Sensors and Materials, Vol. 38, No. 9 (2026) 5137–5152
MYU Tokyo

S & M 4614

Explainable Multimodal Detection of Cerebral Atrophy Using Computed Tomography and Psychological Risk Factors

Xiaohong Yu,[1,2] Jaehoon (Paul) Jeong,[1*] Yoosoo Chang,[3,4,5] Ria Kwon,[3,6] Jeonggyu Kang,[3] Ga-Young Lim,[3,6] and Seungho Ryu[3,4,5]

[1]Department of Computer Science & Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[2]Graduate School of Information and Communication Technology, Ajou University, Suwon, Republic of Korea
[3]Center for Cohort Studies, Total Healthcare Center, Kangbuk Samsung Hospital, School of Medicine, Sungkyunkwan University, Seoul, Republic of Korea
[4]Department of Occupational and Environmental Medicine, Kangbuk Samsung Hospital, School of Medicine, Sungkyunkwan University, Seoul, Republic of Korea
[5]Department of Clinical Research Design & Evaluation, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea
[6]Institute of Medical Research, Sungkyunkwan University, Seoul, Republic of Korea

(Received April 1, 2026; accepted September 2, 2026)

Keywords: cerebral atrophy, computed tomography, multimodal model, explainable AI, psychiatric risk screening

Pathological cerebral atrophy can arise from various neurological and systemic conditions, including depressive disorders, and early identification is clinically important for improving care. However, existing neuroimaging studies primarily rely on magnetic resonance imaging (MRI), which may be less accessible in routine clinical practice. To address this gap, we present an explainable deep learning framework for cerebral atrophy classification from widely available brain computed tomography (CT) images, while incorporating psychological risk factors as auxiliary input features in a multimodal model. The proposed method integrates a convolutional neural network (CNN) to capture structural brain features from CT images with a multilayer perceptron to incorporate psychological risk factors, including depressive symptoms and suicidal ideation. This multimodal design is intended to provide a broader representation of patient characteristics. In addition, gradient-based class activation mapping is employed to provide visual explanations by highlighting anatomically relevant regions that contribute to model predictions. In the current evaluation setting and at a predefined operating threshold, the image-only model achieved 83.01% accuracy, and the multimodal model achieved 87.25% accuracy. The multimodal approach therefore showed improved performance in this cohort, and attention maps were generally consistent with clinically relevant patterns such as ventricular enlargement and sulcal widening. These findings provide preliminary evidence for the feasibility of CT-based multimodal atrophy assessment, while threshold-independent analyses and broader external validation remain necessary before deployment-oriented conclusions.

*Corresponding author: e-mail: pauljeong@skku.edu
https://doi.org/10.18494/SAM6361

ISSN 0914-4935

1. Introduction

Brain volume in healthy adults declines with age, and pathological cerebral atrophy can be accelerated by neurological and psychiatric disorders.[1–3] Structural brain-volume loss has been reported in conditions such as Alzheimer's disease and major depressive disorder (MDD), and these alterations are clinically relevant for diagnosis and longitudinal risk assessment.[4–6] Clinically, cerebral atrophy is a structural manifestation rather than a single disease entity, and it can arise from several causes, such as neurodegenerative, vascular, inflammatory, and stress-related mechanisms. As these mechanisms often overlap, reliable detection of atrophy patterns can support earlier risk stratification and intervention, especially in psychiatric populations where structural changes are subtle and frequently under-recognized.

Most established atrophy studies are based on magnetic resonance imaging (MRI).[4,7,8] Although MRI provides high soft-tissue contrast, routine deployment is constrained by cost, scan time, and accessibility. Computer tomography (CT) is faster and more widely available, and recent studies indicate that automated CT analysis can quantify atrophy-related markers with good agreement to MRI.[1,9] These findings support CT as a practical modality for scalable atrophy assessment. CT is also routinely performed in health-check and emergency workflows, creating opportunities for opportunistic screening using existing scans. However, CT-based atrophy assessment in practice remains largely qualitative and heavily dependent on reader interpretation. Machine learning (ML)-based learning analysis can reduce this variability by extracting reproducible features and producing consistent patient-level predictions.

Despite this progress, AI-based atrophy research remains concentrated on older or neurodegenerative cohorts, and evidence in psychiatric-risk populations is limited. In particular, studies that integrate CT imaging with psychological risk information remain limited. This gap is important because depression and suicidal ideation are associated with structural brain alterations, yet clinically accessible imaging models for these populations remain scarce.[3,10,11] Another practical challenge is limited high-quality training data for psychiatric-risk-associated structural tasks. Annotation requires expert review, inter-rater variability is nontrivial in borderline cases, and privacy constraints restrict cross-institutional sharing. These constraints motivate multimodal models that use both imaging and nonimaging information rather than relying only on subtle CT visual cues.

To address this gap, we propose an explainable multimodal framework for CT-based cerebral atrophy classification in psychiatric-risk settings. The image branch uses a convolutional neural network (CNN) to learn structural features from CT slices, and the clinical branch uses a multilayer perceptron (MLP) to encode psychological risk factors. The two branches are fused to improve classification performance, and gradient-based class activation mapping (Grad-CAM) provides visual explanations of model decisions. This design reflects clinical workflow, where radiological findings and psychiatric assessments are interpreted jointly. The aim of combining both modalities is to improve the sensitivity of the model in subtle atrophy-related cases while maintaining interpretability for clinical adoption.

The main contributions of this paper are as follows.

- We present a CT-based multimodal learning framework that jointly models imaging features and psychological risk factors for cerebral atrophy detection.

