}\Big(\prod_{l=1}^{L} d_l\Big)$, making it feasible for deeper and wider models with numerous paths.

This makes the scoring approach practical for high-dimensional neuroimaging data. For example, a structural MRI volume of size $91 \times 109 \times 91$ yields $d_0 = 902{,}629$ voxels. For a KAN with layer widths $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$, the efficient matrix-based propagation has complexity $\mathcal{O}\Big(\sum_{l=1}^{L} d_l d_{l-1}\Big) = 256 \cdot 902{,}629 + 128 \cdot 256 + 64 \cdot 128 \approx 2.31 \times 10^8$, which is significantly smaller than naive path enumeration, scaling as $\mathcal{O}\Big(\prod_{l=1}^{L} d_l\Big) = 902{,}629 \cdot 256 \cdot 128 \cdot 64 \approx 1.89 \times 10^{12}$, an increase of more than four orders of magnitude, making exhaustive path computation infeasible. This numerical comparison highlights the necessity of the efficient matrix-based formulation for voxel-level global attribution in high-dimensional neuroimaging settings. Once the inputs have been attributed and scored, the scores are then converted to percentage-wise contribution to highlight individual contributions of each input node, at input-level granularity. This resolution is especially important in brain imaging due to the requirement of a high granularity such that individual brain regions are highlighted and accurate medical inferences can be carried out.

4 Experimental Setup

4.1 Dataset Formulation Details

The datasets utilized in this study include the volumetric public OASIS-1 dataset Marcus et al. (2007) to evaluate the framework's performance on 3D-Volumetric data, the middle sagittal slice from the ADNI dataset Jack Jr et al. (2008), selected to analyze the volume on 2D-slice data, and the Mendeley coronal-slice MRI collection Yakkundi (2023). We choose the middle sagittal slice as it passes through key midline brain structures (e.g., corpus callosum, cingulate gyrus, ventricles), which are highly informative for Alzheimer's pathology, while reducing dimensionality compared to full 3D volumes Odimayo et al. (2024),Hoang et al. (2023). All three datasets comprise sMRI scans of cognitively normal individuals, subjects with mild cognitive impairment (MCI), and patients diagnosed with Alzheimer's disease (AD). These cohorts provide a progression of disease states, and are established benchmarks for Alzheimer's prediction on neural networks. OASIS consists of 436 volumes of shape 91 x 109 x 91. ADNI consists of 592 volumes, from which we extract the sagittal slice and crop it to 184 x 152. The Mendeley data cohort consists of 6400 samples of coronal slices. Class imbalance is handled by the usage of weighted Binary Cross-Entropy loss. Standard preprocessing procedures such as skull-stripping Smith (2002)Smith et al. (2004), intensity normalization Tustison et al. (2010), and affine registration Avants et al. (2009), were applied to ensure anatomical consistency and comparability across scans. We use these datasets with well-defined diagnostic categories to rigorously test how KANs capture structural differences associated with cognitive decline. The path-based attribution framework is applied to generate global attribution maps, highlighting the contribution of specific imaging regions towards the models decision. This is quantitatively assessed using Insertion AUC, Deletion AUC, and Sufficiency metrics. Insertion AUC measures how the model's confidence increases as the most important features (voxels) identified by the attribution are progressively added to a blank input. Deletion AUC assesses how quickly the model's confidence decreases as these important features are removed, indicating the extent to which the model relies on these features for its decision Petsiuk et al. (2018). Sufficiency measures whether the top-ranked features alone are sufficient to maintain the model's predictive performance DeYoung et al. (2020). This allows for an assessment of whether the network identifies clinically meaningful features for differentiating between healthy and diseased individuals.

4.2 Implementation Details

All experiments were conducted on a device running the Nvidia A100 GPU using PyTorch 2.5.1+cu121 as the primary deep learning framework. Supporting data processing and analysis were performed using NumPy 1.26.4, Pandas 2.2.2, and scikit-learn 1.2.2. Neuroimaging data was handled using NiBabel 5.3.2. We use a simple feedforward deep KAN, each with about 6-10 hidden layers and run the attributions on the fully trained model. To assess and compare our methods against other attribution methods that capture

importance across training data, we utilize Insertion and Deletion AUC metrics. Insertion AUC shows the increase in model confidence as the most relevant regions ranked by the attribution map are progressively added to a blank input, while the Deletion AUC measures the decrease in confidence as these same regions are incrementally removed from the input. Additionally, we use the Sufficiency metric to assess whether the top-ranked features alone are adequate to preserve performance DeYoung et al. (2020). As a Proof-of-Concept, we also test the attribution frameworks on MNIST classification.

To interpret model attributions and further assess their stability, we performed region-of-interest (ROI) analyses on the attribution maps generated across 15 independent runs. Two anatomical atlases were used for comprehensive grey and white matter coverage: the Brainnetome (BN) Atlas Fan et al. (2016) and the Johns Hopkins University (JHU) ICBM-DTI White Matter Atlas Mori et al. (2006).

The BN Atlas provides a 246-region parcellation of the grey matter at a 2 mm isotropic resolution. It contains both structural and functional connectivity information to define connectivity-informed cortical and subcortical ROIs. This atlas enables detailed inference across diverse grey matter regions and aligns well with OASIS T1-weighted MRI volumes ($91 \times 109 \times 91$ voxels) after standard MNI registration. ROI-level intensity measures and region-based features were extracted for each subject in native space following co-registration.

The JHU ICBM-DTI Atlas offers a 48-region parcellation of the brain's white matter based on high-quality diffusion tensor imaging acquired in the ICBM-152 template space, also at a 2 mm isotropic resolution. Each region corresponds to a major white matter tract manually delineated using fiber orientation information, ensuring standardized tract identification across subjects. For every OASIS MRI, the JHU atlas was co-registered to native space after MNI transformation to extract regional white matter features. This facilitates tract-based ROI inferences for comparative and cross-sectional evaluation of attribution stability within white matter structures.

Together, the BN and JHU atlases provide complementary insights into grey and white matter attribution patterns, allowing for a comprehensive region-level interpretation of the model's behavior across both tissue domains and insights into regions affected by Alzheimer's.

