Intelligent Information Systems

UDC 004.934:141.1 **doi:** https://doi.org/10.20998/2522-9052.2026.2.09

Alexandra Čižmárová[1], Kristína Dostálová[1], Patrik Hrkut[1], Anton Poroshenko[2]

[1] Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia
[2] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

EXPLAINABLE ARTIFICIAL INTELLIGENT METHOD GRAD-CAM IN MEDICAL IMAGES PROCESSING

Abstract. The reliability of modern deep learning models in the medical domain is frequently questioned due to their black-box nature. Post-hoc explainability techniques from the field of explainable artificial intelligence (XAI) offer a means to improve transparency and assess the reliability of predictions produced by convolutional neural networks. **The research aims** to investigate how XAI methods, specifically Gradient-weighted Class Activation Mapping (Grad-CAM), can provide reliable explanations for medical image classification. For this purpose, MRI images of brain were used to train a convolutional neural network to categorize the four stages of dementia in Alzheimer's disease. To make each prediction transparent, the areas of the brain which the trained network used to make the categorization on were highlighted using Grad-CAM. The resulting relevance maps, heatmaps, were evaluated using two approaches: spatial comparison with anatomically defined brain regions associated with Alzheimer's disease using atlas overlay, and quantitative faithfulness assessment using a deletion-based metric, where highly influential regions identified by Grad-CAM were progressively removed and the impact on classification confidence was measured.

Keywords: explainable AI; convolutional neural network; Grad-CAM; Alzheimer's disease; heatmap; gradients.

Introduction

Relevance and an overview of scientific works. One of the modern trends in medicine is precision, or smart, medicine, which entails the widespread application of deep learning models [1], particularly convolutional neural networks (CNNs) [2]. CNN demonstrates strong performance in medical image analysis tasks such as disease detection, classification, and disease staging [3]. In neuroimaging, CNN-based approaches have been widely applied to brain MRI for the analysis of Alzheimer's disease and related neurodegenerative disorders [4]. Despite their predictive accuracy, the deployment of such models in clinical environments remains limited by concerns about transparency, trust, and the reliability of automated decision-making [5].

From a broader perspective, reliability and trustworthiness are central requirements for healthcare decision-support systems, where incorrect or poorly understood model outputs may have serious consequences. Recent studies in reliability engineering for healthcare emphasise that model performance alone is insufficient and must be complemented by mechanisms that enable practitioners to understand, assess, and validate system behaviour under uncertainty [6]. In this context, explainabilitvy [5, 7] and importance analysis [8] are increasingly recognised as key components of reliable intelligent systems in medicine.

Several lines of research have addressed interpretability and importance analysis using transparent or semi-transparent models. For example, fuzzy-based classification methods and fuzzy decision trees have been successfully applied in smart and precision medicine to provide both competitive predictive performance and explicit information about the relative importance of decision-making factors [2, 8]. These approaches highlight the value of explanations that are not only accurate but also interpretable and reliable, enabling domain experts to reason about model outputs and their underlying drivers.

However, many state-of-the-art medical imaging systems rely on deep neural networks whose internal representations are inherently opaque. Explainable Artificial Intelligence (XAI) has therefore emerged as a complementary approach, providing post-hoc explanations for complex models without substantially compromising their predictive performance [9, 10]. Among post-hoc XAI techniques, Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely adopted methods for visualising class-discriminative regions in CNN-based image classifiers [11, 12]. Grad-CAM produces spatial heatmaps that highlight image regions contributing most strongly to a model's prediction, making it particularly attractive in medical imaging scenarios where spatial localisation is critical [11].

Despite its popularity, the practical reliability of Grad-CAM explanations in medical applications remains an open challenge. In many published studies, Grad-CAM heatmaps are evaluated primarily through qualitative visual inspection, with explanations considered acceptable if they appear visually plausible or align with expected anatomical patterns. However, recent research has demonstrated that visual plausibility alone does not guarantee that a saliency map faithfully reflects the features actually driving a model's prediction. In medical imaging, widely used methods such as Grad-CAM have been shown to produce visually convincing explanations that nonetheless exhibit poor localization accuracy, limited robustness, and weak faithfulness, raising concerns for clinical use [13, 14].

Setting objectives. From a reliability engineering perspective, this problem can be viewed as a gap between

model output and the decision justification. While Grad-CAM provides a form of importance visualisation, it does not inherently guarantee that the highlighted regions are either anatomically meaningful or causally relevant to the prediction. Consequently, there is a growing consensus that explanations generated by post-hoc XAI methods should be treated as validation artefacts rather than self-evident justifications, and should be systematically assessed using complementary evaluation strategies.

In this work, we focus on evaluating Grad-CAM explanations for Alzheimer's disease stage classification from brain MRI. Rather than proposing a new explainability method, we investigate how the reliability of Grad-CAM can be assessed within a structured evaluation pipeline inspired by reliability engineering and importance analysis principles. The proposed approach combines two complementary perspectives. First, anatomical grounding is achieved by aligning Grad-CAM activation maps with a unified bank of brain atlases, enabling interpretation of model attention at the level of named anatomical structures. Second, faithfulness is quantitatively evaluated using a perturbation-based deletion metric that measures the model's confidence sensitivity to the removal of regions identified as important.