- We design and compare two image-based classification strategies for adjacent CT slices, slice-channel fusion (SCF), and slice-wise multibranch fusion (SMBF). These strategies provide a practical approximation using limited slice context.
- We provide interpretable prediction evidence via Grad-CAM and demonstrate improved performance over image-only baselines.

Note that this study is an enhancement of our previous work.[12] Compared with our previous work,[12] in this study, we provide expanded methodological details, such as CT image Hounsfield Unit (HU) conversion, brain tissue windowing, image-based classification strategies, and multimodal fusion formulations, as well as more comprehensive comparative experiments, and enhanced clinical interpretation and discussion.

The remainder of this paper is organized as follows. In Sect. 2, we review prior studies on cerebral atrophy and AI-based neuroimaging. In Sect. 3, the dataset and cohort characteristics are described. The proposed framework, including preprocessing, multimodal model design, and visual explanation, is presented in Sect. 4. The details of the implementation settings and evaluation metrics are given in Sect. 5. In Sect. 6, we report comparative performance results. Finally, in Sect. 7, we conclude the paper and mention our future work.

2. Related Work

Most neuroimaging studies of cerebral atrophy and neurodegeneration are MRI-based, especially in Alzheimer's disease and related dementia research.[4,8] By contrast, CT-based brain analysis has been more limited and has often focused on other diagnostic tasks such as hemorrhage detection.[13] Nevertheless, CT remains attractive in routine clinical practice because it is widely available, inexpensive, and fast. Recent work has shown that CT can support meaningful quantitative atrophy analysis: Gao *et al.*[13] used deep learning for CT-based brain image classification, Cauley *et al.*[1] developed an automated method to quantify brain parenchymal fraction and related measures from clinical head CT, and Kaipainen *et al.*[9] reported that CT-based automated analysis can provide results comparable to MRI for measuring brain atrophy and white matter lesions.

Conventional atrophy assessment has commonly relied on structural markers such as ventricular enlargement and sulcal widening. Linear and volumetric indices, including Evans' Index and the brain volume/cerebrospinal fluid ratio, have been used to characterize global brain atrophy.[7,14] Chrzan *et al.*[2] further showed on CT scans that multiple atrophy-related parameters increase with age, confirming that quantitative CT markers can capture progressive structural degeneration even in very elderly adults.

Several studies have also linked brain atrophy to psychiatric or neuropsychiatric conditions. Opel *et al.*[5] reported that reduced hippocampal volume is strongly associated with childhood maltreatment in patients with major depressive disorder (MDD), while Lee *et al.*[10] identified regional gray matter volume differences between suicide attempters and suicide nonattempters with MDD. Banasr *et al.*[11] further described depression as being associated with structural alterations in limbic regions, and Espinoza Oyarce *et al.*[3] summarized broad volumetric differences in depression through a large meta-analysis. Koolschijn *et al.*[6] also reported structural brain-volume abnormalities in MDD from meta-analytic MRI evidence.

Importantly, evidence relevant to young groups has started to emerge. Cauley *et al.*[1] analyzed adult clinical CT data from individuals and provided age-stratified quantitative atrophy references, including early-adult ranges. In addition, Ragan *et al.*[14] evaluated ventricular linear indices in a pediatric cohort, supporting the utility of imaging-based structural markers outside late-life populations. In CT-based neuropsychiatric imaging, Backman *et al.*[15] showed that early cortical atrophy is associated with clinical depression in parkinsonism patients. Related recent studies have expanded atrophy analysis to other conditions, including post-stroke cognitive impairment, Alzheimer's disease subtypes, dementia, and type 2 diabetes mellitus.[16–20]

Deep learning has become a dominant approach for medical image analysis, with architectures such as U-Net supporting robust representation learning in biomedical imaging.[21] More recently, large-scale models have been explored for 3D medical imaging.[22] At the same time, explainability has become increasingly important for clinical deployment, and Grad-CAM has been widely adopted to highlight image regions that drive model predictions.[23] However, most prior studies on brain atrophy were focused on MRI, single-modal imaging, or diseases other than psychiatric-risk populations. Few groups have addressed CT-based cerebral atrophy detection by jointly modeling imaging findings and psychological risk factors while also providing interpretable visual evidence for the prediction. This gap was the motivation behind the explainable multimodal framework proposed in this paper.

3. Dataset

The dataset was collected from the Kangbuk Samsung Hospital health-check cohort, including adults aged 18 years or older. The source repository was collected between 2010 and 2020. For this study, we constructed a cohort of 518 subjects for cerebral atrophy classification using CT imaging and psychiatric-risk information including 271 atrophy cases and 247 controls. Each subject had 20 to 50 axial CT slices. Because the number of CT slices varied across subjects, we adopted a fixed three-slice input to standardize the model across the cohort. We selected the three middle axial slices because they generally preserve more stable and diagnostically informative parenchymal features for atrophy assessment (e.g., ventricular configuration and sulcal widening), whereas superior and inferior slices are often less consistent informative context. This strategy balances information effectiveness and computational efficiency, enabling robust and feasible model training while preserving clinically meaningful structural signals.

Intermediate axial slices capture representative structural changes associated with cerebral atrophy, including gray and white matter volume loss, ventricular enlargement, and sulcal widening. Ground-truth labels were established from expert radiologist interpretation of brain CT scans in the routine clinical workflow at Kangbuk Samsung Hospital. In this retrospective health-check cohort, the available reference standard followed routine radiology reporting categories, so each subject was annotated in binary form (i.e., atrophy versus no atrophy). Therefore, we used the same binary labels for consistent and reliable model development and evaluation.

