UNIVERSITY OF ŽILINA
Faculty of Management Science and Informatics

Faculty of management science and informatics

Analysis of Medical Data by XAI with a Grad-CAM method

Diploma Thesis

BC. ALEXANDRA ČIŽMÁROVÁ

Study Program: Biomedical Informatics
Study Division: Informatics
Study Institution: University of Žilina, Žilina
Thesis Supervisor: prof. Ing. Elena Zaitseva, PhD.
Žilina 2026

ŽILINSKÁ UNIVERZITA V ŽILINE, FAKULTA RIADENIA A INFORMATIKY.

ZADANIE TÉMY DIPLOMOVEJ PRÁCE.

Študijný program: Biomedicínska informatika

Meno a priezvisko

Alexandra Čižmárová

Osobné číslo

563191

Názov práce v slovenskom aj anglickom jazyku

Analýza lekárskych údajov pomocou XAI s metódou Grad-CAM

Analysis of Medical Data by XAI with a Grad-CAM method

Zadanie úlohy, ciele, pokyny pre vypracovanie (Ak je málo miesta, použite opačnú stranu)

Cieľ diplomovej práce:
Cieľom práce je implementácia a vyhodnotenie algoritmov na analýzu medicínskych údajov (medicínskych obrazov).

Obsah:

- Analýza problému a štúdium dostupnej literatúry.
- Analýza existujúcich riešení, knižníc pre zvolený typ obrázkov.
- Návrh algoritmu pre získane údaje.
- Implementácia algoritmu vo vhodnom vývojovom prostredí.
- Overenie a validácia vybratého algoritmu.

Meno a pracovisko vedúceho DP: prof. Ing. Elena Zaitseva, PhD., KI, ŽU
Meno a pracovisko tútora DP:

garant štud. programu
(dátum a podpis)

Zadanie zaregistrované dňa 31. 10. 2025 pod číslom 3390/2025 podpis _____

Declaration of Honor

I declare that I have prepared the assigned diploma thesis independently, under the professional guidance of the thesis supervisor/advisor, and that I have used only the literature cited in the thesis.

During the preparation of this thesis, I used generative AI tools (ChatGPT by OpenAI and Claude by Anthropic) for language proofreading, stylistic correction and minor formatting of the text. All outputs were reviewed and edited by me, and I take full responsibility for the content of this thesis.

Žilina, August 5, 2026

signature

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, prof. Ing. Elena Zaitseva, PhD., for her support, guidance, and valuable advice throughout the development of this thesis. Her direction and constructive feedback helped shape the work and contributed significantly to its completion.

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. However, in many medical imaging datasets only image-level diagnostic labels are available, and pixel-level annotations of pathological regions are not provided, limiting direct quantitative evaluation of explanation maps. 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 artificial intelligence; convolutional neural network; Grad-CAM; Alzheimer's disease; heatmap; gradients.

Content

List of Figures

List of Tables

List of Abbreviations

Abbreviation	English meaning
AD	Alzheimer's disease
AI	Artificial intelligence
XAI	Explainable artificial intelligence
DL	Deep learning
DNN	Deep neural network
CNN	Convolutional neural network
CAM	Class activation mapping
Grad-CAM	Gradient-weighted class activation mapping
MRI	Magnetic resonance imaging
T1	T1-weighted (MRI sequence)
CT	Computed tomography
PET	Positron emission tomography
CSF	Cerebrospinal fluid
NIA-AA	National Institute on Aging and Alzheimer's Association
WHO	World Health Organization
MTA	Medial temporal atrophy
AUC	Area under the curve
MI	Mutual information
MNI	Montreal Neurological Institute space (standard brain space)
MNI152	MNI152 brain template
AAL	Automated Anatomical Labeling atlas

Abbreviation	English meaning
HO	Harvard–Oxford atlas
ReLU	Rectified linear unit
RGB	Red, green, blue (image channels)
IoU	Intersection over union (Jaccard overlap)
GPU	Graphics processing unit
CPU	Central processing unit
TF	TensorFlow
OpenCV	Open-source computer vision library
CXR	Chest X-ray

Glossary

Affine transformation: A geometric transformation that preserves lines and parallelism, allowing translation, rotation, scaling, and shearing. Used here to warp atlas labels onto an MRI slice.

Anatomical atlas: A labelled reference map of brain regions used to assign anatomical names to locations in an image.

Area under the curve (AUC): A single numeric summary of a curve. In this thesis it summarizes the deletion curve of confidence versus fraction removed.

Attention map: A heatmap-style visualization indicating which parts of an input image are most influential for a model decision, typically produced by an XAI method.

Benchmarking: Systematic comparison of methods or models under shared conditions to assess performance, reliability, or limitations.

Class activation mapping (CAM): A family of methods that produce class-specific localization maps from convolutional feature maps to highlight regions relevant to a class prediction.

Class imbalance: A dataset property where some classes have many more samples than others, which can bias training and evaluation.

Classification: A machine learning task where the model assigns an input to one of several discrete categories.

Confidence score: The model's predicted probability for a class, often taken as the maximum softmax probability for the predicted class.

Convolutional neural network (CNN): A neural network architecture designed for image data, using convolutional layers to extract spatial features.

Canny edge detection: An image processing method that detects edges. Used here to compute boundary overlap between atlas and MRI images.

Deletion test: A faithfulness evaluation method where regions ranked as important are progressively removed and the change in prediction confidence is measured.

Deterministic split: A dataset partitioning approach that produces the same train, validation, and test sets across runs by using fixed ordering and seeds.

Diffuse pattern: A spatial pattern that is spread out rather than localized, common in neurodegenerative imaging where abnormalities may not form clear boundaries.

Explainable artificial intelligence (XAI): Methods and tools that aim to

make model decisions more transparent and interpretable to humans.

Faithfulness: The extent to which an explanation reflects features that truly influence the model's prediction, often tested through perturbations.

Feature map: The output of a convolutional layer that encodes learned spatial features of the input.

Global average pooling: A layer that compresses each feature map into a single value by averaging across spatial dimensions, producing a feature vector for classification.

Grad-CAM: A CAM method that uses gradients and convolutional feature maps to produce a class-specific heatmap showing regions that contribute positively to a prediction.

Ground truth: The reference "correct" labels used for evaluation. In this thesis, pixel-level ground truth for pathology localization is not available.

Heatmap: A coloured visualization that represents values across an image. Used here to display Grad-CAM relevance intensity.

Histogram-based mutual information: A similarity measure computed from joint intensity histograms. Used here to score atlas-to-image alignment.

Image-level label: A label that applies to the entire image or study, such as diagnosis or disease stage, without specifying spatial regions.

Image registration: The process of aligning one image to another so that corresponding anatomical structures match.

Intersection over union (IoU): A similarity measure between two sets, defined as intersection divided by union. Used here as Jaccard overlap for edge maps.

Iso-contour: A boundary line connecting points of equal value. Used here to outline high-activation regions in a Grad-CAM heatmap.

Jaccard overlap: A set similarity metric equivalent to IoU. Used here to quantify overlap between edge maps.

Keras: A high-level deep learning API used here with TensorFlow for model training and inference.

Label map: An image where each pixel stores an integer region label, used in atlases to represent anatomical structures.

Mutual information (MI): A measure of statistical dependence between two signals. Used here to quantify alignment similarity between images with different intensity distributions.

Patch-wise deletion: Deletion performed by removing small blocks of pixels rather than individual pixels,

improving stability of perturbation-based evaluation.

Pixel-level annotation: A spatial label defining which pixels belong to a region of interest, such as a segmentation mask of pathology.

Plausibility: Whether an explanation appears anatomically or clinically reasonable, typically judged by domain knowledge rather than causal testing.

Post-hoc explanation: An explanation generated after model training, without changing the model architecture or parameters.

ReLU: A non-linear function that sets negative values to zero. Used in Grad-CAM to keep only positive contributions.

ResNet50: A convolutional neural network architecture with residual connections, commonly used as a transfer learning backbone for image classification.

Saliency method: An XAI approach that highlights input regions important for a prediction, often producing a heatmap over the image.

Softmax: A function that converts model outputs into a probability distribution over classes.

Stratified split: A dataset split that preserves the class distribution across train, validation, and test sets.

T1-weighted MRI: A type of MRI sequence that produces contrast useful for anatomical structure visualization.

Transfer learning: Training strategy where a model pretrained on one dataset is adapted to a new task, often by fine-tuning.

Weak supervision: Training with limited label detail, typically using image-level labels without pixel-level annotations.

INTRODUCTION

The reliability of deep learning models in medical applications is often questioned due to their black-box nature, which limits transparency and trust in clinical decision-making. Although convolutional neural networks have demonstrated high accuracy in medical image analysis, the lack of interpretability remains a significant obstacle to their widespread adoption in healthcare. As a result, there is growing interest in explainable artificial intelligence (XAI) methods that aim to make model predictions more transparent and reliable. [1]

This research focuses on the use of deep learning for classifying stages of Alzheimer's disease from brain MRI scans and on interpreting the model's decisions using Gradient-weighted Class Activation Mapping (Grad-CAM). By highlighting brain regions that contribute most to individual predictions, Grad-CAM provides intuitive visual explanations. The reliability of these explanations is evaluated through anatomical atlas comparison and a deletion-based metric that quantitatively assesses the faithfulness of the highlighted regions. [2], [3]

A major practical limitation in medical imaging is that many datasets provide only image-level labels, while pixel-level ground truth masks of pathological regions are missing. That means explanations cannot be directly compared against "correct" annotated regions in a fully quantitative way. This is not unique to brain MRI. [4] That's what makes Grad-CAM attractive in real-world medical datasets, but it also increases the risk that explanations are accepted without strong verification.

When Grad-CAM is used to see how neural networks make decisions, people just look at the heatmaps and assume they're right if they look reasonable. That doesn't mean the heatmap is correct. Heatmaps can look convincing even when the model is focusing on shortcuts, noise, or irrelevant regions. This is especially risky in medical imaging, where "looking plausible" can easily be confused with being correct and clinically meaningful.

It is important to know for sure that the highlighted areas truly reflect the model's thinking. This work tackles that problem by stress-testing Grad-CAM to see if it holds up, particularly for medical scans. The atlas-based comparison checks whether the highlighted regions align with anatomically meaningful areas linked to Alzheimer's disease, while the deletion-based evaluation tests whether the highlighted regions are actually important for the model's confidence. Together, these two views aim to provide a more reliable and defensible way to judge Grad-CAM explanations than visual inspection alone.

The goal of this work is to evaluate the efficiency of Grad-CAM explanations in the context of MRI classification for Alzheimer's disease. This goal is achieved by the implementation of the next steps (objectives):

1. Train a CNN to classify dementia stages.
2. Apply Grad-CAM for visualisation.

3. Implement two approaches for validating explanations (based on an atlas and a deletion test).
4. Compare the results of these approaches.

1 SCIENTIFIC AND TECHNICAL BACKGROUND

This chapter lays the foundation for the thesis. It starts with medical images as data and with the main ways they are used in machine learning. It then explains how medical datasets are labelled in practice, and why pixel-level ground truth is often missing – especially for conditions where imaging evidence is diffused rather than sharply localized.

The chapter then frames the problem through reliability in healthcare, where hidden failure modes matter. Finally, it introduces explainable artificial intelligence, specifically focusing on saliency-based methods such as Grad-CAM, which are popular because they work under image-level supervision, even if they still require careful validation. [1]

1.1 Medical images in the Healthcare Domain

Medical imaging covers a wide range of modalities, such as X-ray, CT, MRI, ultrasound, and functional imaging methods. These images are used for many activities such as diagnosing, monitoring, screening, treatment planning and many more. From a machine learning perspective, they are also diverse in terms of resolution, noise, acquisition protocols, and what "signal" looks like. Some findings are localized and visually distinct, while others are subtle or spread across multiple regions. This matters because it influences what models learn and what, in the end, counts as a meaningful explanation. [5]

A common way to use medical images in deep learning is classification. Classification serves for predicting a diagnosis or disease stage from an image or 2D slice or 3D volume. In practice, this is attractive because it matches what is available in many real datasets – one label per image (or per study), often produced through clinical workflow rather than created with machine learning research in mind. Large-scale pipelines based on image-level labels are common mostly in radiology datasets and have enabled strong predictive models even without detailed region annotations. [6]

For this thesis, the chosen medical image type is brain MRI, which scans showcase level of dementia of the patients. This thesis will mainly focus on brain MRI data, which is a type of data that is often used in the clinic for diagnosing dementia. In contrast to other types of data, MRI gives us much more structural information than other types of medical images, while it also has more practical challenges. It is a 3D volume, which comes with multiple MRI sequences, many variations among different MRI scanners and processing methods. And with dementia related imaging, the evidence we seek to detect are often about patterns of changes, such as atrophy or other types of structural changes that occur to the brain, which is often not about finding specific lesions that are clearly defined and easy to identify. This problem is therefore particularly illustrative of the reasons why explanation validation is challenging, especially when the pixel-level ground truth is not clear. [7]

1.1.1 Annotation Types and Why Are They Often Missing

Medical image datasets can be labelled at different levels, depending on the task and the availability of time from clinical experts. Common levels of annotations include:

- Image-level labels (e.g., “pneumonia present” or “dementia stage”),
- Region-level labels (bounding boxes or rough regions of interest),
- Pixel/voxel-level labels (segmentation masks),
- Landmarks and measurements (key points, volumes, thickness),
- Textual reports linked to images.

