Multi-output lesion-symptom mapping using deep learning and explainable AI in small vessel disease

Ana San Román Gaitero
a.sanromangaitero@students.uu.nl
MSc Medical Imaging, Graduate School of Life Sciences, Utrecht University
Image Sciences Institute, University Medical Center Utrecht (UMCU)

***Abstract*—Cerebral small vessel disease causes cognitive impairment, dementia, and stroke and is characterized by white matter hyperintensities (WMH) and other brain lesions. Lesion-symptom mapping (LSM) aims to understand the relationship between brain lesion location and cognition by identifying strategic lesion locations. This study presents a multi-output deep learning lesion-symptom mapping (DL−LSM) approach using explainable artificial intelligence (XAI). This approach is validated in a simulation study using WMH segmentations of 821 memory clinic patients and artificial cognitive scores. The study comprised three experiments. The first involved generating artificial cognitive scores based on the lesion load within three predefined regions of interest (ROIs). The second experiment studied the impact of adding noise to these scores on the DL and XAI methods. The third experiment explored whether intercorrelations between different ROIs in the artificial cognitive scores could be detected. Two convolutional neural networks (CNN) were developed to predict the artificial cognitive scores, and XAI was used to identify the locations that influenced these predictions. The methods were evaluated by quantifying the model's predictive performance, identifying the ROIs in the XAI's attribution maps, and quantifying the intercorrelation of the detected ROIs. This study demonstrates that DL models can predict multiple artificial cognitive scores based on WMH segmentations, and that XAI can identify the ROIs associated with the simulated cognitive scores. Additionally, the results demonstrate that DL−LSM is robust to low levels of noise in the artificial cognitive scores and can detect intercorrelations between ROIs. These findings indicate that DL and XAI can be used to perform LSM in order to predict cognitive scores and determine their relationship with specific lesion locations.**

***Index Terms*—Small Vessel Disease, Lesion-symptom mapping, Deep Learning, Explainable AI, Neuroimaging, MRI, Simulation Study**

I. Introduction

Cerebral small vessel disease (SVD) is a common microvascular disorder that manifests during aging and can lead to stroke, cognitive impairment, as well as behavioral or functional problems [1]. It often co-occurs with vascular cognitive impairment (VCI), causing long-term disabilities and worsening the quality of life of the patients [2]. In addition, SVD can result in a variety of brain lesions, including lacunes, infarcts, microbleeds, and white matter hyperintensities (WMH) seen in magnetic resonance imaging (MRI) and computed tomography (CT) scans [1, 2]. The diagnosis can be challenging and, it is often unclear which factors cause these brain lesions and how they relate to the clinical symptoms. However, it is also known that the severity of cognitive impairment is related to the location of brain tissue damage and, it has been shown that these locations are more correlated to cognition than to the total lesion volume [3, 4]. For instance, one study has demonstrated that WMH directly affects specific cognitive domains and global cognitive function according to their position in the brain [5].

Hence, lesion-symptom patterns play a significant role in understanding the cognitive impact of vascular lesions. Lesion-symptom mapping (LSM) is a technique applied to identify the specific areas of brain lesions that have the most impact on cognition. This technique is carried out by analyzing lesion maps [6]. These maps are lesion segmentations previously obtained by a neurologist or automatic procedures that derive from MRI or CT scans. By computing LSM, an attribution map is generated; this map displays the relative importance of the lesion map's voxels that are most related to the cognitive outcomes. Therefore, further research on LSM can provide more information about the underlying brain mechanisms that contribute to SVD and improve those maps that are already used in clinical practice [7].

Several LSM methods have been developed over the years, including the traditional overlap-subtraction approaches, the conventional voxel-based lesion-symptom mapping (VLSM) method, and the current machine learning state-of-art method called support vector regression lesion-symptom mapping (SVR−LSM) [6]. VLSM compares patients with lesions to those without lesions on a voxel-by-voxel basis [8], performing as a univariate method that cannot asses intervoxel relationships. SVR-LSM overcomes this issue as a multivariate approach evaluating all voxel intercorrelations within an entire lesion map, rather than considering each voxel independently [9]. Nonetheless, the main limitation of SVR−LSM is that it can only identify these correlations between a single lesion and a single cognitive measure, and not on multiple lesion types and cognitive domains simultaneously. Given that the human brain is a complex system of interconnected neurons where different brain areas contribute to one or more cognition

domains, this becomes a problem [7, 10]. In this regard, when understanding the cognitive impact of vascular lesions, it is important to consider the relationship between multiple lesion types and cognitive scores simultaneously.

The use of deep learning (DL) models for LSM to understand the relationship between brain lesion location and cognition has not yet emerged. Current research papers only focus on predicting cognitive performance from MRI lesion images using DL models, specifically, convolutional neural networks (CNNs) [11, 12]. CNNs are the most commonly used neural networks in medical image analysis. Their model's complexity and flexibility allow CNNs to extract features from any kind of image data [13]. In addition, CNNs can capture the correlation between multiple inputs and multiple outputs simultaneously. By bringing DL to LSM, studies can potentially overcome the main limitation of SVR−LSM and identify the relation between multiple lesion types and multiple cognitive outcomes. Nevertheless, DL models have one drawback, called the "black box" problem. The black box problem refers to the lack of the model's interpretability, where the understanding of how the model has chosen the respective predictions is not evident. For this, explainable artificial intelligence (XAI) methods are applied to comprehend the underlying decision-making process of the model, providing saliency maps (attribution maps in LSM) that reflect the voxel's contribution to the model's decision [14].

When it comes to LSM, research papers perform simulation studies to validate their chosen methods [15, 16, 17, 9]. Usually, this is done by obtaining simulated behavioral scores that follow the patients' real lesion load of specific brain areas. Most of the studies perform linear correlations, whereas in clinical practice the brain-behavior relationship is not only not linear but also noisy. Pustina et al. (2018) first introduced this relationship by injecting an error in the artificial cognitive scores to match the notoriously noisy brain-behavior relationships [15].

The current study wishes to expand upon the SVR−LSM algorithm by introducing a novel approach of deep learning lesion-symptom mapping (DL−LSM). More specifically, this project proposes a multi-output 3D CNN regression model to predict multiple artificial cognitive scores based on a 3D MRI brain lesion image in patients with SVD. It also proposes implementing XAI techniques to reflect the attribution of each voxel in the input image to each output value, identifying the brain lesion locations responsible for the artificial cognitive scores. This framework will be validated using a simulation study across a range of potential brain-cognition relationships, assessing the capabilities as well as the robustness of the XAI methods and the deep learning model.

II. Methods

This section describes the research methodology and provides a comprehensive understanding of the dataset, simulations, models, experiments, and analyses conducted. It begins with a description of the dataset and continues with an explanation of the artificial cognitive score simulations, providing insight into the process of obtaining lesion-cognition relationships. Next, it describes the architectural framework and principles of the DL models. Following this, a detailed explanation of the experimental approaches is provided, including the proof-of-concept experiment, as well as the noise and intercorrelation experiments. The last section summarizes the evaluation metrics used to analyze the robustness and performance of the experimental results.

A. Dataset

This study used 821 patients from the TRACE-VCI cohort study, which was conducted between 2009 and 2013 [18]. The TRACE-VCI dataset contains 861 memory clinic patients with evidence of vascular brain injury on MRI, regardless of the severity of cognitive impairment. Table I presents the clinical and demographic characteristics of the study sample in the current research work. Patients with a presumed primary etiology other than vascular brain injury or neurodegeneration were excluded. Moreover, each patient underwent physical and neurological examination, laboratory testing, extensive neuropsychological and cognitive testing, and an MRI scan of the brain [19].

Specifically, this study uses WMH segmentations derived from the FLAIR sequences of the TRACE-VCI MRI scans. These WMH segmentations were obtained by applying the k-nearest neighbor classification considering tissue type priors method. All WMH segmentations, also called lesion maps, were registered to the T1 1-mm MNI-152 (Montreal Neurological Institute) brain template [19] using the Elastix toolbox [20]. Lastly, the 821 lesion maps were cropped to only brain-containing regions of the MNI-152 space to minimize memory usage, downsizing the image to $152 \times 179 \times 142\ mm^3$.

TABLE I
Demographic and clinical characteristics of the study sample.

Characteristics	**Study Sample (n=821)**
Demographic Characteristics	
Female, n (%)	382 (46.5%)
Age, mean ± SD	67.5 (8.5)
Imaging Characteristics	
WMH volume in milliliters[a], median (IQR)	8.02 (3.25-21.34)
Cognitive Characteristics	
MMSE, median (IQR)	25.00 (22-28)
CDR, median (IQR)	0.50 (0.50-1)

CDM, Clinical Dementia rating; MMSE, Mini-Mental State Examination; SD, standard deviation; IQR, interquartile range.

[a]Standardized WMH volumes were calculated from lesion maps after transformation to the MNI-152 brain template.

Fig. 1 shows a heatmap of the total lesion prevalence across patients. This was achieved by aggregating the lesion maps of each patient, allowing for a clear visualization of the distribution of the WMH lesions throughout the brain.