5 Results and Discussions

We train a KAN on the datasets and achieve test accuracies of 93.24%, 81.85% and 91.25% respectively on the OASIS-1, ADNI and Mendeley test sets, respectively. We also train the KAN on MNIST classification as a PoC and achieve a test accuracy of 99.4%. Using the proposed path-based attribution with L2 edge scoring, we obtain an Insertion AUC of 0.865 ± 0.049, Deletion AUC of 0.282 ± 0.020, and Sufficiency of 0.068 ± 0.064 on the OASIS dataset as indicated in Table 1. On ADNI, the method achieves an Insertion AUC of 0.758 ± 0.044, Deletion AUC of 0.380 ± 0.051, and Sufficiency of 0.209 ± 0.143. For the Mendeley cohort, we report an Insertion AUC of 0.791 ± 0.016, Deletion AUC of 0.463 ± 0.099, and Sufficiency of 0.226 ± 0.033. On MNIST, the results are Insertion AUC 0.964 ± 0.003, Deletion AUC 0.393 ± 0.022, and Sufficiency 0.072 ± 0.015. These values indicate that the L2-scored path attributions consistently rank features that affect model confidence under both insertion and deletion perturbations. We compare against global attribution methods in GAM, KAN and KAN 2.0, we also use our path-attribution method in tandem with the localized edge scoring approach from KAN to prove the efficacy of our method. It is to be noted that, although this framework has been applied for the pathology of Alzheimer's in this study on 3D-volumetric and 2D-slice data, it be applied to other neurological conditions where global, population-level pattern identification across subjects is required.

5.1 Comparative Analysis

Table 1 compares the global attribution and edge scoring approaches from GAM, KAN and KAN 2.0 and two of the best performing edge scoring approaches paired with the path-based attribution frameworks. GAM was implemented with local attribution scoring of Integrated Gradients (IG), which performed better than simple gradients Ibrahim et al. (2019). As the second best edge-scoring method was found to be

the maximum of the base and spline contribution of the KAN edge-function, $E_{i,o}^{\max} = \max_{1 \le b \le B} \left| A_{b,i,o} \right|$, with $A_{b,i,o} = \sum_{c=1}^{C} B_{b,i,c} W_{o,i,c}^{\text{spline}} + W_{o,i}^{\text{base}} \, \sigma(x_{b,i})$, this metric is also included in our comparison and is denoted as *Path Attr. + max*. The proposed path attribution framework, in tandem with the L2 norm as the edge scoring function, yielded the highest Insertion AUC values on all datasets (OASIS: 0.865 ± 0.049, ADNI: 0.758 ± 0.044, Mendeley: 0.791 ± 0.016, MNIST: 0.964 ± 0.003), which indicates that it is the best suited to identify the most informative regions for model predictions. Lower Deletion AUC scores show that removing high-score regions leads to a pronounced decrease in model confidence, supporting the faithfulness of its attributions.

While GAM occasionally achieves slightly lower Sufficiency scores (OASIS), the proposed approach closely matches or surpasses external methods in showing that top-ranked features alone are often adequate for maintaining performance. The path-attribution in tandem with the L2 norm or max of base and spline edge scoring methods performing comparable to or better than GAM, KAN1, and KAN2 on Deletion AUC and Sufficency metrics, with the framework performing the best or a close second in the table. This also supports the framework's ability in balancing attribution concentration and model reliability. On MNIST, due the dataset simplicity and high base model accuracy, we can observe that the path attribution approach achieves the most compact Sufficiency value, proving that it is capable of performing with minimal attribution noise and maximal feature relevance.

Across all datasets and benchmarks, our proposed approach consistently provides accurate, faithful, and clinically interpretable attribution maps. Its low variance in metric values further indicates reproducibility and robustness, making it the overall best choice for model interpretability in both neuroimaging-focused and general scenarios. These results show that path-based multilayer attributions produce reliable feature explanations across multiple experimental contexts, provided an acceptable model accuracy and performance.

Table 1: Comparison Table of InsAUC, DelAUC, and Sufficiency metrics

Dataset (Baseline)	**Category**	**Method**	**InsAUC ↑ (± std)**	**DelAUC ↓ (± std)**	**Suff ↓ (± std)**
OASIS (Acc: 93.24%, AUC: 95.74)	OURS	Path-Attr. + max	0.837 ± 0.015	0.331 ± 0.080	0.103 ± 0.076
	OURS	Path-Attr. + l2	**0.865 ± 0.049**	**0.282 ± 0.020**	0.068 ± 0.064
	External	GAM + IG	0.860 ± 0.048	0.366 ± 0.161	**0.046 ± 0.071**
	External	KAN2	0.845 ± 0.030	0.354 ± 0.101	0.087 ± 0.073
	External	KAN1	0.839 ± 0.042	0.287 ± 0.024	0.072 ± 0.058
ADNI (Acc: 81.85%, AUC: 91.30)	OURS	Path-Attr. + max	0.749 ± 0.059	**0.364 ± 0.041**	**0.201 ± 0.134**
	OURS	Path-Attr. + l2	**0.758 ± 0.044**	0.380 ± 0.051	0.209 ± 0.143
	External	GAM + IG	0.746 ± 0.026	0.506 ± 0.048	0.269 ± 0.136
	External	KAN2	0.716 ± 0.033	0.439 ± 0.034	0.326 ± 0.124
	External	KAN1	0.666 ± 0.016	0.376 ± 0.075	0.221 ± 0.029
Mendeley (Acc: 91.25%, AUC: 96.62)	OURS	Path-Attr. + max	0.773 ± 0.012	0.513 ± 0.067	0.265 ± 0.093
	OURS	Path-Attr. + l2	**0.791 ± 0.016**	0.463 ± 0.099	**0.226 ± 0.033**
	External	GAM + IG	0.767 ± 0.037	**0.399 ± 0.090**	0.293 ± 0.113
	External	KAN2	0.781 ± 0.019	0.554 ± 0.027	0.301 ± 0.034
	External	KAN1	0.710 ± 0.080	0.401 ± 0.053	0.444 ± 0.102
MNIST (Acc: 99.41%, AUC: 99.89)	OURS	Path-Attr. + max	0.944 ± 0.004	0.269 ± 0.009	0.201 ± 0.134
	OURS	Path-Attr. + l2	**0.964 ± 0.003**	0.293 ± 0.022	**0.052 ± 0.015**
	External	GAM + IG	0.956 ± 0.004	0.380 ± 0.008	0.103 ± 0.013
	External	KAN2	0.911 ± 0.010	0.326 ± 0.111	0.412 ± 0.164
	External	KAN1	0.924 ± 0.012	**0.240 ± 0.012**	0.111 ± 0.015