The pixel-level annotation of images is extremely costly and time-consuming and usually needs to be done by highly skilled clinicians. In many settings, it is not part of the standard clinical workflow, so it is simply not available at scale. As a result, many practical medical imaging models are trained in a weakly supervised way: the model learns from image-level labels and no explicit ground truth exists for “where” the disease evidence should be in the image. [5]

This is especially relevant for diseases where the pathological signal is not naturally represented as a clear, sharp region that can be outlined with high agreement. Alzheimer’s disease is one example where imaging evidence can be diffuse and related to broader structural patterns rather than one localized abnormality. Because of that, many datasets used in research provide disease stage labels but no pixel-level annotation of “the region the model should focus on.” This creates the core motivation of the thesis: heatmaps are used as explanations, but there is no direct ground truth to validate them against, so validation must be handled in different ways. [6]

1.1.2 Alzheimer’s Disease and Brain MRI Imaging

Alzheimer’s disease is a progressive neurodegenerative condition and a world’s leading cause of dementia. Since dementia affects a large part of population worldwide, Alzheimer’s represents a major healthcare burden and serves as a typical example for situations where a late diagnosis or a misleading one can have serious consequences for the patients and their families. [7]

From a clinical and research perspective, Alzheimer’s disease can be described by symptoms and by underlying biological processes. Modern frameworks state that the disease is linked to characteristic pathological changes and biomarkers, especially those related to amyloid and tau, and that these processes develop over time before the most severe clinical decline appears. [6]

Brain MRI is often used in dementia assessment because it provides structural information and helps characterize patterns of neurodegeneration. Another advantage is that MRI also supports differential diagnosis. In Alzheimer’s, the imaging evidence is often not a single “lesion” that could be outlined with a precise boundary. Instead, it shows gradual changes in the brain including shrinkage of certain areas and these changes can be different from one person to another. This makes MRI a realistic example of a setting where models may learn meaningful patterns, but where pixel-level ground truth annotations for “what the model should look at” are not available. [8]

A commonly discussed structural marker in dementia research is medial temporal lobe atrophy, which is strongly associated with dementia even when accounting for underlying neuropathologies. This is relevant for MRI-based learning because it suggests that the discriminative signal may be tied to subtle structural differences rather than a clear pathological region. [9] Beyond the medial temporal lobe, Alzheimer’s is also associated with changes in broader brain systems. For example, thalamic nuclei

changes have been studied in the context of early and late onset Alzheimer's, and the thalamus is discussed as a key structure for cognitive integration in dementia-related decline. This again reinforces the idea that Alzheimer's-related imaging patterns can be multi-region and network-like rather than localized. [10], [11]

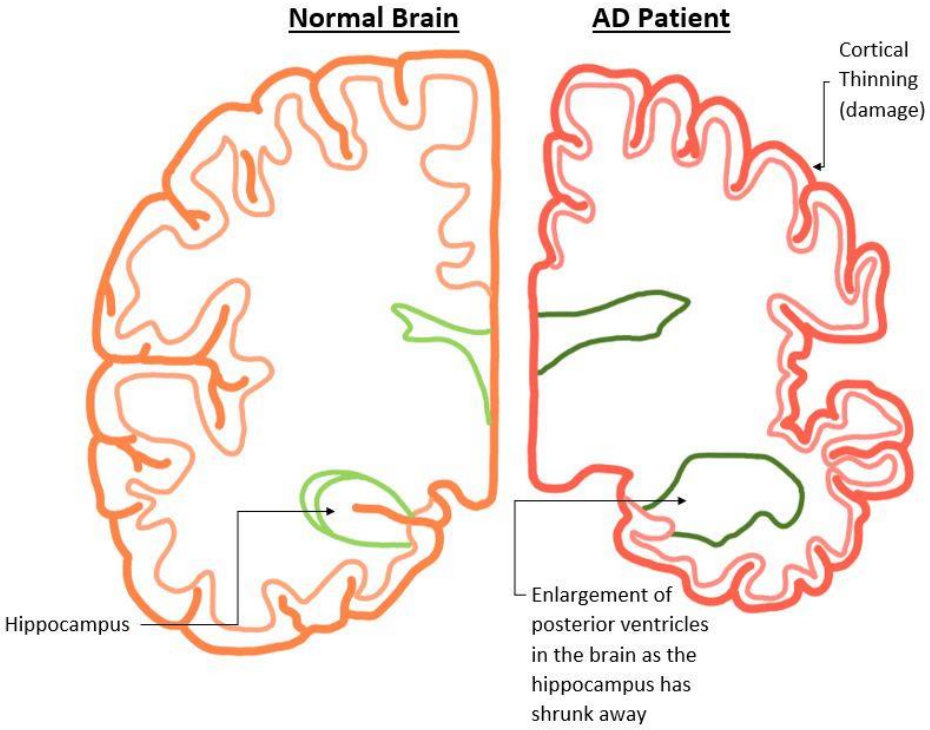

Figure 1: Healthy brain vs. Alzheimer's brain [12]

For this thesis, Alzheimer's disease serves as a practical case study of a common limitation in medical imaging data. In many Alzheimer MRI datasets, the available supervision is limited to an image-level label, such as diagnosis or disease stage, while pixel-level annotations of pathological regions are not provided. This is not only a matter of missing manual effort. In neurodegenerative diseases, the imaging signal is often expressed through gradual and distributed structural change rather than a single focal abnormality with a clear boundary that could be consistently outlined. As a result, there is typically no well-defined pixel-level "ground truth" region that could be used to directly verify where disease evidence is in the image. MRI of brain with Alzheimer's disease is therefore a representative example of medical imaging scenarios where labels exist at the study level, but spatial annotations are absent and difficult to define in a clinically meaningful way. [6], [8]

1.2 Reliability Considerations in Healthcare and in Medical Imaging

In healthcare, reliability is not about average performance. It concerns whether a system behaves consistently, whether its failure modes are understood, and whether the risks are acceptable for the intended use. Reliability engineering in healthcare emphasizes that healthcare systems are complex socio-technical environments, where technology interacts with processes, staff, and patient variability, and where safety and accountability requirements are higher than in many other domains. [13]

For medical imaging algorithms, this reliability framing has several concrete implications. Firstly, evaluation must go beyond a single number such as accuracy. The robustness of data shifts must be considered, as well as realistic operational conditions. Secondly, the absence of detailed ground truth in many datasets (for example, missing pixel-level annotations in neurodegenerative MRI) limits what can be verified directly, which increases the need for careful validation design and discussion of uncertainty. Thirdly, because healthcare decisions are high impact, the standard for "good enough" evidence is stricter. A system should be assessed not only on what it predicts, but also on whether its behaviour can be justified and trusted within clinical workflows. [13]

This thesis is modelled with this logic in mind. It draws its focus on why validation matters and simply trusting what can be seen without evidence because the imaging appears correct and the accuracy is high cannot and is not an acceptable mindset in healthcare field.

1.2.1 Reliability Risks Specific to Medical Imaging Models

Medical imaging models inherit the general reliability challenges of healthcare, but imaging adds its own failure variables. Images are produced by different scanners, protocols, and institutions, with subtle differences in acquisition settings and preprocessing. This causes for models to face distribution shifts that may not be obvious to humans but can change model's behaviour. This is a big risk for deployment, because the training set rarely captures the full operational variability. [14]

Another reliability issue is supervision quality. Many imaging datasets are labelled at the study or image level because this matches clinical workflow since it is faster than detailed annotation. These labels can be noisy or indirect proxies of the underlying pathology, and they provide limited information about where the relevant signal should be. In conditions where the imaging evidence is diffuse and not naturally defined as a bounded region, pixel-level ground truth is typically unavailable, which restricts what can be verified directly. [8], [9]

Medical imaging models are exposed to shortcut learning, which are correlations unrelated to disease biology (scanner signatures, artifacts, text markers, or site-specific patterns). Those can become so called predictive signals. These shortcuts can preserve high test-set accuracy while undermining reliability in real conditions when the evaluation protocol is not designed to realistic shifts. [14]

1.2.2 Why Reliability Motivates Interpretability and Validation

Since medical imaging has critical reliability, it is not enough to know that a model performs well on average. It matters whether its decisions are stable and defensible when conditions change. This creates pressure for transparency as clinicians and developers need ways to detect when a model is relying on spurious cues and to understand failure cases rather than treating the model as a black box. [13]

However, interpretability is only useful if it can be evaluated with some discipline. Interpretable machine learning is explicitly framed as a field where methods and claims should be tested rigorously, not accepted because outputs look plausible. In medical imaging, this point is especially sharp because explanation artefacts can appear visually convincing even when they are not trustworthy. [15], [16]

Empirical work in medical imaging has shown that saliency-style explanations can be unreliable and that their behaviour depends on method choice, setup, and evaluation design. It is important to note that explanations must be treated as objects that require validation, not as self-evident proof. The next section therefore introduces explainable artificial intelligence and the main family of explanation approaches used in this thesis. [17], [18]

1.3 Explainable Artificial Intelligence

Deep learning models can achieve impressive results on medical imaging analysis, but explainability is a major reliability problem. While it is difficult to investigate and justify the workings of the network, that makes it difficult to identify workarounds and fail modes, which is precisely the goal of Explainable Artificial Intelligence (XAI) which introduces new aspects in the model behaviour aiming at explaining what the model does and how. This means explaining every prediction given by the model as well as identifying general patterns emerging from the model, as showed in reference. [19]

As introduced in the definition section, XAI methods can be categorized into several groups. Some models are intrinsic explanatory, which means they are designed to be interpretable. Some others are post-hoc, which are methods to explain a pre-trained model. The XAI results can be global or local, depending on the level of explanation required. The global explanation reveals the characteristics of the whole model that the doctor or patient can rely on. The local explanation focuses more on a specific image and the related output, by explaining the reasoning behind the output prediction for that image. This type of explanation is particularly prevalent in the medical imaging domain, as strong pre-trained CNN models can be explained using post-hoc local XAI methods without modifying the underlying model architecture. [16], [19]

Interpretability does not necessarily imply correctness. A good explanation can still be wrong. Hence, to ensure that a model is properly understood, one must be on the lookout for fallacies inherent in the explanation process. In other words, explanations are outputs that must be validated. They must not be accepted because they look reasonable. [15], [16]

1.3.1 Saliency Methods Overview

Saliency methods are one of the most common XAI approaches used with medical images. They produce a heatmap over the input image which highlights pixels or regions that are considered important for the model's prediction. Since it overlayers the images, this format is popular because it matches how clinicians already read them. Because of that the visualization and comparison across cases is easier. [16], [20]

There are different families of saliency methods. Some are gradient-based and use derivatives of the output with respect to the input or intermediate feature maps. Others are perturbation-based and measure how the prediction changes when parts of the input are removed or altered. A widely used subfamily in computer vision and medical imaging is the CAM family, which produces class-specific activation maps. These methods are practical because they can be generated from a trained CNN using only the input image and the model. For this family a no pixel-level annotations are required. [21], [22]

At the same time, saliency maps come with reliability risks. They can be sensitive to method choices and implementation details. They can also highlight regions that are not clinically meaningful. Heatmaps can look reasonable even when the model is using shortcuts and the mistake can therefore be missed. Several studies have shown that saliency methods need careful evaluation and that trust should not be based on appearance alone. [17], [18], [23], [24]

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

Gradient-weighted Class Activation Mapping (Grad-CAM) is a saliency method designed to produce a class-specific heatmap that highlights image regions which are most relevant to a model's prediction. It is widely used with convolutional neural networks because it can be applied post-hoc to an already trained classifier. It then generates explanations in a visual form that is easy to interpret on medical images. [21]

In practice, Grad-CAM uses the feature representations from a late convolutional layer, where spatial information is still preserved but the features are already high-level. The method links these feature maps to the model's predicted class score by computing an importance weight for each feature map. These weights are then used to form a weighted combination of the feature maps, followed by a non-linearity to keep only positive contributions. The result is a coarse localization map, which is upsampled to the input resolution and typically overlaid on the original image as a heatmap. [21], [22]

The overall workflow of how the method works is summarized in the diagram included in this section. An MRI image is passed through the CNN to obtain feature maps and class scores. Grad-CAM then assigns weights to feature maps for the target class, and the weighted map is converted into a heatmap aligned with the input image. This makes Grad-CAM practical in weakly supervised imaging settings, because it does not require any pixel-level annotations to be generated. [21]