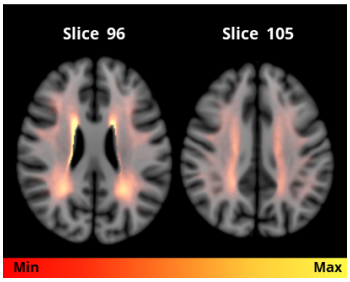

Fig. 1. Total lesion prevalence across all patients represented in the MNI-152 space.

B. Artificial cognitive scores

The present study validates the proposed DL−LSM approach in a simulation study that creates synthetic lesion-cognition relations, based on the procedure described in Zhang et al. (2014) [9]. Therefore, it serves as a validation to determine the ability of DL models to localize predefined lesion-symptom regions of interest (ROIs).

The simulation study was developed using three ROIs based on the real WMH lesion maps from the 821 patients. The ROIs were defined in brain areas where at least 10 patients contained a lesioned voxel. This was done by creating a lesion mask containing the occurrence of lesions within each voxel of all patients. Each ROI was then defined as a $5 \times 5 \times 5$ mm^3 cube. The first ROI was randomly located inside the lesion mask and its corresponding area was removed to ensure that, in the following repetitions, all of the ROIs were positioned at different locations. This resulted in three non-overlapping origins for ROI 1, ROI 2, and ROI 3 that were located at (94, 112, 94), (106, 90, 45), and (105, 116, 90), respectively (seen in Fig. 2 A). To further assess the ROIs, the study conducted a descriptive analysis by calculating the lesion prevalence within each ROI and its correlation with the total lesion volume of the WMH lesion maps.

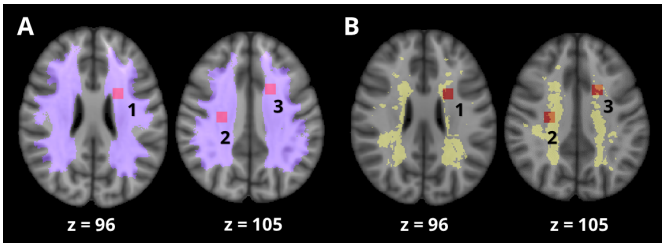

Fig. 2. Overlap (A) between the three ROIs (red) and the lesion mask (purple), as well as the overlap (B) between the ROIs and the lesion map of one patient (yellow). The z slices are in the MNI-152 space.

The artificial cognitive scores were generated based on the sum of lesion volume within the ROIs (overlap seen in Fig. 2 B) using the following equations:

$$\text{score}_i = \sum_{k}^{3} (\text{lesion-load of ROI}_k)_i \quad (1)$$

$$\text{score}_{i,j} = (\text{lesion-load of ROI}_j)_i \quad (2)$$

where i denotes the i-th subject, and j refers to the j-th score corresponding to the j-th ROI ($j = 1, 2, 3$).

Equation 1 was used to assign each patient a simulated cognitive score based on the sum of the lesion load present in the three ROIs of the WMH lesion maps. This single-score simulation was initially generated to establish a simple one-to-one lesion-cognition relationship and demonstrate the applicability of DL to LSM. The following DL-LSM approach was developed to identify relationships between multiple lesions and cognitive scores. *Equation* 2 was applied to establish the following relationships based on the lesion load in a specific ROI of the WMH lesion maps. Accordingly, each patient was assigned three different artificial cognitive scores related to the corresponding ROI: Score 1 was associated with ROI 1, Score 2 with ROI 2, and Score 3 with ROI 3. These scores ranged from 0 to 1, with a higher score indicating a greater lesion load within the ROI.

C. LSM: Deep Learning and XAI

To develop a DL−LSM approach two procedures have to be considered: (i) designing the neural network architecture responsible for predicting cognitive scores from the lesion map input images and, (ii) implementing the XAI technique to generate the attribution maps that highlight the specific lesion areas that influenced these predicted cognitive scores.

1) CNN architectures

To overcome this regression task, two different 3D CNN models were designed. CNNs use 3D convolutional kernels to analyze the relationship between multiple voxels and the output, enabling them to identify multi-voxel correlations.

The aim of this study was to compare two CNN models and determine whether increasing model complexity is more effective in identifying multiple lesion-behavior relationships. The CNN architecture initially consists of two blocks of 3D convolutional layers, which are then followed by batch normalization and a ReLU activation function. Both convolutional layers use 3D filters with kernel sizes of 10 and a stride of 5. The number of output channels increases from 50 to 100 in the second block. These two blocks lead to a fully connected layer that is applied for the final regression task. The architecture of the second model, referred to as Residual CNN in this study, aims to implement a residual learning framework. The Residual CNN consists of a first convolutional block with the same layers and parameters as the previous CNN, followed by two sets of residual blocks (as seen in Fig. 3). Each residual block consists of two 3D convolutional layers with batch normalization and ReLU activation. The first convolutional layers within each block have a kernel size of 5, a stride of 3, and a padding of 2. In the second convolutional layer,

the stride is changed to 1. The number of output channels increases from 50 to 100 in the second layer. Finally, the residual connection is introduced by downsampling the input and output dimensions with a $1 \times 1 \times 1$ convolution. Following the residual blocks, the model proceeds to an adaptive 3D average pooling layer that reduces the spatial dimensions to $1 \times 1 \times 1$. Subsequently, the outputs are processed by a fully connected layer.

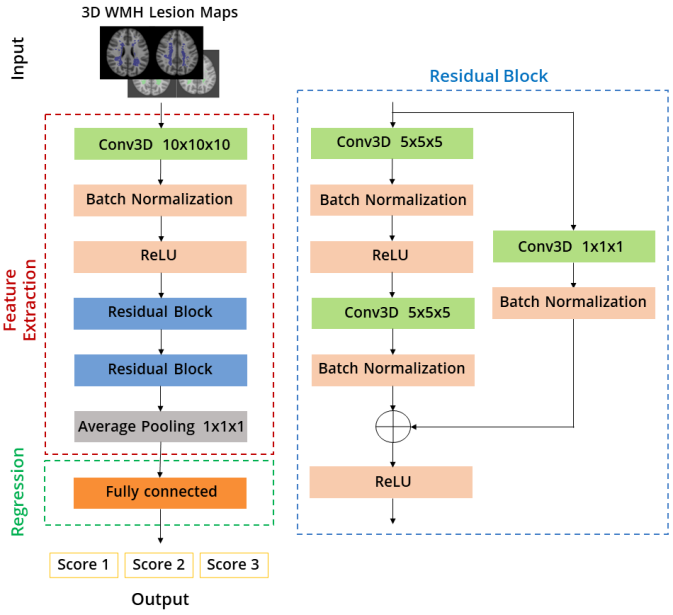

Fig. 3. Residual CNN model architecture.

Both models were trained using 5-fold cross-validation. In this case, the validation fold was used as a test set to ensure a more robust evaluation of the methods, and the remaining folds were used to train the model. Therefore, predictions of the artificial cognitive scores were obtained for all test patients at each fold. As a result, five different models were trained and tested using the complete dataset.

The cross-validation process, as well as the architecture models, were implemented with the Pytorch library using Python 3.10.12 [21]. Both models were trained using default parameters, including a batch size of 4, the AdamW optimizer, and the Mean Squared Error (MSE) loss function. The learning rate was set to 0.001 and the weight decay to 0.01.

2) XAI

Neural networks typically consist of many layers connected through numerous non-linear relations, making it unfeasible to fully comprehend how the neural network came to its decision. Several researchers in medical imaging are increasingly using XAI to explain the results of their algorithms [22]. In this regard, to comprehend the underlying mechanism of the proposed models, XAI was the chosen approach. Multiple methods from the *captum.ai* library [23] were evaluated, of which two were used in this study, Gradient Shap, and Occlusion.

Gradient SHAP uses the respective gradients of the model output to the input features to approximate Shapley values [24]. It computes the gradients by randomly sampling from the distribution of a baseline (reference) input and integrates these along all possible path combinations from the baseline to the input features. This provides a fair distribution of the contribution of each feature towards the prediction for a specific instance, considering all possible combinations and interactions.

Occlusion is a perturbation-based method that changes the input image to assess the importance of certain areas of that image for the task under consideration [25]. This is done by using a sliding window over the input image and replacing it with baseline values. After this replacement the model's prediction is recalculated in the trained model to detect changes in the output [22].

The XAI methods were applied with different parameters. Each method uses a different mathematical calculation, and the parameters selected strongly influence the generated attribution maps. Gradient SHAP used a zero image baseline, 30 randomly generated examples per input, and a 0.01 standard deviation of Gaussian noise added to the inputs. For occlusion, a sliding window of size 5 was used with strides of 3 and 10 perturbations per batch.

The XAI methods were applied to compute 3D attribution maps using the model weights of the last epoch at each fold. Attribution maps were created for all test patients at each fold to later obtain a group-level attribution map by adding these individual attribution maps. The attribution map for the entire dataset was created by combining all fold group-level maps.

D. Experiments

The project conducted several experiments to evaluate the predictive performance of DL models and the ROI identification of XAI methods under different simulation scenarios. The experiments included a proof-of-concept experiment, a noise experiment, and an intercorrelation experiment.