The brain CT scans were stored in the Digital Imaging and Communications in Medicine (DICOM) format. Each DICOM object contains both imaging data and metadata (e.g., patient identifier and study information), enabling reliable matching between images and clinical records. To protect patient privacy, all personal identifiers were removed through anonymization and pseudonymization before analysis.

4. Methods

Figure 1 illustrates the overall pipeline of the proposed cerebral atrophy classification framework. The framework consists of four components: (i) data preprocessing, (ii) image-based classification, (iii) multimodal fusion with psychological variables, and (iv) Grad-CAM-based visual explanation.

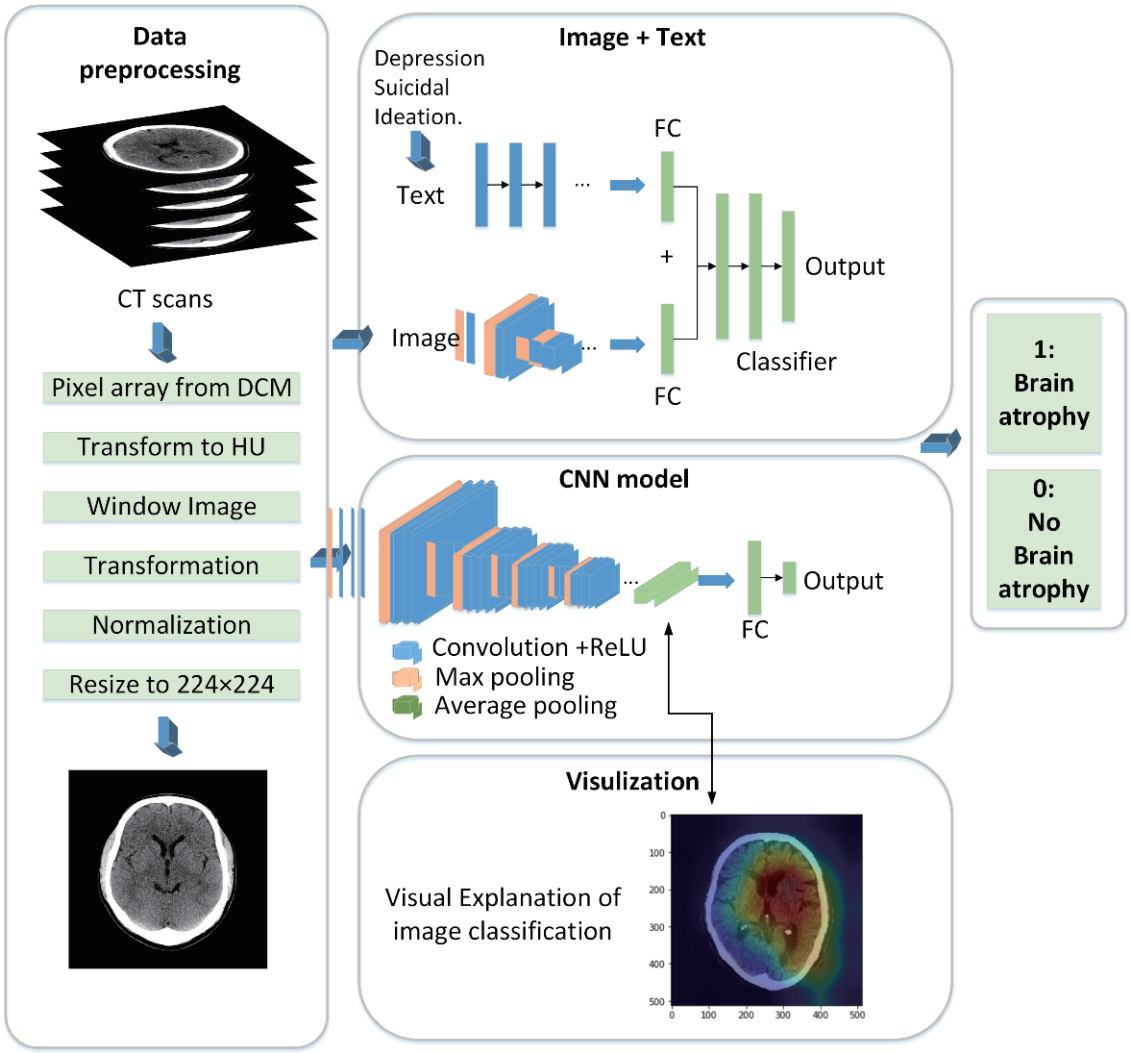

Fig. 1. (Color online) Framework of cerebral atrophy classification.

We first evaluate image-only classification using one middle slice and three middle slices per subject to assess the effect of slice selection. We then use the three-slice setting and combine CT-derived image features with psychological risk information in a multimodal model. Finally, we generate class activation heatmaps to visualize image regions contributing to model predictions.

4.1 Data preprocessing

Image pixel arrays were extracted from brain CT scans and preprocessed before model training. We converted pixel arrays to Hounsfield Unit (HU)-scaled images to preserve radiodensity information:[24]

$$I_{HU} = I_{raw} \times RescaleSlope + RescaleIntercept\,, \tag{1}$$

where I_{raw} is the original pixel value and I_{HU} is the converted HU value.

On the basis of brain-tissue radiodensity characteristics, a brain window was applied to each slice:

$$I_{brain} = min\left(\max\left(I_{HU}, WC - \frac{WW}{2}\right), WC + \frac{WW}{2}\right), \tag{2}$$

where WC and WW denote the window center and window width, respectively. After windowing, slices were converted to Python Imaging Library (PIL) image objects for transformation and normalization. Finally, each image was resized to a 224 × 224 tensor for further model training.