In addition, we analyze practical factors that influence explanation quality, including the choice of convolutional layer and the design of visualization representations. Through experiments on a CNN trained to classify four stages of Alzheimer's disease, we demonstrate that Grad-CAM explanation reliability is not uniform across classes and that visually plausible heatmaps do not necessarily imply faithful model reasoning. These findings reinforce the broader message from reliability-focused research in healthcare: explainability mechanisms must themselves be evaluated and validated before they can be relied upon in safety-critical decision-support systems.

Gradient-weighted Class Activation Mapping (Grad-CAM) method

Explainable Artificial Intelligence refers to a suite of methods that aim to make the decision-making processes of AI systems more understandable to human users, without substantially compromising their performance. XAI methods seek to provide insights into how a model decides on specific outputs by exposing which image regions or features led to the prediction. Regions that have the greatest influence on a model's prediction can be identified using various explainability techniques, such as saliency maps, attention mechanisms, or example-based explanations. XAI enables users to validate predictions, detect biases, and overall improve trust in AI-based systems [2].

To interpret the model's classificational decisions, the Gradient-weighted Class Activation Mapping (Grad-CAM) method was applied to test MRI images. It belongs to the family of post-hoc explainability techniques in the field of Explainable Artificial Intelligence (XAI), which aim to make the inner workings of complex models more transparent and understandable to human users. There are multiple categories of post-hoc explainability techniques, each with its own methods. Grad-CAM belongs in saliency-based techniques, which share a single core principle – they produce a spatial map that quantifies how much each pixel (or voxel) contributes to the network's output for a specific class. By back-propagating information from the prediction to the input, they translate the abstract decision-making process of a deep model into a visual heatmap that clinicians can interpret [5, 7].

Grad-CAM computes the gradient of the target-class score with respect to the last convolutional feature maps, pools these gradients to obtain channel-wise importance weights, and forms a weighted sum of the feature maps to produce a class-discriminative heatmap. The importance of each convolutional feature map is determined by computing the gradient of the class score y^c with respect to the feature map activations A^k. These gradients $\frac{\partial y^c}{\partial A_{ij}^k}$, obtained via backpropagation, measure how sensitive the model's prediction for class c is to changes at each spatial location (i, j) in feature map k. The gradients are then spatially aggregated using global average pooling to produce channel-wise importance weights as:

$$\alpha_k^c = \frac{1}{H \cdot W} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^c}{\partial A_{ij}^k}.$$

H and W denote the spatial dimensions of the selected convolutional feature maps, i.e., the feature map height (H, number of rows) and width (W, number of columns). This equation quantifies each feature map's contribution to the prediction. Because it requires only the model's gradients, it works with any CNN architecture and adds no overhead, making it the default visual explainer in many radiology prototypes [7, 10].

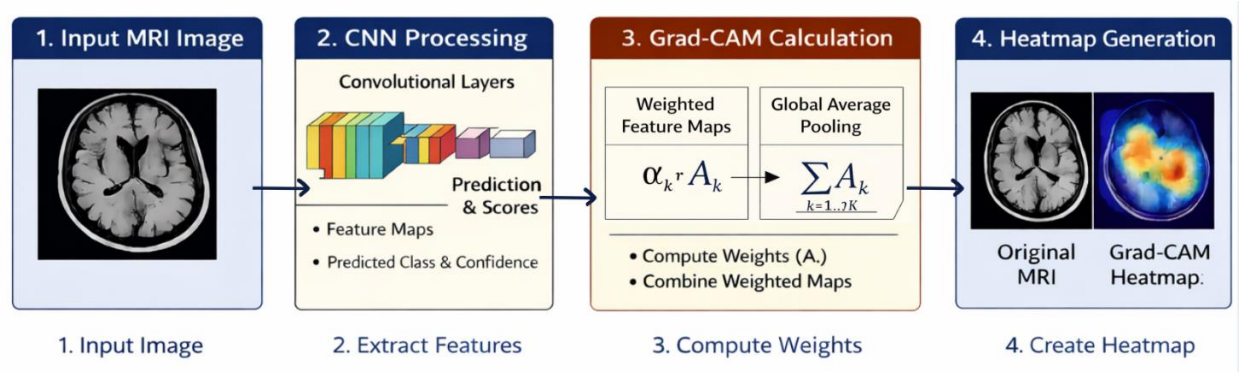

Fig. 1. Grad-CAM heatmap generation pipeline

The diagram on Fig. 1 illustrates the step-by-step process used to compute Grad-CAM explanations. First, an input brain MRI is passed through the convolutional neural network to extract hierarchical feature maps and produce a class prediction with an associated confidence score. Grad-CAM then computes the gradients of the target-class score with respect to the feature maps of the selected convolutional layer and aggregates these gradients to obtain channel-wise importance weights. The weighted feature maps are combined through global average pooling to form a class-discriminative activation map, which is finally upsampled and overlaid on the original MRI to produce the Grad-CAM heatmap highlighting regions that most influenced the model's decision.

Model Architecture and Performance

The model architecture for this research's purpose is based on a fine-tuned ResNet50, a convolutional neural network, trained on the dataset from Kaggle with focus on Alzheimer's. The dataset consisted of 6 200 pre-processed T1-weighted brain MRI scans. It includes four classes: *Non-Demented*, *Very Mild Demented*, *Mild Demented*, and *Moderate Demented*. The dataset was split into training, validation, and test sets using a deterministic approach to ensure correct reproducibility and class balance. The MRI images were pre-processed and normalised before used in training. The base model was modified by replacing the final dense layers to match the four output classes. All intermediate convolutional layers were initially frozen during the early training phases to allow the new classification head to adapt to the medical domain [15].