1) Proof-of-concept experiment

The aim of the proof-of-concept experiment is to replicate linear brain-behavior relations, ensuring that each score depends solely on the lesion load present in the ROIs. The experiment serves as an introduction to DL−LSM, aiming to explore the ability of DL models to localize the predefined lesion-symptom ROIs and capture their relationships. The first part of this experiment focused on a 3D CNN single-output model that used the artificial cognitive scores from the single-score simulation previously mentioned (refer to *Section B, Artificial cognitive scores*: *Equation* 1) to demonstrate that DL models can capture one-to-one lesion-cognition relationships. The 3D CNN model was trained for 60 epochs, and Occlusion was selected as the explainability method to obtain an attribution map for each patient.

The second part of the experiment was to test the ability of two DL models to predict multiple cognitive scores and map these to multiple lesion locations. Therefore, it involved designing two 3D multi-output models: the CNN and the Residual CNN architectures. Each patient contained three artificial cognitive scores, as shown in *Equation* 2 of *Section B, Artificial cognitive scores*. Both models were compared to investigate whether increasing model complexity is more effective in identifying strategic lesion locations that are associated with artificial cognitive scores. The CNN was trained for 70 epochs while the Residual CNN was trained for 90 epochs due to the deeper network. Attribution maps were obtained using Gradient SHAP and Occlusion to analyze the differences between using two XAI methods, with one map generated for each score, and therefore three maps per patient.

2) Noise experiment

The purpose of the noise experiment was to replicate the noisy nature observed in brain-behavior relationships and generate a more realistic validation of the lesion-cognition interactions. This can provide insight into the robustness of DL models to increasing levels of noise in the artificial cognitive scores.

To conduct this experiment, it was necessary to introduce noise into the artificial cognitive scores. The first step involved creating noise distributions, which were then added to the artificial cognitive scores. Each score was associated with a distinct noise distribution. Therefore, before obtaining these distributions, three standard deviations were calculated based on each proof-of-concept artificial cognitive score to ensure that the generated noise accurately reflected the inherent variability within each score. Each noise distribution was then obtained following a Gaussian distribution as shown in *Equation* 3, with a zero mean and its corresponding standard deviation.

$$\text{Noise}_j \sim \mathcal{N}(0, \sigma_j), \quad i = 1, 2, \ldots, K \tag{3}$$

where $\mathcal{N}$ is the Gaussian distribution with 0 mean and σ standard deviation, K is the number of patients in the dataset, and $j = 1, 2, 3$ is the distribution of the j-th score.

Next, noisy artificial cognitive scores were generated based on the formula described in Pustina et al. (2018) [15] and normalized to a range of 0-1, where a higher score indicated more lesion load inside the ROI. Each patient was then assigned three different artificial scores, one for each corresponding ROI.

$$\text{score}_{i,j} = (1 - a) \times (\text{ROI}_j \text{ lesion-load})_i + a \times \text{Noise}_{j,i} \tag{4}$$

where $0 \leq a \leq 0.5$ denotes the noise weight, i represents the i-th subject, $j = 1, 2, 3$ is the index for the score and ROI, and Noise is the noise distribution of the j-th score.

Five noise simulations produced five datasets of noisy artificial cognitive scores using *Equation* 4, characterized by different noise levels. The noise levels were injected into the scores by setting the noise weight 'a' to values between 0.1 and 0.5 in steps of 0.1. The noise weight was applied to the noise distribution and the remaining unity portion to the lesion load of the ROI. For instance, in the simulation with a noise level of 0.3, 30% was assigned to the noise distribution, while the remaining 70% to the ROI lesion load. This experiment was conducted for both 3D multi-output models, CNN and Residual CNN. Each architecture was trained and tested independently per noise level leading to a total of 10 models, five per architecture. The multi-output CNNs were trained for 90 epochs, while the Residual CNNs were trained for 110 epochs. GradientSHAP was used as the XAI method to obtain one attribution map for each score.

3) Intercorrelation experiment

The intercorrelation experiment was conducted to simulate a more plausible behavior of lesion-cognition interactions. It is important to note that lesions in different areas of the brain are not completely independent of each other, but rather exist interdependencies or correlations between them. Accordingly, this experiment was designed to investigate interdependencies between lesion loads in different ROIs and to evaluate whether a 3D multi-output DL model can identify these cognitive intercorrelations. Specifically, the purpose was to evaluate the minimal correlation the model could detect between ROIs and whether the measured attribution could serve as a linear predictor of the predefined ROI intercorrelations.

To conduct this experiment, the intercorrelated artificial scores were generated by introducing dependency between the lesion load of different ROIs. Score 1 always depended solely on ROI 1 as a baseline check. In this way, each patient was also assigned three artificial cognitive scores:

$$\begin{aligned} \text{s}_1 &= \text{lesion load of ROI}_1 \\ \text{s}_2 &= b \times \text{ROI}_2 \text{ lesion-load} + (1 - b) \times \text{ROI}_3 \text{ lesion-load} \\ \text{s}_3 &= b \times \text{ROI}_3 \text{ lesion-load} + (1 - b) \times \text{ROI}_1 \text{ lesion-load} \end{aligned} \tag{5}$$

where $0.6 \leq b \leq 0.9$ is the ROI-contribution weight.

Four intercorrelation simulations were conducted to produce four datasets of intercorrelated artificial cognitive scores using *Equation* 5. The correlations were injected into the scores by setting the ROI-contribution weight 'b' to values between 0.6 and 0.9 in steps of 0.1. Note that the score's ROI-contribution weight is always set to the corresponding ROI. Therefore, when generating the scores with a 0.6 ROI-contribution weight, Score 2 will contain 60% of the lesion load from ROI 2 and the remaining 40% from ROI 3. Similarly, Score 3 will have 60% from ROI 3 and the remaining 40% from ROI 1. The DL model selected for this experiment was the CNN 3D multi-output architecture. It was trained and tested independently for each ROI-contribution level, resulting in four models. All models were trained for 70 epochs, and the attribution maps were computed using Gradient SHAP.

E. Analysis and evaluation

To evaluate each model, the score predictions and ground truths across all test folds and patients were concatenated, and the coefficient of determination or R^2 was used to quantify the predictive performance. R^2 was calculated using the *r2_score* function from *sklearn.metrics*, measuring the DL model's ability to predict artificial cognitive scores.

The next step involved evaluating the ability of XAI techniques to highlight the relevant brain regions that are critical for these artificial behaviors. Hence, for the proof-of-concept and noise experiments, ROI identification was evaluated by precision and recall (PR) in the final attribution map of the entire dataset. To achieve this, attribution maps were compared to a binary image of the corresponding ROI. PR curves were generated by binarising the attribution maps with a thousand different thresholds from maximum to minimum image value. For each attribution map, PR curves were obtained by plotting precision against recall obtained at every possible threshold. Precision and recall were calculated by [26]:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

with true positives (TP), false positives (FP), and false negatives (FN).

On top of that, the area under the curve (AUC) was calculated for each PR curve to provide a more direct indication of performance. Lastly, when multiple XAI methods were computed for the same experiment, the Pearson correlation was calculated between the attribution maps to evaluate their similarity.

The PR evaluation was not applicable for the intercorrelation experiment due to the interdependence between the ROIs in the artificial cognitive scores. In this experiment, XAI maps should contain attribution values across multiple ROIs. To capture this interdependent behavior, an ROI-contribution rate (RC) was calculated for each ROI with respect to each artificial cognitive score. The first step was to obtain the RC score, which involved identifying the common voxels between the attribution map and the ROI under examination, as shown in *Equation* 8. This involves a simple multiplication between the XAI map and the ROI binary mask. Secondly, in order to mitigate the influence of the background voxels present in the attribution map, the relative contribution of each ROI was calculated. As a result, this normalization procedure yielded the final three RC rates for each artificial score, providing a quantitative measure of overlap that accurately reflects the proportion of each ROI's contribution to the artificial cognitive score, and thus the interdependence between ROIs.

$$\text{RC score}_j = \frac{\sum_i (\text{Attribution map}_j \times \text{ROI mask}_j)}{\sum_i \text{Attribution map}_j + \sum_i \text{ROI mask}_j} \tag{8}$$

where $j = 1, 2, 3$ is for the j-th ROI and j-th score.

III. Results

Before conducting each experiment, a simple analysis was performed on the ROIs, shown in Table II. The analysis involved calculating the lesion prevalence within each ROI and correlating it with the total lesion volume to comprehend the experimental results. ROI 2 presented the lowest lesion prevalence and weakest correlation, while ROI 3 exhibited the highest correlation with total lesion volume, and ROI 1 had the highest lesion load prevalence.

TABLE II
ROIs Specifications

ROI	Spearman Correlation	Lesion prevalence
ROI 1	0.77	70614
ROI 2	0.71	51142
ROI 3	0.78	67549

Spearman correlation between the ROI's lesion load and the total lesion volume, and the lesion prevalence in the ROIs.

A. Proof-of-concept experiment

The CNN single-output model demonstrated a strong predictive performance ($R^2 = 0.94$) when using the proof-of-concept artificial cognitive scores. Fig. 4. A displays two slices of the attribution map computed using Occlusion. The PR curve was then calculated, resulting in an AUC of 0.62 (Fig. 4. B).

