Received: 24 February 2022 | Accepted: 26 September 2022

DOI: 10.1002/epi4.12655

Epilepsia Open® Open Access

ORIGINAL ARTICLE

Interpretable deep learning-based hippocampal sclerosis classification

Dohyun Kim[1] | **Jungtae Lee**[2] | **Jangsup Moon**[3,4] | **Taesup Moon**[5,6]

[1]Department of Artificial Intelligence, Sungkyunkwan University, Suwon, South Korea

[2]Application Engineering Team, Memory Business, Samsung Electronics Co., Ltd., Suwon, South Korea

[3]Department of Neurology, Seoul National University Hospital, Seoul, South Korea

[4]Department of Genomic Medicine, Seoul National University Hospital, Seoul, South Korea

[5]Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea

[6]ASRI/INMC/IPAI/AIIS, Seoul National University, Seoul, South Korea

Correspondence
Taesup Moon, Department of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea.
Email: tsmoon@snu.ac.kr

Jangsup Moon, Department of Genomic Medicine, Department of Neurology, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
Email: jangsup.moon@gmail.com

Funding information
Institute for Information and Communications Technology Promotion, Grant/Award Number: IITP-2021-0-02068 and 2021-0-01343; National Research Foundation, Grant/Award Number: NRF-2021M3E5D2A01024795; Seoul National University, Grant/Award Number: New Faculty Startup Fund

Abstract

Objective: To evaluate the performance of a deep learning model for hippocampal sclerosis classification on the clinical dataset and suggest plausible visual interpretation for the model prediction.

Methods: T2-weighted oblique coronal images of the brain MRI epilepsy protocol performed on patients were used. The training set included 320 participants with 160 no, 100 left and 60 right hippocampal sclerosis, and cross-validation was implemented. The test set consisted of 302 participants with 252 no, 25 left and 25 right hippocampal sclerosis. As the test set was imbalanced, we took an average of the accuracy achieved within each group to measure a balanced accuracy for multiclass and binary classifications. The dataset was composed to include not only healthy participants but also participants with abnormalities besides hippocampal sclerosis in the control group. We visualized the reasons for the model prediction using the layer-wise relevance propagation method.

Results: When evaluated on the validation of the training set, we achieved multiclass and binary classification accuracy of 87.5% and 88.8% from the voting ensemble of six models. Evaluated on the test sets, we achieved multiclass and binary classification accuracy of 91.5% and 89.76%. The distinctly sparse visual interpretations were provided for each individual participant and group to suggest the contribution of each input voxel to the prediction on the MRI.

Significance: The current interpretable deep learning-based model is promising for adapting effectively to clinical settings by utilizing commonly used data, such as MRI, with realistic abnormalities faced by neurologists to support the diagnosis of hippocampal sclerosis with plausible visual interpretation.

KEYWORDS

convolutional neural network, hippocampal sclerosis, interpretable AI, MRI

1 | INTRODUCTION

Temporal lobe epilepsy (TLE) is the most common type of focal epilepsy in adults and might be cured by surgical treatment. Hippocampal sclerosis (HS) is the histopathological hallmark and the essential underlying etiology of TLE.[1] Surgical treatment is the most effective method for alleviating the symptoms in patients with medial TLE with HS.[2,3] The typical MRI features of HS are hippocampal volume loss, increased signal intensity on T2-weighted imaging, and distortion of internal architecture.[4] However, in many cases, the structural hippocampal abnormalities in an MRI can be very subtle; therefore, it could be very challenging to accurately identify the epileptic hippocampus, even by expert epileptologists and neuroradiologists. It is well-known that interobserver agreement on HS is not perfect, ranging from approximately 70% to 90%.[5-7] Therefore, the presence of HS is often judged by combining the results of other imaging tests, such as FDG-PET and SPECT.

Machine learning is increasingly being applied to the medical field, especially to aid systematic diagnosis based on image analysis. In line with this trend, many attempts have been made to apply machine learning techniques for MRI-based HS classification.[8-11] Although some studies have demonstrated noticeable results, most of them used simple and conventional algorithms that led to a few drawbacks. First, they require complex and time-consuming data preprocessing steps to extract manually crafted features from the MRI images. Second, most previous work mainly focused on achieving high predictive accuracy but lacked useful interpretation or evidence for the prediction. Finally, all previous work used only healthy individuals with normal MRIs as a control group; such a setting may lead to an overestimation of the prediction accuracy when considering actual clinical applications. For example, discriminating HS from other abnormalities in the medial temporal region would be a critical challenge in practice.

To overcome the above limitations, we applied an end-to-end deep learning framework[12-14] to MRI-based HS classification. We took the entire brain MRI image with minimal preprocessing as an input to the model, letting the deep neural network learn useful features, as well as the classifier itself in an end-to-end manner. Second, we applied the state-of-the-art neural network interpretation method to highlight the most relevant regions for the prediction and visualized the learned features to validate the reliability and effectiveness of the applied deep learning framework. Finally, to reflect the real clinical setting we composed a more challenging and larger dataset that consists of a control group not only with normal participants but also with those having other structural abnormalities in the brain, including in the hippocampus.

Key points

- Deep learning-based HS classification achieved significant performance on a large dataset with a minimum preprocessing requirement.
- Practical applications are anticipated through the clinical control group and dual utilization of multiclass and binary classification.
- Visual interpretation suggested plausible grounds, which exhibited well correlated agreement with the known literature.