The final model achieved a test accuracy of 99.3%, with strong per-class performance reflected in F1-scores, macro-averaged precision and recall, as the latest two both achieved 99.4%. A confusion matrix, shown in Fig. 2, confirmed that most misclassifications occurred between neighbouring disease stages, suggesting the model learned clinically plausible distinctions.

In Fig. 2 the number representation is as follows: Non-Demented – 0, Moderate Demented – 1, Mild Demented – 2, and Very Mild Demented – 3.

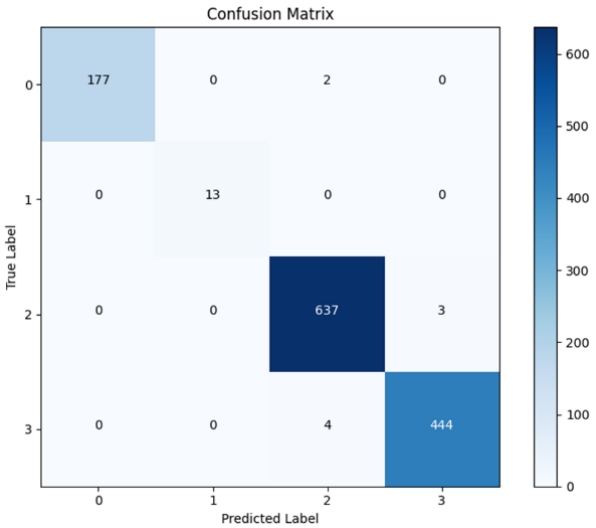

Fig. 2. Confusion matrix of the trained model

Results and Visualization

As first step after the model has finished training and its performance has been evaluated, Grad-CAM was applied to randomly selected MRI scans from the test pool for each of the four categories. Each MRI scan is pre-processed by resizing it to the input resolution specified in the configuration file and normalizing pixel intensities to the [0,1] range.

Grad-CAM is computed with respect to the predicted class using the final convolutional layer (conv5_block3_out). A gradient model is constructed to output the activations of this convolutional layer and the final prediction scores. Gradients of the predicted class score with respect to the convolutional feature maps are computed using GradientTape from TensorFlow (a Python-based deep learning library), which automatically differentiates by recording operations during the forward pass. Those are gradients that quantify the sensitivity of the prediction to changes in each feature map.

The gradients were globally averaged across spatial dimensions to obtain channel-wise importance weights, which were then combined with the corresponding feature maps through a weighted sum. Finally, a ReLU operation is applied to isolate the positive contributions, producing a coarse localisation map. The resulting Grad-CAM heatmap was normalised, resized to the original image resolution, and overlaid on the MRI image to visually highlight regions most influential for the model's decision.

This can be seen in Fig. 3.

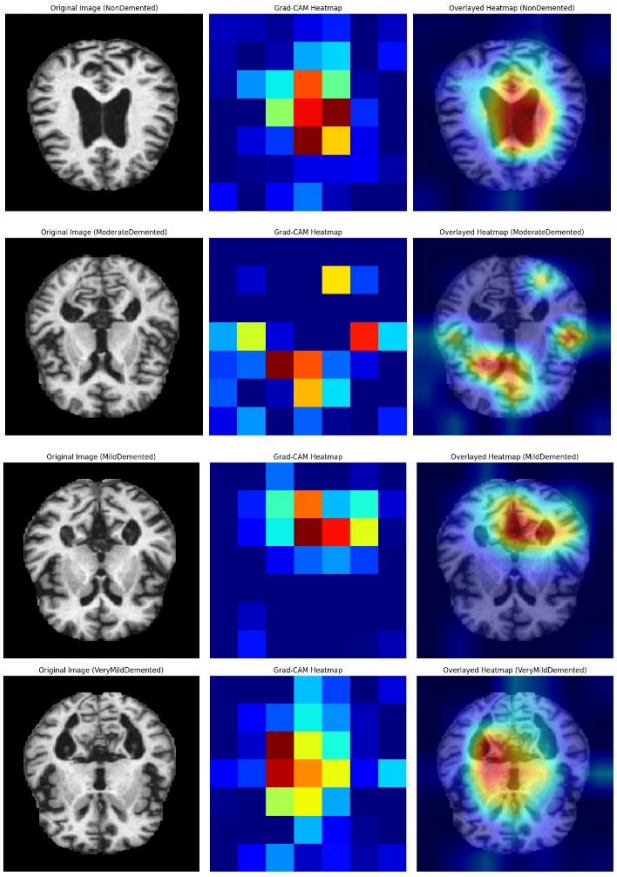

Fig. 3: Grad-CAM heatmaps for Alzheimer's classes

Heatmap activations outside the brain region

In a few cases, heatmaps produced misaligned or ambiguous activations, with heatmaps emphasising the background of the MRI scan and not the area of the brain (Fig. 4).

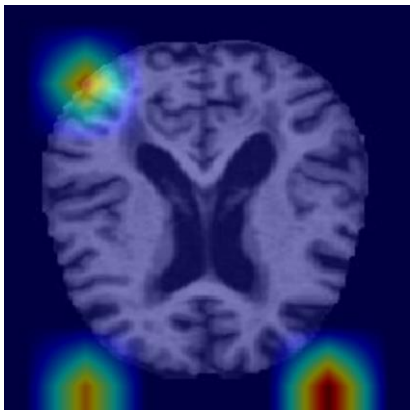

Fig. 4: Activation centred outside the brain in MRI scan

There are multiple hypotheses that may explain these misaligned activations, one of them being that the trained model is overfitting to non-diagnostic cues, such as pixel intensities in specific locations unrelated to brain anatomy. As to a possible cause, if looked at the Grad-CAM limitations, as it is known that sometimes it produces noisy explanations, particularly when activations are weak, or gradients are unstable.