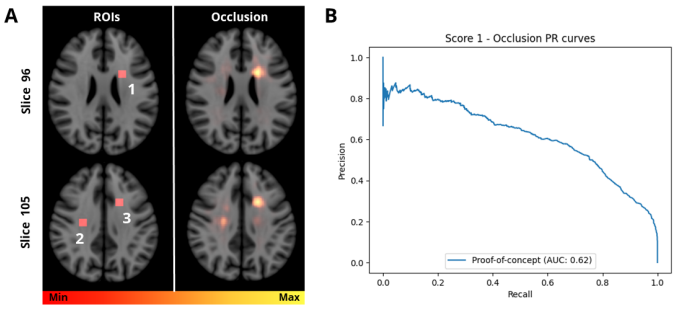

Fig. 4. (A) Attribution map obtained from the single-output CNN model using Occlusion and (B) PR curve of the corresponding attribution map. Slices are in the MNI-152 space.

Fig. 5 illustrates the attribution maps for both the CNN and Residual CNN architectures, comparing Occlusion and Gradient Shap methods for each score in the multi-output proof-of-concept experiment. The attribution maps produced by Gradient Shap were more patchy and noisy than those produced by Occlusion. The PR curves obtained from the attribution maps, along with their respective AUC, consistently showed lower values when using Gradient Shap, as shown in Fig. A.1 (see *Appendix* A). Furthermore, the CNN had a slightly lower sensitivity to Score 1 compared to

the Residual CNN, but a much higher sensitivity to Score 2 when identifying their corresponding ROIs. ROI 3 was the most accurate location identified and was similar for both XAI methods. Although Gradient Shap produced slightly worse results when quantifying the attribution maps with the PR curves and AUC, it was substantially faster than Occlusion. Additionally, the attribution maps from the CNN were similar between the two XAI methods for each score, with map-wise correlation coefficients of 0.846, 0.862, and 0.902 for Score 1, Score 2, and Score 3, respectively. In contrast, the correlation between both XAI methods for the Residual CNN was somewhat lower, with a coefficient of 0.721, 0.781, and 0.870, respectively. Due to the time-consuming computation of Occlusion, the attribution maps in the following experiments were generated using Gradient Shap.

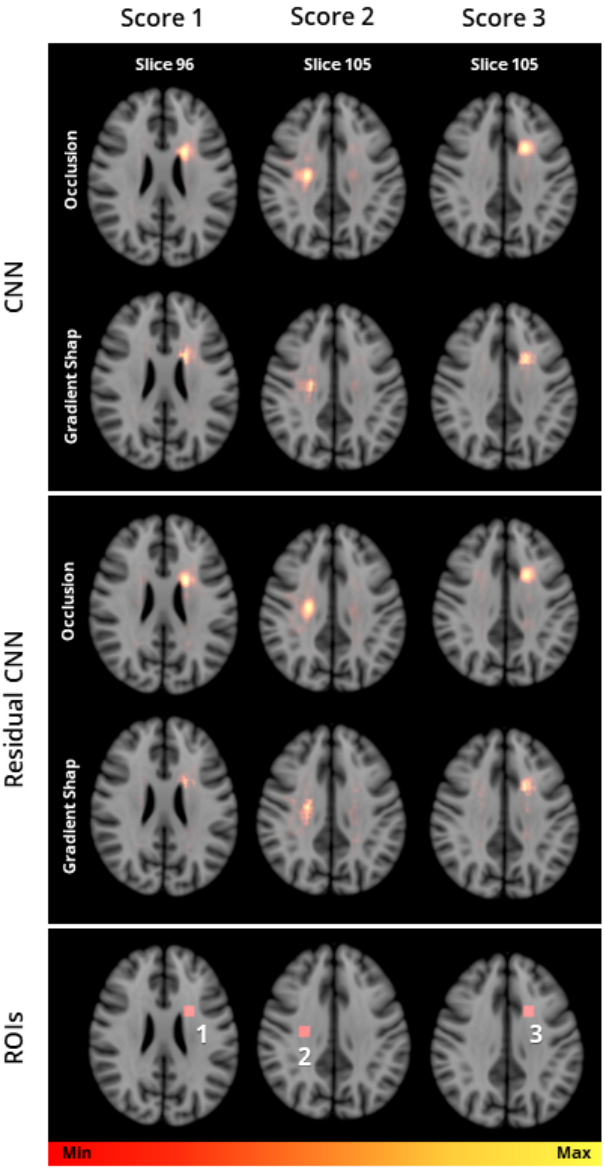

Fig. 5. Attribution maps obtained from the multi-output CNN and Residual CNN models using Occlusion and Gradient Shap in the proof-of-concept experiment. Slices are shown in the MNI-152 space.

Table III presents the predictive performance for the proof-of-concept experiment, quantified by the R^2. The CNN demonstrated superior predictive performance for each score compared to the Residual CNN. Additionally, the attribution maps obtained for each of the scores can be seen in the first row of Fig. A.2 and Fig. A.3 (refer to *Apendix* A) for Residual CNN and CNN, respectively. Finally, the PR curves calculated based on the attribution maps are shown in blue in Fig. 6. The results of the proof-of-concept attribution maps are consistent with the predictive performance of the model, as they show an accurate overlap between the model's attributions of important regions and the actual ROIs.

TABLE III
PREDICTIVE PERFORMANCE FOR PROOF-OF-CONCEPT (POC) AND NOISE EXPERIMENTS

		CNN			**Residual CNN**		
		Score 1	Score 2	Score 3	Score 1	Score 2	Score 3
Simulations	POC	0.909	0.907	0.941	0.881	0.847	0.934
	Noise 10%	0.815	0.798	0.842	0.848	0.805	0.838
	Noise 20%	0.416	0.550	0.645	0.522	0.528	0.626
	Noise 30%	-0.041	-0.281	-0.742	-0.012	-0.133	-0.697
	Noise 40%	-0.538	-1.326	-0.126	-0.590	-1.269	-0.125
	Noise 50%	-3.376	-3.138	-0.920	-3.346	-3.305	-0.960

Predictive performance quantified by R^2. Each noise model is trained on a different simulation of artificial cognitive scores, obtained by injecting the percentage of noise represented.

B. Noise experiment

Table III also presents the predictive performance of the noise experiment at different noise levels. The results show a decreasing trend in predictive performance as the percentage of noise within the artificial cognitive scores increases. Additionally, attribution maps for the Residual CNNs can be seen in Fig. A.2, and for the CNN in Fig. A.3 (see *Appendix* A). The results show that the attribution maps become more noisy as the noise level of the score increases. Fig. 6 shows the PR curves and AUC values for all attribution maps obtained in the noise experiments. The PR curves reflect this decreasing trend and show the effect of the noise on the AUC, which decreases to almost a zero AUC at 50% noise in each of the scores.

C. Intercorrelation experiment

Table IV presents the predictive performance of each model on the intercorrelation experiment for the three scores, quantified by R^2. All CNNs demonstrate high predictive performance, with Score 1, Score 2, and Score 3 having approximate R^2 values of 0.92, 0.93, and 0.95 respectively, at each ROI-contribution level.

Figure 7 displays the RC rates of relative ROI contribution to each score, calculated for the corresponding