2 | METHODS

2.1 | Participant selection

Epilepsy patients who underwent a brain MRI epilepsy protocol between 1999 and 2020 at Seoul National University Hospital were considered for inclusion. A brain MRI epilepsy protocol was performed in patients when epilepsy was suspected or to exclude the possibility of epilepsy. The diagnostic criteria for selecting patients with HS used in the training data set were (a) patients diagnosed with ipsilateral TLE to the abnormal hippocampus and (b) when the neuroradiologist and the epileptologist in charge agreed to the presence of HS on brain MRI. Among the patients with HS used for the training data set, HS was pathologically confirmed by surgery in 38.1% (44.0% in left HS and 28.3% in right HS) of cases (Table S1). One hundred left TLE patients with HS (57 females; mean age 45.3 ± 13.3; range 18-78 years) and 60 right TLE patients with HS (42 females; mean age 47.4 ± 13.3; range 22-78 years) were included as the training Left and Right HS groups, respectively. Right HS patients were relatively less common compared with Left HS patients; hence, we had an imbalanced dataset. Furthermore, 160 participants (84 females; mean age 41.7 ± 15.7; range 19-86 years) were included in the No HS (control) group. Our No HS group was composed of three different types of participants: (a) with normal MRI findings (66 participants), (b) with extratemporal abnormalities (66 participants), and (c) with abnormal findings in the medial temporal region besides HS (28 participants). Among the participants with extratemporal abnormalities (66 participants), unidentified bright objects (focal white matter T2 hyperintensities) (n = 16) were most common followed by postoperative parenchymal defect (n = 12), encephalomalacia (n = 8), vascular malformation (n = 7), and others. Among the participants with abnormal findings in the medial temporal region besides HS (28 participants), glioneuronal

tumors (n = 8) and benign cysts (n = 8) were most common followed by vascular malformation (n = 4), cortical dysplasia (n = 3), and others. By adding such abnormal cases, achieving accurate classification in our setting became much more challenging and was more realistic in a clinical sense compared with the setting in other recent works,[8,10,11] where only the participants with normal MRI were included in the control group.

To accurately evaluate the generalizability of our trained model, we constructed a separate test set that consisted of (a) pathologically confirmed HS cases or (b) HS supported by either EEG or nuclear imaging (PET of SPECT). Among the patients with HS used for the test data set, HS was pathologically confirmed by surgery in 66.0% (60.0% in left HS and 72.0% in right HS) of cases. The majority of pathologically confirmed cases belonged to ILAE classification type I (28/33, 84.9%) while one case was classified as type II (1/33, 3.0%). Due to the poor orientation of the hippocampus specimen, ILAE classification could not be applied in four cases (4/33, 12.1%) (Table S2). Twenty-five left TLE patients with HS (14 females; mean age 42.6 ±14.5; range 19-80 years) and 25 right TLE patients with HS (15 females; mean age 46.8 ±12.2; range 27-68 years) were included as the test Left HS and Right HS groups, respectively. In the test No HS group, we included 252 participants (128 females; mean age 41.3 ±16.4; range 19-92 years), which consisted of 168 participants with normal findings, 69 participants with abnormalities in the extratemporal region and 15 participants with abnormal findings in the hippocampus besides HS.

This study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. 1906-106-1041). Informed consents from all participants were also obtained.

2.2 | MRI acquisition and image processing

T2-weighted oblique coronal MR images were obtained from each patient using various scanners in Digital Imaging and Communications in Medicine (DICOM) format. Scanning conditions for each scanner are available in Table S3. The DICOM images were converted to Neuroimaging Informatics Technology Initiative (NifTI) format to combine multiple slices of the brain into a single three-dimensional brain file. Then, we unified the brain orientations, removed the shading artifacts, and stripped the skull through Analysis of Functional NeuroImages (AFNI) software.[15] We registered the skull-stripped brain using FMRIB Software Library (FSL) software[16] onto the Montreal Neurological Institute (MNI). 152 standard brain template[17] with a spatial resolution of $1\times1\times1\,mm^3$ to regulate the brain structure differences among the participants (Figure 1A). Finally, we cropped the backgrounds without loss of any potentially informative voxels. Therefore, we were able to unify all the brains to have an identical size of $160\times200\times170$.

2.3 | Convolutional neural network development

Among various deep learning models, we implemented a 3D convolutional neural network (CNN) that consisted of four convolution layers followed by three fully connected layers. Each convolution layer was constituted by 3D convolutional filters, a rectified linear unit (ReLU) activation function and a pooling layer. The number of convolutional filters for each layer was 5, 10, 20, and 40. For the fully connected layers, when all of the nodes in a layer were connected to the nodes of the previous and next layers, we had 64, 64, and 3 nodes in each layer. As each group was defined as the class that the classifier was trying to predict, the softmax layer was implemented at the last layer of the classifier to compute the probabilities of the given input brain MRI belonging to each group. The specific implementation details could be found in the "Data Availability" section.

For training, the stochastic gradient descent (SGD) optimizer[18] was used to update the model weight parameters with a batch size of 42. Using a hyperparameter optimization framework called Optuna,[19] the initial learning rate was set as 0.0221 and decayed by 0.124 if the validation loss had stopped improving for 25 epochs during a total of 150 epochs. As an important training detail, we applied data augmentation by creating horizontally flipped training data and modified their labels by swapping the Left and Right HS groups. Such augmentation that was possible specifically for HS classification as the brain and hippocampus were more or less symmetric, effectively doubled the training data and was found to substantially help improve prediction accuracy. We also noted that neither the validation nor the test data were horizontally flipped. The model was trained on a single GPU (Tesla V100 SXM2 32GB), and took 20 hours to train with our implementation in Python 3.6.5 with PyTorch 1.4.0. The overall model structure is shown in Figure 1B.

We also employed a voting ensemble method that was commonly used to improve the prediction accuracy. We trained six distinct models that were randomly initialized with different random seeds before training, and performed the majority voting among the predictions of the models to obtain an ensemble prediction.