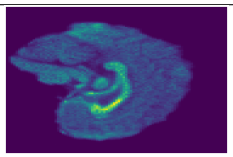
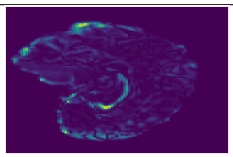
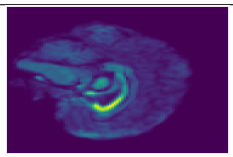
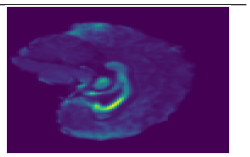
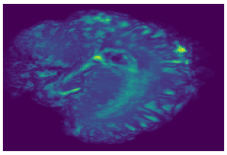
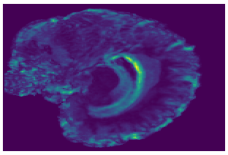
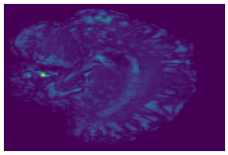
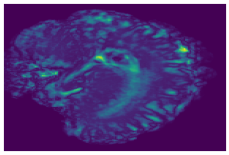
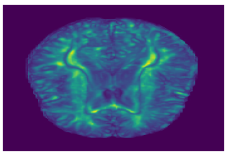
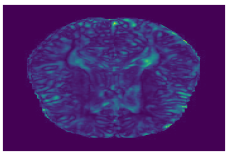
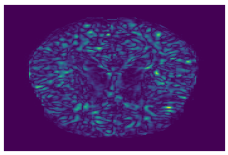
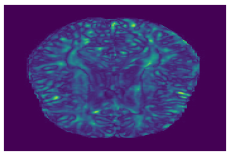
5.2 Visual Comparisons

The side by side heatmaps of the attribution maps are presented in Table 2, where the middle slice is plotted along the sagittal axis for OASIS and ADNI and the coronal slice is presented for the Mendeley Dataset, along with the MNIST global attribution maps. The KAN, KAN 2.0, GAM plots are presented alongside the proposed path-based attribution approach. Although the general regions identified by all approaches appears to be similar there are differences in the magnitude and number of regions highlighted. The KAN approach has a lot of area highlighted as downstream path interactions are not considered. KAN 2.0 sees some regions that are highlighted by our approach and GAM not highlighted as some functions that contribute equally without a large standard deviation are suppressed.

From the ADNI attribution maps it can be observed that the slice provides a visualization of midline structures such as the corpus callosum, cingulate gyrus, thalamic region, and the ventricular cavity, all of which exhibit different tissue intensity patterns. These structures are what are identified to degenerate for Alzheimer's. The smooth high-intensity band near the center corresponds to dense white-matter tracts of the corpus callosum Walterfang et al. (2014), while surrounding regions with moderate intensity reflect gray-matter cortical areas. The darker cavities near the center and lower portions of the image represent cerebrospinal fluid (CSF) spaces, primarily the lateral and third ventricles. Even though volumetric and ROI-level analyses are not feasible on this single slice, it remains structurally informative due to the inclusion of midline anatomy that captures the general morphological organization of the brain.

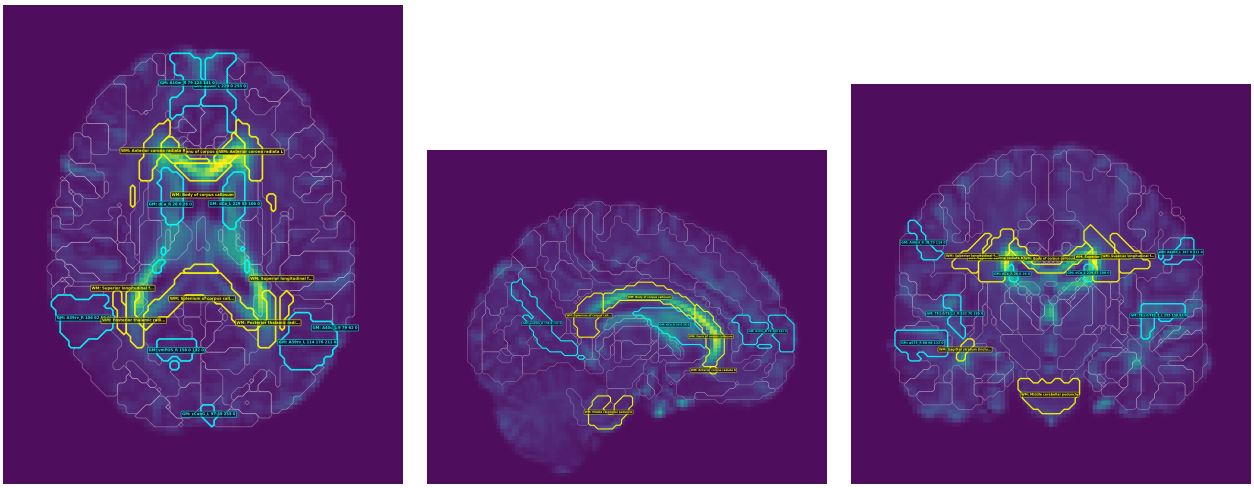
The Mendeley attributions show the coronal mid-slice showing pronounced activations along the medial temporal regions, particularly surrounding the hippocampal and parahippocampal areas Echávarri et al. (2011), and extending around the periventricular zones. Highlighted regions are also visible across the cortical ribbon, outlining the gyral-sulcal boundaries along the outer cerebrum. The central midline region-corresponding primarily to the thalamus, corpus callosum, and basal ganglia structures-appears comparatively subdued, with minimal activation intensity. This lower response suggests that the discriminative emphasis in this slice arises more from cortical and medial temporal morphometry than from deep gray matter contrast.