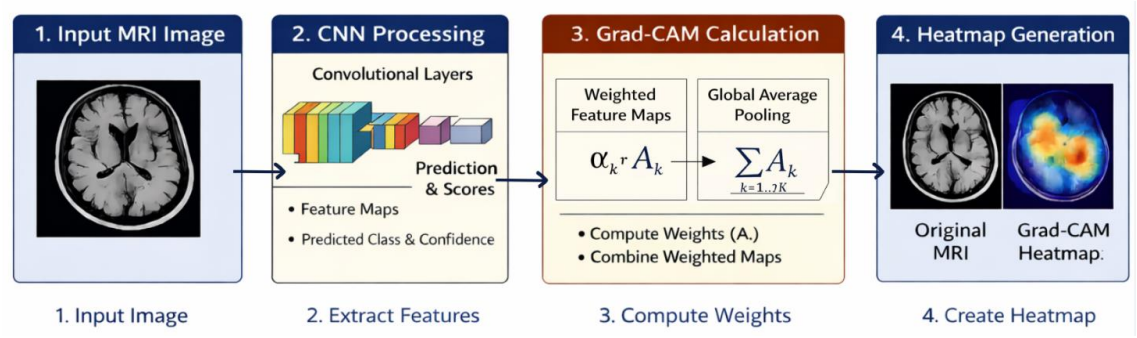

Figure 2: Grad-CAM heatmap generation pipeline

1.3.2.1 How Grad-CAM Computes Gradients

Grad-CAM creates a class-discriminative heatmap by calculating the gradient of the target-class score with respect to the most recent convolutional feature maps, pooling these gradients to obtain channel-wise importance weights, and creating a weighted sum of the feature maps. By calculating 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:

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

The spatial dimensions of the chosen convolutional feature maps are indicated by *H* and *W*, which stand for the feature map's height (H, number of rows) and width (W, number of columns). Each feature map's total contribution to the prediction is measured by this formula. It is the default visual explainer in many radiology prototypes because it only needs the model's gradients, works with any CNN architecture, and adds no overhead. [19], [20], [21], [25]

While Grad-CAM can be in comparison to other XAI methods considered simple it is widely adopted. It is important to note its outputs should not be treated as self-validating evidence. The resulting heatmap is an interpretation of what the model is using, and its reliability depends on the model, the layer choice, and the evaluation setup. [24]

This chapter's goal is to provide the scientific and technical context for the thesis by outlining the nature of medical imaging data, the practical absence of pixel-level annotations in many clinical datasets, and why pixel-level ground truth is often unavailable, especially in neurodegenerative conditions such as Alzheimer's disease. It also introduced reliability as a key requirement in healthcare systems and motivated the need for explainability methods, with a focus on saliency-based explanations specifically Grad-CAM. [13], [19], [21]

2 RELATED WORK

This chapter reviews prior work that motivates the thesis focus on validation of explanation heatmaps in medical imaging. It summarizes how weakly supervised image-level labels became the standard in many medical datasets, why saliency-style heatmaps, specifically Grad-CAM, are widely used as post-hoc explanations, and what recent studies show about their limitations and trustworthiness. The chapter then outlines how explanation maps are evaluated in the literature and explains why neurodegenerative brain MRI, where pixel-level ground truth is typically unavailable, is a representative setting for studying explanation validation.

2.1 Weak Supervision and Standardized Heatmaps

A large part of medical imaging research in deep learning is built on datasets where supervision is provided at the image or study level. This includes a dataset design choice and a reflection of clinical workflow with regularly recorded labels such as diagnosis, finding presence, or severity. On the other hand, pixel-level delineations are costly and rarely available at scale. [26]

An example of this setup is CheXNet, where a deep learning model is trained to detect pneumonia on chest X-rays using image-level labels rather than pixel-level pathology annotations. The work demonstrates that strong diagnostic classification performance can be achieved without segmentation masks. It also reflects how medical data is commonly available at scale. [27]

This weakly supervised setup creates a problem that shows up immediately once the model is meant to be used beyond a benchmark. If a model outputs a label, it is natural to ask what in the image drove the decision. Since there is usually no pixel-level ground truth to compare it against, researchers frequently use post-hoc explanation methods that can generate a spatial visualization directly from the model. In the standard XAI taxonomy, this corresponds to post-hoc local explanations. The goal is not to redesign the model, but to generate an explanation for a specific input and prediction. [16], [19]

Heatmap-style explanations fit this need well. They can be generated for a single image, they produce an intuitive overlay, and they look like a reasonable form of evidence in an imaging context. That is why saliency methods became a default choice in medical imaging papers. The difficulty is that these heatmaps are often treated as self-explanatory – if the highlighted regions look plausible, the explanation is accepted.

However, there is strong evidence that plausibility is not enough. Arun et al. explicitly study the trustworthiness of saliency maps for localizing abnormalities in medical imaging and show that saliency maps may not reliably point to the true abnormal regions, even when the underlying classifier performs well. This directly challenges the common assumption that a visually convincing heatmap implies that the model is using clinically meaningful evidence. [17]

Saporta et al. reinforce this point through benchmarking saliency methods for chest X-ray interpretation. Their results show that different saliency approaches can behave very differently, and that explanation quality is not guaranteed simply because a method

is popular. In weakly supervised settings it matters, because when there is no pixel-level ground truth, it becomes easy to “validate” explanations by visual inspection alone. [18]

2.2 Grad-CAM and CAM-Family Methods in Healthcare Practice

Among heatmap-based explanation techniques, CAM-family methods are especially common in medical imaging since they match how convolutional networks process images. Instead of attributing importance directly to pixels, these methods generate a spatial relevance map from convolutional feature maps, which makes the output easy to visualize as an overlay on the original image. In practice, this is one of the main reasons CAM-style methods became a default choice in healthcare papers. [22]

Grad-CAM is the best-known method in this family and is often treated as a baseline explanation technique across tasks. In the original Grad-CAM work, the method is presented to connect a model’s class prediction to the spatial structure preserved in the final convolutional layers, producing a class-specific heatmap that highlights regions contributing positively to the predicted class. The output is coarse but interpretable, and it can be computed for a trained model without changing the architecture or training procedure. [21]

This practicality explains why Grad-CAM appears so frequently in medical imaging studies. Once a CNN classifier is trained, Grad-CAM can be applied to any input image to generate an explanation map, including in weakly supervised settings where only image-level labels exist. This makes it attractive for domains where pixel-level pathology masks are missing or hard to define, since it offers a visual link between the prediction and image regions without requiring additional supervision.

At the same time, its popularity has created a false sense of reliability. Because Grad-CAM produces visually intuitive heatmaps, it is easy to treat the overlay as evidence that the model “looked in the right place.” But as shown in prior work on saliency trustworthiness and benchmarking, the existence of a heatmap does not guarantee that it is stable, clinically meaningful, or causally connected to the prediction in the way users assume. This is why Grad-CAM should be seen as a tool that still requires careful validation not as a final result. [17], [18]

2.3 Explanation Heatmaps Evaluated in Prior Work

The literature makes a clear distinction between showing a heatmap and evaluating it. Showing is easy but evaluation is hard, because an explanation map is not a prediction target that comes with a label in most datasets. In many medical imaging problems, the model is trained from image-level labels, so there is no direct “correct” pixel-level answer for where the evidence must be. This is why evaluation methods in prior work tend to fall into two broad groups – ground-truth based and behaviour-based. [17], [18]

Ground-truth based evaluation compares a heatmap to known annotated regions. When pixel-level masks exist, researchers can measure overlap, ranking quality, or localization performance and use that as a proxy for explanation quality. For example, studies that explicitly compare saliency maps against ground truth can quantify how different algorithms behave when there is a well-defined target region available. This is useful, but it relies on the presence of reliable spatial annotations, and it does not transfer

cleanly to domains where the pathology is diffuse or where boundaries are not clinically meaningful. [23]

Because of those limits, many influential medical imaging papers focus on behaviour-based evaluation instead. Here, the heatmap is treated as a hypothesis about what the model uses, and evaluation tests whether this hypothesis is consistent with model behaviour. Arun et al. emphasize trustworthiness questions directly and show that saliency maps can be unreliable for abnormality localization, meaning that heatmaps should not be treated as automatically faithful just because they look plausible. [17]

2.4 MRI Alzheimer Context as a Representative Weak-Supervision Use Case

Neuroimaging is a common source of data for studying dementia and Alzheimer's disease, and deep learning approaches for diagnostic classification and prognosis using brain imaging have been widely explored. This confirms that MRI is a realistic and relevant for building predictive models, but it also highlights a limitation that is central for this thesis – the supervision available in typical datasets is rarely spatial. In most cases, the label describes the patient or study (diagnosis, stage, clinical status), not a pixel-level map of pathology. [28]

MRI is also a particularly good example of why pixel-level ground truth is often unavailable in a meaningful form. In many imaging tasks, a clinician can point to a localized abnormality. In contrast, dementia-related imaging evidence is commonly linked to structural patterns that are gradual and distributed. A key example is medial temporal atrophy, which is strongly associated with dementia even after accounting for other neuropathologies. This type of signal is clinically important, but it does not naturally translate into a crisp segmentation target that could serve as ground truth for validating explanation heatmaps. [9]

For that reason, Alzheimer MRI is used here as a demonstration domain. The goal is not to claim that this specific dataset solves Alzheimer diagnosis, but to use a realistic neuroimaging setting to study how explanation heatmaps behave and how they can be evaluated when pixel-level ground truth is absent. The analysis of this subject shows that (1) visual explanations can be unreliable, (2) there are different ways to evaluate them, (3) there is no pixel-based mapping for Alzheimer's disease. This creates a research opportunity, which is addressed by proposing a combined validation approach.

3 METHODOLOGY

This chapter describes the methodological design of the thesis. It defines the dataset handling and preprocessing pipeline, the selected neural network approach for image-level classification, and the procedure used to generate explanation heatmaps. It also specifies the evaluation logic used to assess the behaviour of the explanations under weak supervision, where pixel-level ground truth annotations are not available.

3.1 Dataset Characteristics and Preprocessing Pipeline

The project uses the *Augmented Alzheimer MRI Dataset* published by Uraninjo on Kaggle. The dataset consists of T1-weighted brain MRI images organized into four classes: *Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented*. The data is provided as individual 2D image file, so the pipeline works with single 2D slices rather than full 3D MRI volumes. Each image is loaded in RGB, resized to 224×224, and passed to the model as a (224, 224, 3) input. [29]

The dataset contains two parts – original images and a pre-augmented subset that is used only in training. The validation and test sets are constructed from the original images only following the standard practice. Augmentation is used to improve learning during training, while validation and test performance should be measured on original, unmodified images to reflect real evaluation conditions and to keep the estimate of generalization comparable across experiments. [29]

To keep the planned experiments reproducible, the original part of the dataset is split into training (60%), validation (20%), and test (20%) sets with class stratification and fixed random seeds. Before splitting, all file paths are collected programmatically and sorted deterministically to avoid filesystem ordering effects. After the split is created, an integrity check verifies that the three subsets are mutually exclusive. The split is then saved to a serialized file and reused in all later steps (training, evaluation, and visualization), so that results are always computed on the same partitions. In preprocessing, pixel values are scaled to [0, 1] by dividing by 255.0 and no additional normalization is applied beyond this rescaling.

3.2 Model Architecture

The classifier is based on ResNet50 adapted to a four-class output. ResNet-style backbones are commonly used in medical imaging pipelines because transfer learning allows a model pretrained on large natural image datasets to be adapted to smaller medical datasets with relatively stable performance. [14]

Architecturally, the network can be described as a feature extractor followed by a lightweight classification head. The input image is processed through the ResNet50 backbone, which transforms the image into high-level convolutional feature maps using stacked residual blocks. These feature maps are then reduced through global average pooling, producing a compact representation that is passed into a small dense head. The final layer uses a softmax activation to output probabilities over the four target classes.

3.3 Training Configuration

Training is performed as continuation fine-tuning from an existing model checkpoint. The training scripts load a previously saved model checkpoint and continue training from that point, rather than rebuilding the architecture and training from scratch. All layers are set to trainable, and the model is optimized end-to-end using a small learning rate to refine representations while reducing the risk of destabilizing previously learned features.

The training uses the Adam optimizer with categorical cross-entropy loss and accuracy as the primary metric. A batch size of 32 is used. Fine-tuning runs for up to 10 epochs with a learning rate of 1e-5. Training is controlled through validation performance using early stopping (monitoring validation accuracy with patience 3 and restoring the best weights) and model checkpointing (saving the best-performing model by validation accuracy). No explicit regularization is applied in the saved model configuration, and class weights are not used.

Hyperparameter values were selected based on preliminary experiments and common transfer-learning practice for fine-tuning pretrained CNN models. The purpose of this study was not hyperparameter optimization but reliable explanation analysis under a fixed model configuration.

3.4 Application of Grad-CAM for Visual Interpretation

After training, each image classified by the model is interpreted using Grad-CAM. Grad-CAM was selected because it is one of the most widely used saliency methods for identifying the image regions that are most important to a model's decision, which is precisely the question this thesis asks of the classifier. It produces a class-specific heatmap showing which spatial regions contribute most to a selected class score, it can be applied post-hoc to an already trained convolutional network without changing its architecture, and it requires no pixel-level annotations. These properties make it well suited to the weakly supervised setting of this work, where only image-level labels are available. It also builds the explanation from late convolutional feature maps, so spatial structure is preserved in the resulting localization map. [21], [22]

As already mentioned, in this thesis, Grad-CAM is computed using the last convolutional layer of the ResNet50 backbone (*conv5_block3_out)* and the explained target is the model's predicted class. The resulting relevance map is passed through ReLU to retain positive contributions only and is normalized by its maximum value for visualization. The heatmap is then resized to the input resolution and overlaid on the original image using a fixed transparency setting. [21]

4 IMPLEMENTATION

This chapter documents the practical implementation of the proposed pipeline. It summarizes the development environment and hardware setup, the software stack and libraries used, and how the codebase is organized to support reproducible experiments. The chapter also outlines how training, inference, heatmap generation, and evaluation are executed in the implemented system.