4.2 Image-based classification

For image-based classification, we used a basic CNN as a from-scratch baseline and compared it with multiple pretrained backbones for transfer learning. The basic CNN is composed of three convolutional blocks with Rectified Linear Unit (ReLU) activation and max-pooling, followed by a fully connected classifier. The pretrained candidates include ResNet50 and Alexnet. These models were selected on the basis of their strong performance in general image classification tasks and their varying architectural designs, which allow us to evaluate the impact of different feature extraction strategies on cerebral atrophy classification.

To capture feature changes across adjacent CT slices, we further designed two three-slice construction strategies. In the first strategy, termed SCF, three consecutive axial slices were stacked as a three-channel image and directly fed into a pretrained backbone. In the second strategy, termed SMBF, each slice was converted to a color image and passed into a separate pretrained model block; the three block outputs were then concatenated and processed by a linear layer to produce the final classification output. Formally, for three input slices x1, x2, and x3, the branch features are written as

$$f_i = g_i(x_i),\ i \in \{1,2,3\}, \qquad f_{cat} = [f_1; f_2; f_3], \tag{3}$$

where $g_i(\cdot)$ denotes the i-th pretrained feature extractor and $[\cdot\ ;\ \cdot]$ denotes feature concatenation. The final prediction is obtained by a linear classifier:

$$\hat{y} = \sigma(Wf_{cat} + b), \tag{4}$$

where $\sigma(\cdot)$ is the output activation function for binary classification.

To ensure a fair comparison, all models were trained with the same subject-level data split, one-slice and three-slice input settings, optimizer configuration, and stopping criteria. Model selection was based on validation performance, and the best-performing image backbone was used in the subsequent multimodal framework.

4.3 Multimodal model

We further developed a multimodal model to integrate imaging and nonimaging features. The image branch uses the best-performing pretrained backbone described in Sect. 4.2 to encode CT slices, while the clinical branch uses an MLP to encode psychological variables. The psychological variables include depression status and suicidal ideation status, both encoded as binary values (i.e., 1: positive, 0: negative). Let x denote the CT input and c denote the psychological feature vector. The fused representation is defined as

$$z_{img} = \phi_{img}(x),\ z_{cli} = \phi_{cli}(x),\ \ z = [z_{img}; z_{cli}], \tag{5}$$

where $\phi_{img}(\cdot)$ and $\phi_{cli}(\cdot)$ are the image and clinical encoders, respectively. The final prediction is obtained by

$$\hat{y} = \sigma(Wz + b), \tag{6}$$

where $\sigma(\cdot)$ is the output activation function for binary classification. The entire multimodal network is jointly optimized end to end.

4.4 Visual explanation

We used Grad-CAM[23] to generate localization maps highlighting regions that most strongly affect classification decisions. Grad-CAM was computed from the final convolutional block of the best-performing pretrained image backbone using the atrophy-class score as the target class. Convolutional feature maps retain spatial information, and the last convolutional layer provides a strong balance between semantic discrimination and localization. Grad-CAM computes a class-

specific localization map $L^c_{Grad-CAM} \in R^{u \times v}$ for class c using the feature maps of the last convolutional layer. First, the neuron importance weights are computed as

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \ , \tag{7}$$

where A^k is the activation map of channel k in the target convolutional layer, and $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient of the class-c score with respect to spatial location (i, j). Then, the gradient-weighted class activation map is generated as

$$L^c_{Grad-CAM} = ReLU\left(\sum_k a_k^c A^k \right). \tag{8}$$

The gradient weights are obtained by globally averaging gradients over spatial dimensions, then combining them with forward activation maps, followed by a ReLU operation, as shown in Eq. (8). The resulting Grad-CAM map is min–max normalized to [0, 1], resized to the CT resolution input, and overlaid on the original image to generate an interpretable heatmap.

5. Implementation

5.1 Experiment setup

We evaluated image-only models under one-slice and three-slice input settings (i.e., SCFSCF and SMBF strategies), and then trained the multimodal model using the best-performing image backbone. Grad-CAM maps were generated from that backbone for visual interpretation.

To prevent data leakage, we used a subject-level protocol such that all samples from each participant were confined to a single subset. The dataset was split into training, validation, and test sets. For the CNN single-slice and multimodal models, we used a 6:2:2 split, while the other models used 8:1:1 split. The test set remained fully held out during model development, while model selection was conducted via subject-level threefold cross-validation within the training and validation subset. Final performance was then reported on the independent subject-level test set. Hyperparameters were selected using a validation-based tuning strategy under a shared training protocol for fair comparison across models. The final configuration was the Adam optimizer, a learning rate of 1×10^{-6}, a batch size of 8, and up to 100 epochs. Across the folds, model ranking trends were consistent, indicating the limited sensitivity of the main comparative conclusions to split variation. The final hyperparameter setting is summarized in Table 1.

Table 1
Model hyperparameters.