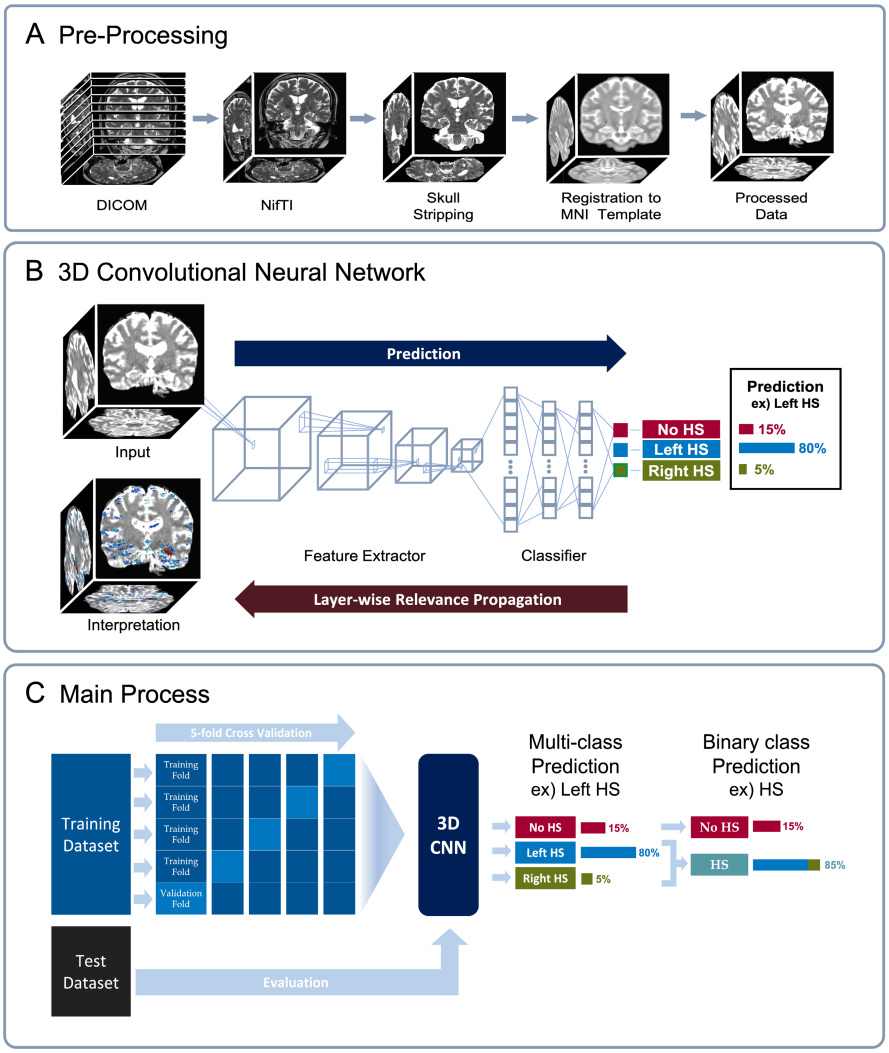

FIGURE 1 Schematic description of the current study. A, Brief preprocessing of the neuroimaging data, including format conversion, skull-stripping, and registration to Montreal neurological institute (MNI) standard brain template. B, 3D convolutional neural network (CNN) architecture for the classification among the No, left, and right hippocampal sclerosis (HS) groups. We used layer-wise relevance propagation (LRP) to highlight the brain regions relevant to the prediction. C, the main process of the study. We aggregated the predictions for the left and right HS groups and compared the sum with the No HS group prediction, which became the binary classification.

2.4 | Evaluation protocol

While the model was initially designed to classify the input into one of the No HS, Left HS, and Right HS groups, we also considered an easier binary classification setting in which the goal was to simply determine whether or not a person has HS. This would be done by combining the prediction probabilities for the Left and Right HS groups and comparing them with the probability for the No HS group. The accuracies of the models were evaluated through both five-fold stratified cross validation (CV) and test set. The stratification assures that each group had been included proportionally in each fold, and the cross validation was repeated 10 times with a different random seed each time.

After determining an optimal epoch for which the model achieved the highest accuracy on average among the validation folds, we trained a new model utilizing the entire training set until the epoch and evaluated its performance on a separate test set. As the test set is imbalanced, we adopted a balanced accuracy which is the average of the accuracies for each group. The overall model development and evaluation processes are given in Figure 1C.

2.5 | Interpretation of the model prediction

One of the typical criticisms for deep learning models is that the neural networks are very complex; hence, it is difficult to provide proper explanations for the models' predictions. Several interpretation methods[20-22] had recently been developed to provide more reliable explanations for deep learning models, and among those, we adopted layer-wise relevance propagation (LRP).[20,23]

The reason for choosing LRP was that it was computationally efficient and had been shown to be effective in obtaining reliable and sparse interpretations in various application areas.[23-27] Once the model made a prediction for a given individual 3D brain image, LRP recursively ran the relevance propagation step to decompose and distribute the final prediction score to each input voxel.

The decomposed score, dubbed as the relevance score of LRP, represented the importance of each voxel for the given prediction. In our case, the positive and negative scores for the prediction were gathered separately during the process so that each value would be preserved without being canceled-out and represent the relevant regions better.[23] We obtained the saliency map of the important voxels by visualizing only the positive scores. The interpretation was given as feedback to confirm which regions of the MRI were referenced and give confidence for the prediction. The clinicians would decide whether to trust the model by referring to the region of high relevance in the MRI.

2.6 | Statistical analysis

To verify that the participants in both datasets have similar features, the demographic variables of the participants among the validation and test sets were compared using Mann-Whitney U test to denote a *P* value of .01.

The multiclass classification performance was evaluated by computing the accuracy for each group on both the validation set and the test set. We also showed the binary classification performances for the same datasets, and the following four evaluation criteria were used: accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC).

3 | RESULTS

3.1 | HS classification performance

Tables 1 and 2 summarize the overall classification accuracy obtained for multiclass and binary classification settings, respectively. In each table, both five-fold CV and test accuracy results are shown for the single model and voting ensemble. For the single model, we displayed the performance of the model with the highest accuracy among the six models used for the ensemble method. The numbers in parentheses represent the number of correct predictions and the total number of participants for each group.