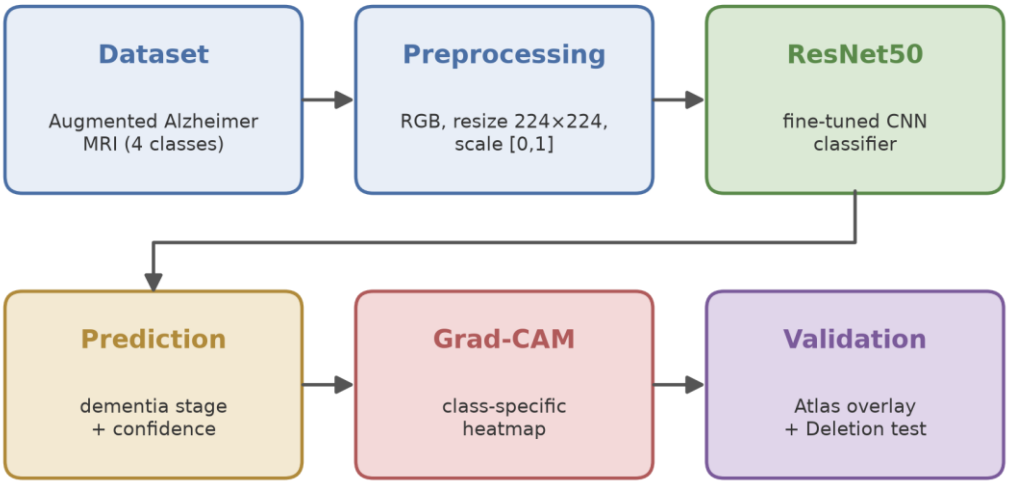

Figure 3: Overview of the implemented pipeline

4.1 Environment, Configuration, and Data Management

The implementation is built in Python (version 3.11.0) using TensorFlow and Keras libraries for model loading, continuation training, and inference (specifically TensorFlow 2.16.2, Keras 3.13.2). In this setup, TensorFlow runs as a CPU build. To keep experiments consistent across the pipeline, key parameters are centralized and reused across training, evaluation, and visualization. This includes dataset paths, the fixed input resolution of 224×224 for slices, the number of output classes, the mapping between class indices and class names, and the batch size used during training. Centralizing these settings reduces accidental mismatches between stages, such as different class orderings or different input sizes at evaluation time.

The dataset is organized into class-specific folders, and images are ingested by collecting file paths and assigning labels based on their class directory. To avoid randomness introduced by filesystem ordering while loading the images, the collected file list is sorted deterministically before any split is created. The original dataset portion is then split into training, validation, and test subsets using stratification to preserve class balance and fixed random seeds to ensure reproducibility. The pipeline also includes an explicit leakage check to confirm that the splits are mutually exclusive. The resulting split is serialized into a reusable file and loaded in all later stages, which prevents accidental re-splitting and ensures that training, evaluation, and Grad-CAM visualization always refer to the same partitions.

The atlas-based validation and visualization components were implemented using standard scientific Python tooling. Array processing and metric computation were handled with NumPy, while OpenCV was used for key image operations, including affine warping for label projection, edge extraction with Canny detection, and heatmap colour mapping and overlay rendering. Matplotlib was used to generate the comparison figures. This makes the atlas overlay reproducible at the level of operations applied to each slice, while keeping the methodological description independent of any specific software package.

The deletion-based faithfulness evaluation was implemented with patch-wise operations on NumPy arrays. Grad-CAM heatmaps were resized to the input resolution, partitioned into fixed 8×8 blocks, ranked by mean heatmap intensity, and progressively replaced using a blur baseline over a fixed number of steps. The resulting confidence trajectory was recorded and summarized using area under the curve.

4.2 Data Loading, Preprocessing, and Batching

As mentioned, at runtime, images are loaded as 2D RGB inputs and resized to the fixed resolution of 224×224 to match the model input. Each image is converted to a numeric array and normalized by scaling pixel values to the [0, 1] range through division by 255.0. The pipeline keeps the same preprocessing for training, evaluation, and Grad-CAM generation so that explanation maps are produced for the same input representation that the model receives during prediction.

Batching is handled through a generator-style approach that loads images on demand and assembles them into mini-batches of size 32. This avoids holding the full dataset in memory and keeps the preprocessing logic consistent. The training data is constructed by combining the original training subset with the dataset's pre-augmented training-only subset, while validation and test batches are built exclusively from original (non-augmented) images to keep evaluation conditions clean and comparable.

4.3 Model Fine-Tuning, Training and Evaluation

Model training is implemented as continuation fine-tuning from an existing saved checkpoint. The previously trained model is loaded, and all layers are set to trainable, so optimization proceeds end-to-end rather than training a separate head-only stage. The model is compiled with the Adam optimizer using a small learning rate of 1e-5 and categorical cross-entropy loss, with accuracy tracked as the primary metric.

Training is controlled using validation performance to reduce overfitting. Early stopping monitors validation accuracy stops training when performance no longer improves for three consecutive epochs and restores the best-performing weights. In parallel, model checkpointing saves the best model according to validation accuracy. Fine-tuning runs for up to 10 epochs, but in practice training may terminate earlier depending on validation behaviour.

Evaluation is performed on the fixed test split loaded from the stored partition file. Test images are pre-processed as described and then passed through the trained model to obtain predicted class probabilities. From these outputs, predicted labels are derived using the argmax over class probabilities and compared to ground-truth test labels. The

evaluation pipeline is designed to be consistent across runs by always using the same test subset and the same preprocessing.

4.4 Grad-CAM Generation and Visualization Pipeline

Grad-CAM visualizations are generated as a post-hoc step using the trained model and the same test images used for evaluation. For a selected image, the pipeline computes the predicted class and generates a Grad-CAM heatmap for that predicted label using the final convolutional layer of the ResNet50 backbone. The heatmap is normalized, resized to the input resolution, converted into a colour representation using a standard colormap, and blended with the original image using a fixed transparency factor ($\alpha = 0.5$). For qualitative inspection, the implementation displays three views side by side: the original input, the raw heatmap, and the heatmap overlay. A small number of test samples are selected to demonstrate typical cases as well as potential failure modes.

As stated, the most important reproducibility control is the stored dataset split. This ensures that performance and visualization comparisons are not affected by accidental re-splitting. However, one part of the pipeline remains intentionally variable – when Grad-CAM examples are chosen randomly, the exact set of displayed images can differ between runs unless a fixed random seed is set for the sampling step.

5 MODEL EVALUATION AND EXPLAINABILITY ANALYSIS

This chapter reports results of evaluation for the trained classifier and it inspects Grad-CAM heatmaps. Classifier is evaluated as a standard image-level model, using test-set performance metrics and a confusion matrix to see how well the four classes are separated. Grad-CAM heatmaps are generated on test images to understand how the model behaves in practice, including typical patterns and visible failure modes. The goal here is not to claim spatial correctness of the heatmaps, but to establish what the model achieves, what the explanations look like, and which issues motivate the dedicated validation chapter that follows.

5.1 Model’s Performance

To evaluate the effectiveness of the trained model in classifying Alzheimer's disease stages, a set of standard classification metrics was employed. These include overall accuracy, precision, recall, F1-score, and a confusion matrix. Together, these metrics are supposed to provide both an overview and insights into the model’s predictive capabilities, revealing its strengths as well as areas for improvement. Special attention is given to per-class performance and error patterns, which are particularly important in medical diagnosis tasks where different types of misclassifications may carry different clinical consequences.

5.1.1 Overall Accuracy

Overall accuracy measures the proportion of correct predictions across the entire test set. It is a useful first check, but it can be misleading if the dataset is imbalanced,

because a model can achieve a high accuracy by performing well on the dominant classes while underperforming on smaller ones.

On the test set, the final model achieved an overall accuracy of 0.9930 (99.30%) on 1280 test images. This result indicates that the model correctly classifies many MRI inputs into their corresponding Alzheimer's stage categories. However, to ensure that this performance is not driven mainly by the larger classes, it is necessary to examine results of another metrics.

5.1.2 Precision, Recall, and F1-Score

Accuracy does not show whether performance is balanced across classes. For that reason, the evaluation also reports precision, recall, and F1-score. Precision reflects how often the model's predictions for a class are correct, while recall reflects how many true instances of a class are detected. The F1-score combines both and is especially useful when class sizes differ.

Table 1 summarizes the classification report across all four classes. The results show consistently strong performance, with high scores for all categories. The *Non-Demented* class reaches perfect precision (1.000) with recall slightly below 1, indicating that the few errors present are primarily missed detections rather than false positives. The *ModerateDemented* class shows perfect scores as well, but the support is only 13 images, so this result should be interpreted cautiously. *MildDemented*, the largest class in the test set, remains very strong, which suggests that performance is not limited to small, "easy" subsets. *VeryMildDemented* also shows balanced precision and recall, indicating stable behaviour on another large class. [29]

	Precision	Recall	F1-Score	Support
Non Demented	1	0.988827	0.994382	179
Moderate Demented	1	1	1	13
Mild Demented	0.990669	0.995313	0.992985	640
Very Mild Demented	0.993289	0.991071	0.992179	448
Accuracy	0.992969	0.992969	0.992969	0.992969
Macro Average	0.995989	0.993803	0.994886	1280
Weighted Average	0.992985	0.992969	0.99297	1280

Table 1: Precision, Recall, and F1-Score for classes

5.1.3 Confusion Matrix

The confusion matrix provides a direct breakdown of predicted versus true classes in a multi-class setting. While accuracy and F1-scores summarize performance, the confusion matrix shows *where* the errors happen and whether they follow a meaningful pattern (for example, confusion between neighbouring stages rather than random jumps).

Table 2 lists the misclassification counts in a readable form. The same information is easier to interpret visually in Figure 4, where correct predictions appear on

the diagonal and errors appear off diagonal. In this case, the matrix is strongly diagonal, confirming that most test images are classified correctly. The few errors are concentrated between clinically adjacent stages, specifically between *MildDemented* and *VeryMildDemented*. This is a more reasonable failure pattern than confusion between distant severity levels.

True Class	Predicted Class Breakdown
Non-Demented	177 correctly classified, 2 misclassified as *MildDemented*
Moderate Demented	All 13 correctly classified
Mild Demented	637 correctly classified, 3 misclassified as *VeryMildDemented*
Very Mild Demented	444 correctly classified, 4 misclassified as *MildDemented*

Table 2: Predicted Class Breakdown

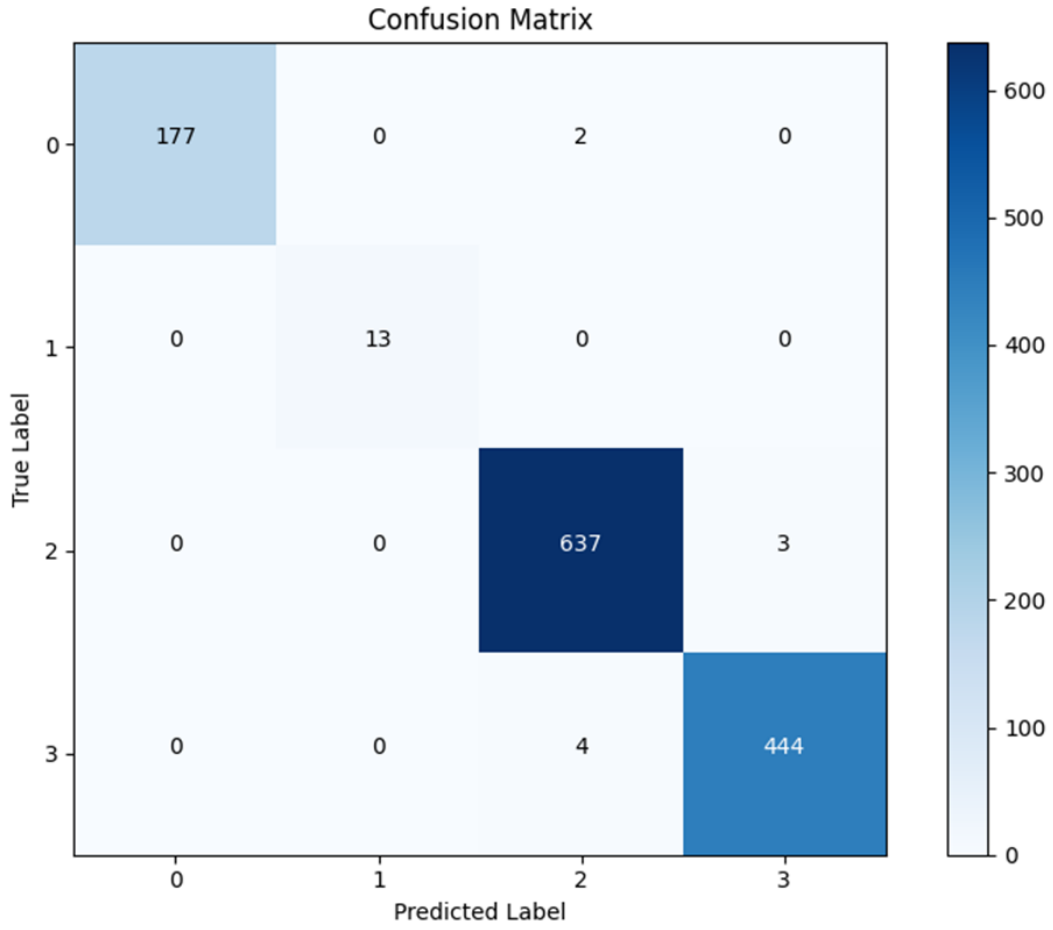

Figure 4: Confusion matrix of the trained model

Class *ModerateDemented* showed perfect performance with precision, recall, and F1-score of 1.000. However, this class also has the smallest support in the test set, only 13 images. This low support reflects the overall underrepresentation of the *ModerateDemented* category in the dataset, so the estimate of performance for this class is less stable than for the other classes. In this situation, it is possible that the model's apparent success is influenced by memorization of a small set of visually similar samples rather than robust learning of generalizable patterns. For this reason, the

ModerateDemented results should be interpreted cautiously and would benefit from validation on a larger and more balanced sample.