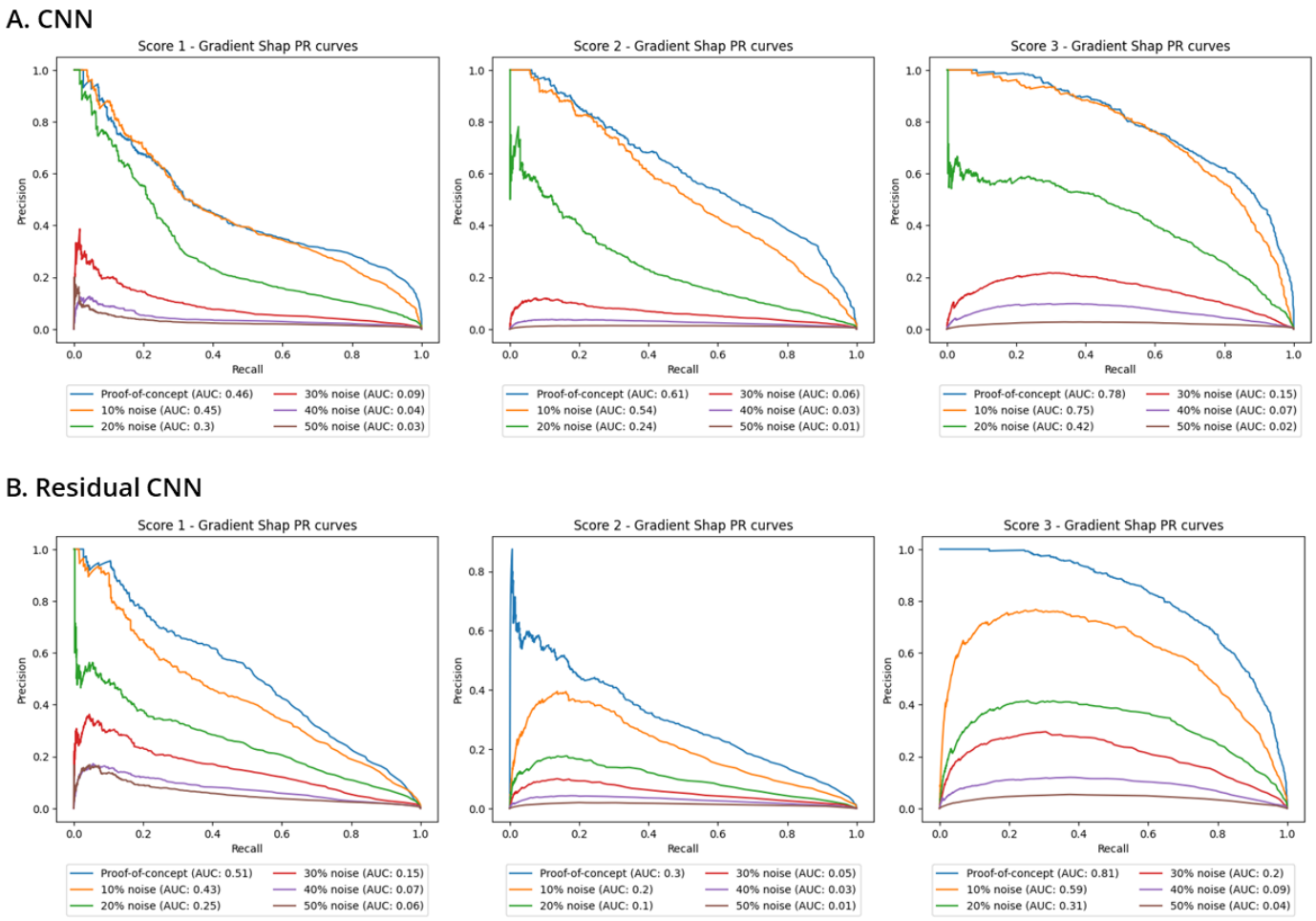

Fig. 6. PR curves for the proof-of-concept and noise experiment models for the (A) CNN and (B) Residual CNN; AUC per PR curve is shown in the legend.

attribution maps at each model. The results show that interdependencies are identified at almost all levels of ROI-contribution. The intercorrelation between ROIs in Score 3 was accurately identified at each level; each ROI was correctly assigned its relative contribution. However, identifying the intercorrelation between the detected ROIs in Score 2 was the most difficult. The RC rates measured for each ROI differed from the predefined ROI-contribution weights in Score 2. For instance, at an 80% level, ROI 2 had an RC rate of approximately 60%, ROI 3 had 32%, and ROI 1 had 8%, which should have been 80% contribution from ROI 2 and 20% from ROI 3. In addition, at an ROI-contribution level of 60%, the relative contributions were reversed, with ROI 3 having more influence on Score 2 than ROI 2. Furthermore, Score 1 exhibited a minor dependence on ROI 3, despite the fact that it is only dependent on ROI 1 and not on any other ROIs. The attribution maps for each of the CNN models are presented in Fig. A.4 (see Appendix A). These maps accurately identify the corresponding ROIs that influence each of the scores. Additionally, it is apparent that the Score 2 attribution map, at 60%, mainly highlights the three ROIs.

TABLE IV
PREDICTIVE PERFORMANCE FOR INTERCORRELATION EXPERIMENT

	Intercorrelation Simulations			
	90%	80%	70%	60%
Score 1*	0.920	0.921	0.925	0.935
Score 2†	0.927	0.912	0.939	0.929
Score 3‡	0.951	0.943	0.956	0.953

Predictive performance quantified by R^2. Each noise model is trained on a different simulation of artificial cognitive scores, characterized by the percentage of ROI-contribution 'b' represented. *Score 1 = ROI 1; †Score 2 = $b \times ROI2 + (1-b) \times ROI3$; and ‡Score 3 = $b \times ROI3 + (1-b) \times ROI1$.

IV. DISCUSSION

In this study, a novel approach to lesion-symptom mapping is introduced, which was validated across a range of potential brain-cognition relationships in a simulation study. The simulation study incorporates noise and cognitive dependence on multiple lesion locations. The approach aims to expand upon state-of-the-art techniques such as

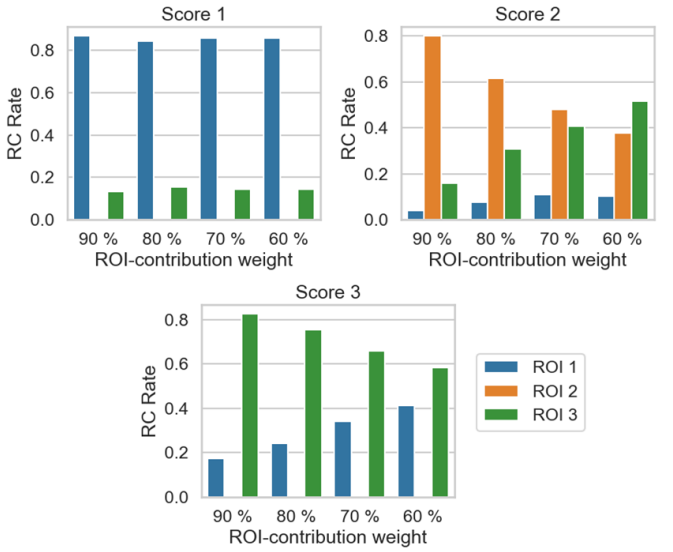

Fig. 7. RC rates for the intercorrelation experiment CNNs. The x axis represents each of the models by the percentage of their ROI-contribution weight b, used to generate their respective artificial cognitive scores. Score 1 is only dependent on ROI 1, Score 2 = $b \times ROI2 + (1-b) \times ROI3$ and Score 3 = $b \times ROI3 + (1-b) \times ROI1$

SVR−LSM, allowing the simultaneous consideration of multiple cognitive scores and lesion types when studying the cognitive impact of vascular lesions. The study has demonstrated that explainable artificial intelligence can identify specific brain lesion locations responsible for multiple simulated scores from a multi-output deep learning model in 3D WMH segmentations of patients with SVD. One major advantage of using DL−LSM over SVR−LSM is that the attribution maps produced by the XAI methods are aggregations of the attribution maps of individual patients, whereas the SVR-LSM method directly produces group-based β maps. Therefore, DL−LSM provides personalized insights for diagnosis and prognosis, enabling targeted analysis and post-processing to reveal patterns that may be missed by group-level analysis.

The initial step involved conducting a proof-of-concept single-output experiment to validate that LSM can be achieved through DL and XAI. Accordingly, the CNN model demonstrated strong predictive performance with an R^2 value of 0.94 in predicting artificial cognitive scores based on the 3D WMH segmentations. The attribution map supports the findings, as the highlighted areas corresponded precisely to the ROIs. However, the attribution map showed fewer attribution values in Score 2, possibly due to its lower lesion prevalence across all ROIs (Table II). Furthermore, some noise was detected on the contralateral side of the brain where the ROIs are located. This could be attributed to the high degree of symmetry of the WMH lesions that can be seen in Fig. 1. In addition, the differences in the PR curve, which quantifies the attribution map, compared to the visually observed overlap between the XAI map and the ROIs, are attributed to the sensitivity of this metric to false positives. The reason for this is that the attribution maps contain a considerable number of false positives in the voxels adjacent to the ROIs.

The second step was to expand the proof-of-concept experiment to a multi-output model using two CNN regression architectures. The CNN showed better predictive performance for each score than the Residual CNN architecture when comparing the R^2 values. However, when analyzing the PR curves (refer to Fig. 5 in *Appendix* A), the CNN had lower sensitivity to Score 1 compared to the Residual CNN. Additionally, both models performed best in Score 3, despite ROI 3 being located in proximate areas to ROI 1. When using the CNN, Score 2 showed similar predictive performance as Score 1. ROI 2 was accurately identified by both XAI methods, with AUC values of 0.61 and 0.7 in Gradient Shap and Occlusion, respectively. However, it is worth noting that the Residual CNN had the lowest predictive performance for Score 2. As a result, the PR curves of the Score 2 attribution maps showed a quick decay, with an AUC of 0.3 and 0.51 for Gradient Shap and Oclusion respectively. Based on the specifications of the ROIs in Table II and the findings of both models, it appears that DL models may be more sensitive to Score 3 due to ROI 3 having the highest correlation with the total lesion volume. Furthermore, the lower sensitivity of Residual CNN to Score 2 may be attributed to the difficulty in identifying strategic lesion locations with lower lesion prevalence, such as ROI 2. Instead, the model may be focusing more on the total lesion volume.