Hyperparameter	Value
Cross-validation folds	3
Batch size	8
Epochs	100
Optimizer	Adam
Learning rate	1×10^{-6}

5.2 Evaluation metrics

We evaluated classification performance using accuracy, precision, recall, and *F*1-*score*. For each selected class, *TP*, *TN*, *FP*, and *FN* were computed by treating that class as positive and the other class as negative. These metrics are defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$Precision = \frac{TP}{TP + FP}, \tag{10}$$

$$Recall = \frac{TP}{TP + FN}, \tag{11}$$

$$F1\_score = \frac{2*Precision*Recall}{Precision + Recall}. \tag{12}$$

Recall (sensitivity) and *F*1-*score* are clinically important in this task because both missed atrophy cases (false negatives) and misclassification of nonatrophy cases (false positives) can lead to inappropriate clinical decisions. Therefore, to account for performance in both classes under class imbalance, we report macro-averaged precision, macro-averaged recall, and macro-averaged *F*1-*score*. The macro-averaged metrics are then obtained by taking the mean of the two class-wise values, as shown in the following equations:

$$Macro - Precision = \frac{Precision_0 + Precision_1}{2}, \tag{13}$$

$$Macro - Recall = \frac{Recall_0 + Recall_1}{2}, \tag{14}$$

$$Macro - F1 = \frac{F1_0 + F1_1}{2} . \quad (15)$$

6. Performance

Table 2 summarizes classification performance across image-only and multimodal settings. To ensure transparent interpretation, we note that the CNN single-slice and multimodal models were evaluated with a participant-wise 6:2:2 split, whereas the other models were evaluated with an 8:1:1 split; therefore, absolute metric comparisons across all models should be interpreted with caution due to different test-set sizes. Within the single-slice setting, the transfer-learning baseline outperformed the from-scratch CNN baseline, supporting the benefit of pretrained feature extraction under the current dataset. Extending to three slices generally improved performance, although SCF did not consistently improve over single-slice settings, suggesting that direct channel stacking may not always capture interslice information efficiently. Under the SMBF design, where slices are encoded in separate branches before feature fusion, both AlexNet and ResNet50 showed stronger balanced performance than simpler image-only baselines within the same protocol. In this setting, ResNet50 (SMBF) showed a slightly higher *F*1-*score* than AlexNet (SMBF), indicating a modest advantage in feature representation under the current experiment conditions. When image features were combined with psychological risk variables in the multimodal model, performance improved within its evaluation protocol, including higher recall, which is clinically relevant because false negatives (missed atrophy cases) are associated with greater diagnostic risk. Overall, these results are encouraging but should be interpreted as preliminary rather than deployment-ready evidence, given the limited cohort size, nonuniform split protocols across all model groups, and the absence of prospective multi-center external validation and threshold-calibration analysis.

6.1 Confusion matrix

Figure 2 presents row-normalized confusion matrices for all evaluated models. In each matrix, rows denote the true class labels and columns denote the predicted class labels; thus, each row sums to 1 and represents the distribution of predictions within that true class. Because two evaluation protocols were used, the row totals differ by model group. For the CNN single-

Table 2
Classification performance of image-only and multimodal models on the cerebral atrophy dataset.

Model	Input	Accuracy (%)	Precision-macro (%)	Recall-macro (%)	*F*1 *scor*e-macro (%)
CNN	1 slice	74.14	74.25	74.19	74.13
ResNet50	1 slice	79.25	80.38	78.64	78.76
CNN	3 slices	79.25	81.58	78.42	78.48
ResNet50 (SCF)	3 slices	77.36	77.42	77.5	77.78
AlexNet (SMBF)	3 slices	83.01	87.83	82.00	82.10
ResNet50 (SMBF)	3 slices	83.01	84.39	82.43	82.62
Multimodal	3 slices + 2 clinical	87.25	91.58	84.84	88.08