To investigate this issue closer, a custom interactive tool was developed to explore Grad-CAM outputs across different convolutional layers. This revealed that while the default conv5_block3_out layer was most effective in general, earlier layers sometimes yield more precise spatial focus in difficult cases.

This comparison indicates that the quality and the most influential pixels of Grad-CAM explanations depend on the convolutional layer used for visualisation. While deeper layers generally capture more semantically meaningful features, they may also produce misaligned activations in challenging cases. Exploring multiple layers, therefore, provides valuable insight into the stability of Grad-CAM explanations.

Limited samples in *ModerateDemented* class leading to unstable explanations. During mentioned Grad-CAM analysis, the *ModerateDemented* class posed challenges, with several samples producing weak or nearly uniform heatmaps that lacked clear spatial focus. Some heatmaps shoved no influential areas at all yet the model categorization was correct. This observation suggests that the model may have overfitted to this class due the class imbalance, as *ModerateDemented* contained the fewest samples.

Interactive tool for layer-wise and class-wise Grad-CAM inspection. The layer comparison performed for cases with activations outside the brain showed that the default “last convolutional layer” is not always the most suitable choice. In some cases, an earlier layer can produce a more anatomically plausible focus since the activations were focused in the brain area. This raised a related question for the second issue mentioned - heatmaps with weak or seemingly absent activations, particularly in the underrepresented ModerateDemented class. Can the lack of visible focus be inherent to the model’s evidence, or can a different convolutional layer reveal a more informative activation pattern? To investigate both problems systematically, an interactive visualization tool was created to enable fast, consistent switching between layers and comparison of Grad-CAM outputs under different settings.

The tool was implemented as a lightweight web application around the trained classifier, allowing potential users/clinicians to browse test images by selecting a true class and a specific MRI slice. For the chosen image, the application displays the predicted class, confidence score, and the full probability distribution across all four Alzheimer categories. Grad-CAM can then be generated for any selected target class (not only the predicted one) and for any chosen convolutional layer, with the original MRI and the heatmap overlay shown side-by-side. This makes two comparisons straightforward - layer-wise changes in localization for the same prediction, and class-wise changes in what the model “looks for” on the same MRI when the target class is varied.

Overall, the tool adds practical value to the explainability workflow by making Grad-CAM inspectable rather than static. In addition, it supports rapid identification of failure modes (background attention, diffuse maps), provides evidence for a justified layer choice, and enables controlled comparisons across classes on the same image, by which it is improving the transparency of the evaluation without claiming clinical validity on its own.

Validation of Grad-CAM Explanations

To assess the reliability of Grad-CAM explanations beyond qualitative visual inspection, two complementary validation strategies were selected. First, a spatial comparison with an anatomical brain atlas was employed to evaluate whether regions highlighted by Grad-CAM correspond to anatomically meaningful structures. Second, a deletion-based faithfulness metric was applied to quantitatively measure the impact of removing highly relevant regions on the model’s prediction confidence. Together, these approaches were chosen to capture both anatomical plausibility and causal relevance of the generated explanations.

Anatomical Atlas Comparison. It is important to note that the atlas overlay is not treated as ground truth, but as an interpretative aid that supports anatomical plausibility assessment rather than definitive localisation. To move beyond purely qualitative inspection, an atlas-based spatial validation was implemented to test whether Grad-CAM highlights fall into anatomically meaningful brain structures. A unified atlas “bank” was created by combining three publicly available brain atlases: two Harvard–Oxford atlases (covering cortical and subcortical structures) and the Automated Anatomical Labelling atlas, which assigns an integer code to each anatomical region. All atlases were brought into the same standard reference space (the MNI152 brain template) and reduced to a set of representative axial slices. Each atlas slice contains both a reference anatomy image and

a corresponding label map, together with a lookup table that converts label IDs into region names [16–18].

For each input MRI slice, the atlas bank was aligned to the image using image registration, and the best-matching atlas slice was selected automatically. The main registration method uses feature-based matching (detecting stable points in both images and estimating an affine transform with outlier rejection), while a fallback intensity-based registration is used when feature matching is unreliable. Candidate alignments were scored using a combination of mutual information (a similarity measure robust to intensity differences) and edge overlap (agreement of anatomical boundaries). The highest-scoring alignment was then used to warp the atlas labels into the MRI slice coordinate system [16–18].

As a result, final visualization is a heatmap that is not presented as an abstract blob of colour, but as an anatomically grounded explanation: after aligning the atlas to the patient slice, the highlighted pixels can be interpreted in terms of named structures.

In the example shown on Figure 5, the model prediction for the class chosen, *VeryMildDemeneted*, was extremely confident (100%), and the atlas-to-image alignment quality was strong based on the high mutual information and a high combined atlas score, with a successful affine registration. Practically, that means the region boundaries in the overlay are reliable enough that the anatomical labels are worth interpreting.