5.2 Baseline Grad-CAM Visualizations

Figure 5 presents a baseline set of Grad-CAM overlays for representative test images from each of the four classes. The purpose of this figure is not to claim anatomical correctness, but to give a first qualitative impression of whether the model's highlighted regions appear stable and concentrated, and whether the visual patterns differ across classes in a consistent way. Grad-CAM is used here as a standard class-specific heatmap explanation, following its common use in medical imaging studies. [21]

Across many samples, the heatmaps appear visually plausible at first glance, which is exactly why Grad-CAM is often adopted as an explanation tool. However, plausibility alone is not evidence of reliability. Prior work has shown that saliency maps can look convincing while still being untrustworthy, motivating the need to treat these visualizations as inspection artifacts rather than proof. [17]

The baseline examples also make it easier to spot issues that are not visible from classification metrics alone. The next section therefore focuses on failure modes observed during inspection, including cases where the heatmap focus shifts away from brain tissue or becomes weak and diffuse.

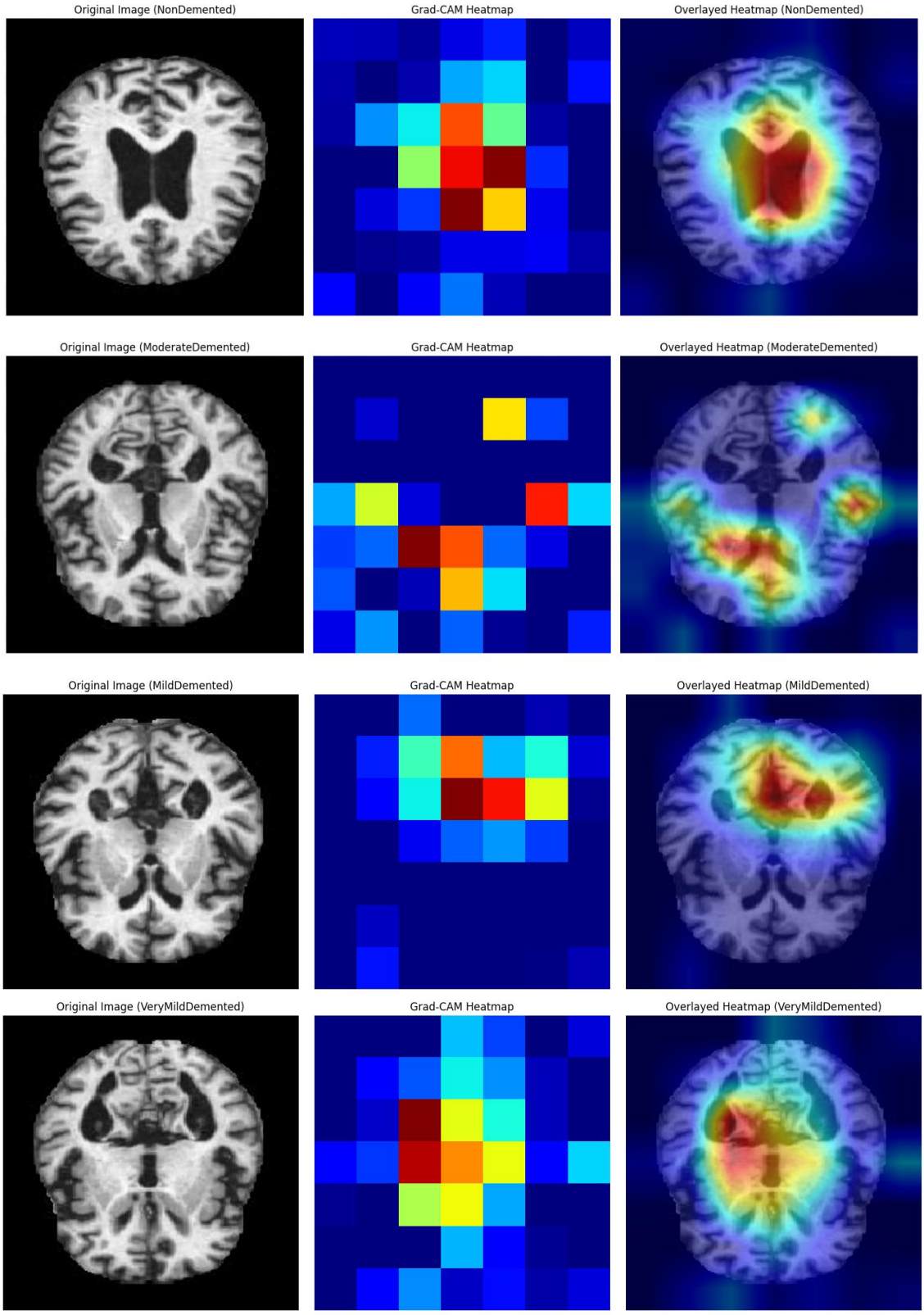

Figure 5: Grad-CAM visualisation for all classes of Alzheimer's dataset

5.2.1 Misaligned Attention Maps and Limitations

While most Grad-CAM visualizations focused on relevant brain regions, a minority of cases produced misaligned or anatomically implausible heatmaps. In these instances, the highlighted areas appeared outside the brain, sometimes concentrated on image borders, background artifacts, or empty space surrounding the head.

One such example is shown in Figure 6**Error! Reference source not found.**, where the overlayed heatmap from a test image classified as *MildDemented* revealed activation centred in a region outside the cranial structure. Although the prediction was correct, the attention map raises concerns about what features the model truly relied on to reach that decision.

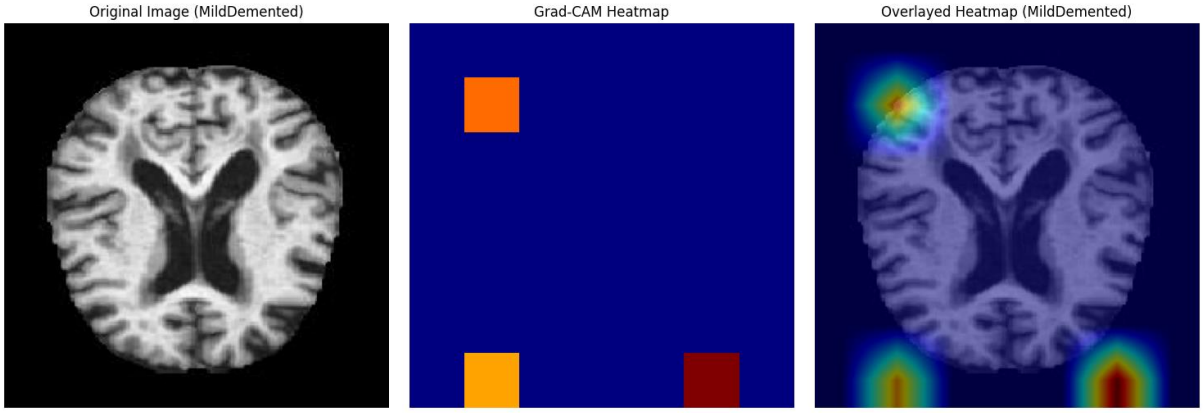

Figure 6: Activation centred outside the brain in MRI scan

This phenomenon was rare, but consistent enough to warrant further investigation. Several hypotheses may explain these misaligned activations:

- Sensitivity to image borders or background noise, especially if such patterns were inadvertently correlated with class labels during training.
- Overfitting to non-diagnostic cues, such as pixel intensities in specific locations unrelated to brain anatomy.
- Grad-CAM limitations, as it is known to sometimes produce noisy or diffuse explanations, particularly when activations are weak, or gradients are unstable.

At this point, it is not possible to determine which of these factors is responsible for the observed behaviour, and visual inspection alone is not sufficient to identify the true cause.

5.2.2 Limited Interpretability in the ModerateDemented Class

During the Grad-CAM analysis, the *ModerateDemented* class presented a distinct challenge. In multiple instances, the generated heatmaps for images from this class were largely uniform and faintly blue-tinted, indicating very low gradient activations. These results suggest that the model either struggled to identify strongly discriminative features for this class, or that the class itself lacked distinct spatial patterns that the model could learn.

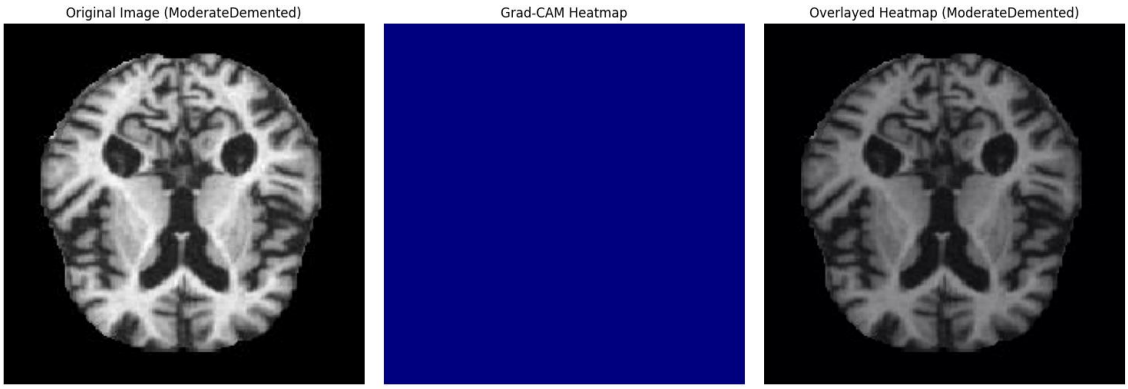

Figure 7: Generated heatmap for ModerateDemented class with low gradients activations

This effect is likely compounded by class imbalance, as *ModerateDemented* had the smallest number of samples in the dataset. In scenarios where the model is uncertain or the gradients are weak, the normalization process within Grad-CAM may produce almost flat or visually uninformative heatmaps.

5.2.3 Response to Limitations

To address the challenges – low gradient activations or activations outside of the brain structure – and to better understand them, a custom interactive visualization tool was developed. This application allows users to load any MRI image and observe the model's predicted class.

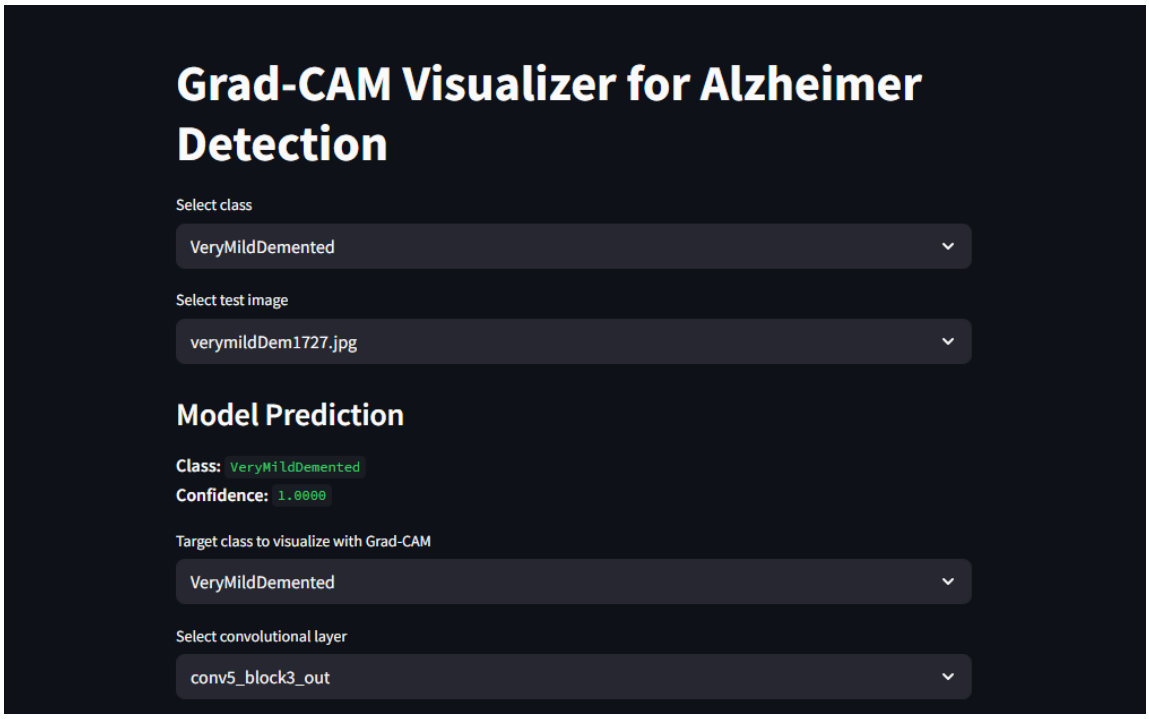

Figure 8: Grad-CAM visualizer user interface

It also displays the corresponding confidence score and enables manual navigation through different convolutional layers used for Grad-CAM generation. The goal was to investigate whether certain layers produce more meaningful heatmaps than others – particularly in cases where the default layer gave questionable outputs.