Table 2: Visual comparisons with example attribution maps

Dataset	KAN	KAN 2.0	GAM	Ours
OASIS				
ADNI				
Mendeley				

5.2.1 Medical Insights from the OASIS Attribution Maps

The mapped ROI inferences on the OASIS volume on grey matter from the Brainnetome atlas were as follows. The bilateral dorsal caudate nuclei (dCa_R and dCa_L) was found to be the most important region pair seen in Fig. 2c, as indicated by Jiji et al. (2013), with mean attribution importance of 0.759±0.144 (Coefficient of Variation (CV) = 18.91%) and 0.687±0.109 (CV = 15.87%) respectively, with the regions being fully consistent across all runs. This finding aligns with prior OASIS studies identifying caudate



(a) Axial view (b) Sagittal view (c) Coronal view

Figure 2: Top ROIs identified through attribution analysis in (a) axial, (b) sagittal, and (c) coronal views, showing gray matter (GM) and white matter (WM) regions most contributive to model predictions. The regions primarily highlighted in the maps were body and genu of the Corpus Callosum, Bilateral posterior thalamic radiation, the prefrontal, parietal, temporal and posterior cortex

involvement in AD Kim et al. (2024). As seen in Fig. 2, the bilateral medial prefrontal cortex (A10m_R: 0.571±0.016, CV = 2.76%; A10m_L: 0.568±0.019, CV = 3.39%), left angular gyrus (A39rv_L: 0.487±0.026, CV = 5.38%), and left cuneus (cCunG_L: 0.458±0.031, CV = 6.85%) all show perfect consistency. The medial prefrontal regions showed the highest stability among 100%-consistent ROIs, indicating highly robust attribution patterns. This convergence of default mode network regions with frontostriatal structures is consistent with established AD neurobiology.

Bilateral superior temporal regions (TE1.0/TE1.2_L: 0.672±0.063; TE1.0/TE1.2_R: 0.531±0.040) also were fully consistent, indicating temporal lobe involvement in AD Castellano et al. (2024). Near-perfect consistency was observed for right inferior parietal cortex (A40c_R: 0.461±0.025) and left orbital frontal region (A13_L: 0.448±0.026). The right hippocampus (cHipp_R) appeared in 66.7% of runs with exceptional stability when present (CV = 2.90%, range 0.433-0.476).

The top ROIs comprised prefrontal cortex (4 ROIs: bilateral A10m, A10l_R, A13_L), subcortical regions (2 ROIs: bilateral caudate), parietal cortex (4 ROIs: A39rv_L, A40c_R/L, A40rv_R), temporal cortex (2 ROIs: bilateral TE1.0/TE1.2), posterior cortex (1 ROI: cCunG_L), medial temporal lobe (1 ROI: cHipp_R), and orbitofrontal cortex (1 ROI: A11l_R). This distribution reflects distributed network pathology encompassing default mode, frontoparietal, temporal, and frontostriatal systems. Independent OASIS studies identified overlapping anatomical regions with our findings: Castellano et al. (2024) identified temporal and frontal regions as critical for classification; Kim et al. (2024) identified visual cortices, caudate and hippocampus as AD contributors. This convergence on frontostriatal, temporal, and medial network regions across independent OASIS analyses supports the biological validity of our attributed ROIs.

The top white matter ROIs identified from the JHU atlas were as follows. The body of corpus callosum (importance = 2.331) and genu of corpus callosum (importance = 2.167) emerged as the highest-ranked commissural structures, as seen in Fig. 2b. Corpus callosum atrophy in Alzheimer's disease reflects interhemispheric commissural fiber degeneration secondary to cortical neuronal loss Tomimoto et al. (2004). The splenium of corpus callosum (importance = 1.236) showed third-highest importance; posterior callosal regions show particular vulnerability in AD due to their connections with temporal and parietal cortices Kamal et al. (2021).

Association fiber tracts demonstrated substantial importance scores. Bilateral posterior thalamic radiation (left: 0.866, right: 0.847) ranked among the top white matter regions of interest, consistent with established

involvement of thalamic relay structures in AD populations Biesbroek et al. (2024). Bilateral anterior corona radiata (left: 0.848, right: 0.762) and bilateral superior corona radiata (left: 0.709, right: 0.632) were also identified among the ROIs seen in Fig. 2a. The corona radiata comprises projection fibers connecting cortical regions through the internal capsule to subcortical structures; compromised white matter microstructural integrity within corona radiata tracts, as measured by reduced fractional anisotropy and elevated mean diffusivity.

The middle cerebellar peduncle (importance = 0.660) emerged with moderate importance, reflecting potential cerebellar involvement in cognitive processing deficits. Bilateral posterior corona radiata (left: 0.546, right: 0.531) completed the set of significant white matter tracts identified. Collectively, these white matter tract findings reflect both primary axonal degeneration and secondary Wallerian degeneration consequent to gray matter neuronal loss, with progressive global microstructural deterioration characterizing the transition from cognitively normal aging to mild cognitive impairment and ultimately mild dementia. This convergence supports the biological validity of our attributed ROIs and their relevance to established AD neurobiology.

5.3 Limitations

To analyze the performance of our framework on 2D-slices, ADNI experiments utilized only the middle sagittal slice rather than full 3D volumes. While this slice captures structures relevant to Alzheimer's pathology, future work should extend path-based attribution to volumetric KANs using efficient patch-based processing or varied architectures or training methodologies. A caveat of the approach is model accuracy. The attribution maps would be accurate only if the model is able to accurately perform the task at hand. The drawback of our proposed path-based approach is its inability to produce class-wise attribution maps, the attribution maps produced are cumulative of all classes representing the complete learning of the model in one singular map. Finally, while our ROI validation demonstrates strong concordance with prior literature, prospective clinical validation with domain experts would further strengthen the medical utility of the generated attribution maps.

6 Conclusion

We introduced a path-based attribution framework for Kolmogorov-Arnold Networks that generates global, population-level importance maps by tracing information flow through the network's explicit edge structure. By combining L2 edge scoring with path-based propagation, our method produces faithful attribution maps that reflect intrinsic network properties learned during training. The efficient matrix-based implementation scales linearly with network edges, making the approach practical for high-dimensional neuroimaging applications. Our evaluation across three publicly available Alzheimer's disease classification datasets (OASIS, ADNI, Mendeley) and MNIST demonstrates that the proposed method consistently outperforms existing global attribution approaches in attribution metrics. The attribution maps identify clinically meaningful brain regions including the bilateral caudate nuclei, medial prefrontal cortex, and default mode network structures, with high consistency across multiple runs and strong concordance with established AD neurobiology literature. The superior Insertion AUC and lower Deletion AUC values confirm that our method accurately identifies the most informative features for model predictions.

This work addresses a critical gap in medical AI interpretability by providing voxel-level global attributions for KAN-based neuroimaging models. The ability to generate population-level importance maps as network-intrinsic properties enables identification of biomarkers across patient populations, supporting clinical validation and regulatory approval of AI systems. By bridging the gap between model interpretability and medical knowledge, our framework facilitates trust and transparency in AI-assisted diagnosis.