For Fig. 5 plausibility for Alzheimer's is as follows - the emphasis on temporal regions is directionally consistent with Alzheimer related degeneration [19], and attention near the ventricles can be indirectly consistent with atrophy-driven ventricular enlargement [20]. However, the strongest activations in this slice also concentrate around deep central structures (e.g., thalamus/adjacent areas), which are not the most typical primary focus in very early Alzheimer's [21, 22]. So, this example looks partly reasonable but is not perfectly specific. A clinician could use the atlas-labelled overlay to judge whether the highlighted targets are medically meaningful or whether the model may be relying on less diagnostic cues.

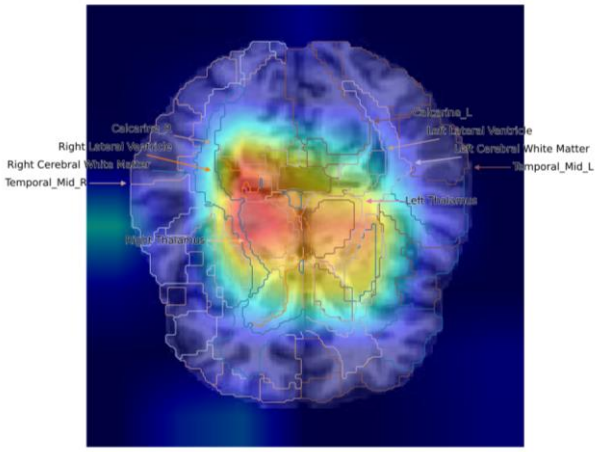

Fig. 5. MRI scan with generated atlas overlay

The atlas overlay makes Grad-CAM easier to judge in practice because it links the heatmap to named brain regions instead of leaving the explanation at the level of coloured pixels. This helps spot cases where the model is focusing on implausible areas (e.g., outside the brain or in unexpected structures) and gives clinicians a clearer basis for deciding whether the highlighted regions make medical sense, which can support faster review and triage when assessing many scans.

Boundary Extraction via Iso-Contours. To make Grad-CAM outputs easier to validate, the final visualization was redesigned so that it explicitly separates where the model focuses from which anatomical regions are implicated. Instead of relying only on a continuous heatmap, the Grad-CAM map is converted into a high-activation boundary meaning the heatmap is lightly smoothed to reduce speckle and then threshold is fixed at an iso-level 0.6. The connected high-activation area is outlined as a contour, creating a clear "focus perimeter" that is typically easier to assess than a fuzzy colour gradient. This kind of abstraction is also consistent with findings that visualization design strongly affects how reliably humans judge whether a saliency map aligns with meaningful structures.

Because Grad-CAM localization depends on the chosen convolutional layer, several candidate ResNet50 layers are evaluated and the one producing the most usable explanation is selected automatically. The selection combines model confidence with simple heatmap characteristics (overall intensity and how concentrated the activation is), aiming to avoid explanations that are either hard to interpret or trivially saturated. This aligns with the broader medical-imaging XAI literature emphasizing that saliency maps should be treated as validation objects (requiring quantitative and qualitative checks), not as self-evident explanations.

After atlas warping, the anatomical overlay is then filtered by overlap with the iso-contour region. Each anatomical label is kept only if a minimum fraction of its area lies inside the high-activation boundary 0.05. Practically, this yields a cleaner and more clinically readable output since the figure reports only the atlas regions that the model's strongest evidence actually intersects, reducing annotation clutter.

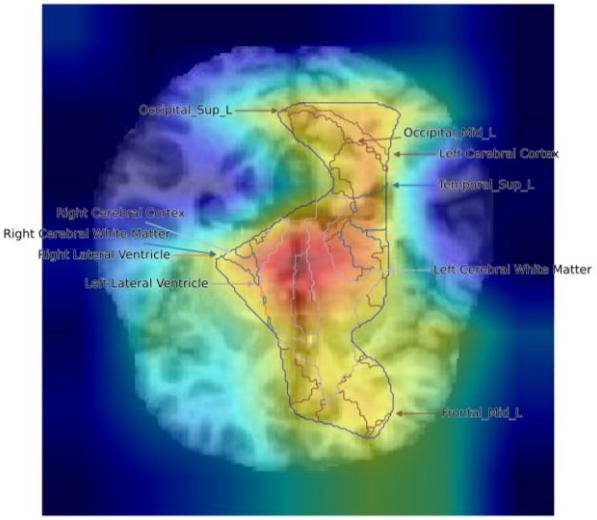

Fig. 6. Filtered Grad-CAM–atlas overlay for a VeryMildDemented MRI slice with iso-contour

By turning the heatmap into a clear boundary and showing only atlas regions that truly overlap with the model's strongest evidence, the visualization becomes faster to read and less cluttered. In practice, this supports quicker clinical review and triage because the clinician can immediately see the core focus area and the few anatomical regions it intersects, instead of interpreting a diffuse heatmap across the whole slice.

Deletion-Based Faithfulness Evaluation. The second validation strategy uses a perturbation-based faithfulness metric, commonly referred to as the deletion test. The underlying assumption is straightforward: if Grad-CAM truly highlights the evidence the model relies on, then removing the most "important" regions (according to Grad-CAM) should quickly reduce the model's confidence for the target class. Concretely, the image is progressively corrupted in the order given by the Grad-CAM ranking (from highest-activation regions to lowest), and the model's predicted probability is recorded after each step. The resulting curve (confidence vs. fraction removed) is summarized by its area under the curve (AUC): a lower deletion AUC indicates that confidence drops earlier, which is interpreted as a more faithful explanation [23, 24].