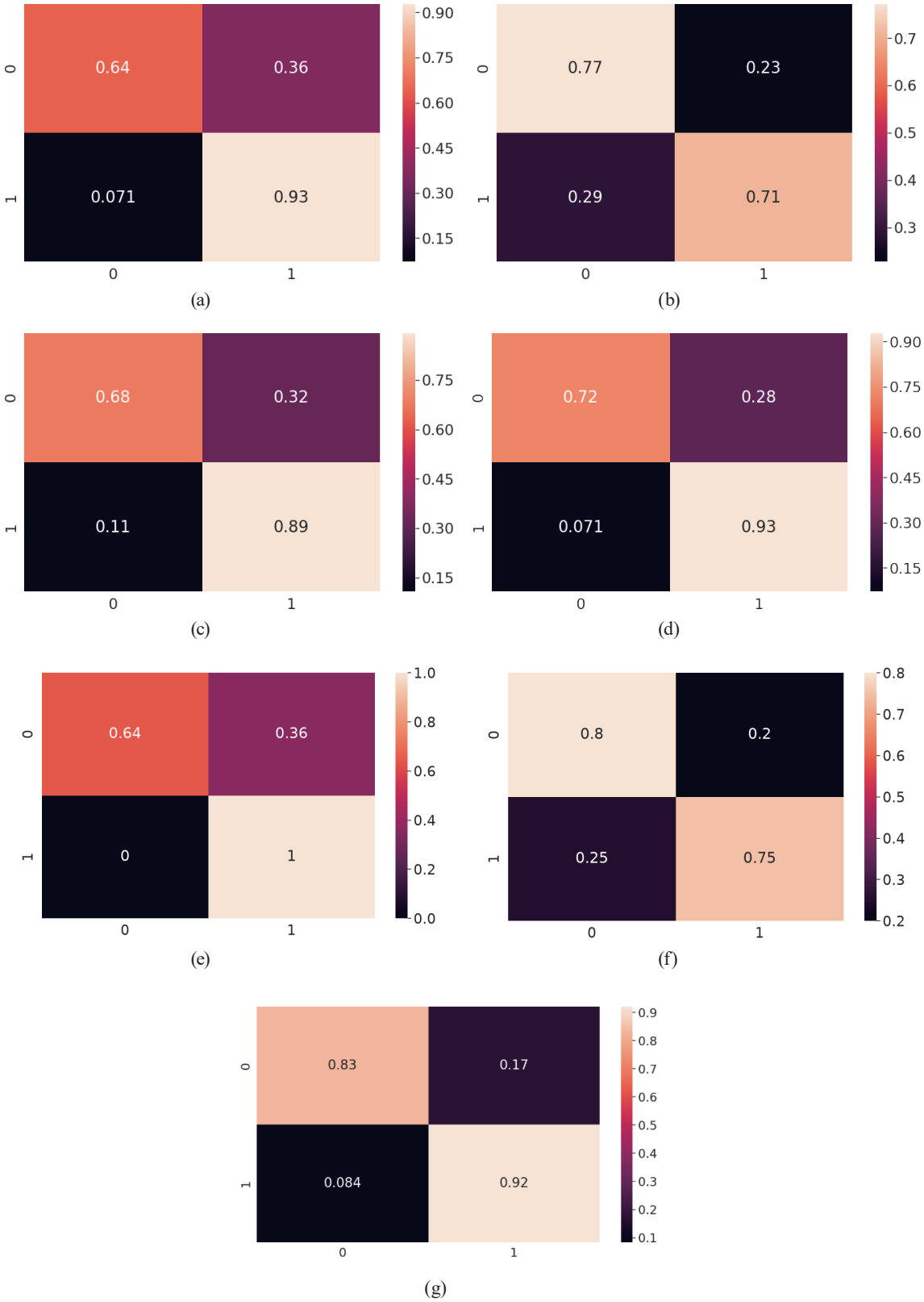

Fig. 2. (Color online) Confusion matrices. (a) CNN - 1 slice, (b) CNN - 3 slices, (c) ResNet50 - 1 slice, (d) ResNet50 - 3 slices - SMBF, (e) Alexnet - 3 slices - SMBF, (f) ResNet50 - 3 slices - SCF, and (g) Multimodal - 3 slices + psychological risk factors.

slice and multimodal models (6:2:2 split), the test set contained 57 negative samples and 59 positive samples. For the other models (8:1:1 split), the test set contained 25 negative samples and 28 positive samples. Accordingly, raw confusion-matrix counts are recovered by multiplying each row-normalized value by the corresponding row total for that model group.

Panels (a) and (b) correspond to the CNN from-scratch model with one-slice and three-slice inputs, respectively. Panels (c) and (d) show the pretrained ResNet50 model under the same input settings, with panel (d) using the SMBF strategy for three-slice fusion. Panel (e) shows the pretrained AlexNet model with 3 slices of input under the SMBF strategy. Panel (f) shows the pretrained ResNet50 model with 3 slices of input under the SCF strategy. Panel (g) presents the multimodal model (i.e., three CT slices plus psychological risk factors), enabling direct visual comparison with image-only baselines. Under row normalization, a lower *FN* proportion and a higher *TP* proportion in the positive-class row indicate better sensitivity (i.e., macro-averaged recall). Compared with image-only models, panel (g) shows improved class-based detection, consistent with the quantitative gains in Table 2 and supporting the added value of psychological risk factors for borderline atrophy cases.

Clinically, false negatives (missed atrophy cases) are of particular concern because undetected atrophy in high-risk psychiatric populations may delay further assessment and appropriate intervention. Figure 2 presents row-normalized confusion matrices for all evaluated models and visualizes class-wise prediction patterns corresponding to Table 2. Because two evaluation protocols were used (6:2:2 for the CNN single-slice and multimodal models; 8:1:1 for the other models), confusion-matrix patterns and summary metrics should be compared primarily within the same protocol. Within the image-only settings, multi-slice configurations showed fewer false negatives than single-slice baselines in the corresponding protocol, with associated improvements in recall. The multimodal model also showed improved balance between false negatives and false positives within its own evaluation protocol, consistent with higher overall performance metrics in that setting. Overall, these findings suggest that adjacent-slice information and multimodal features may improve classification performance in this cohort, while requiring further confirmation under a unified split protocol and external validation.

6.2 Grad-CAM visualization

Figure 3(a) shows a representative Grad-CAM heatmap generated from the best-performing image backbone. The left image shows the original scan used for classification, and the right image shows a heatmap overlay that combines feature-attention gradients with the original image to highlight regions that most influenced the prediction. This case is from a patient with brain atrophy. The most influential features include ventricular enlargement and widening of the cortical sulci, both of which are characteristic imaging findings of brain atrophy.

In the false-negative case shown in Fig. 3(b), the true label was atrophy, but the model predicted nonatrophy. Grad-CAM showed a focal hotspot near the frontal peripheral region, with relatively limited activation across broader regions typically considered in atrophy assessment.