From both multiclass classification and binary classification, the voting ensemble achieved the best performance for both the CV and test sets (Tables 1, 2). The overall accuracy for the test set showed little difference compared with that of CV, which reflected the true generalization performance of our model. We also observe the discrepancy of the validation accuracy across each group, which may be an artifact of the imbalanced training dataset. Furthermore, the test set accuracy for the control group participants with normal findings was significantly higher than that of the other classes within the control group in both the single model and voting ensemble model (Table S4). Considering that the overall accuracy would be higher if the control group only consisted of the participants with the normal MRI findings, we can claim that distinguishing HS from nonHS disease controls with abnormalities is more challenging than distinguishing only from the normal, healthy controls. Moreover, the True Positive Rates (TPR) only for the pathologically confirmed HS cases (Table S5) in the test set were equally high, again showing good generalization capability of our model. As a summary, we plot the accuracy comparison among several single models and the voting ensemble method (Figure S1A) as well as the ROC curve of the voting ensemble method (Figure S1B).

We also stress that the data augmentation using the horizontally flipped MRI data as described in the Methods section, plays a critical role in achieving a high accuracy. For the multiclass classification, the model without data augmentation had accuracies of 79.4% and 75.0% for the CV and test sets, respectively (Table S6), which were decreased to 8.1% and 16.5% compared with those in Table 1. Likewise, for the binary classification, the model without data augmentation only achieved accuracies of 83.4% and 82.3% for the CV and test sets, respectively (Table S7), resulting in a reduction of 5.4% and 7.4% from Table 2.

TABLE 1 The multiclass classification performance for the five-fold cross validation and test set

		Multiclass			
No HS/left HS/right HS	**Method**	**Total accuracy**	**No HS**	**Left HS**	**Right HS**
5 fold CV Performance n = 160, 100, 60	Single model	0.869 (278/320)	0.888 (142/160)	0.85 (85/100)	0.85 (51/60)
	Voting ensemble	0.875 (280/320)	0.913 (146/160)	0.87 (87/100)	0.783 (47/60)
Test Performance n = 25, 25, 25	Single model	0.897	0.81 (204/252)	0.88 (22/25)	1.0 (25/25)
	Voting ensemble	0.915	0.865 (218/252)	0.92 (23/25)	0.96 (24/25)

Note: The numbers of evaluated participants for the No, left and right HS groups are indicated in the first column. We separately measured the accuracies for each group with the number of correctly predicted participants over the total participants, shown in parentheses. The total accuracy for the test set was evaluated using the balanced accuracy metric which is an average of the accuracies for each group for the imbalanced dataset.

TABLE 2 The binary classification performance for the five-fold cross validation and test set

		Binary class			
No HS/HS	**Method**	**Accuracy**	**Sensitivity**	**Specificity**	**AUROC**
Fivefold CV Performance n = 160, 100, 60	Single Model	0.884 (283/320)	0.881 (141/160)	0.887 (142/160)	0.942
	Voting Ensemble	0.888 (284/320)	0.875 (140/160)	0.9 (144/160)	0.950
Test Performance n = 50, 25, 25	Single Model	0.883	0.96 (48/50)	0.806 (203/252)	0.945
	Voting Ensemble	0.897	0.94 (47/50)	0.853 (215/252)	0.961

Note: The correctly predicted participants over the total participants for accuracy, sensitivity, and specificity are shown in the parentheses. The setting for the test set is identical as in Table 1.

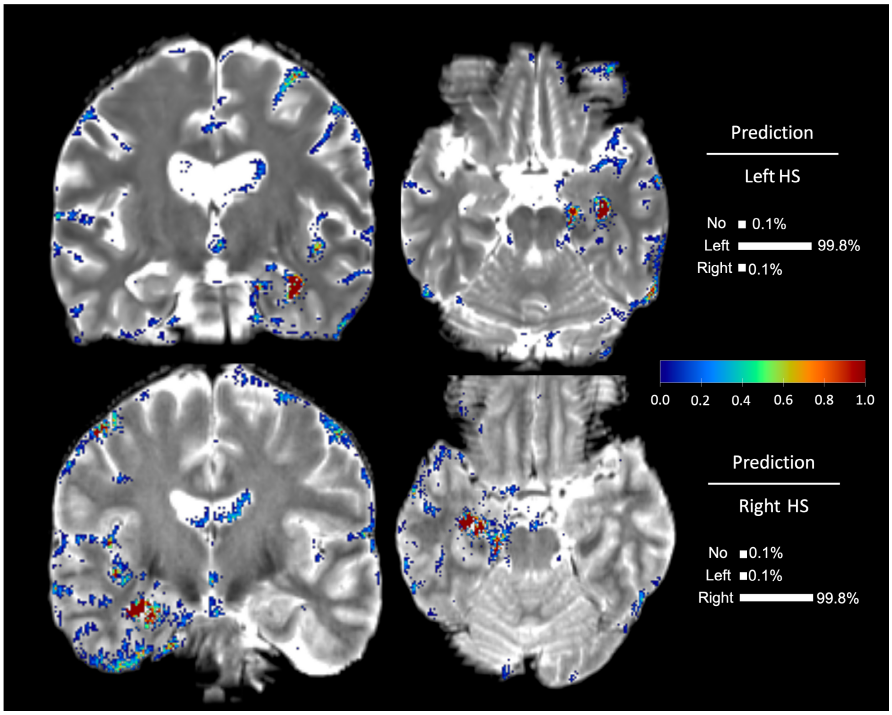

FIGURE 2 Individual interpretations of the predictions for the participants in the left and right HS groups. The model predicted each participant correctly with 99.8% confidence. Each interpretation was projected onto the MRI of the corresponding participants. The colors represented the influential brain regions in the prediction, and they indicated higher contributions as the color became close to red. The visualization revealed that most information was obtained from the medial temporal area and hippocampus; however, some information was also obtained from the subarachnoid space adjacent to the ipsilateral cortex. For visual clarity, we cut off the relevance scores under the thresholds.