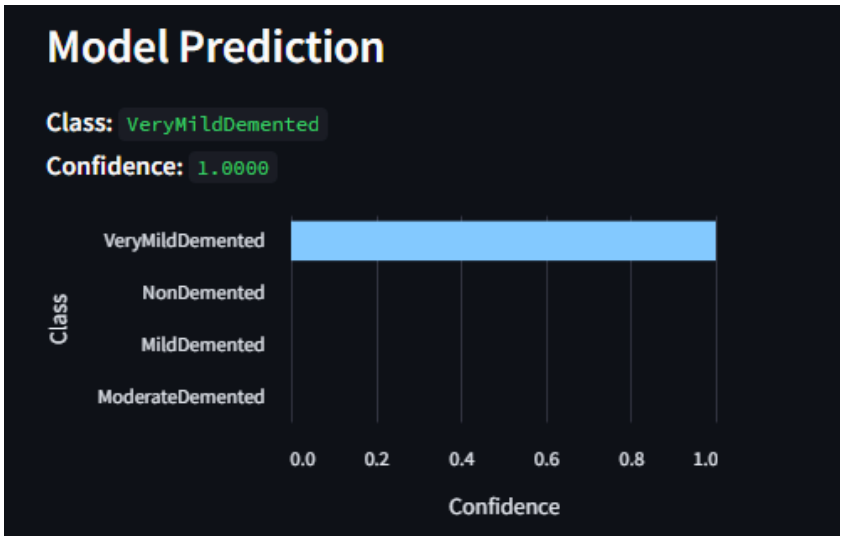

Figure 9: Model prediction and confidence display

Figure 10 illustrates the results of this layer-wise exploration for a test image. The original input is shown on the left, followed by Grad-CAM heatmaps and overlayed activations from three different convolutional layers: conv3_block4_out, conv4_block6_out, and conv5_block3_out. These layers were selected based on experimental testing, as they consistently produced interpretable and spatially coherent heatmaps across multiple images.

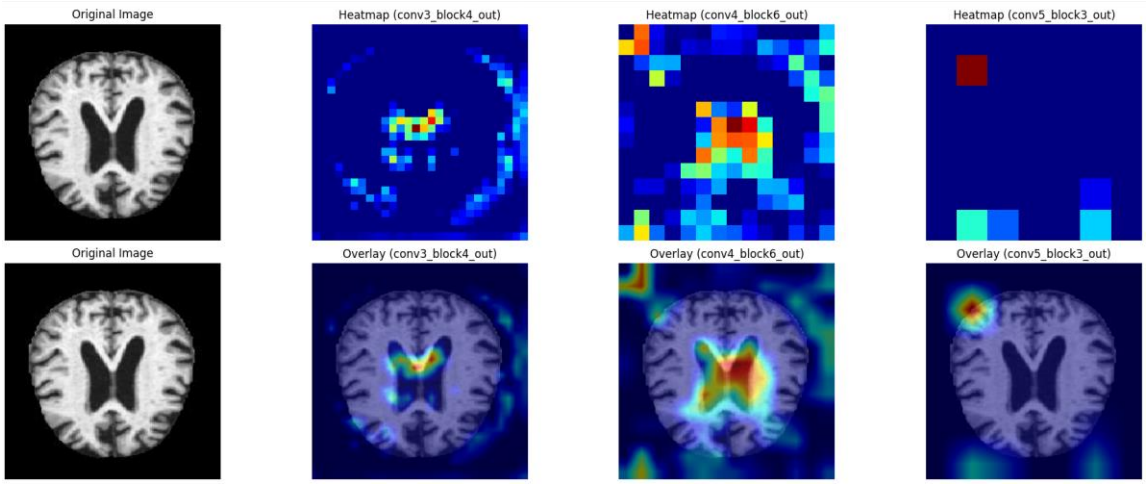

Figure 10: Comparison of different layers on MRI scan

What the layers in Figure 10 describe:

- The conv3_block4_out output highlights a compact and localized area near the brain's ventricles. The overlay suggests that although the heatmap is slightly noisy, the focus remains within the cranial structure.
- The conv4_block6_out layer provides the most precise activation in this case, with clear emphasis on central brain regions. The heatmap here aligns closely with clinically relevant anatomical markers and is well-distributed across the medial temporal area.
- The conv5_block3_out layer, which had been the default in earlier experiments, once again shows attention outside the brain region, like previously discussed

limitations. This supports the hypothesis that deeper layers can sometimes lose spatial specificity, especially in complex or underrepresented classes.

This example shows that changing the convolutional layer can substantially change the resulting Grad-CAM focus, and in some cases an earlier layer may produce a more anatomically plausible localization than the default choice. However, this behaviour is not consistent across all images: different samples can benefit from different layers, and no single layer is universally optimal. For consistency across the study, the conv5_block3_out layer remained the default because it most often produced the most concentrated and interpretable heatmaps overall.

Overall, the tool was valuable during development because it made it easier to understand how Grad-CAM behaves across layers and why certain failure modes appear. As a practical clinical tool, however, its layer-navigation feature is not well matched to typical medical workflows, since selecting and interpreting CNN layers is not part of clinical expertise. The confidence display and prediction summary may be useful to clinicians, but the main value of the layer-wise control in this thesis was exploratory rather than intended for routine professional use.

Qualitative Grad-CAM inspection is included because it provides a fast way to understand how the model behaves on individual images and to identify obvious failure patterns that are not visible from performance metrics alone. Examples such as attention outside the brain or weak, diffuse activations are important warning signs, even when the predicted class is correct. At the same time, qualitative inspection cannot determine whether highlighted regions are truly responsible for the prediction, nor can it establish correctness when no pixel-level ground truth exists.

6 EXPLANATION RELIABILITY EVALUATION

In this thesis, explanation maps cannot be validated in the usual way because the chosen dataset does not provide pixel-level ground truth annotations of pathology since the imaging signaling of Alzheimer's disease is diffused and does not naturally form a single bounded region that could be consistently segmented. As a result, explanation quality cannot be assessed by direct overlap with "correct" regions, and visual plausibility alone is not sufficient evidence of reliability. This makes validation a necessary step rather than an optional addition, especially when heatmaps could otherwise be over-interpreted as proof of what the model used. [17], [18]

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.

6.1 Anatomical Atlas Comparison

The atlas overlay is not treated as ground truth. Its purpose is to support an anatomical plausibility check by translating highlighted pixels into named brain regions, rather than leaving the explanation as an unlabelled heatmap. This is especially useful in neuroimaging settings where pixel-level pathology annotations are not available and where the relevant imaging patterns are often diffuse rather than sharply bounded. [30]

6.1.1 Atlas Bank Construction

To enable consistent region interpretation across slices, a unified atlas "bank" was created by combining three publicly available atlases: two Harvard–Oxford atlases (cortical and subcortical coverage) and the Automated Anatomical Labelling atlas. All atlas resources were brought into the same standard reference space (MNI152) and reduced to a set of representative axial slices. Each slice contains a reference anatomy image and a corresponding label map with a lookup table that maps label IDs to region names, so the setup would follow the general goal of standardizing brain region definitions to make anatomical reporting consistent and comparable. [31]

6.1.2 Atlas-to-Image Alignment and Label Warping

For each input MRI slice, the atlas bank is aligned to the image using image registration and the best-matching atlas slice is selected automatically. The system evaluates a set of candidate axial atlas slices rather than assuming a fixed slice in advance. Candidate alignments are ranked using a single combined similarity score that captures both intensity agreement and boundary agreement. Intensity agreement is measured using mutual information computed from a 64-bin joint histogram over intensity values scaled to the [0, 1] range. Boundary agreement is measured using the Jaccard overlap of Canny edge maps (with thresholds 40 and 120). These two terms are combined with a higher weight on mutual information than on edge overlap, and the

highest-scoring candidate is selected and used to warp the atlas label map into the MRI slice coordinate system via an affine transform. No fixed acceptance threshold is applied meaning the method always selects the best available match from the atlas bank. [32]

6.1.3 Example Interpretation with Generated Atlas Mask

Figure 11 shows an example of the resulting Grad-CAM atlas overlay for a test slice predicted as *VeryMildDemented* with confidence 1.000. In this case, the selected atlas slice index, with score 81, corresponds to the candidate that achieved the highest combined alignment score among the available atlas slices. The reported mutual information (2030.703) indicates strong intensity-based agreement under the chosen similarity definition, while the edge overlap term (0.114) provides additional boundary support. The final overlay uses an affine registration, and the reported atlas score summarizes the ranking used to select this match. Because the selection is based on a best-of-candidates rule rather than an absolute quality threshold, these values should be interpreted as evidence that this alignment is the strongest available match in the atlas bank for this slice, not as a guarantee of perfect anatomical correspondence. [30]

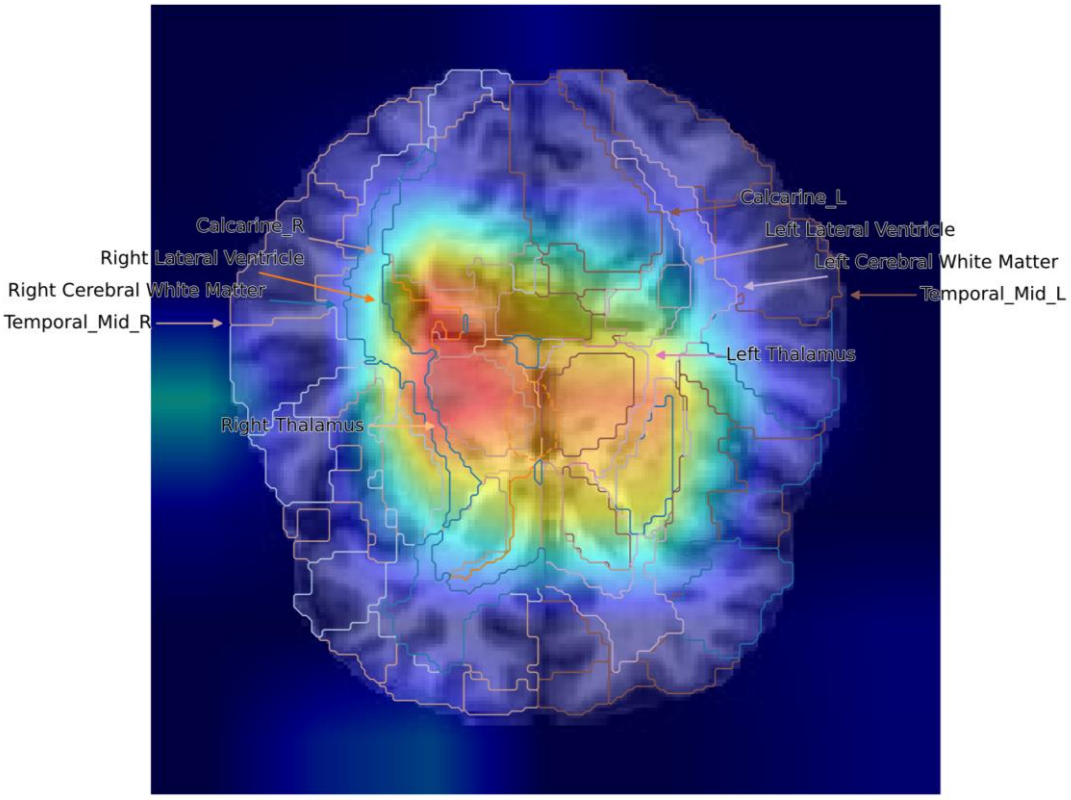

Figure 11: Grad-CAM heatmap overlaid with the anatomically registered atlas mask for a test MRI slice predicted as Very Mild Demented

Once aligned, the overlay helps convert the heatmap into a region-level interpretation. In this example, high activations intersect temporal and ventricular-adjacent regions, which can be directionally consistent with dementia-related structural change, while a substantial portion of activation also overlaps deep central structures such as thalamic regions. This makes the example partly plausible but not perfectly specific, and it illustrates the intended role of the atlas overlay as an interpretative aid rather than a ground-truth localisation tool. [9], [10], [11]

6.1.4 Boundary Extraction via Iso-Contours and Filtered Atlas Overlay

A practical limitation of using Grad-CAM for plausibility checks is that a raw heatmap is often difficult to judge consistently. The colour gradient can be dispersed, speckled, or visually dominated by mid-level activations, which makes it harder to decide what the model's core evidence is. This is one reason why saliency maps are frequently over-interpreted as the visualization looks informative, but the boundary between "signal" and "background" is unclear and strongly influenced by presentation choices. [17], [18]

To make the output easier to validate, the visualization was redesigned to explicitly separate two questions. Firstly, where the model focuses, and secondly which anatomical regions this focus intersects. Instead of relying only on a continuous heatmap, the Grad-CAM map is converted into a high-activation boundary. Concretely, the heatmap is lightly smoothed to reduce speckle and then thresholder is fixed at an iso-level of 0.6. The connected high-activation region is outlined as an iso-contour, producing a clear "focus perimeter" that is typically faster to assess than a fuzzy gradient. This representation makes it easier to spot cases where activations are scattered, misplaced, or overly broad, and it also supports more consistent comparisons across images. [18]

Because Grad-CAM localisation depends on the selected convolutional layer, several candidate layers are evaluated and the one producing the most usable explanation is selected automatically. The selection combines the model's confidence with simple heatmap characteristics, such as overall intensity and how concentrated the activation is, to avoid explanations that are either weak and uninformative or trivially saturated. After the atlas label map has been warped into the MRI slice coordinate system, the anatomical overlay is filtered using the iso-contour region. A region label is kept only if at least 5% of its area overlaps the high-activation boundary. Practically, this removes label clutter and ensures that the final overlay reports only the regions that intersect the model's strongest evidence, rather than listing structures that appear in the slice but are not meaningfully implicated by the heatmap.