Zhang et al. (2014) [9] performed a multivariate SVR−LSM using different approaches. One part of the study aimed to compare the results of applying total lesion volume control to the lesion images with those obtained without it. Total lesion volume is already predictive of cognition, but there are areas in the brain that have a stronger relationship with cognition than the total volume alone [3]. Therefore, the aim was to minimize the impact of the total lesion volume, making it easier to identify strategic lesion locations. Accordingly, the ROC curves calculated on the β maps showed an AUC of 0.9405 when the correction was applied, compared to an AUC of 0.7574 when it was not applied. This indicates that the sensitivity to identify the specific lesioned areas responsible for the scores was increased with the total lesion volume correction. Moreover, it suggests that when using SVR-LSM this type of correction needs to be applied to obtain more accurate results. In this work, the total lesion volume control was not applied, as it is hypothesized that DL models are capable of learning this correction themselves. The attribution maps show that the highest attribution values always overlap within the areas of the ROIs. This reduces false positives in other WMH lesion areas that are more correlated to the total lesion volume and have a higher lesion prevalence across patients (refer to Figure 1). The results suggest that the total lesion volume may have less impact on the Dl−LSM results. However, as previously mentioned, the multi-output model is more sensitive to Score 3 than the other two scores, possibly

due to its stronger correlation with total lesion volume. This could be partly due to the lack of total lesion volume control on the WMH segmentations, which causes the model to base part of its decision on total lesion volume. To improve the sensitivity of the model to Score 1 and 2, future research could explore the implementation of lesion volume control before applying the DL−LSM technique. It would be valuable to explore how this modification affects the outcome and whether there are significant differences in ROI identification when performing DL−LSM with and without correction.

When comparing the XAI techniques, both demonstrated their ability to identify whether a lesion region is directly related to a cognitive score. It is clear that Gradient Shap produces more patchy and noisy images due to its sensitivity to small variations in the input image (WMH lesions are highly variable between patients). In contrast, Occlusion maps provide a clearer representation of the ROI location, but at the cost of a much longer computation time. It should be noted that both methods use different mathematical calculations and the resulting attribution maps are strongly influenced by the parameters chosen. In this study, Gradient Shap was chosen as the XAI method due to its significantly lower computational time compared to Occlusion. However, future experiments should also consider Occlusion since Score 2 showed a notable difference in the Residual CNN performance when both XAI methods were used. Occlusion achieved an AUC of 0.51, while Gradient Shap only achieved 0.3, indicating a substantial difference in accuracy.

In the following experiment, noise levels were added to the synthetic lesion-behavior relationships. Both architectures exhibited a decrease in predictive performance as the noise level in the artificial cognitive scores increased. At a noise level of 10%, the PR curves of the CNNs remained the same as the proof-of-concept PR curves. This shows that XAI can still identify accurately the ROIs responsible for the scores, indicating robustness to low levels of noise. In contrast, Residual CNN showed the opposite pattern, with a preserved predictive performance at lower noise levels but decreased accuracy in attribution maps and PR curves, particularly in Score 3, where the AUC dropped a third of its value. These findings suggest that, at low levels of noise, XAI methods are more effective in identifying strategic lesion locations when using the CNN, despite its lower predictive performance compared to the Residual CNN, possibly due to its simpler architecture. Both architectures lost half of their predictive performance at a level of 20% of noise, and, as expected, the model's predictive performance dropped drastically at noise levels above 30%, with negative R^2 values. These negative values indicate that the regression models do not follow the trend of the data and that the predictions are worse than those obtained by simply using the average of the artificial scores as the predictor. Attribution maps support this behavior, with the highest attribution values decreasing in ROI regions and appearing in other areas. The lesion prevalence of all patients shown in Fig 1 indicates that the patients' WMH lesions are more frequently distributed around the periventricular areas and parietal lobes, which correspond to those identified by the XAI method at higher noise levels. This suggests that the model relies more on total lesion volume at higher noise levels and loses its ability to accurately identify the ROIs.

The intercorrelation experiment was conducted using the CNN model architecture because it showed greater stability across all scores compared to the Residual CNN in previous experiments. The attribution maps for Score 1 and Score 3 accurately identified the specific ROIs that contributed to their respective scores. However, the attribution maps highlighted the correct ROIs associated with Score 2 up to a 70% contribution, where the ROI 1 becomes visible. In relation to the RC rates that quantify the attribution maps, it was found that in Score 1, which depends only on ROI 1, there was a slight correlation with ROI 3. However, in Score 3, which depends on both ROIs, the CNN accurately identified the minimal contribution of ROI 1 and ROI 3 at each level. The RC rates for Score 2 did not match the predefined ROI-contribution weights, but they captured more dependency in the correct ROI. The contributions were slightly different from the actual values, and there was always a contribution of ROI 1, which may be due to its higher lesion prevalence. The RC rate only showed a greater dependency on ROI 3 than ROI 2 at the 60% level, suggesting inaccuracies in the CNN's decision-making process at that level. These findings are in line with previous results, indicating that Score 3 consistently outperforms other scores. In this case, it has a higher level of dependency due to the strongest correlation of ROI 3 to the total lesion volume. Meanwhile, Score 2 is the most challenging due to the lowest prevalence of lesion load in ROI 2. All things considered, these results demonstrate that DL−LSM can identify intercorrelations between strategic lesion locations. The model detected intercorrelations between ROIs from an ROI-contribution level of 70% up to 90%, meaning that these measured RC rates can serve as linear predictors of the predefined ROI intercorrelations.

One limitation of the study concerns the XAI maps. Although they provide valuable insights into the model's decision-making, they can be misleading if not appropriately validated. In this case, the performance of the XAI methods could be assessed due to the presence of a ground truth. Therefore, the study was certain that the identified areas by the XAI methods corresponded with the predefined ground truth ROIs. In situations where the ground truth is not available in real patient data, the reliability and robustness of XAI should be evaluated by examining the consistency across different models or datasets, performing sensitivity analyses, or relying on clinician experience. Additionally, the XAI method itself must be chosen carefully, as the multi-output proof-of-concept experiment with Gradient Shap and Occlusion showed substantial differences in LSM.

Future research should consider applying these methodologies to real patient data, specifically cognitive outcomes, to study if these types of DL model architectures are suitable for performing LSM in real scenarios. This could provide insight into the accuracy of these simulations and test whether they reflect real lesion-cognition relationships. Moreover, future research should also consider developing a multi-input multi-output DL model to better capture the cognitive impact of vascular lesions caused by SVD. SVD often causes multiple brain lesions, which are intercorrelated and can affect various cognitive domains. Furthermore, future work may include assigning varying weights to the ROIs, altering their morphologies, or applying nonlinear functions to the lesion-behavior relationships, such as logarithmic or exponential functions. These simulations might better reflect the complex relationship between WMH and cognitive scores. Additionally, a potential experiment could be to test various noise simulations of artificial cognitive scores within the same DL model. In clinical settings, each patient may present varying relations between their cognitive scores and MR-visible vascular lesions, thus it is important to ensure that models are robust enough to handle various scenarios. All in all, these research directions aim to improve the ability of DL−LSM to understand the complex mechanisms that cause cognitive impairment associated with SVD and to make DL−LSM models more practical in clinical settings by ensuring their resilience to different scenarios.

V. Conclusion

Cerebral small vessel disease presents a challenge in clinical neuroscience due to its complex manifestation and impact on cognitive function. The aim of this study was to introduce a novel approach to LSM using DL and XAI techniques. The study demonstrates that DL algorithms and XAI techniques are suitable for predicting multiple artificial cognitive scores and identifying the strategic lesion locations that affect these scores in 3D WMH lesions of patients with SVD. The findings reveal that the DL models remain robust to artificial cognitive scores with up to 20% noise, after which their performance declines. Furthermore, it demonstrates that intercorrelations between different lesioned areas associated with artificial cognitive scores can be identified using these types of models. DL−LSM can help in understanding the underlying brain mechanisms that lead to neurological dysfunction and provide improved LSM techniques. Additionally, it could serve as personalized medicine, allowing clinicians to tailor interventions to the needs of individual patients through the ability of XAI methods to generate attribution maps of individual patients. In conclusion, DL−LSM has the potential to improve our understanding and management of cognitive impairment associated with SVD.

References

[1] Marco Duering et al. “Neuroimaging standards for research into small vessel disease—advances since 2013”. In: *The Lancet Neurology* 22.7 (2023), pp. 602–618.

[2] Joanna M Wardlaw, Colin Smith, and Martin Dichgans. “Small vessel disease: mechanisms and clinical implications”. In: *The Lancet Neurology* 18.7 (2019), pp. 684–696.

[3] Fanny Munsch et al. “Stroke location is an independent predictor of cognitive outcome”. In: *Stroke* 47.1 (2016), pp. 66–73.

[4] J Matthijs Biesbroek, Nick A Weaver, and Geert Jan Biessels. “Lesion location and cognitive impact of cerebral small vessel disease”. In: *Clinical Science* 131.8 (2017), pp. 715–728.

[5] J Matthijs Biesbroek et al. “High white matter hyperintensity burden in strategic white matter tracts relates to worse global cognitive performance in community-dwelling individuals”. In: *Journal of the Neurological Sciences* 414 (2020), p. 116835.

[6] Dorian Pustina and Daniel Mirman. *Lesion-to-symptom mapping: principles and tools*. Vol. 180. Springer Nature, 2022.

[7] Nick A Weaver et al. “Strategic infarct locations for post-stroke cognitive impairment: a pooled analysis of individual patient data from 12 acute ischaemic stroke cohorts”. In: *The Lancet Neurology* 20.6 (2021), pp. 448–459.

[8] Elizabeth Bates et al. “Voxel-based lesion-symptom mapping”. In: *Nature neuroscience* 6.5 (2003), pp. 448–450.

[9] Yongsheng Zhang et al. “Multivariate lesion-symptom mapping using support vector regression”. In: *Human brain mapping* 35.12 (2014), pp. 5861–5876.