Thus, due to the symmetry of the brain and hippocampus, horizontal flip-based data augmentation became an essential recipe in improving the generalizability of our model.

3.2 | Visualizations and interpretations

To make our model more clinically interpretable, we visualized the most relevant region on the brain MRI template. As small relevance scores may scatter across the input and could distract from highlighting the most relevant region, we set a threshold for each individual participant and only displayed values over the threshold for visual clarity. Figure 2 shows two examples of the LRP visualizations for the individual participants from the Left and Right HS groups. The red voxels were of higher relevance than the blue voxels. From both figures, we observed that the visualized regions displayed variable patterns. Most of the visualized regions were located in the medial temporal area; however, a substantial extent of other regions was also identified to be useful for HS prediction. In particular, the subarachnoid spaces adjacent to the ipsilateral temporal and frontal cortex and lateral sulcus turned out to be important, but this demonstrated different patterns among patients.

To obtain group-level interpretations that are more robust to individual variability, we averaged the relevance scores among the correctly classified participants for each group. The group-level interpretation for the Left and Right HS groups mainly converged to the left and right medial temporal areas, respectively (Figure 3). At the same time, we observe the relevance around the subarachnoid spaces adjacent to the ipsilateral temporal and frontal cortex and lateral sulcus were highlighted more significantly, which reflects that patients with HS are likely to be accompanied by neocortical atrophy of the ipsilateral temporal and frontal cortex. More details on the color scales in Figures 2 and 3 are given in Appendix S1.

In addition to the individual- and group-level interpretations, we visualized the learned feature embeddings for the participants of each group to show if they were well clustered, corroborating the good classification performance of our model. For both validation and test data, we applied a dimension reduction method called uniform manifold approximation and projection (UMAP)[28] to the output of the first fully connected layer in our 3D CNN (Figure 4). As UMAP projects the high dimensional features preserving the relative distances, we would assert that the clearly clustered feature embeddings reflects the discriminative capability of our model. The group-level LRP interpretations for the selected samples in each group clearly demonstrated relevant brain regions for each group.

4 | DISCUSSION

Accurate diagnosis of HS is particularly important because it is a crucial factor in determining whether to perform surgical resection in epilepsy patients. In TLE patients with HS, a seizure-free state can be achieved after surgical resection.[2,3] However, in some cases, the diagnosis of HS can be very tricky, and only experienced experts could make an accurate diagnosis.[29] Many efforts have been made, such as applying the HS scoring system[7,30] or providing automated quantitative MRI reports,[5,6] to overcome the inter-rater diagnostic discrepancies. However, a substantial number of HS is still confirmed only by pathological examinations after surgery. To that end, we believe our results showed potential in developing a reliable HS diagnosis assistant based on deep learning.

Our proposed framework has several strengths. First, our end-to-end deep learning framework dramatically

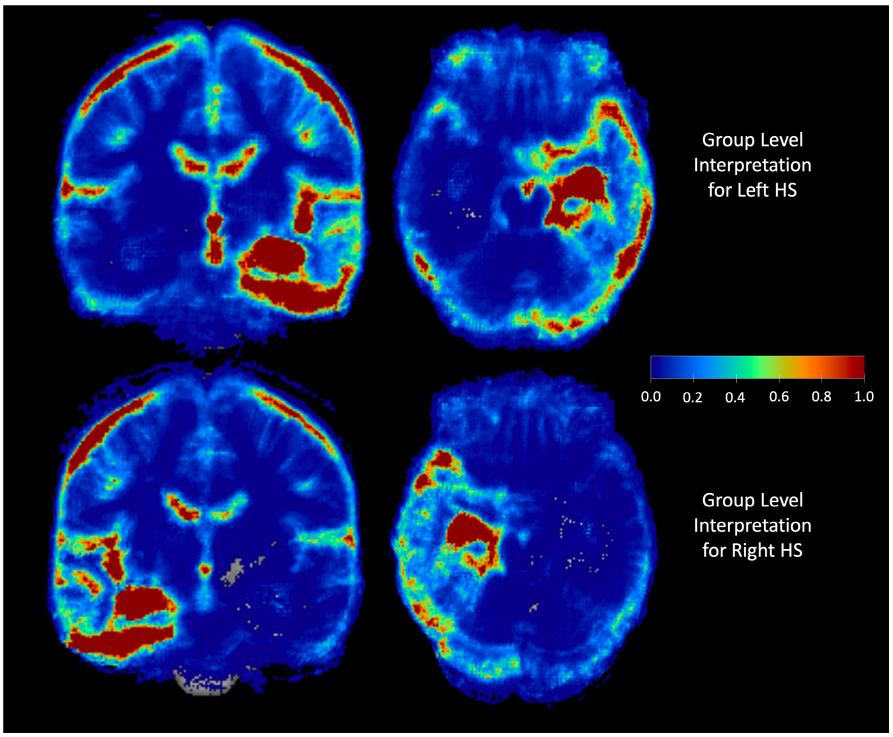

FIGURE 3 Group-level interpretation for the correctly predicted participants with left and right HS group. We averaged the interpretations of the correctly predicted participants, which were 85 out of 100 for the left HS group in the validation folds and 51 out of 60 for the right HS group in the validation folds. A group-level interpretation was projected onto the MNI brain template so that we could check which regions of the brain the model had referred to predict HS correctly. The subarachnoid spaces adjacent to ipsilateral temporal and frontal cortex and lateral sulcus were highlighted along with the medial temporal area, which reflects that patients with HS are likely to be accompanied by neocortical atrophy of the ipsilateral temporal and frontal cortex.