Several promising directions emerge from this work. Extending the framework to multi-class classification and survival analysis would broaden its applicability. Applying the method to other medical imaging modalities (fMRI, CT, PET) and domains beyond medicine would validate its generalizability. Exploring ways to generate classwise attribution maps and bettering KAN performance on a larger scale data is one of our ongoing works. Finally, developing interactive visualization tools that allow clinicians to explore path-level attributions could enhance clinical adoption and trust in KAN-based diagnostic systems.

References

Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365): 1–35, 2009.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European conference on computer vision*, pp. 618–634. Springer, 2020.

J Matthijs Biesbroek, Marieke G Verhagen, Stefan van der Stigchel, and Geert Jan Biessels. When the central integrator disintegrates: A review of the role of the thalamus in cognition and dementia. *Alzheimer's & Dementia*, 20(3):2209–2222, 2024.

Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. *Frontiers in aging neuroscience*, 11:194, 2019.

Giovanna Castellano, Andrea Esposito, Eufemia Lella, Graziano Montanaro, and Gennaro Vessio. Automated detection of alzheimer's disease: A multi-modal approach with 3d mri and amyloid pet. *Scientific Reports*, 14(1):5210, 2024.

Ho Tung Jeremy Chan and Eduardo Veas. Importance estimate of features via analysis of their weight and gradient profile. *Scientific Reports*, 14(1):23532, 2024.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in neural information processing systems*, 33:17212–17223, 2020.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Liang Dong, Tianqi Ding, and Keith Evan Schubert. Gcn-kan: Graph kolmogorov-arnold networks for interpretable alzheimer's disease diagnosis from structural mri. In *IEEE-EMBS International Conference on Biomedical and Health Informatics 2025*, 2025.

Carmen Echávarri, P Aalten, Harry BM Uylings, HIL Jacobs, Pieter Jelle Visser, EHBM Gronenschild, FRJ Verhey, and S Burgmans. Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease. *Brain Structure and Function*, 215(3):265–271, 2011.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, 26(8):3508–3526, 2016.

Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. Concept attribution: Explaining cnn decisions to physicians. *Computers in biology and medicine*, 123:103865, 2020.

Derek LG Hill. Ai in imaging: the regulatory landscape. *British Journal of Radiology*, 97(1155):483–491, 2024.

Gia Minh Hoang, Ue-Hwan Kim, and Jae Gwan Kim. Vision transformers for the prediction of mild cognitive impairment to alzheimer's disease progression using mid-sagittal smri. *Frontiers in Aging Neuroscience*, 15:1102869, 2023.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 279–287, 2019.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Sudevan Jiji, Karavallil Achuthan Smitha, Arun Kumar Gupta, Vellara Pappukutty Mahadevan Pillai, and Ramapurath S Jayasree. Segmentation and volumetric analysis of the caudate nucleus in alzheimer's disease. *European journal of radiology*, 82(9):1525–1530, 2013.

Sadia Kamal, Ingyu Park, Yeo Jin Kim, Yun Joong Kim, and Unjoo Lee. Alteration of the corpus callosum in patients with alzheimer's disease: Deep learning-based assessment. *PloS one*, 16(12):e0259051, 2021.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019.

Sunghun Kim, Mansu Kim, Jong-eun Lee, Bo-yong Park, and Hyunjin Park. Prognostic model for predicting alzheimer's disease conversion using functional connectome manifolds. *Alzheimer's Research & Therapy*, 16(1):217, 2024.

William Knottenbelt, William McGough, Rebecca Wray, Woody Zhidong Zhang, Jiashuai Liu, Ines Prata Machado, Zeyu Gao, and Mireia Crispin-Ortuzar. Coxkan: Kolmogorov-arnold networks for interpretable, high-performance survival analysis. *Bioinformatics*, 41(8):btaf413, 2025.

Ziming Liu, Pingchuan Ma, Yixuan Wang, Wojciech Matusik, and Max Tegmark. Kan 2.0: Kolmogorov-arnold networks meet science. *arXiv preprint arXiv:2408.10205*, 2024.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=Ozo7qJ5vZi`.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift fuer medizinische Physik*, 29(2):102–127, 2019.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

S Mori, S Wakana, LM Nagae-Poetscher, and PCM Van Zijl. Mri atlas of human white matter. *American Journal of Neuroradiology*, 27(6):1384, 2006.

Lindsay Munroe, Mariana da Silva, Faezeh Heidari, Irina Grigorescu, Simon Dahan, Emma C Robinson, Maria Deprez, and Po-Wah So. Applications of interpretable deep learning in neuroimaging: A comprehensive review. *Imaging Neuroscience*, 2:1–37, 2024.

Simisola Odimayo, Chollette C Olisah, and Khadija Mohammed. Structure focused neurodegeneration convolutional neural network for modelling and classification of alzheimer's disease. *Scientific Reports*, 14 (1):15270, 2024.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Shangran Qiu, Matthew I Miller, Prajakta S Joshi, Joyce C Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang, Ileana De Anda-Duran, Phillip H Hwang, Justin A Cramer, et al. Multimodal deep learning for alzheimer's disease dementia assessment. *Nature communications*, 13(1):3404, 2022.

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. In Jessica Zosa Forde, Francisco Ruiz, Melanie F. Pradier, and Aaron Schein (eds.), *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pp. 21–32. PMLR, 12 Dec 2020. URL `https://proceedings.mlr.press/v137/rosca20a.html`.

Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

Stephen M Smith, Mark Jenkinson, Michael W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Paul R Bannister, Maria De Luca, Ilija Drobnjak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, 23(S1):S208–S219, 2004.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Hidekazu Tomimoto, Jin-Xi Lin, Akinori Matsuo, Masafumi Ihara, Ryo Ohtani, Masunari Shibata, Yukio Miki, and Hiroshi Shibasaki. Different mechanisms of corpus callosum atrophy in alzheimer's disease and vascular dementia. *Journal of neurology*, 251(4):398–406, 2004.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6): 1310–1320, 2010.

Mark Walterfang, Eileen Luders, Jeffrey CL Looi, Priya Rajagopalan, Dennis Velakoulis, Paul M Thompson, Olof Lindberg, Per Östberg, Love E Nordin, Leif Svensson, et al. Shape analysis of the corpus callosum in alzheimer's disease and frontotemporal lobar degeneration subtypes. *Journal of Alzheimer's Disease*, 40 (4):897–906, 2014.