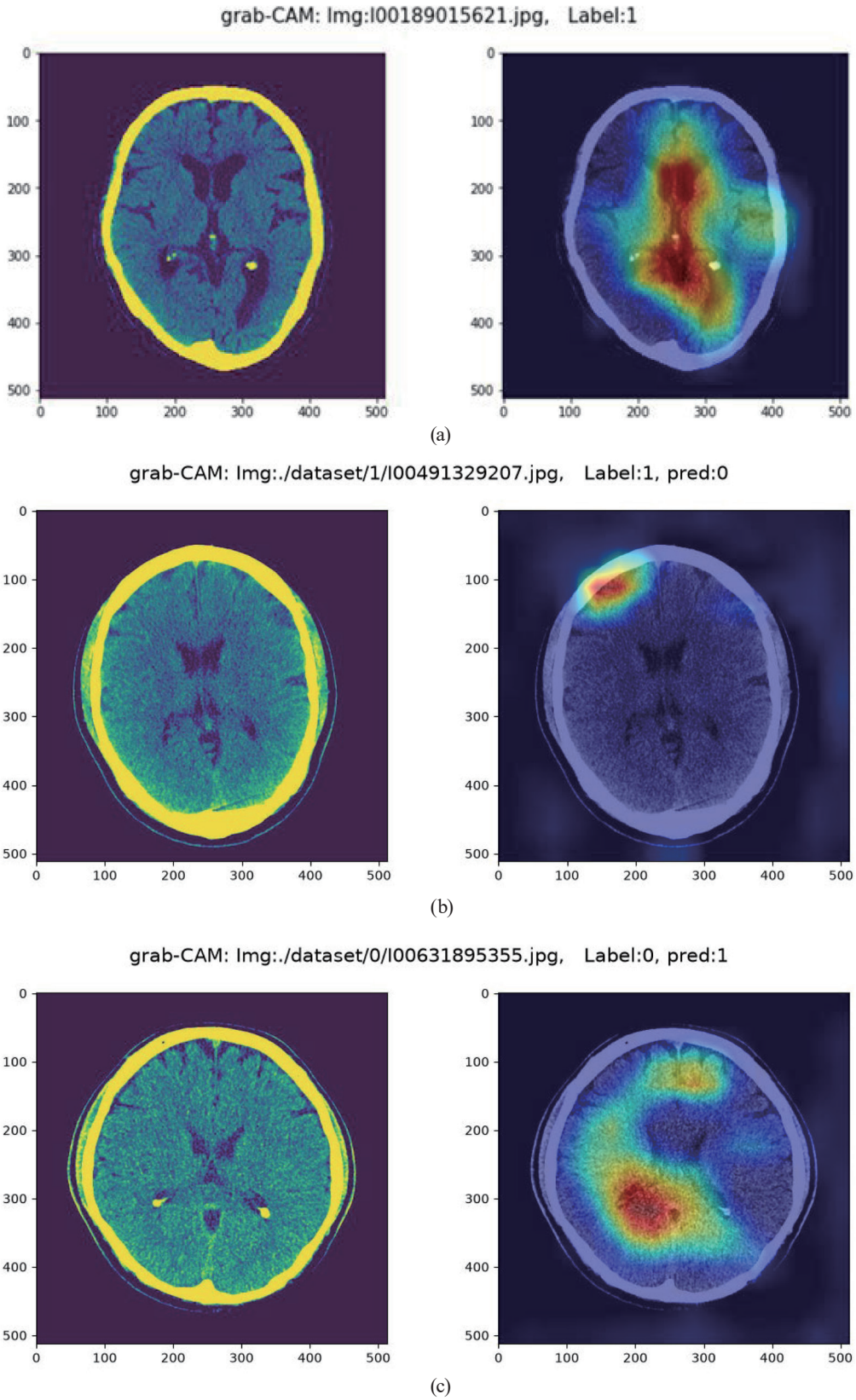

Fig. 3. (Color online) Visual explanation. (a) *TP* samples, (b) *FN* samples, and (c) *FP* samples.

This localized attention pattern suggests that subtle or spatially distributed atrophic features were not sufficiently captured, providing a case-based explanation for the missed detection.

In the false-positive case shown in Fig. 3(c), the true label was nonatrophy, but the model predicted atrophy. Grad-CAM demonstrated strong activation in central intracranial and superior cortical regions, with diffuse emphasis that was not confined to a clearly lesion-relevant pattern. This suggests sensitivity to nonspecific morphologic or intensity characteristics that can mimic atrophy-related cues, providing a data-based explanation for false alarms and reduced specificity.

Taken together, these error cases suggest that misclassification is more likely when ventricular enlargement is subtle or when age-related morphologic variability overlaps with mild atrophy patterns. Under such conditions, Grad-CAM tends to show broader and more diffuse attention around ventricular and sulcal regions, indicating weaker structural contrast for class separation. This pattern suggests that the current model may underutilize distributed atrophy cues and may be influenced by nonspecific intensity or shape variation. To improve robustness, future work should incorporate stronger augmentation strategies and 3D contextual modeling.

7. Conclusions

We presented an explainable multimodal framework for cerebral atrophy classification in psychiatric-risk populations by integrating CT imaging with psychological risk factors. In this cohort, pretrained image models showed higher performance than from-scratch CNN baselines, and the multimodal model achieved the best threshold-dependent results among the evaluated settings. Grad-CAM visualizations provided supportive interpretability by highlighting regions commonly associated with atrophy-related structural change, including ventricular enlargement and sulcal widening. These findings suggest the feasibility of CT-based multimodal assessment, particularly in contexts where MRI accessibility is limited.

Despite these encouraging results, several limitations should be noted. First, we used a relatively small cohort, adopted a slice-based 2D design for an inherently volumetric process, and did not include external multi-center validation. Second, threshold-independent evaluation (for example, ROC-AUC with confidence intervals) could not be fully reported in this work, and augmentation and segmentation-guided analysis remained limited. Third, reproducibility and additional reanalysis are constrained by current data-access conditions: although the preserved final experiment package supports verification of the reported confusion-matrix-based metrics, trained model weights, and representative Grad-CAM cases, it does not support newly requested reanalysis in the present cycle. In line with this constraint, we restricted our conclusions to verifiable results only. Misclassification further indicates that errors can arise when subtle ventricular/sulcal patterns overlap with age-related morphologic variability, with diffuse or nonspecific attention patterns observed in representative false-negative and false-positive cases. Therefore, the current findings should be interpreted as preliminary evidence rather than deployment-ready performance.