[10] Christopher M Filley. “White matter and behavioral neurology”. In: *Annals of the New York Academy of Sciences* 1064.1 (2005), pp. 162–183.

[11] Sucheta Chauhan et al. “A Comparison of Shallow and Deep Learning Methods for Predicting Cognitive Performance of Stroke Patients From MRI Lesion Images”. In: *Frontiers in Neuroinformatics* 13 (2019). URL: https://api.semanticscholar.org/CorpusID:198352944.

[12] Sandra Vieira, Walter Hugo Lopez Pinaya, and Andrea Mechelli. “Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications”. In: *Neuroscience & Biobehavioral Reviews* 74 (2017), pp. 58–75. URL: https://api.semanticscholar.org/CorpusID:207093716.

[13] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.

[14] Bas HM van der Velden. “Explainable AI: current status and future potential”. In: *European Radiology* (2023), pp. 1–3.

[15] Dorian Pustina et al. “Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations”. In: *Neuropsychologia* 115 (2018), pp. 154–166.

[16] Maria V Ivanova et al. “An empirical comparison of univariate versus multivariate methods for the analysis of brain-behavior mapping”. In: *Human Brain Mapping* 42.4 (2021), pp. 1070–1101.

[17] Christoph Sperber, Chloé Nolingberg, and Hans-Otto Karnath. *Post-stroke cognitive deficits rarely come alone: Handling comorbidity in lesion-behaviour mapping*. Tech. rep. Wiley Online Library, 2020.

[18] Jooske Marije Funke Boomsma et al. “Vascular cognitive impairment in a memory clinic population: rationale and design of the “Utrecht-Amsterdam clinical features and prognosis in vascular cognitive impairment”(TRACE-VCI) study”. In: *JMIR research protocols* 6.4 (2017), e6864.

[19] Nick A Weaver et al. “Cerebral amyloid burden is associated with white matter hyperintensity location in specific posterior white matter regions”. In: *Neurobiology of aging* 84 (2019), pp. 225–234.

[20] Stefan Klein et al. “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1 (2009), pp. 196–205.

[21] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[22] Bas HM Van der Velden et al. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Medical Image Analysis* 79 (2022), p. 102470.

[23] Narine Kokhlikyan et al. “Captum: A unified and generic model interpretability library for pytorch”. In: *arXiv preprint arXiv:2009.07896* (2020).

[24] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[25] Matthew D Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*. 2013. arXiv: 1311.2901 [cs.CV].

[26] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer Science & Business Media, 2013, pp. 1–595. ISBN: 978-1-4614-6848-6. DOI: 10.1007/978-1-4614-6849-3. URL: http://link.springer.com/10.1007/978-1-4614-6849-3.

Appendix A

Supplementary data associated with this project.

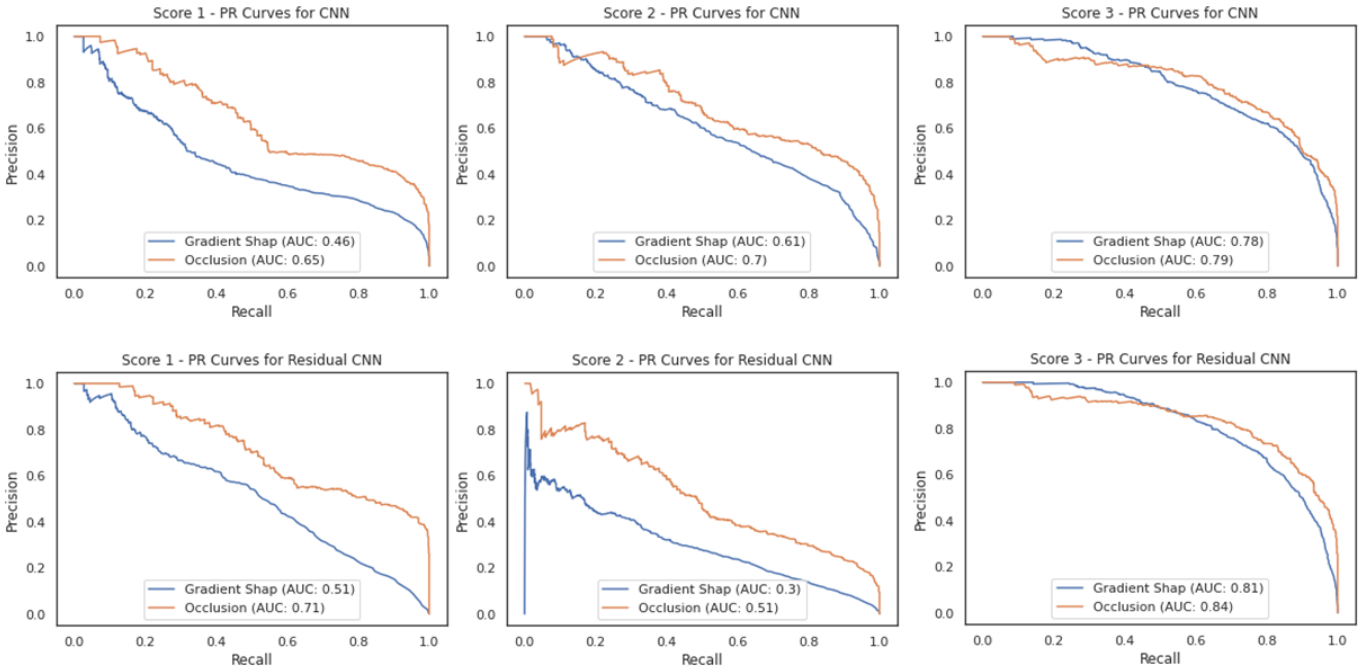

Fig. A.1. PR curves for each of the scores obtained from the CNN and Residual CNN models in the proof-of-concept experiment, using Occlusion and Gradient Shap for the attribution maps.

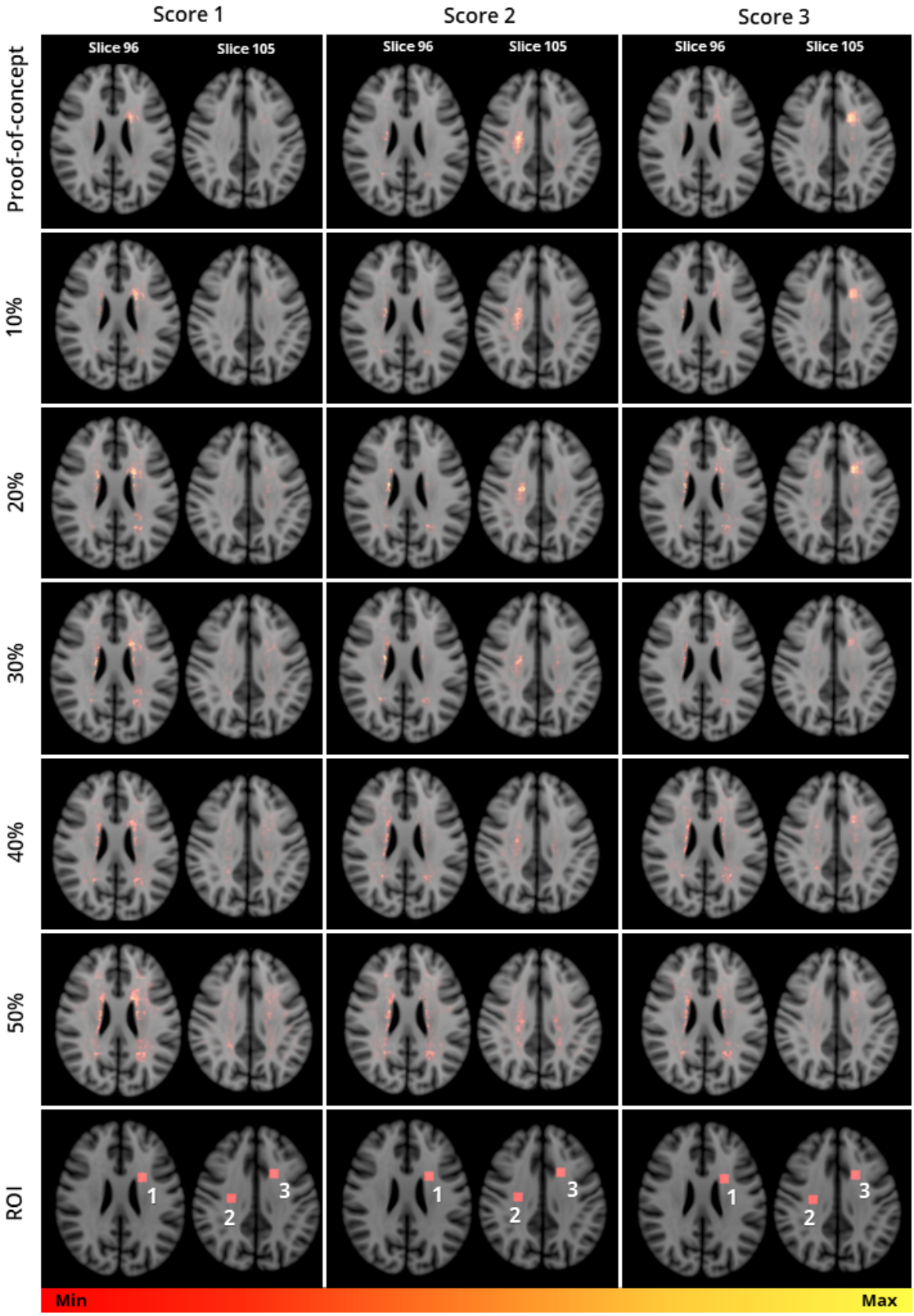

Fig. A.2. Attribution maps for each score generated by Gradient Shap for each Residual CNN model. The rows correspond to different models, starting with the proof-of-concept at the top, followed by the five noise experiment models. The percentages refer to the level of noise added to the artificial cognitive scores used to train the corresponding model. Slices are shown in the MNI-152 space