In this research, deletion is done patch-wise which means removing small blocks rather than individual pixels, to make the process more stable and less sensitive to pixel-level noise. Removed regions are replaced by a blurred baseline (instead of hard-zero masking), which reduces fine detail while preserving coarse structure and avoids introducing overly artificial artefacts that could distort the model's behaviour. The deletion AUC is computed per image and then aggregated across classes to compare how reliably Grad-CAM behaves for different Alzheimer categories [23, 25].

The deletion-metric results are summarized in Fig. 7 as class-wise distributions of deletion AUC (lower values indicate that removing Grad-CAM–selected regions reduces the model's confidence faster, i.e., a more faithful localization).

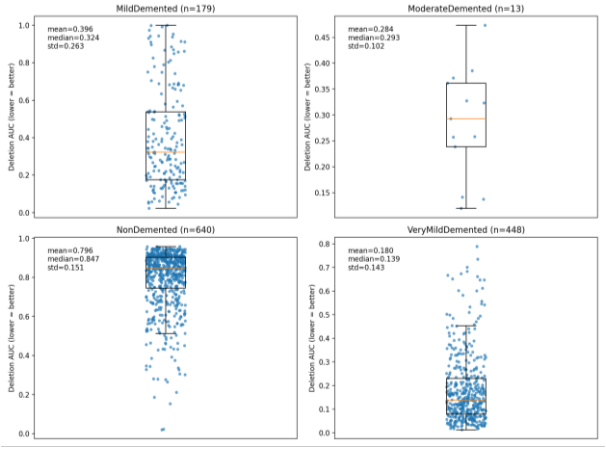

Fig. 7. Grad-CAM Deletion AUC by True Class

Using 30 deletion steps with 8×8 patch removal and a blur baseline, *VeryMildDemented* shows the lowest AUC values overall (mean ≈ 0.180, median ≈ 0.139), suggesting that for this class the model's prediction depends strongly on a relatively compact set of image regions captured by Grad-CAM. *ModerateDemented* class also yields comparatively low AUC, but the interpretation is limited by the small sample size, only 13 scans.

In contrast, *MildDemented* exhibits a broader spread and higher central tendency with many outliers reaching high AUC values, indicating that Grad-CAM faithfulness is less consistent across cases: for some images, deleting the highlighted regions reduces confidence quickly, while for others the confidence remains relatively stable even after substantial deletion. The most striking pattern appears in *NonDemented*, where deletion AUC is substantially higher and clustered near the upper range. This implies that removing the regions emphasized by Grad-CAM often has only a limited effect on the *NonDemented* confidence score, consistent with either more global evidence for this class or a mismatch between Grad-CAM's highlighted regions and the features the classifier actually relies on.

Overall, the deletion analysis suggests that Grad-CAM explanations are most faithful for *VeryMildDemented*, moderately faithful for *MildDemented* and *ModerateDemented* (with higher variability for *MildDemented*), and least faithful for *NonDemented* under the tested perturbation setup.

Unlike purely visual inspection, the deletion test provides a quantitative check of whether the highlighted regions matter for the prediction. This makes it possible to compare Grad-CAM reliability across Alzheimer classes and to flag cases where a heatmap looks plausible but does not meaningfully influence the model's confidence–supporting more robust validation and faster identification of unreliable explanations.

Conclusions

This work shows that visually "reasonable" Grad-CAM heatmaps are not enough for medical use unless their reliability is tested. A CNN was used to classify Alzheimer's stages from brain MRI, and Grad-CAM was applied to explain individual predictions. To validate these explanations, two complementary checks were introduced: an anatomical atlas comparison to see whether highlighted regions correspond to meaningful brain structures, and a deletion metric to quantify whether removing highlighted regions actually reduces the model's confidence.

Together, these validations provide a more defensible way to use Grad-CAM: clinicians can judge whether activations align with plausible anatomy, while the deletion score indicates whether the explanation is faithful to the model's decision. The results also suggest that explanation quality is not uniform across classes, reinforcing that Grad-CAM should be treated as supportive evidence rather than a guaranteed explanation.

All experiments were conducted using standard deep learning libraries (TensorFlow/Keras), and the evaluation procedures rely on commonly used Grad-CAM implementations and perturbation-based metrics. The proposed pipeline is fully reproducible given access to the dataset and model checkpoints.

The principal contributions of this study are:

- The introduction of a structured explainability pipeline for medical image classification that combines Grad-CAM with complementary validation strategies,

addressing the common limitation of relying solely on qualitative inspection of saliency maps.

- An anatomically grounded interpretation of Grad-CAM explanations by aligning activation maps with a unified bank of brain atlases, enabling region-level analysis of model attention in clinically meaningful terms.
- A visualisation refinement based on iso-contour extraction and automatic convolutional layer selection, improving the interpretability and stability of Grad-CAM explanations across heterogeneous MRI samples.
- The development of a class-wise evaluation of Grad-CAM's faithfulness using a deletion-based perturbation metric demonstrates that explanation reliability varies substantially across Alzheimer's disease stages.
- The empirical evidence from experiments on Alzheimer's MRI classification shows that a visually plausible explanation does not necessarily imply faithful model reasoning, reinforcing the need for systematic validation of XAI methods in medical imaging.