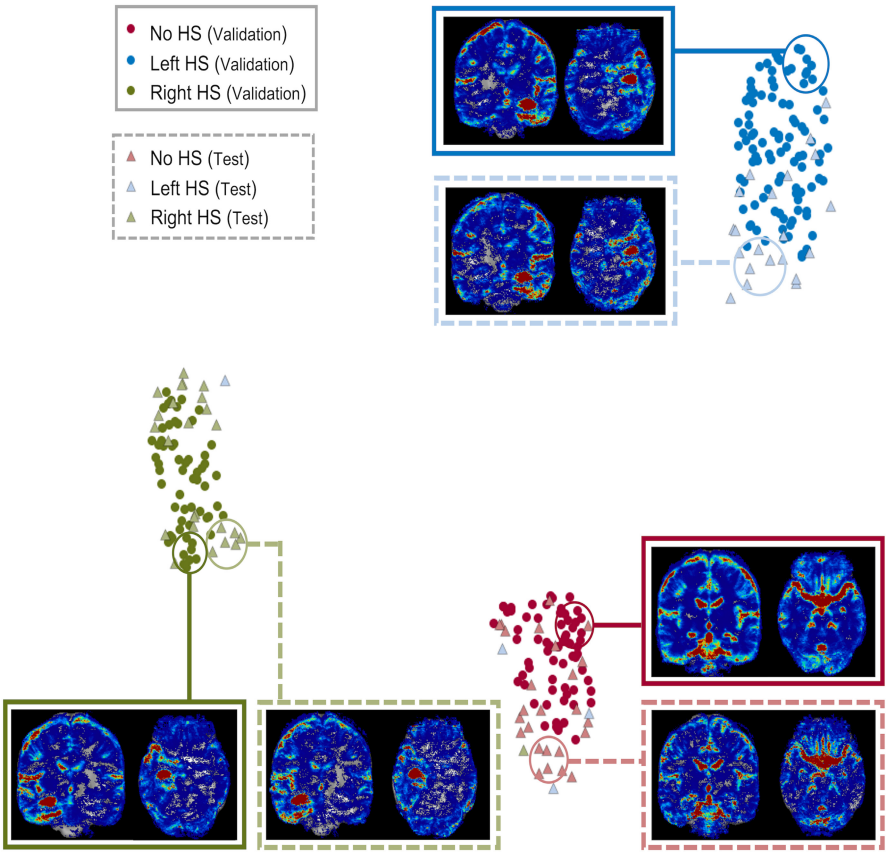

FIGURE 4 Dimensionality reduction of the feature embeddings through UMAP. We projected the last convolutional layer output into two-dimensional space using UMAP. The solid circles and light triangles represent the validation and test data samples, respectively, and the No, left, and right HS groups are in red, blue, and green, respectively. For the No HS group, only the samples with normal findings were visualized to prevent any adverse effects in the interpretations. As UMAP projected a high-dimensional feature into low dimensions, preserving the relative distances between each data point, we assumed that the data located at the tip of a group cluster were most representative. We plotted the group-level interpretations for those data points to observe the distinctive features for each group and similar interpretations among the validation and test data.

reduced the computation time compared with the conventional machine learning approach[9,10] that usually required complex and expensive data preprocessing for extracting handcrafted features. For example, data preprocessing for a model[10] required approximately 50 minutes per image (on CPU) for the extraction and segmentation of the brain features. In contrast, our method only performed skull-stripping and registration onto the MNI template as preprocessing, which took only 5 minutes per image on a single CPU.

Second, our model was developed under conditions similar to the real-world practice of epilepsy patients by including various types of control group examples. HS can be present in epilepsy patients with "dual pathology", which refers to the coexistence of HS with additional epileptogenic lesions.[31-33] When HS and focal cortical dysplasia coexist, patients who receive hippocampal resection are more likely to have a favorable surgical outcome.[34] Similarly, temporal lobectomy was related to better surgical outcomes in posttraumatic epilepsy patients with HS.[35] These results implied that the identification of HS was also very important in epilepsy patients with extratemporal abnormalities. Nevertheless, most of the previous machine-learning approach studies conducted thus far have only included normal brain MRIs as the control group[1,23,27,36] where the model may have been simply distinguishing abnormal brain MRIs from normal MRIs.

Third, our model achieved high test accuracy that was comparable with or better than previous studies conducted in much simplified settings, underscoring a good generalization capability of our model despite being evaluated on a larger test set compared with other previous work. Namely, Chen et al.[8] developed an SVM for 37 participants for binary classification between HS and healthy controls (HC) and achieved 83.8% ± 3% validation accuracy. Zhou et al.[11] developed a SVM model for 148 participants and achieved a validation accuracy of 72.3% in the binary classification between participants with mesial TLE and HC. Mo et al.[1] utilized SVM and logistic regression models, and achieved fairly high

accuracy on the 100 training set participants and 60 test set participants. However, the insufficient test set participants raises doubts about whether the model performance will be maintained even with a large dataset or in practice.

Finally, our LRP-based interpretations were informative and could be obtained much more efficiently than previous work that utilized deep learning-based classifiers with brain MRI input. Namely, the interpretation methods used in the previous studies had a few drawbacks. Some of them were too dense and noisy[36]or plotted on the automatically generated asymmetry maps, not on the MRI itself,[37] hence the clinician had to further match the interpretations with the anatomical brain regions. In another work,[38] the regions identified as important also included irrelevant regions like background or cranium, which cannot be truly associative with the prediction. In contrast, our interpretation based on LRP successfully highlighted the focused relevant regions directly on the individuals' and template MRI.