Future work should prioritize larger and more diverse multi-center datasets, direct comparison with extended multi-slice and volumetric approaches, segmentation-aware modeling, stronger augmentation, broader clinical feature integration, and more rigorous subject-level validation protocols, together with complete threshold-independent performance reporting.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00398199) and also by AI Convergence Research Fund, Sungkyunkwan University, 2022.

References

1 K. Cauley, Y. Hu, and S. Fielden: Am. J. Neuroradiol. **41** (2020) 809. https://www.ajnr.org/content/41/5/809
2 R. Chrzan, A. Gleń, A. Bryll, and A. Urbanik: Int. J. Environ. Res. Public Health **16** (2019) 3659. https://www.mdpi.com/1660-4601/16/19/3659
3 D. A. Espinoza Oyarce, M. E. Shaw, K. Alateeq, and N. Cherbuin: J. Psychiatry Neurosci. **45** (2020) 406.
4 F. Zhang, B. Pan, P. Shao, P. Liu, S. Shen, P. Yao, and R. X. Xu: arXiv preprint arXiv:2107.13200 (2021).
5 N. Opel, R. Redlich, P. Zwanzger, D. Grotegerd, V. Arolt, W. Heindel, C. Konrad, H. Kugel, and U. Dannlowski: Neuropsychopharmacology **39** (2014) 2723. https://doi.org/10.1038/npp.2014.145
6 P. C. M. P. Koolschijn, N. E. M. van Haren, G. J. L. M. Lensvelt-Mulders, H. E. Hulshoff Pol, and R. S. Kahn: Hum. Brain Mapp. **30** (2009) 3719.
7 C. Orellana, D. Ferreira, J.-S. Muehlboeck, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, L.-O. Wahlund, and E. Westman, for the AddNeuronMed consortium, and for the Alzheimer's Disease Neuroimaging Initiative: Neurodegenerative Dis. **16** (2016) 77. https://doi.org/10.1159/000442443
8 J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, and O. Colliot: Med. Image Anal. **63** (2020) 101694. https://www.sciencedirect.com/science/article/pii/S1361841520300591
9 A. L. Kaipainen, J. Pitkänen, F. Haapalinna, O. Jääskeläinen, H. Jokinen, S. Melkas, T. Erkinjuntti, R. Vanninen, A. M. Koivisto, J. Lötjönen, J. Koikkalainen, S.-K. Herukka, and V. Julkunen: Neuroradiology **63** (2021) 2035.
10 Y. J. Lee, S. Kim, A. R. Gwak, S. J. Kim, S.-G. Kang, K.-S. Na, Y.-D. Son, and J. Park: Compr. Psychiatry **67** (2016) 59.
11 M. Banasr, J. M. Dwyer, and R. S. Duman: Curr. Opin. Cell Biol. **23** (2011) 730.
12 X. Yu, J. P. Jeong, Y. Chang, R. Kwon, J. Kang, G.-Y. Lim, and S. Ryu: Proc. 15th Int. Conf. Mobile Computing and Ubiquitous Networking (ICMU, 2025) 1–4.
13 X. W. Gao, R. Hui, and Z. Tian: Comput. Methods Programs Biomed. **138** (2017) 49. https://www.sciencedirect.com/science/article/pii/S0169260716305296
14 D. K. Ragan, J. Cerqua, T. Nash, R. C. McKinstry, J. S. Shimony, B. V. Jones, F. T. Mangano, S. K. Holland, W. Yuan, and D. D. Limbrick: J. Neurosurgery: Pediatrics **15** (2015) 547. https://thejns.org/pediatrics/view/journals/j-neurosurg-pediatr/15/6/article-p547.xml
15 E. A. Backman, L. Luntamo, R. Parkkola, J. Koikkalainen, M. Gardberg, and V. Kaasinen: J. Neurol. Sci. **455** (2023) 122804. https://www.sciencedirect.com/science/article/pii/S0022510X23022657
16 W. Jia, Y. Zhou, L. Zuo, T. Liu, and Z. Li: Brain Res. **1822** (2024) 148635. https://doi.org/10.1016/j.brainres.2023.148635
17 R. Mohanty, D. Ferreira, and E. Westman: Cereb. Circ. - Cognit. Behav. **6** (2024) 100294. https://doi.org/10.1016/j.cccb.2024.100294

18 N. Gao, C. Ye, H. Chen, X. Hao, and T. Ma: Hum. Brain Mapp. **45** (2024) e26715. https://doi.org/10.1002/hbm.26715
19 S. R. Syed and M. A. Saleem Durai: Front. Neurosci. **17** (2023) 1291753. https://doi.org/10.3389/fnins.2023.1291753
20 J. Cui, C. Chen, and S. Hilal: Cereb. Circ. - Cognit. Behav. **9** (2025) 100463. https://doi.org/10.1016/j.cccb.2025.100463
21 O. Ronneberger, P. Fischer, and T. Brox: arXiv preprint arXiv:1505.04597 (2015). http://arxiv.org/abs/1505.04597
22 J. Wu, Y. Wang, Z. Zhong, W. Liao, N. Trayanova, Z. Jiao, and H. X. Bai: npj Artif. Intell. **1** (2025) 17.
23 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra: Int. J. Comput. Vision **128** (2020) 336. https://doi.org/10.1007/s11263-019-01228-7
24 D. Mishra, R. K. Ghimire, R. B. Chand, N. Thapa, and O. B. Panta: J. Inst. Med. Nepal **38** (2016) 70.