This study has several limitations. The experiments were conducted on a single public dataset with pre-processed MRI slices, and no external clinical validation was performed. The atlas-based alignment operates at the slice level and does not capture full 3D anatomical context. In addition, the deletion-based faithfulness results depend on the chosen perturbation strategy and baseline. These limitations do not invalidate the findings, but highlight that the proposed evaluation pipeline should be interpreted as a methodological assessment rather than a clinical validation.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Acknowledgment

This work was supported by the Ministry of Education, Science, Research, and Sport of the Slovak Republic under Grant VEGA 1/0090/25 "Creation of new methods and algorithms of eXplainable Artificial Intelligence based on Importance Analysis".

REFERENCE

1. Uhryn, D. I., Ushenko, Y. O.,; Karachevtsev, A. O. and Halin, Y. O. (2026), "Comparative Evaluation of Deep Neural Networks for Brain Tumor Classification from Magnetic Resonance Imaging", *Herald of Advanced Information Technology*, vol. 9, no. 1, pp. 100–113, doi: https://doi.org/10.15276/hait.09.2026.08
2. Zaitseva, E., Levashenko, V., Rabcan, J. and Kvassay, M. (2023), "A New Fuzzy-Based Classification Method for Use in Smart/Precision Medicine", *Bioengineering*, vol. 10(7), no. 838, doi: https://doi.org/10.3390/bioengineering10070838
3. Chen, C., Isa, N.A.M. and Liu, X. (2025), "A review of convolutional neural network based methods for medical image classification", *Computers in Biology and Medicine*, vol. 185, doi: https://doi.org/10.1016/j.compbiomed.2024.109507
4. Lien, W.-C., Yeh, C.-H., Chang, C.-Y., Chang, C.-H., Wang, W.-M., Chen, C.-H. and Lin, Y.-C. (2023), "Convolutional Neural Networks to Classify Alzheimer's Disease Severity Based on SPECT Images: A Comparative Study", *Journal of Clinical Medicine*, vol. 12, article number 2218, doi: https://doi.org/10.3390/jcm12062218
5. Purwono, P., Wulandari, A. N. E. and Nisa, K. (2025), "Explainable artificial intelligence (XAI) in medical imaging: Techniques, applications, challenges, and future directions", *Advanced Mechanical and Mechatronic Systems*, vol. 1(1), pp. 52–66, doi: https://doi.org/10.53623/amms.v1i1.692
6. Zaitseva, E. and Levashenko, V. (2026), "Reliability engineering in healthcare: Opportunities and challenges", *Reliability Engineering and System Safety*, vol. 267, article no. 111933, doi: https://doi.org/10.1016/j.ress.2025.111933
7. Cheng, Z., Wu, Y., Li, Y., Cai, L. and Ihnaini, B. (2025), "A comprehensive review of explainable artificial intelligence (XAI) in computer vision", *Sensors*, vol. 25, article no. 4166, doi: https://doi.org/10.3390/s25134166
8. Zaitseva, E., Rabcan, J., Levashenko, V. and Kvassay, M. (2023), "Importance analysis of decision-making factors based on fuzzy decision trees", *Applied Soft Computing*, vol. 134, article no. 109988, doi: https://doi.org/10.1016/j.asoc.2023.109988
9. Ortigossa, E. S., Gonçalves, T. and Nonato, L. G. (2024), "Explainable artificial intelligence (XAI) - From theory to methods and applications", *IEEE Access*, vol. 12, pp. 80799–80840, doi: https://doi.org/10.1109/ACCESS.2024.3409843
10. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Chatila, R. and Herrera, F. (2020), "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, vol. 58, pp. 82–115, doi: https://doi.org/10.1016/j.inffus.2019.12.012
11. Fayyaz, A.M., Abdulkadir S,J., Talpur, N., Al-Selwi, S. M., Hassan, S. U. and Sumiea, E. H. (2025), "Grad-CAM (Gradient-weighted Class Activation Mapping): A systematic literature review", *Computers in Biology and Medicine*, vol. 198, Part B, 111200, doi: https://doi.org/10.1016/j.compbiomed.2025.111200
12. Ozer, C., Guler, A., Cansever, A. T. and Oksuz, I. (2026), "Consistent explainable image quality assessment for medical imaging". *Health Information Science and Systems*, vol. 14, 31, 2026, doi: https://doi.org/10.1007/s13755-025-00411-0
13. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Matthew, J. A. and Kalpathy-Cramer, J. (2021), "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging." *Radiology: Artificial Intelligence*. vol. 3, no. 6, article no. 200267, 2021, doi: https://doi.org/10.1148/ryai.2021200267
14. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P. and Rajpurkar, P. (2022), "Benchmarking saliency methods for chest X-ray interpretation", *Nature Machine Intelligence*, vol. 4, pp. 867–878, doi: https://doi.org/10.1038/s42256-022-00536-x
15. Uraninjo. (2022). *Augmented Alzheimer MRI Dataset for Better Results on Models*, Kaggle, available at: https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset
16. Lawrence, R. M., O'Toole, C. M., Duffy, B., G. C. Arvapalli, S. C. Ramachandran, D. A. Pisner, P. F. Frank, A. D. Lemmer, A. Nikolaidis and J. T. Vogelstein (2021), "Standardizing human brain parcellations", *Scientific Data*, vol. 8, no. 98, doi: https://doi.org/10.1038/s41597-021-00849-3