Interestingly, our LRP-based interpretations also suggested additional important features useful for the identification of HS. From the interpretations in Figures 2 and 3, our model seemed to use the information of subarachnoid spaces adjacent to ipsilateral temporal and frontal cortex and lateral sulcus in the diagnosis of HS. It has been reported that progressive brain damage caused by the gliotic process may occur in the paralimbic cortices of patients with prolonged epilepsy.[39-44] Progressive neocortical atrophy had been reported in pharmacoresistant TLE patients[45] and in seizure-free TLE patients.[46] The pattern of neocortical atrophy have provided plenty of clinically meaningful information in epilepsy patients. The pattern of cortical thinning was different for TLE patients with and without HS,[42] and neocortical atrophy of certain areas had been reported to be associated with disease progression in TLE. However, assessing neocortical atrophy by visual inspection and considering it in the diagnosis of HS was an extremely difficult task for humans. Based on recent rapid progress, the deep learning model showed the potential to surpass human capabilities in terms of comprehensively analyzing the morphology of the entire brain.[47,48] Considering these characteristics, our 3D CNN model could be useful for predicting several important properties of epilepsy, such as pharmacoresistance, disease progression, and surgical outcome.

4.1 | Limitations

There are several future research directions worth pursuing to improve our work. One is to further enlarge the datasets, particularly the separated test set, and more reliably evaluate the generalizability of our 3D CNN model. Moreover, we could also consider merging multimodal data obtained from different structural and functional neuroimaging studies, or using features from T1-weighted, FLAIR, and T2-weighted images as Caldairou et al.[37] to utilize the association among different data sources. Furthermore, although the MRI images we used were collected over 21 years with 12 different scanners and have sufficient variation in image quality and distribution, our work lacks the cross-site validation as the data were collected from a single institution. We could assert the generalizability by evaluating our model with the data from other institutions in the future. Finally, comparing our model's accuracy with the neurologist validation would objectively verify whether it can serve as an informative assistant for neurologists in HS diagnosis.

5 | CONCLUSIONS

We have developed an interpretable deep learning-based model for HS classification on MRI. Our model utilized the largest amount of MRI data with minimal preprocessing for developing the HS classification model, including various types of abnormal MRIs that mimicked the real clinical setting. With the high accuracy and capability of visual interpretation, we anticipate that our model could aid in diagnosing HS as a third opinion in practice and extend the scope of hippocampal pathology assessment in temporal lobe epilepsy.

ACKNOWLEDGMENTS

This work was supported in part by the New Faculty Startup Fund from Seoul National University, NRF grants [NRF-2021M3E5D2A01024795] and IITP grants [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)][No.2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)] funded by the Korean government. The corresponding authors would also like to thank Sang H. Moon and Woon O. Cha for their helpful discussion and encouragement for this project.

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

DATA AVAILABILITY STATEMENT

The data that supported the findings of this study were available from the corresponding author on reasonable request. Python scripts were available on github.[49]

ORCID

Jangsup Moon https://orcid.org/0000-0003-1282-4528
Taesup Moon https://orcid.org/0000-0002-9257-6503

REFERENCES

1. Mo J, Liu Z, Sun K, Ma Y, Hu W, Zhang C, et al. Automated detection of hippocampal sclerosis using clinically empirical and radiomics features. Epilepsia. 2019;60(12):2519–29.
2. Thom M, Mathern GW, Cross JH, Bertram EH. Mesial temporal lobe epilepsy: how do we improve surgical outcome? Ann Neurol. 2010;68(4):424–34.
3. Tugcu B, Gungor A, Akpinar A, Kinay D, Kuscu DY, Gül G, et al. Outcome of surgical treatment of hippocampal sclerosis from relatively new epilepsy surgery center. J Neurosurg Sci. 2016;60(2):159–68.
4. Malmgren K, Thom M. Hippocampal sclerosis—origins and imaging. Epilepsia. 2012;53:19–33.
5. Goodkin O, Pemberton HG, Vos SB, Prados F, Das RK, Moggridge J, et al. Clinical evaluation of automated quantitative MRI reports for assessment of hippocampal sclerosis. Eur Radiol. 2021;31(1):34–44.
6. Hu W-h, Liu L-n, Zhao B-t, Wang X, Zhang C, Shao X-Q, et al. Use of an automated quantitative analysis of hippocampal volume, signal, and glucose metabolism to detect hippocampal sclerosis. Front Neurol. 2018;9:820.
7. Ver Hoef LW, Paige AL, Riley KO, Cure J, Soltani M, Williams FB, et al. Evaluating hippocampal internal architecture on MRI: inter-rater reliability of a proposed scoring system. Epilepsy Res. 2013;106(1–2):146–54.
8. Chen S, Zhang J, Ruan X, Deng K, Zhang J, Zou D, et al. Voxel-based morphometry analysis and machine learning based classification in pediatric mesial temporal lobe epilepsy with hippocampal sclerosis. Brain Imaing Behav. 2020;14(5):1945–54.
9. Riederer F, Seiger R, Lanzenberger R, Pataraia E, Kasprian G, Michels L, et al. Voxel-based morphometry—from hype to Hope. A study on hippocampal atrophy in mesial temporal lobe epilepsy. Am J Neuroradiol. 2020;41(6):987–93.
10. Wang H, Ahmed SN, Mandal M. Computer-aided detection of mesial temporal sclerosis based on hippocampus and cerebrospinal fluid features in MR images. Biocybern Biomed Eng. 2019;39(1):122–32.
11. Zhou B, An D, Xiao F, Niu R, Li W, Li W, et al. Machine learning for detecting mesial temporal lobe epilepsy by structural and functional neuroimaging. Front Med. 2020;14(5):630–641.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
13. Hinton G. Deep learning-a technology with the potential to transform health care. JAMA. 2018;320:1101–2.
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
15. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 1996;29(3):162–73.
16. Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, et al. Bayesian analysis of neuroimaging data in FSL. Neuroimage. 2009;45(1):S173–S86.
17. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, et al. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage. 2011;54(1):313–27.
18. Ruder S. An overview of gradient descent optimization algorithms. arXiv. 2016. https://doi.org/10.48550/arXiv.1609.04747
19. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery, 2019; p. 2623–31.
20. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One. 2015;10(7):e0130140.
21. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc, 2017; p. 4768–77.
22. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Gradcam BD. Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision. Manhattan, NY: IEEE, 2017; p. 618–26.
23. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194.
24. Arras L, Horn F, Montavon G, Müller K-R, Samek W. "What is relevant in a text document?": an interpretable machine learning approach. PloS One. 2017;12(8):e0181142.
25. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F, et al. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Rep. 2020;10(1):1–12.
26. Park Y, Kwon B, Heo J, Hu X, Liu Y, Moon T. Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. Environ Pollut. 2020;256:113395.
27. Thomas AW, Heekeren HR, Müller K-R, Samek W. Analyzing neuroimaging data through recurrent deep learning models. Front Neurosci. 2019;13:1321.
28. Leland M, John H, Nathaniel S, Lukas G. UMAP: uniform manifold approximation and projection. J Open Source Softw. 2018;3(29):861.
29. Louis S, Morita-Sherman M, Jones S, Vegh D, Bingaman W, Blumcke I, et al. Hippocampal sclerosis detection with NeuroQuant compared with neuroradiologists. Am J Neuroradiol. 2020;41(4):591–7.
30. Dou W, Zhao L, Su C, Lu Q, Liu Q, Guo J, et al. A quantitative MRI index for assessing the severity of hippocampal sclerosis in temporal lobe epilepsy. BMC Med Imaging. 2020;20:1–7.
31. Miyata H, Sudo S, Kuwashige H, Miyao S, Nakamoto H, Kubota Y, et al. Dual pathology in a patient with temporal lobe epilepsy associated with neocortical glial scar after brain abscess and end folium sclerosis/hippocampal sclerosis type 3. Neuropathology. 2021;41(1):42–8.
32. Moon H-J, Chung CK, Lee SK. Surgical prognostic value of epileptic aura based on history and electrical stimulation. J Epilepsy Res. 2019;9(2):111–8.
33. Osawa S-i, Iwasaki M, Suzuki H, Nakasato N, Tominaga T. Occult dual pathology in mesial temporal lobe epilepsy. Neurol Sci. 2015;36(9):1743–5.