Alan Q Wang, Batuhan K Karaman, Heejong Kim, Jacob Rosenthal, Rachit Saluja, Sean I Young, and Mert R Sabuncu. A framework for interpretability in machine learning for medical imaging. *IEEE Access*, 12:53277–53292, 2024.

Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661, 2020.

Ananya Yakkundi. Alzheimer's disease dataset. Mendeley Data, V1, 2023. URL `https://doi.org/10.17632/ch87yswbz4.1`.

A Appendix

A.1 Training Details for Models

All experiments were carried out using the Kolmogorov–Arnold Network with weight decay architecture and trained on a CUDA-A100. Each model is built using a sequence of `KANLinear` layers, where the connection weights are represented using B-spline functions. Unless stated otherwise, the following spline and training settings were kept the same across all datasets:

- grid size: 5 points per input dimension,
- spline order: cubic (order 3),
- base activation function: SiLU,
- noise scale: 0.1,
- linear and spline scaling factors: 1.0,
- grid epsilon: 0.02,
- initial spline grid range: $[-1, 1]$.

Input features for all datasets were linearly scaled to the $[-1, 1]$ range to match the spline grid. These choices were kept fixed to maintain consistency across experiments and to avoid unnecessary hyperparameter variation.

The layer widths were selected to reflect the input dimensionality and complexity of each dataset. For OASIS-1, the network width configuration was

$$[D, 20, 10, 10, 10, 10, 10, 10, 10, 5, 1],$$

where D denotes the number of input features. This produced an 11-layer network with progressively narrower layers for binary classification.

For ADNI, a larger first hidden layer was used to accommodate its higher input dimensionality. The width configuration was

$$[D, 800, 20, 15, 15, 15, 10, 10, 10, 5, 1].$$

For the Mendeley dataset, a slightly shallower network was used:

$$[D, 800, 20, 15, 15, 10, 10, 10, 1].$$

All models were optimized using AdamW with a learning rate of 10^{-3}. An exponential learning rate scheduler was applied by a factor of 0.8. The same optimizer and scheduler settings were used for all datasets for reproducibility.

These configurations were sufficient for stable training across datasets and provided a controlled setting for comparing model behaviour.

A.2 Detailed Analysis Table

Comprehensive comparison of explanation quality metrics across four datasets (OASIS-1, ADNI, Mendeley, and MNIST), evaluating 14 attribution methods applied to KAN and Path-Attr. models. Metrics include Insertion/Deletion AUC (InsAUC/DelAUC), Sufficiency/Comprehensiveness (Suff/Comp), Average Positive Correlation (AOPC), Precision@5, and Faithfulness (Faith). Best performance per dataset and metric appears in bold. Model accuracies and AUCs are reported parenthetically for each dataset. Lower L0 values indicate sparser explanations. Path-Attr. + std consistently achieves superior sparsity (lowest L0) across datasets while maintaining competitive faithfulness, particularly on ADNI ($6{,}836 \pm 422$) and MNIST (84 ± 0.4). KAN + max excels in AOPC on OASIS-1, while GAM leads InsAUC but suffers high L0 norms. To be noted that GAM is in combination with Integrated Gradients. The base KAN paper uses L1 norms at the input node, and we test it with different scoring functions, listed as KAN + scoring function. This table also emperically justifies our choice of L2 norms alongside our path attribution framework, as it consistently outperforms other methods.

Table 3: Comparison of explanation metrics across datasets (mean ± SD). This table emperically justifies our choice of L2 norms alongside our path attribution framework, as it consistently outperforms other methods.