17. Nowinski, W. L. (2020), “Evolution of Human Brain Atlases in Terms of Content, Applications, and Visualization. Brain Imaging and Behavior”, *Neuroinformatics*, vol. 19 (1), pp.1–22, doi: https://doi.org/10.1007/s12021-020-09481-9
18. Sengupta, D., Gupta, P. and Biswas, A. (2022), “A survey on mutual information based medical image registration”, *Neurocomputing*, vol. 486, pp. 174-188, doi: https://doi.org/10.1016/j.neucom.2021.11.023
19. Woodworth, D. C., et al. (2022), “Dementia is strongly associated with medial temporal atrophy even after accounting for neuropathologies.” *Brain Communications*, vol. 4(2), doi: https://doi.org/10.1093/braincomms/fcac052
20. Rosa-Neto, P. (2021), “Chapter 9. Brain imaging using CT and MRI”, *Alzheimer’s Disease International*, available at: https://www.alzint.org/u/World-Alzheimer-Report-2021-Chapter-09.pdf
21. Forno, G., Saranathan, M., Contador, J., Guillen, N., Falgàs, N., Tort-Merino, A., Balasa, M., Sanchez-Valle, R., Hornberger, M. and Lladó, A. (2023), “Thalamic nuclei changes in early and late onset Alzheimer’s disease”, *Current Research in Neurobiology*, vol. 4, article number 100084, doi: https://doi.org/10.1016/j.crneur.2023.100084
22. Biesbroek, J. M., Verhagen, M. G., van der Stigchel, S. and Biessels, G. J. (2024), “When the central integrator disintegrates: A review of the role of the thalamus in cognition and dementia”, *Alzheimer’s & Dementia*, vol. 20, pp.2209–2222, doi: https://doi.org/10.1002/alz.13563
23. Gomez, T., Fréour, T. and Mouchère, H. (2022), “Metrics for saliency map evaluation of deep learning explanation methods”, arXiv preprint arXiv:2201.13291, doi: https://arxiv.org/abs/2201.13291
24. Hedström, A., Weber, L., Bareeva, D., Krakowczyk D., Motzkus F., Samek W., Lapuschkin S. and Höhne M. M.-C. (2023), “Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations”, *Journal of Machine Learning Research*, vol. 24, pp. 1–11, doi: https://doi.org/10.48550/arXiv.2202.06861
25. Nieradzik, L., Zięba, M., & Wróbel, K. (2024), “Reliable evaluation of attribution maps in CNNs: A Perturbation-Based Approach”, *International Journal of Computer Vision*, vol. 133, pp. 2392–2409, doi: https://doi.org/10.48550/arXiv.2411.14946

Received (Надійшла) 21.12.2025
Accepted for publication (Прийнята до друку) 18.03.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Чижмарова Олександра – аспірантка кафедри інформатики, факультет управлінських наук та інформатики, Жилінський університет, Жиліна, Словаччина;
Alexandra Čižmárová – Postgraduate Student at the Department of Informatics, Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia;
e-mail: alexandra.cizmarova@st.fri.uniza.sk; ORCID Author ID: https://orcid.org/0009-0007-1021-4722.

Досталова Крістіна – аспірантка кафедри інформатики, факультет управлінських наук та інформатики, Жилінський університет, Жиліна, Словаччина;
Kristína Dostálová – Postgraduate Student at the Department of Informatics, Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia;
e-mail: kristína.dostalova@st.fri.uniza.sk; ORCID Author ID: https://orcid.org/0009-0002-8567-716X.

Хркут Патрік – доцент, завідувач кафедри програмних технологій, факультет управлінських наук та інформатики, Жилінський університет, Жиліна, Словаччина;
Patrik Hrkut– Associate Professor, Head of the Department of Software Technologies, Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia;
e-mail: patrik.hrkut@fri.uniza.sk; ORCID Author ID: https://orcid.org/0000-0002-8747-9194;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55927516000.

Порошенко Антон Ігорович – доктор філософії, старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Anton Poroshenko – PhD, Senior Lecturer of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: anton.poroshenko@nure.ua; ORCID Author ID: https://orcid.org/0000-0001-7266-4269;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57250025600.

Пояснюваний метод штучного інтелекту GRAD-CAM в обробці медичних зображень

О. Чижмарова, К. Досталова, П. Хркут, А. І. Порошенко

Анотація. Надійність сучасних моделей глибокого навчання в медичній галузі часто ставиться під сумнів через їхню «чорну скриньку». Методи постфактумної пояснювальності з галузі пояснювального штучного інтелекту (XAI) пропонують засоби для підвищення прозорості та оцінки надійності прогнозів, отриманих згортковими нейронними мережами. **Метою дослідження** є вивчення того, як методи XAI, зокрема градієнтно-зважене картування активації класів (Grad-CAM), можуть забезпечити надійні пояснення для класифікації медичних зображень. Для цього МРТ-зображення мозку були використані для навчання згорткової нейронної мережі для категоризації чотирьох стадій деменції при хворобі Альцгеймера. Щоб зробити кожне прогнозування прозорим, області мозку, які навчена мережа використовувала для категоризації, були виділені за допомогою Grad-CAM. Отримані карти релевантності, теплові карти, були оцінені за допомогою двох підходів: просторове порівняння з анатомічно визначеними областями мозку, пов'язаними з хворобою Альцгеймера, за допомогою накладання атласу та кількісна оцінка достовірності за допомогою метрики на основі видалення, де високо впливові області, визначені Grad-CAM, були поступово видалені, а вплив на достовірність класифікації вимірювався. .

Ключові слова: пояснювальний ШІ; згорткова нейронна мережа; Grad-CAM; хвороба Альцгеймера; теплова карта; градієнти.