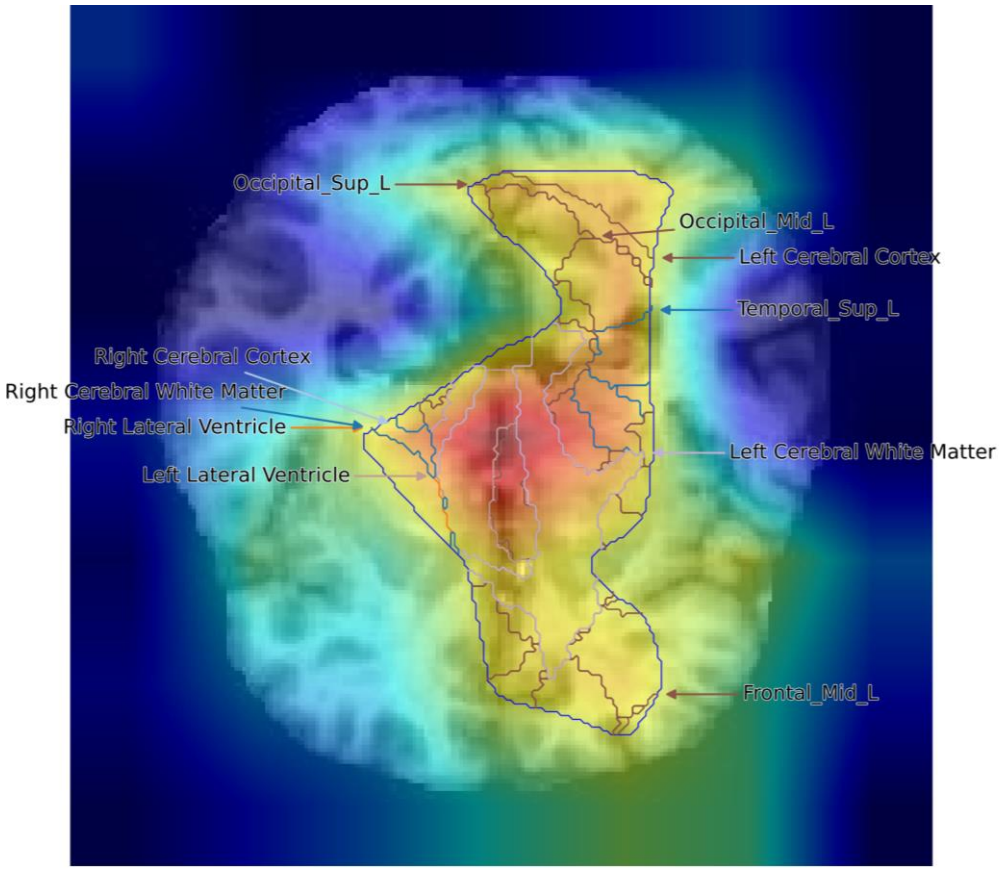

Figure 12: Filtered Grad-CAM atlas overlay for a VeryMildDemented MRI slice using an iso-contour boundary

By turning the heatmap into a clear boundary and reporting only overlapping atlas regions, the visualization becomes less ambiguous and faster to read. It does not turn the overlay into ground truth, but it makes the plausibility check more structured by explicitly showing the model's primary focus area and the anatomical regions it intersects.

6.2 Deletion-based Faithfulness Evaluation

The second validation strategy evaluates faithfulness using a perturbation-based metric commonly referred to as the deletion test. The core idea is that if a heatmap truly highlights evidence the model relies on, then removing the most relevant regions first should cause the model's confidence for the target class to decrease quickly. The image is progressively perturbed in the order induced by the heatmap ranking, and the model's predicted probability is recorded after each deletion step. The resulting curve is summarized by its area under the curve (AUC), where lower AUC values indicate that confidence drops earlier and the explanation is therefore treated as more faithful under this perturbation scheme. [17], [18]

6.2.1 Perturbation Design

In this work, deletion is performed patch-wise rather than pixel-wise. Instead of removing individual pixels, the method removes small blocks of the image, which makes the procedure more stable and less sensitive to pixel-level noise. Removed regions are replaced with a blurred baseline instead of hard masking. This reduces fine detail while

preserving coarse structure and avoids introducing extremely artificial patterns that could distort model behaviour in a way unrelated to the explanation itself. The deletion AUC is computed per image and then aggregated by true class to compare how faithfulness behaves across the four categories. [17]

Deletion order is derived directly from the Grad-CAM heatmap by ranking fixed-size patches according to their mean activation. The resized heatmap is partitioned into an 8×8 patch grid, and each patch receives a score equal to the average heatmap intensity inside that block. Patches are then sorted in descending order and removed from highest to lowest score. Deletion is applied progressively over 30 steps, where each step removes the next batch of top-ranked patches by replacing them with the blur baseline.

6.2.2 Class-wise Results and Interpretation

Figure 13 summarizes deletion AUC distributions by true class. Lower values indicate that removing regions selected by Grad-CAM reduces the model's confidence faster, which is interpreted as stronger faithfulness under the tested setup. With the chosen configuration described in the previous section, the *VeryMildDemented* class shows the lowest AUC values overall (mean ≈ 0.180, median ≈ 0.139). This suggests that for many images in this class, the model's prediction depends on a relatively compact set of regions that Grad-CAM can capture. The *ModerateDemented* class also shows comparatively low AUC, but interpretation is limited by the small number of test samples (13 images), so these results should be treated cautiously.

In contrast, *MildDemented* exhibits a broader spread and a higher central tendency with many outliers, indicating less consistent faithfulness across cases. For some images, deleting highlighted regions reduces confidence quickly, while for others the confidence remains relatively stable even after substantial deletion. The most distinct pattern appears in *NonDemented*, where deletion AUC is higher and clustered closer to the upper range. This indicates that removing the regions emphasized by Grad-CAM often has only a limited effect on the *NonDemented* confidence score, which is consistent with either more distributed evidence for this class or a mismatch between the highlighted regions and the features the model relies on. [17], [24]

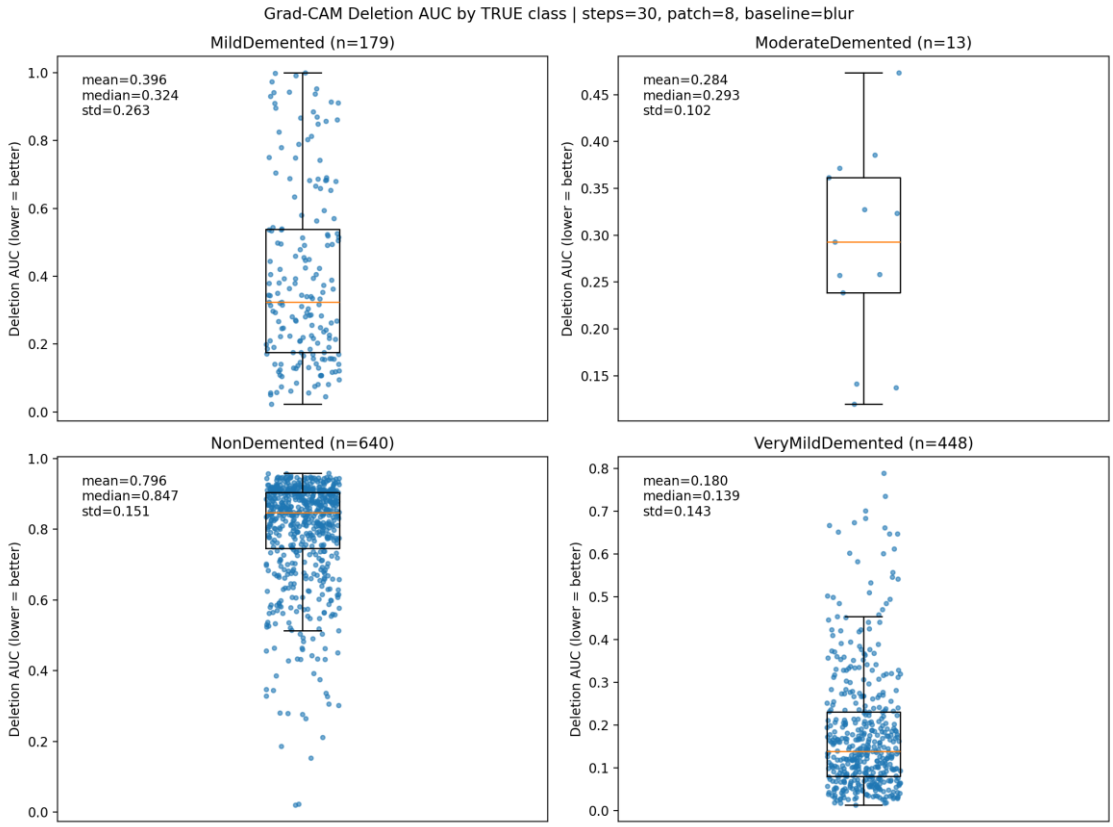

Figure 13: Grad-CAM deletion AUC by True class

Overall, under the tested perturbation setup, the deletion analysis suggests that Grad-CAM behaves as most faithful for *VeryMildDemented*, less consistently for *MildDemented* and *ModerateDemented* (with strong uncertainty due to low support), and least faithful for *NonDemented*.

Unlike visual inspection, the deletion test provides a quantitative check on whether the highlighted regions have measurable influence on the model output. This makes it possible to compare explanation behaviour across classes and to flag cases where a heatmap may look plausible but does not meaningfully affect the model's confidence when removed. [17], [18]

6.3 Complementarity of Plausibility and Faithfulness Validation

The two validation strategies used in this thesis address different questions, which is why they are useful together. The atlas overlay is designed to support anatomical plausibility as it helps interpret where the model's strongest activations fall by mapping them onto named brain regions. This makes it easier to judge whether the highlighted areas form a medically reasonable story, or whether the model appears to focus on unexpected structures or non-brain regions.

The deletion test addresses a different dimension, faithfulness. Instead of asking whether the highlighted regions look anatomically meaningful, it tests whether they matter for the model's prediction by measuring how quickly the model's confidence drops when those regions are removed. This provides a behavioural check that visual inspection alone cannot provide. [17]

Importantly, plausibility and faithfulness do not have to agree. A heatmap can overlap plausible anatomical regions and still be weakly connected to the prediction, and conversely, a heatmap can be highly influential under deletion while highlighting areas that are difficult to justify anatomically. For that reason, using both checks together provides a more balanced assessment than relying on a single validation signal. [18]

7 DISCUSSION

This chapter interprets the results of the previous chapters and discusses what they imply for the reliability of Grad-CAM explanations. It first relates the model's classification performance to explanation quality, then examines the two validation strategies separately in terms of plausibility and faithfulness and considers what their agreement or disagreement reveals. It closes with the main limitations and threats to validity and outlines directions for future work.

7.1 Performance

The results of this thesis show that strong image-level classification performance does not automatically translate into reliable explanation maps. The trained model achieved strong performance metrics on the selected dataset, yet the Grad-CAM inspection still revealed cases where attention was misaligned, including activations outside brain tissue and weak or diffuse heatmaps in specific cases. This supports the central motivation of the work, which is that explanation maps must be treated as outputs that require evaluation rather than as visual proof of correct reasoning. The validation chapter therefore focused on methods that remain possible even when pixel-level ground truth does not exist, which is common in neurodegenerative imaging where the signal is often diffuse and not naturally represented as a bounded region. [17], [18]

High classification performance and questionable heatmaps can coexist because they measure different things. The classification metrics only show that the model can separate the dataset labels under the given preprocessing and split. They do not show whether the model is relying on clinically meaningful features, nor whether it would behave similarly under small changes in acquisition or background cues. A model can therefore achieve strong test performance while still using shortcut signals that are correlated with labels in the dataset. In such cases, Grad-CAM may reveal attention patterns that look anatomically implausible, not because Grad-CAM is necessarily wrong in its own terms, but because it is exposing that the model's evidence may not be the evidence a human expects. This is one reason why visual explanations should not be treated as proof of correct reasoning and why validation needs to focus on what can be tested under weak supervision rather than on appearance alone.

7.2 Plausibility

A first important observation is that qualitative Grad-CAM inspection is useful but limited. It helps detect obvious failure patterns that would be invisible from classification metrics alone, such as attention drifting to background regions. At the same time, visual plausibility cannot establish whether the highlighted regions are truly responsible for the prediction, and it cannot explain why certain failures occur.