Method	InsAUC	DelAUC	Suff	Comp	AOPC	Prec@5	Faith
OASIS (Acc: 93.24%, AUC: 95.74)							
KAN + max	0.823 ± 0.004	0.703 ± 0.035	0.099 ± 0.080	0.635 ± 0.036	$\mathbf{0.555 \pm 0.043}$	0.896 ± 0.072	0.782 ± 0.021
Path-Attr. + max	0.837 ± 0.015	0.331 ± 0.080	0.103 ± 0.076	0.521 ± 0.178	0.386 ± 0.134	0.891 ± 0.063	$\mathbf{0.623 \pm 0.112}$
GAM	0.860 ± 0.048	0.366 ± 0.161	0.046 ± 0.071	0.279 ± 0.285	0.276 ± 0.225	0.915 ± 0.094	0.067 ± 0.379
KAN2	0.845 ± 0.030	0.354 ± 0.101	0.087 ± 0.073	0.626 ± 0.017	0.451 ± 0.021	0.922 ± 0.069	0.384 ± 0.208
Path-Attr. + l1	0.853 ± 0.029	$\mathbf{0.274 \pm 0.049}$	0.041 ± 0.057	0.008 ± 0.006	0.011 ± 0.005	0.921 ± 0.064	0.029 ± 0.539
Path-Attr. + l2	$\mathbf{0.865 \pm 0.049}$	0.282 ± 0.020	0.068 ± 0.064	0.151 ± 0.139	0.109 ± 0.097	0.891 ± 0.063	0.336 ± 0.692
Path-Attr. + std	0.769 ± 0.033	0.218 ± 0.004	0.042 ± 0.068	0.014 ± 0.009	0.023 ± 0.011	0.896 ± 0.072	0.254 ± 0.090
Path-Attr. + l1_max	0.835 ± 0.036	0.271 ± 0.016	0.082 ± 0.054	0.011 ± 0.008	$\mathbf{0.018 \pm 0.009}$	$\mathbf{0.938 \pm 0.083}$	0.162 ± 0.588
Path-Attr. + recursive	0.847 ± 0.019	0.299 ± 0.070	0.087 ± 0.080	0.067 ± 0.065	0.107 ± 0.083	0.901 ± 0.080	0.360 ± 0.204
KAN + l1	0.839 ± 0.042	0.287 ± 0.024	0.072 ± 0.058	0.009 ± 0.003	0.015 ± 0.012	0.989 ± 0.034	0.153 ± 0.622
KAN + l2	0.846 ± 0.017	0.454 ± 0.075	0.066 ± 0.092	0.026 ± 0.019	0.136 ± 0.038	0.923 ± 0.091	0.656 ± 0.077
KAN + std	0.610 ± 0.070	0.213 ± 0.006	0.348 ± 0.152	0.007 ± 0.010	0.026 ± 0.018	0.947 ± 0.053	0.278 ± 0.162
KAN + l1_max	0.692 ± 0.053	0.188 ± 0.013	0.614 ± 0.031	0.033 ± 0.020	0.038 ± 0.022	0.578 ± 0.190	0.429 ± 0.519
KAN + recursive	0.831 ± 0.008	0.466 ± 0.070	0.068 ± 0.084	0.164 ± 0.150	0.234 ± 0.143	0.906 ± 0.089	0.700 ± 0.072
ADNI (Acc: 81.85%, AUC: 91.30)							
Path-Attr. + l1_max	0.678 ± 0.025	0.347 ± 0.098	0.221 ± 0.029	0.093 ± 0.049	0.096 ± 0.045	0.783 ± 0.029	0.664 ± 0.221
KAN2	0.716 ± 0.033	0.439 ± 0.034	0.326 ± 0.124	0.218 ± 0.038	0.175 ± 0.032	0.598 ± 0.075	0.174 ± 0.348
KAN + l2	0.746 ± 0.060	0.364 ± 0.041	0.175 ± 0.172	0.131 ± 0.029	0.114 ± 0.021	0.639 ± 0.224	0.325 ± 0.568
GAM	0.746 ± 0.026	$\mathbf{0.506 \pm 0.048}$	0.269 ± 0.136	0.207 ± 0.046	$\mathbf{0.218 \pm 0.045}$	0.560 ± 0.084	0.724 ± 0.086
Path-Attr. + l1	0.698 ± 0.028	0.310 ± 0.081	0.222 ± 0.030	$\mathbf{0.043 \pm 0.032}$	0.051 ± 0.021	0.778 ± 0.030	0.431 ± 0.549
Path-Attr. + l2	0.758 ± 0.044	0.380 ± 0.051	0.209 ± 0.143	0.124 ± 0.032	0.115 ± 0.021	0.702 ± 0.259	0.404 ± 0.551
Path-Attr. + std	0.631 ± 0.021	0.385 ± 0.069	0.211 ± 0.029	0.051 ± 0.021	0.070 ± 0.029	$\mathbf{0.804 \pm 0.026}$	0.523 ± 0.222
Path-Attr. + max	$\mathbf{0.749 \pm 0.059}$	0.364 ± 0.041	0.201 ± 0.134	0.153 ± 0.015	0.126 ± 0.020	0.666 ± 0.216	$\mathbf{0.180 \pm 0.491}$
Path-Attr. + recursive	0.740 ± 0.064	0.356 ± 0.039	0.217 ± 0.168	0.132 ± 0.043	0.110 ± 0.030	0.650 ± 0.230	0.416 ± 0.314
KAN + l1	0.666 ± 0.016	0.376 ± 0.075	0.221 ± 0.029	0.089 ± 0.040	0.090 ± 0.038	0.783 ± 0.029	0.285 ± 0.737
KAN + std	0.631 ± 0.021	0.392 ± 0.058	0.267 ± 0.036	0.051 ± 0.036	0.079 ± 0.026	0.747 ± 0.126	0.582 ± 0.258
KAN + max	0.745 ± 0.061	0.354 ± 0.040	0.196 ± 0.165	0.153 ± 0.020	0.130 ± 0.024	0.702 ± 0.250	0.222 ± 0.532
KAN + l1_max	0.658 ± 0.014	0.399 ± 0.077	0.222 ± 0.030	0.094 ± 0.030	0.112 ± 0.036	0.778 ± 0.030	0.499 ± 0.486
KAN + recursive	0.746 ± 072	0.336 ± 0.042	0.229 ± 0.174	0.135 ± 0.034	0.109 ± 0.029	0.550 ± 0.207	0.250 ± 0.350
Mendeley (91.25%, AUC: 96.62)							
KAN2	0.781 ± 0.019	0.554 ± 0.027	0.301 ± 0.034	$\mathbf{0.496 \pm 0.059}$	$\mathbf{0.453 \pm 0.042}$	0.685 ± 0.011	0.086 ± 0.420
Path-Attr. + max	0.773 ± 0.012	0.513 ± 0.067	0.265 ± 0.093	0.376 ± 0.171	0.345 ± 0.130	$\mathbf{0.784 \pm 0.097}$	0.048 ± 0.446
Path-Attr. + std	0.679 ± 0.065	0.332 ± 0.020	0.261 ± 0.073	0.175 ± 0.058	0.165 ± 0.040	0.743 ± 0.057	$\mathbf{0.618 \pm 0.223}$
GAM	0.767 ± 0.037	0.399 ± 0.090	0.293 ± 0.113	0.193 ± 0.100	0.209 ± 0.095	0.653 ± 0.114	0.398 ± 0.656
Path-Attr. + l1	0.649 ± 0.080	$\mathbf{0.296 \pm 0.053}$	0.504 ± 0.103	0.024 ± 0.034	0.047 ± 0.012	0.521 ± 0.153	0.227 ± 0.465
Path-Attr. + l2	$\mathbf{0.791 \pm 0.016}$	0.463 ± 0.099	0.226 ± 0.033	0.331 ± 0.117	0.311 ± 0.110	0.701 ± 0.029	0.070 ± 0.550
Path-Attr. + l1_max	0.671 ± 0.073	0.300 ± 0.047	0.398 ± 0.087	0.129 ± 0.080	0.129 ± 0.066	0.611 ± 0.167	0.269 ± 0.697
Path-Attr. + recursive	0.778 ± 0.026	0.431 ± 0.097	0.247 ± 0.096	0.227 ± 0.124	0.236 ± 0.102	0.695 ± 0.131	0.500 ± 0.302
KAN + l1	0.710 ± 0.080	0.401 ± 0.053	0.444 ± 0.102	0.145 ± 0.024	0.174 ± 0.020	0.450 ± 0.015	0.196 ± 0.599
KAN + l2	0.730 ± 0.011	0.426 ± 0.063	0.484 ± 0.051	0.176 ± 0.011	0.195 ± 0.048	0.501 ± 0.116	0.212 ± 0.582
KAN + std	0.652 ± 0.074	0.320 ± 0.011	0.351 ± 0.132	0.096 ± 0.037	0.082 ± 0.021	0.651 ± 0.120	0.494 ± 0.318
KAN + max	0.758 ± 0.016	0.444 ± 0.064	0.443 ± 0.122	0.160 ± 0.045	0.205 ± 0.055	0.487 ± 0.128	0.630 ± 0.323
KAN + l1_max	0.683 ± 0.093	0.372 ± 0.048	0.501 ± 0.049	0.176 ± 0.023	0.177 ± 0.040	0.456 ± 0.026	0.341 ± 0.437
KAN + recursive	0.755 ± 0.032	0.402 ± 0.124	0.316 ± 0.095	0.199 ± 0.152	0.200 ± 0.137	0.677 ± 0.125	0.085 ± 0.671
MNIST (Acc: 99.41%, AUC: 99.89)							
Path-Attr. + l2	$\mathbf{0.964 \pm 0.003}$	0.293 ± 0.022	0.052 ± 0.015	0.097 ± 0.026	$\mathbf{0.193 \pm 0.019}$	0.797 ± 0.054	0.953 ± 0.007
KAN + std	0.906 ± 0.023	0.168 ± 0.046	0.153 ± 0.035	0.074 ± 0.018	0.066 ± 0.008	0.769 ± 0.048	0.156 ± 0.390
KAN + recursive	0.936 ± 0.020	0.250 ± 0.053	0.111 ± 0.044	0.053 ± 0.011	0.059 ± 0.006	0.823 ± 0.071	0.654 ± 0.313
GAM	0.956 ± 0.004	0.380 ± 0.008	0.103 ± 0.013	0.064 ± 0.023	0.169 ± 0.164	0.705 ± 0.038	0.955 ± 0.013
KAN2	0.911 ± 0.100	0.326 ± 0.111	0.412 ± 0.164	0.039 ± 0.018	0.130 ± 0.063	0.726 ± 0.119	0.949 ± 0.003
Path-Attr. + l1	0.756 ± 0.102	0.159 ± 0.113	0.490 ± 0.003	0.001 ± 0.001	0.033 ± 0.060	0.565 ± 0.111	0.776 ± 0.121
Path-Attr. + std	0.753 ± 0.006	$\mathbf{0.114 \pm 0.007}$	0.109 ± 0.013	0.046 ± 0.009	0.016 ± 0.001	0.512 ± 0.005	$\mathbf{0.550 \pm 0.048}$
Path-Attr. + max	0.944 ± 0.004	0.269 ± 0.009	0.161 ± 0.015	0.058 ± 0.012	0.054 ± 0.013	0.790 ± 0.022	0.826 ± 0.056
Path-Attr. + l1_max	0.917 ± 0.002	0.228 ± 0.007	0.114 ± 0.039	0.048 ± 0.007	0.055 ± 0.010	0.727 ± 0.041	0.538 ± 0.167
Path-Attr. + recursive	0.938 ± 0.014	0.264 ± 0.023	0.143 ± 0.027	0.063 ± 0.014	0.054 ± 0.005	$\mathbf{0.816 \pm 0.062}$	0.765 ± 0.102
KAN + l1	0.924 ± 0.012	0.240 ± 0.012	0.111 ± 0.015	0.042 ± 0.003	0.062 ± 0.016	0.778 ± 0.038	0.506 ± 0.223
KAN + l2	0.941 ± 0.005	0.268 ± 0.006	0.315 ± 0.015	0.054 ± 0.013	0.053 ± 0.002	0.806 ± 0.026	0.873 ± 0.044
KAN + max	0.932 ± 0.014	0.256 ± 0.024	0.128 ± 0.034	0.062 ± 0.016	0.054 ± 0.006	0.769 ± 0.047	0.655 ± 0.194
KAN + l1_max	0.930 ± 0.015	0.248 ± 0.013	0.062 ± 0.016	0.061 ± 0.008	0.061 ± 0.008	0.792 ± 0.052	0.565 ± 0.270