34. Kim DW, Lee SK, Nam H, Chu K, Chung CK, Lee SY, et al. Epilepsy with dual pathology: surgical treatment of cortical dysplasia accompanied by hippocampal sclerosis. Epilepsia. 2010;51(8):1429–35.
35. Hitti FL, Piazza M, Sinha S, Kvint S, Hudgins E, Baltuch G, et al. Surgical outcomes in post-traumatic epilepsy: a single institutional experience. Oper Neurosurg. 2020;18(1):12–8.
36. Qiu S, Joshi PS, Miller MI, Xue C, Zhou X, Karjadi C, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920–33.
37. Caldairou B, Foit NA, Mutti C, Fadaie F, Gill R, Lee HM, et al. MRI-based machine learning prediction framework to lateralize hippocampal sclerosis in patients with temporal lobe epilepsy. Neurology. 2021;97(16):e1583–e93.
38. Ito Y, Fukuda M, Matsuzawa H, Masuda H, Kobayashi Y, Hasegawa N, et al. Deep learning-based diagnosis of temporal lobe epilepsy associated with hippocampal sclerosis: an MRI study. Epilepsy Res. 2021;178:106815.
39. Alhusaini S, Doherty CP, Palaniyappan L, Scanlon C, Maguire S, Brennan P, et al. Asymmetric cortical surface area and morphology changes in mesial temporal lobe epilepsy with hippocampal sclerosis. Epilepsia. 2012;53(6):995–1003.
40. Hocker S, Nagarajan E, Rabinstein AA, Hanson D, Britton JW. Progressive brain atrophy in super-refractory status epilepticus. JAMA Neurol. 2016;73(10):1201–7.
41. Labate A, Cerasa A, Aguglia U, Mumoli L, Quattrone A, Gambardella A. Neocortical thinning in "benign" mesial temporal lobe epilepsy. Epilepsia. 2011;52(4):712–7.
42. Mueller SG, Laxer KD, Barakos J, Cheong I, Garcia P, Weiner MW. Widespread neocortical abnormalities in temporal lobe epilepsy with and without mesial sclerosis. Neuroimage. 2009;46(2):353–9.
43. Lee HM, Fadaie F, Gill R, Caldairou B, Sziklas V, Crane J, et al. Decomposing MRI phenotypic heterogeneity in epilepsy: a step towards personalized classification. Brain. 2022;145(3):897–908.
44. Goubran M, Hammond RR, de Ribaupierre S, Burneo JG, Mirsattari S, Steven DA, et al. Magnetic resonance imaging and histology correlation in the neocortex in temporal lobe epilepsy. Ann Neurol. 2015;77(2):237–50.
45. Bernhardt BC, Worsley K, Kim H, Evans A, Bernasconi A, Bernasconi N. Longitudinal and cross-sectional analysis of atrophy in pharmacoresistant temporal lobe epilepsy. Neurology. 2009;72(20):1747–54.
46. Alvim MK, Coan AC, Campos BM, Yasuda CL, Oliveira MC, Morita ME, et al. Progression of gray matter atrophy in seizure-free patients with temporal lobe epilepsy. Epilepsia. 2016;57(4):621–9.
47. Adeli E, Zhao Q, Zahr NM, Goldstone A, Pfefferbaum A, Sullivan EV, et al. Deep learning identifies morphological determinants of sex differences in the pre-adolescent brain. Neuroimage. 2020;223:117293.
48. Lian C, Liu M, Pan Y, Shen D. Attention-guided hybrid network for dementia diagnosis with structural MR images. IEEE Trans Cybern. 2020;52(4):1992–2003.
49. Kim DH. Source codes for interpretable deep learning based hippocampal sclerosis classification. github.com. Updated October 26, 2021. Accessed November 1, 2021. https://github.com/donny8/Interpretable-Hippocampal-Sclerosis-Classification

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kim D, Lee J, Moon J, Moon T. Interpretable deep learning-based hippocampal sclerosis classification. Epilepsia Open. 2022;7:747–757. https://doi.org/10.1002/epi4.12655