The atlas overlay addresses a different need. Its contribution is not to provide ground truth localisation, but to make plausibility checks more structured by linking highlighted pixels to named anatomical regions. This reduces ambiguity and helps the reviewer reason about what the heatmap implies anatomically. The best-of-candidates selection approach ensures that the overlay corresponds to the strongest match available in the atlas bank for a given slice, but it does not guarantee perfect anatomical

correspondence, and its usefulness depends on registration quality and the suitability of representing 3D anatomy through 2D slice alignment. [30], [32]

7.3 Faithfulness

The deletion results are useful because they test explanation behaviour through the model's output rather than visual plausibility. Interpreting the class-wise differences, the low AUC values for VeryMildDemented are consistent with a situation where the model relies on a relatively compact set of discriminative cues that Grad-CAM captures well as high-intensity regions. In contrast, the higher AUC values for NonDemented suggest that the model's confidence often remains stable even after removing the regions Grad-CAM highlights most strongly. That pattern can occur when the evidence for a class is more distributed across the slice, or when the heatmap does not align with the features that are most causally responsible for the prediction under the tested perturbation scheme.

The broader spread for *MildDemented* also matters. It suggests that the faithfulness signal is not stable across all cases in that class. Some samples behave as expected, with confidence dropping early as the most salient regions are removed, while other samples retain confidence for longer. This variability fits the overall point of the thesis that explanation reliability is not a fixed property of Grad-CAM. It depends on the input and the learned decision logic, and it can differ across classes even when classification metrics are high.

The *ModerateDemented* results remain difficult to interpret due to low support. With only 13 test samples, perfect classification scores and low AUC values can occur even when generalisation is weak, simply because the estimate is based on too few cases. This does not invalidate the result, but it limits what can be concluded from it. For that class, the safest interpretation is that the observed behaviour is consistent with relatively compact evidence under the tested setup, but stronger claims would require more data.

7.4 Disagreement

Plausibility and faithfulness do not have to align, and disagreement between them is informative. A heatmap can intersect plausible anatomical regions while having weak causal influence on the model output, and a heatmap can be highly influential under deletion while remaining difficult to justify anatomically. This is why combining both checks provides a more balanced evaluation than relying on a single signal. The iso-contour boundary and overlap filtering used in the atlas output also improved interpretability by separating the core focus region from the set of anatomical labels, reducing clutter and supporting faster review without claiming clinical localisation validity.

7.5 Limitations

Several limitations should be kept in mind when interpreting these findings. The study uses a pre-processed and augmented dataset that may not reflect the variability of clinical acquisition across sites and scanners. The pipeline operates on 2D slices rather than full 3D volumes, which limits spatial context and may influence both classification and explanation behaviour. In clinical practice, MRI examinations are

typically reviewed as a sequence of slices that together form a volume, so slice-level analysis remains a practical viewpoint, but it cannot capture patterns that depend on consistency across neighbouring slices. For this reason, the validation results in this thesis should be understood as evidence about slice-level explanation behaviour, while full volumetric evaluation remains a natural direction for future work.

Additionally, the *ModerateDemented* class has very low support in the test set, which makes both performance and faithfulness estimates for that class less stable and increases uncertainty. Finally, deletion results depend on the chosen perturbation design, including patch size, number of steps, and baseline replacement, so AUC values should be interpreted as relative evidence under a specific evaluation setup rather than as absolute measures of explanation quality.

Despite these limitations, the work demonstrates that it is possible to evaluate heatmap explanations in a structured way even when pixel-level ground truth is missing. The key practical takeaway is that heatmap explanations should not be accepted based on appearance alone, and that combining anatomically grounded plausibility checks with behaviour-based faithfulness testing provides stronger evidence about explanation reliability than qualitative inspection in isolation.

7.6 Future Directions

A natural next step is to move from slice-level analysis to volumetric evaluation. Extending the pipeline to 3D inputs would allow both the classifier and the explanation maps to use continuity across neighbouring slices, which is especially relevant in neuroimaging where patterns are often distributed and not confined to a single slice. This would also make it possible to validate whether the same regions remain relevant across a volume rather than only within one 2D view.

Another direction is to broaden the explainability comparison. Grad-CAM is widely used, but relying on a single method makes it difficult to know whether observed patterns are method-specific artefacts or more general properties of the model. Comparing Grad-CAM with other attribution approaches would strengthen conclusions about explanation stability and help identify cases where explanations are consistent across methods versus cases where they diverge.

The validation framework itself can also be strengthened by testing robustness. Faithfulness results depend on the perturbation design, so evaluating multiple deletion baselines and patch granularities would clarify how sensitive the conclusions are to the chosen setup. Similarly, explanation stability can be tested under small, clinically irrelevant input changes such as minor intensity shifts or slight spatial perturbations, to evaluate whether explanations remain consistent when predictions remain unchanged.

Future work can also explore using explanation validation as a feedback signal during model development. In weakly supervised diseases where pixel-level annotations are not available, validation outputs such as repeated attention outside the relevant anatomy or consistently weak faithfulness under deletion could be treated as indicators of shortcut learning or poor feature reliance. These indicators could then guide targeted retraining strategies, such as revising preprocessing, reducing background cues, adjusting augmentation, or rebalancing classes, with the goal of encouraging models to rely on more stable and anatomically plausible evidence. This would not replace ground

truth supervision, but it could support iterative improvement of model reliability in settings where traditional pixel-level validation is not feasible.

Finally, the atlas overlay can be refined toward more reliable anatomical grounding. Improvements could include more systematic registration quality control and extending the atlas bank or registration procedure to better handle anatomical variability. These refinements would not turn the overlay into ground truth, but they would increase confidence that the reported region labels reflect the intended anatomical mapping and reduce the chance of misleading overlays in difficult cases.

CONCLUSION

This thesis addressed the problem of evaluating heatmap explanations in medical imaging tasks where pixel-level ground truth annotations are not available. Using Alzheimer brain MRI as a representative weakly supervised setting, a ResNet50-based classifier was trained to predict four diagnostic categories and Grad-CAM was used to generate class-specific explanation heatmaps. The evaluation showed that strong image-level classification performance can still coexist with explanation failure modes, including misaligned attention outside brain tissue and weak or diffuse heatmaps in some cases.

To move beyond visual plausibility, the thesis applied two complementary validation strategies. An atlas-based overlay was used to support anatomical plausibility assessment by mapping heatmap focus to named brain regions, while a deletion-based metric measured faithfulness by testing how model confidence changes when highlighted regions are progressively removed. The results indicated that explanation behaviour is not uniform across classes and that plausibility and faithfulness can diverge, which supports the need for structured validation rather than relying on appearance alone.

Overall, the thesis demonstrates that explanation validation is feasible even under weak supervision, but it must be treated as an explicit part of the workflow. Combining anatomically grounded plausibility checks with behaviour-based faithfulness testing provides stronger evidence about explanation reliability than qualitative inspection in isolation, and it offers a practical foundation for future work on more robust and clinically relevant explanation assessment.

Bibliography

[1] P. Purwono, A. N. E. Wulandari, and K. Nisa, 'Explainable artificial intelligence (XAI) in medical imaging: Techniques, applications, challenges, and future directions', *Advanced Mechanical and Mechatronic Systems*, vol. 1, no. 1, p. 52 66, 2025.

[2] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, 'Explainable artificial intelligence (XAI)—From theory to methods and applications', *IEEE Access*, vol. 12, pp. 80799–80840, 2024, doi: https://ieeexplore.ieee.org/document/10465593.

[3] A. Barredo Arrieta *et al.*, 'Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', in *Proceedings of the IEEE International Conference on Computer Vision (ICCV*, 2017, pp. 618–626.

[5] K. Shahi and A. Bagale, 'Weakly supervised pneumonia localization from chest X-rays using deep neural network and Grad-CAM explanations', *Cluster Computing*, 2025, doi: 10.1007/s10586-025-05281-5.

[6] C. R. Jack *et al.*, 'NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease', *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018, doi: 10.1016/j.jalz.2018.02.018.

[7] World Health Organization, 'Dementia (Fact sheet)'. 2025. Accessed: Mar. 08, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[8] P. Rosa-Neto, 'Brain imaging using CT and MRI. In World Alzheimer Report 2021: Journey Through the Diagnosis of Dementia (Chapter 9', *Alzheimer's Disease International*, 2021, [Online]. Available: https://www.alzint.org/u/World-Alzheimer-Report-2021-Chapter-09.pdf

[9] D. C. Woodworth, 'Dementia is strongly associated with medial temporal atrophy even after accounting for neuropathologies', *Brain Communications*, vol. 4, no. 2, p. 052, 2022, doi: https://academic.oup.com/braincomms/article/4/2/fcac052/6543086.

[10] J. M. Biesbroek, M. G. Verhagen, S. Stigchel, and G. J. Biessels, '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, 2024, doi: 10.1002/alz.13563.

[11] G. Forno, M. Saranathan, and J. Contador, 'Thalamic nuclei changes in early and late onset Alzheimer's disease', *Current Research in Neurobiology*, vol. 4, p. 100084, 2023, doi: https://www.sciencedirect.com/science/article/pii/S2665945X23000128.

[12] Alzheimer's Association, 'Credits and Links (Brain Tour)'. 2021. Accessed: Aug. 26, 2021. [Online]. Available: https://www.alz.org/alzheimers-dementia/what-is-alzheimers/brain-tour/credits

[13] E. Zaitseva and V. Levashenko, ‘Reliability engineering in healthcare: Opportunities and challenges’, *Reliability Engineering & System Safety*, vol. 267, no. Part B, p. 111933, Mar. 2026, doi: 10.1016/j.ress.2025.111933.

[14] C. Chen, N. A. Mat Isa, and X. Liu, ‘A review of convolutional neural network based methods for medical image classification’, *Computers in Biology and Medicine*, vol. 185, p. 109507, Feb. 2025, doi: 10.1016/j.compbiomed.2024.109507.

[15] F. Doshi-Velez and B. Kim, ‘Towards A Rigorous Science of Interpretable Machine Learning’. 2017. doi: 10.48550/arXiv.1702.08608.

[16] X. Li *et al.*, ‘Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond’, *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, 2022, doi: 10.1007/s10115-022-01756-8.

[17] N. Arun *et al.*, ‘Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging’, *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021, doi: 10.1148/ryai.2021200267.

[18] A. Saporta *et al.*, ‘Benchmarking saliency methods for chest X-ray interpretation’, *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022, doi: 10.1038/s42256-022-00536-x.

[19] A. Barredo Arrieta *et al.*, ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[20] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, ‘Explainable artificial intelligence (XAI): from theory to methods and applications’, *IEEE Access*, vol. 12, pp. 80799–80846, 2024, doi: 10.1109/ACCESS.2024.3409843.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization’, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[22] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, ‘Reviewing CAM-Based Deep Explainable Methods in Healthcare’, *Applied Sciences*, vol. 14, no. 10, p. 4124, 2024, doi: 10.3390/app14104124.

[23] K. Szczepankiewicz *et al.*, ‘Ground truth based comparison of saliency maps algorithms’, *Scientific Reports*, vol. 13, no. 1, p. 16887, 2023, doi: 10.1038/s41598-023-42946-w.

[24] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, ‘Is Grad-CAM Explainable in Medical Images?’ 2023. doi: 10.48550/arXiv.2307.10506.

[25] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, ‘A comprehensive review of explainable artificial intelligence (XAI) in computer vision’, *Sensors*, vol. 25, p. 4166, 2025.

[26] P. Rajpurkar *et al.*, ‘CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning’. 2017. doi: 10.48550/arXiv.1711.05225.

[27] Y. Zhao, M. T. Ribeiro, T. Wu, and J. Heer, ‘Graphical perception of saliency-based model explanations’, in *Proceedings of the ACM CHI Conference on Human Factors*

in Computing Systems, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3544548.3581320

[28] T. Jo, K. Nho, and A. J. Saykin, 'Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data', *Frontiers in Aging Neuroscience*, vol. 11, p. 220, 2019, doi: 10.3389/fnagi.2019.00220.

[29] Uraninjo, 'Augmented Alzheimer MRI Dataset'. 2022. Accessed: Mar. 08, 2026. [Online]. Available: https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset

[30] W. L. Nowinski, 'Evolution of Human Brain Atlases in Terms of Content, Applications, and Visualization', in *Brain Imaging and Behavior / Brain Informatics (Springer*, 2021. doi: 10.1007/s12021-020-09481-9.

[31] R. M. Lawrence, C. M. O'Toole, and B. Duffy, 'Standardizing human brain parcellations', *Scientific Data*, vol. 8, p. 98, 2021, doi: https://www.nature.com/articles/s41597-021-00849-3.

[32] D. Sengupta, 'A survey on mutual information based medical image registration'. 2022. [Online]. Available: https://dl.acm.org/doi/10.1016/j.neucom.2021.11.023

ATTACHMENTS

LIST OF ATTACHMENTS

ATTACHMENT A | SOURCE CODE (USB MEDIUM)

The attached USB medium contains the complete source code of the implemented pipeline, including the training, inference, heatmap generation, and evaluation scripts. It also includes generated explanations on images including those featured in the thesis and additional ones.