A.3 Tabulated ROIs identified across 15 runs

Table 4 presents the set of regions retained after applying a data-driven importance threshold derived from the distribution of mean ROI contributions. To determine an appropriate cutoff, we performed a one-dimensional clustering analysis (k-means, $k = 4$) on the mean values across all ROIs. This procedure consistently separated the ROIs into four natural clusters, with a clear boundary between a low-intensity cluster (centered at approximately 0.41) and a higher-intensity cluster beginning near 0.52. The transition point between these clusters occurred between 0.485 and 0.525; accordingly, we adopted a conservative cutoff of Mean ≥ 0.50 to exclude regions whose contributions were statistically indistinguishable from the lowest-value cluster.

The ROIs above this threshold form a coherent set dominated by major commissural and projection white-matter pathways, including the corpus callosum, corona radiata, and posterior thalamic radiations, alongside subcortical and temporal regions such as the dorsal caudate and TE1/TE1.2 auditory cortex. These structures exhibited both high mean importance values and variability patterns consistent with known inter-subject differences in white-matter microstructure. Importantly, the retained pattern aligns with established neuroanatomical vulnerability profiles observed in aging and Alzheimer's-related cohorts, supporting the biological plausibility of the model's feature ranking.

This thresholding procedure therefore ensures that the retained ROIs represent the statistically meaningful and anatomically interpretable contributors to the model, while excluding regions that fall within the noise-dominated portion of the distribution.

Table 4: ROIs retained after applying a data-driven cutoff (Mean ≥ 0.50), sorted in descending order of the Mean value.

Rank	**ROI label**	**Mean**	**SD**	**CV (%)**
1	WM: Body of corpus callosum	2.307	0.751	32.53
2	WM: Genu of corpus callosum	1.986	0.635	31.98
3	WM: Splenium of corpus callosum	1.099	0.333	30.31
4	WM: Anterior corona radiata L	0.886	0.268	30.24
5	WM: Posterior thalamic radiation (optic radiation) R	0.836	0.169	20.20
6	WM: Posterior thalamic radiation (optic radiation) L	0.827	0.209	25.32
7	GM: dCa_R	0.759	0.144	18.91
8	WM: Superior corona radiata L	0.758	0.188	24.86
9	WM: Anterior corona radiata R	0.737	0.154	20.91
10	GM: dCa_L	0.687	0.109	15.87
11	GM: TE1.0/TE1.2_L	0.672	0.063	9.42
12	WM: Superior corona radiata R	0.640	0.109	16.97
13	WM: Middle cerebellar peduncle	0.630	0.136	21.61
14	GM: A10m_R	0.571	0.016	2.76
15	GM: A10m_L	0.568	0.019	3.39
16	WM: Posterior corona radiata L	0.562	0.082	14.67
17	WM: Posterior corona radiata R	0.543	0.095	17.45
18	GM: TE1.0/TE1.2_R	0.525	0.030	5.71