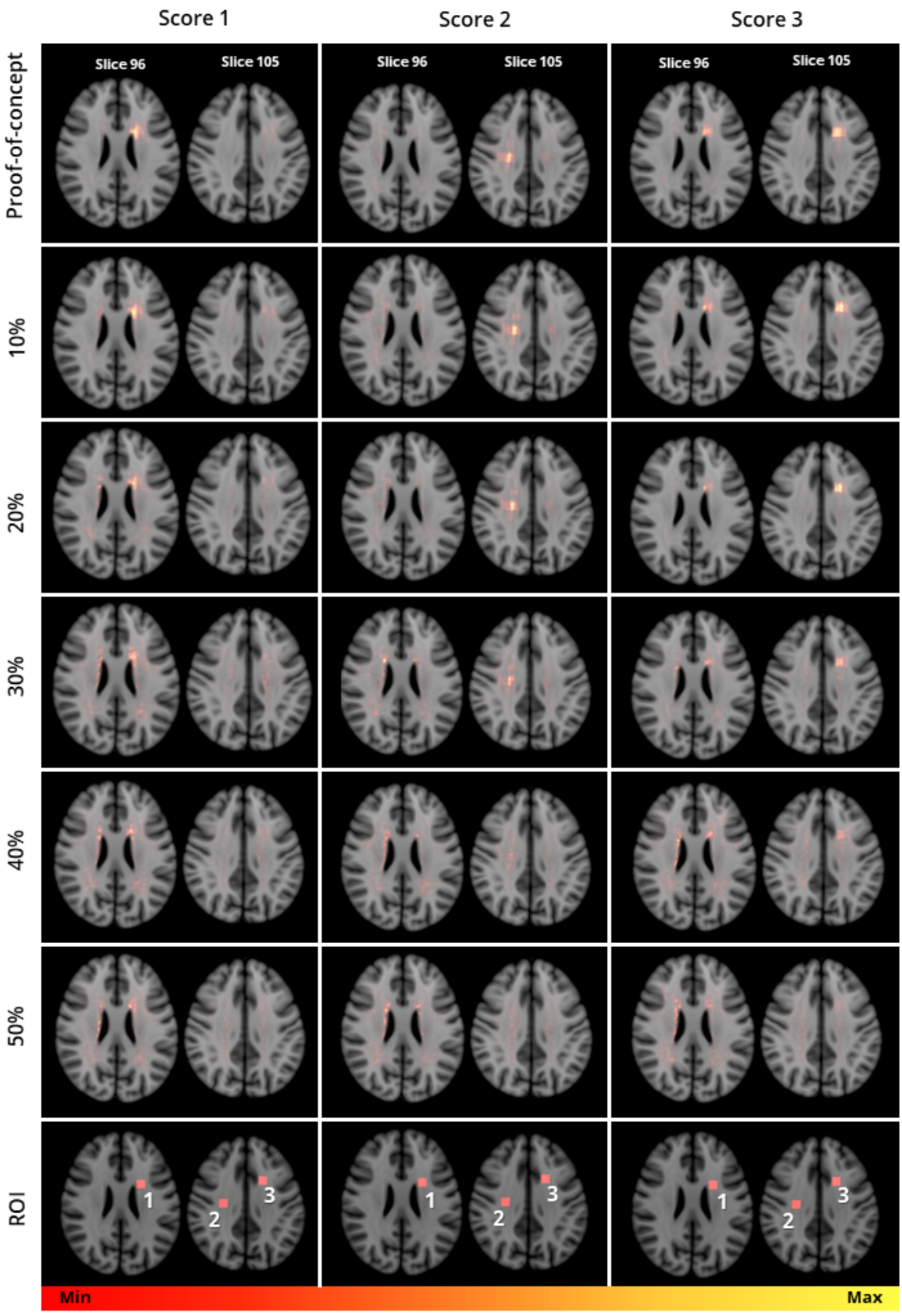

Fig. A.3. Attribution maps are provided for each score generated by Gradient Shap for each CNN model. The rows correspond to different models, starting with the proof-of-concept at the top, followed by the five noise experiment models. The percentages refer to the level of noise added to the artificial cognitive scores used to train the corresponding model. Slices are shown in the MNI-152 space

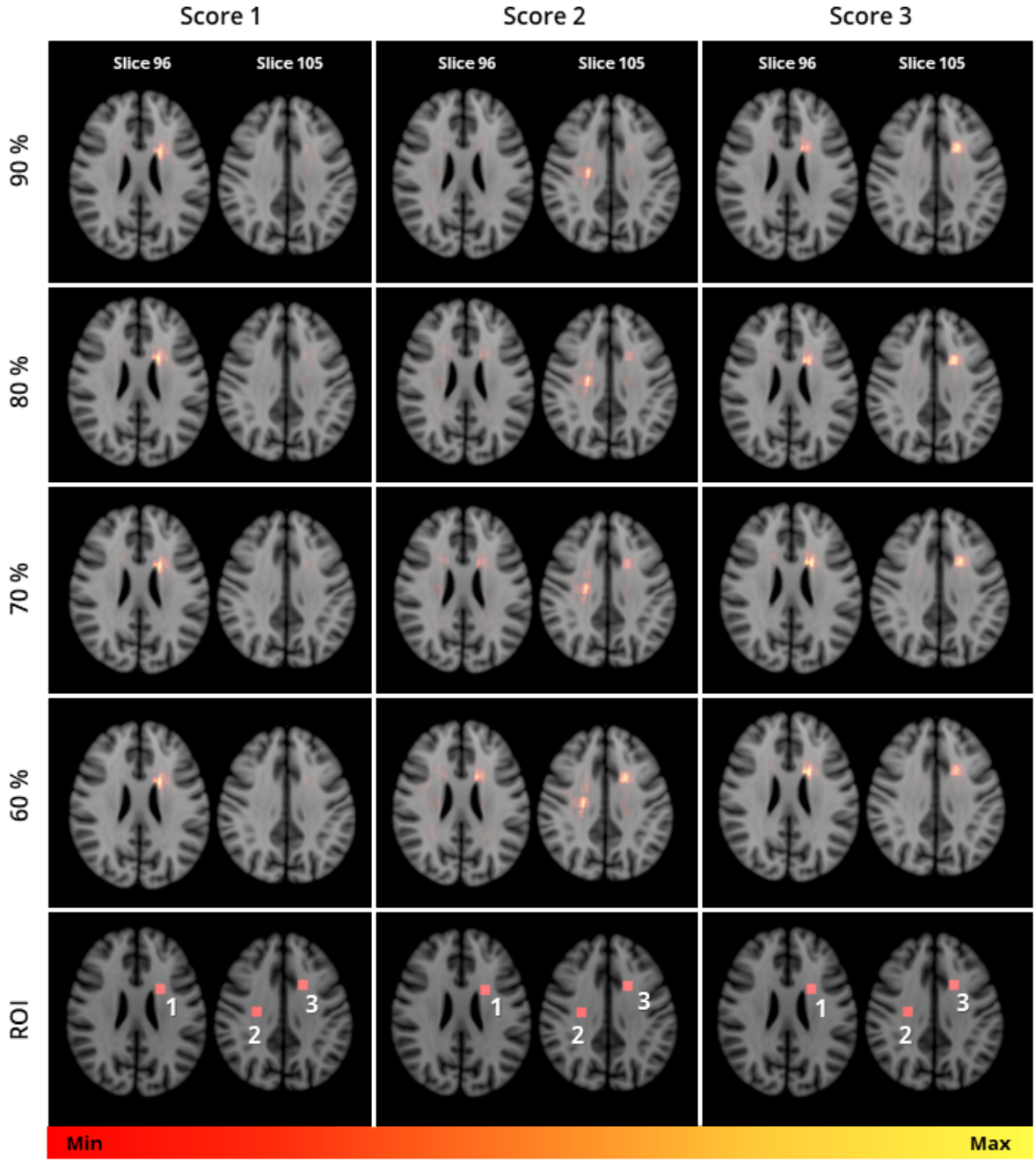

Fig. A.4. Attribution maps for each score generated by Gradient Shap for each CNN model used in the intercorrelation experiment. The percentages refer to each model and indicate the ROI-contribution 'b' used to generate the scores. Note that Score 1 = ROI 1; †Score 2 $= b \times ROI2 + (1 - b) \times ROI3$; and ‡Score 3 $= b \times ROI3 + (1 - b) \times ROI1$. Slices are shown in the MNI-152 space